On the identification of outliers in a simple model

C. S. Wallace*

1 Introduction

We suppose that we are given a set of N observations $\{y_i \mid i = 1, \ldots, N\}$ which are thought to arise independently from some process of known form and unknown (vector) parameter θ. However, we may have reason to suspect that some small fraction of the N observations are in some sense contaminated or erroneous, i.e., that they arise from a process different from the main process. Any such observation is called an "outlier". We will then be interested in methods for identifying or at least estimating the number of the outliers, and for estimating θ in a way which is minimally upset by the outliers.

Freeman [3] has discussed this problem for general linear models, using three different models for the outlier-generating (or "error") process. The resulting analyses of four data sets showed a rather poor ability to find outliers, and a high sensitivity to some parameters of the prior distribution assumed for the error process. In particular Freeman remarked on the inability of the methods he considered to deal with both positive and negative errors.

In this paper, we look at a very simple case, where the main process is a univariate normal distribution, and apply the minimum-message-length estimation technique. The analysis suggests some defects inherent in methods considered by Freeman, and throws some light on their difficulty with bi-directional errors. The estimators developed here are not unduly difficult to calculate for large N, unlike those of Freeman, which suffer an exponential computational complexity.

*This document was written some time in the 1980s by CSW, Professor of Computer Science at Monash University. For more on the Minimum Message Length (MML) principle, see C.S.Wallace, 'Statistical and Inductive Inference by Minimum Message Length,' Springer Verlag, isbn13:978-0387237954, 2005.

2 Two simple outlier models

The main process, by which the majority of the N observations are generated, is independent selection from $N(\mu, \sigma)$, with μ and σ unknown. We assume μ to have a uniform prior in some range of size R, and $\lambda = \ln \sigma$ to have a uniform prior in $(\ln 2\epsilon, \ln R/3)$, where $\epsilon \ll R$ is the measurement or recording error in the data $\underline{x} = \{y_i\}$. The prior for λ is the usual "colourless" prior, bounded below by noting that σ cannot reasonably be supposed to be less than the measurement error of the data. The upper limit of $R/3$ on σ is supposed to reflect the assumption that we have so little prior information about μ that R is actually the range within which measurements of $\underline{x}$ can be made.

Freeman considers three models for the error process, BT due to Box and Tiao [2], AB due to Abraham and Box [1], and GDF due to Guttman, Dutter and Freeman [4]. BT assumes outliers are drawn from $N(\mu, k\sigma)$ with k assumed known. AB assumes outliers are drawn from $N(\mu + \delta, \sigma)$ with δ assumed unknown but constant. GDF assumes outliers are drawn from the convolution of $N(\mu, \sigma)$ with another distribution of known form. Thus GDF assumes that an outlier value is obtained by adding to a value from $N(\mu, \sigma)$ an error drawn from another distribution. The latter distribution may be regarded as the prior distribution for an additional set of parameters, one per outlier, which represent the additive errors included in those outliers. The errors included in different outliers are assured independent.

In all three models, each observation is regarded as having prior probability α of being an outlier, and $(1 - \alpha)$ of coming from the main process. Freeman considers α to be known *à priori* in all cases.

As Freeman observes, model AB is incapable of modelling both positive and negative contamination of values, since it assumes all outliers have been corrupted by the same value δ. As we regard this model as applicable to few real problems, we will not consider it further.

The remaining two models both lead to a model density for the observations of the form

$$t(\underline{x}) = (1 - \alpha)N(\underline{x} \mid \mu, \sigma) + \alpha Q(\underline{x} \mid \phi)$$

where $N(\underline{x} \mid \mu, \sigma)$ is the Normal density, and $Q(\underline{x} \mid \phi)$ is the density produced by the error process with parameter ϕ. For model BT, Q is $N(\underline{x} \mid \mu, k\sigma)$, and for GDF, Q is the convolution of $N(\mu, \sigma)$ with some prior density for additive corruption.

In this paper we address the outlier problem in the above form. That is, we consider the observations to be drawn from a mixed distribution, and are

concerned to estimate the parameters of the mixture.

We will consider two error process distributions.

2.1 Uniform Error Process

In the first, and simpler case, we assume outliers have a uniform distribution over the same range R as the prior for μ. This model approximates the situation where, if an observation is in error, it cannot be expected to have any resemblance to a true (main-process) value. The model is similar to the GDF model if the error distribution of the GDF models is very much broader than the main process.

$$t(\underline{x}) = (1 - \alpha)N(\underline{x} \mid \mu, \sigma) + \alpha/R$$

2.2 Broad Normal Error Process

In the second case, we will adopt the BT model, with an error process which is related to, but broader than, the main process. Then

$$t(\underline{x}) = (1 - \alpha)N(\underline{x} \mid \mu, \sigma) + \alpha N(\underline{x} \mid \mu, k\sigma)$$

In both cases, unlike Freeman, we will assume α to be unknown, with prior distribution uniform in (0, 0.5). The analysis can easily be specialised if α is known.

3 Message length for mixed distribution

The minimum message length estimation method is based on the use of an estimate to allow a concise encoding of the available data given a data vector $\underline{x}$, a known discrete conditional probability function $p(\underline{x} \mid \hat{\theta})$, and a prior density $h(\theta)d\theta$ over the unknown parameter θ, we consider a coded message for encoding the value of $\underline{x}$. The message has two parts: a first part stating an estimated parameter value $\hat{\theta}$, encoded in some suitable code, and a second part which states $\underline{x}$ using a Huffman code based on the distribution $p(\underline{x} \mid \hat{\theta})$. The length of the first part, in units of $\log_2 e$ binary digits, is approximately $-\ln\{h(\theta)s(\hat{\theta})\}$, where $s(\hat{\theta})$ is roughly the expected estimation error of θ, and the length of the second part is $-\ln p(\underline{x} \mid \hat{\theta})$. The details may be found in Wallace [5]. That estimate is preferred which yields the shortest total message length.

If the data $\underline{x}$ comprises a sample of N independent values $\{y_i \mid i = 1, \ldots, N\}$ drawn from a distribution $f(y \mid \theta)$, the second part of the message

may be constructed as the combination of N segments, each encoding one of the observed y values with a code word of length $-\ln f(y_i \mid \hat{\theta})$.

In the outlier model, we are concerned with a sample of N independent value $\{y_i \mid i = 1, \ldots, N\}$ each of which may, with probability $(1 - \alpha)$, be drawn from a main process distribution $f(y \mid \theta)$ or, with probability α, be drawn from an error process distribution $g(y \mid \phi)$. Clearly, we could simply regard the data as being drawn from the mixed distribution

$$t(y \mid \alpha, \theta, \phi) = (1 - \alpha) f(y \mid \theta) + \alpha g(y \mid \phi)$$

and use the minimum message length criterion to estimate the unknown parameters α, θ, ϕ. The length of the message segment for an observed value y_i would then be $-\ln t(y_i \mid \hat{\alpha}, \hat{\theta}, \hat{\phi})$. However, such an approach, while yielding acceptable estimates for α, θ and ϕ, would appear to miss the objective which Freeman, and the others whose work he discusses, were aiming for. That is, the estimation process would appear to yield no useful statement about the identity of the outliers in the sample.

3.1 Message identifying outliers

A different and apparently more "natural" use of the minimum message length approach makes explicit use of the identification of each observation as either an outlier or a main-process value. The message encoding the data $\underline{x} = \{y_i \mid i = 1, \ldots, N\}$ again begins with a first part stating estimates $\hat{\alpha}, \hat{\theta}, \hat{\phi}$ of the parameters of the mixture model, and the second part again comprises a segment for each of the N observation values. However, the segment encoding a value y_i now comprises two phrases. The first states whether the observation is regarded as an outlier or not, and the second gives its value using a Huffman code based on either the outlier distribution or the main process distribution. If the observation is regarded as an outlier, the segment encoding it has total length

$$-\ln \hat{\alpha} - \ln g(y \mid \hat{\phi})$$

and if it is regarded as an uncorrupted observation, the segment has length

$$-\ln(1 - \hat{\alpha}) - \ln f(y \mid \hat{\theta})$$

It would appear that an observation should be classified as an outlier if

$$-\ln \hat{\alpha} - \ln g(y \mid \hat{\phi}) < -\ln(1 - \hat{\alpha}) - \ln f(y \mid \hat{\theta})$$

i.e., if

$$\alpha g(y \mid \hat{\phi}) > (1-\alpha) f(y \mid \hat{\theta})$$

as this assignment rule would appear to minimise the total message length, and, if the outlier/normal status of each observation is regarded as a free parameter, this rule selects the parameter of highest posterior probability.

The length of the segment encoding observation y is then

$$\min\{-\ln(\hat{\alpha} g(y \mid \hat{\phi})), -\ln((1-\hat{\alpha}) f(y \mid \hat{\theta}))\}.$$

However, the above outlier identification rule leads to inconsistent estimates of α, θ and ϕ. This may be easily seen in the case of the $N(\mu, \sigma)$ main process and uniform error process (section 2.1). The minimum message length for given data will be achieved when

(a) each observation y_i is classified as an outlier if $\hat{\alpha}/R > (1-\hat{\alpha}) N(y_i \mid \hat{\mu}, \hat{\sigma})$

(b) The estimate of the fraction of outliers, $\hat{\alpha}$, is just the fraction of observations classified as outliers

(c) The estimates $\hat{\mu}, \hat{\sigma}$ of the mean and standard deviation of the Normal process are the usual estimates calculated from those observations classified as not being outliers.

Suppose that the set of observations indeed consists of $(1-\alpha)N$ values drawn from $N(\mu, \sigma)$, and αN values drawn from a uniform distribution of range R. Let the values be classified into "normal" and "outlier" by test (a) above, using $\hat{\alpha} = \alpha, \hat{\mu} = \mu, \hat{\sigma} = \sigma$. Clearly, a number of members of $N(\mu, \sigma)$ will be misclassified as "outliers" by test (a), namely those members of largest deviation from the mean. Further, a number of members of the uniform distribution which happen to have values close to μ will be misclassified as "normal" values. Overall, the effect is that the standard deviation of those observations classified as "normal" by test (a) will be less than σ. When the estimate $\hat{\sigma}$ is formed as the maximum likelihood (or minimum message length) estimate based on those observations classified as "normal", we must expect to find $\hat{\sigma} < \sigma$. As the effect is independent of N, the estimation process is inconsistent.

Similar effects in a number of other estimation problems suggest that if an estimation process estimates some parameters to higher precision than the data warrants, and if the number of such parameters is proportional to sample size, inconsistent estimates may result.

3.2 Consistent identification of outliers

We now show that, without departing from a message form which asserts an outlier/normal status for each observation, the message length can be reduced further and the inconsistency removed.

Recall that the data comprises an ordered set of observations (y_1, y_2, ..., y_i, ..., y_N). Observation y_i could be coded as an outlier, with segment length $-\ln\{\hat{\alpha} g(y_i \mid \hat{\phi})\}$, or as a "normal" observation, with segment length $-\ln\{(1-\hat{\alpha}) f(y_i \mid \hat{\theta})\}$.

As earlier, define

$$t(y_i \mid \hat{\alpha}, \hat{\theta}, \hat{\phi}) = \hat{\alpha} g(y_i \mid \hat{\phi}) + (1-\hat{\alpha}) f(y_i \mid \hat{\theta}) = t_i \text{ (say)}$$

Also define

$$w_i t_i = (1-\hat{\alpha}) f(y_i \mid \hat{\theta}) \quad v_i = 1 - w_i,$$

where w_i and v_i are of course simply the posterior probabilities of "normal" and "outlier" status for observation y_i, assuming estimated parameter values $\hat{\alpha}$, $\hat{\theta}$, $\hat{\phi}$.

Now construct a Huffman code for the two-state distribution with probabilities (w_i, v_i). (Recall that a Huffman code over a discrete probability distribution $\{p_j, j = 1, 2, \ldots\}, \sum_j p_j = 1$, maps each index j onto a symbol string s_j of length $-\log p_j$. We take the base of logarithms, i.e. the size of the symbol alphabet, to be $e = 2.718\ldots$. The string s_j is called the code work for j. No code word is the prefix of another code word. Every sufficiently long string of symbols has a unique code word as a prefix. If the string is random, the probability that this code word is s_i is p_i.)

Suppose that the code segment for observation y_{i+1} has already been constructed as a string of symbols. Then this string must commence either with the Huffman code word for w_i or the code word for v_i. If the former is the case, encode y_i as a "normal" observation, using a message segment of length $-\ln\{(1-\hat{\alpha}) f(y_i \mid \hat{\theta})\} = -\ln\{w_i t_i\}$. If the latter, encode y_i as an "outlier" using a message segment of length $-\ln\{v_i t_i\}$.

Thus, each observation is classified as an outlier or a normal by a "pseudo-random" choice from its posterior probability distribution. We may call the selection process above "pseudo-random" because the symbol string encoding observation y_{i+1} is not expected to have any statistical or causal relationship to y_i. That is, the symbol string encoding y_{i+1} behaves in this context as a random string. Hence observation y_i has probability w_i of being classified as a "normal", and probability v_i of being classified as an "outlier".

Now consider the information available to someone receiving and decoding a message encoded as above. The first part of the message announces the estimates $\hat{\alpha}, \hat{\theta}, \hat{\phi}$ used in the encoding of observations.

The next segment will concern the first observation, y_1. It may have a part of length $-\ln\hat{\alpha}$ stating that y_1 is an outlier, followed by a part of length $-\ln g(y_1 \mid \hat{\phi})$ giving y_1 encoded as an outlier. Alternatively it may have a part of length $-\ln(1-\hat{\alpha})$ stating y_1 is "normal", and a part of length $-\ln f(y_1 \mid \hat{\theta})$. In either case, the receiver has all the information $(\hat{\alpha}, \hat{\theta}, \hat{\phi})$ needed to decode the message segment.

Having received the segment for y_1, the receiver can of course calculate t_1, w_1 and v_1 as defined above, and hence construct the Huffman code over the distribution (w_1, v_1). Since he knows whether y_1 was encoded as a "normal" or as an "outlier", the receiver can then deduce that the symbol string encoding y_2 has as prefix the appropriate word from this Huffman code.

It is therefore not necessary to include that word in the message segment for y_2, since the receiver can infer its value from the way y_1 was encoded.

The net length of the message segment encoding y_1 is thus:

If encoded as a "normal":

$$-\ln\{(1-\hat{\alpha})f(y_1 \mid \hat{\theta})\} + \ln w_1$$

where the first term is the message length to encode y_1 as a normal observation, and the second subtracts the length of the Huffman code word which can be omitted from the encoding of y_2. Now note that

$$(1-\hat{\alpha})f(y_1 \mid \hat{\theta}) = w_1 t_1$$

Hence, if y_1 is encoded as "normal", the net length of the message segment for y_1 is

$$-\ln\{w_1 t_1\} + \ln w_1 = -\ln t_1$$

The same net length is obtained if y_1 is encoded as an outlier. Hence, although the message form states whether each observation is an outlier or not, the message length behaves as if the total probability distribution $t(y \mid \hat{\alpha}, \hat{\theta}, \hat{\phi})$ had been used. Strictly, this device is applicable to the encoding of all observations save the last, y_N. However, we neglect this end effect.

3.3 Parameter estimation

Our objectives are to estimate the unknown parameters, and to make some inference about the number and identity of outliers. However, we are not really concerned with the exact symbol string which encodes the data, or with the outcomes of the pseudo-random observation classification described in section 3.2. Rather than estimate the parameters α, θ and σ from sets of

observations pseudo-randomly classified as outliers or normal, we let observation y_i contribute to the estimate of the parameters of "normal" observations with weight w_i, and contribute to the estimate of the parameters of "outliers" with weight v_i. Recall that w_i is the posterior probability that y_i is a normal observation. Similarly, the fraction of outliers, α, is estimated from the sum of weights $\sum_i v_i$.

It is easily shown that the estimates so obtained are consistent.

3.4 Minimum message length estimates

Here we consider the uniform error process model of section 2.1.

We have shown elsewhere (Wallace 1984) that, given a sample of size n from a univariate normal distribution $N(\mu, \sigma)$ with μ and $\lambda = \log \sigma$ having uniform priors, the minimum message length estimates of μ and σ are

$$\begin{aligned} \mu' &= \sum y_i/n \\ \sigma' &= \sum (y_i - \mu')^2/(n-1) \end{aligned}$$

The message segments encoding the estimates of μ and σ will optimally state values $\hat{\mu}, \hat{\sigma}$, which are values obtained by rounding μ', σ' to limited precision. The overall message length is minimised when the precision to which $\hat{\mu}$ and $\hat{\sigma}$ are rounded is such that

$$\begin{aligned} E(\mu' - \hat{\mu})^2 &= (\sigma')^2/n \\ E(\lambda' - \hat{\lambda})^2 &= 1/(2(n-1)) \end{aligned}$$

Better precision would increase the length of the specifications of $\hat{\mu}$ and $\hat{\sigma}$ without much reducing the length of the specification of the parameters.

Similarly, the fraction of outliers, α, is best estimated by

$$\alpha' = (\text{number of outliers} + 1/2)/(\text{sample size} + 1)$$

and the estimate stated as a value $\hat{\alpha}$ obtained by rounding α' to limited precision so that

$$E(\alpha' - \hat{\alpha})^2 = \alpha'(1 - \alpha')/N$$

where N is the total sample size.

Strictly, the above results apply to the estimation of α, μ and σ when the classification of each observation is known. The fact that the classification of the observations is unknown means that the estimates should ideally be

stated to slightly less precision than stated above, since the increase in message length arising from rounding the estimates can be in part reduced by reclassifying some of the observations. However, this effect will be ignored for the moment. Later, we will estimate its magnitude and show that it is small.

3.5 Observation code lengths

If the first part of the message encoding the observations asserts an outlier fraction $\hat{\alpha}$, the Huffman code word used to state that an observation is an outlier will have length $-\ln\hat{\alpha}$. However, the value $\hat{\alpha}$ is a rounded version of an estimate α', stated to precision such that

$$E(\hat{\alpha} - \alpha')^2 = \alpha'(1-\alpha')/N$$

Assuming $E(\hat{\alpha}-\alpha') = 0$, i.e., that the rounding is unbiased, the expected length of the code word identifying an outlier value is (to second order)

$$\begin{aligned} E(-\ln\hat{\alpha}) &= -\ln\alpha' + \frac{1}{2}(1/\alpha')^2 E(\hat{\alpha}-\alpha')^2 \\ &= -\ln\alpha' + (1-\alpha')/2N\alpha' \end{aligned}$$

Similarly, the expected length of the code word identifying a main-process value is

$$E(-\ln(1-\hat{\alpha})) = -\ln(1-\alpha') + \alpha'/2N(1-\alpha')$$

Rounding the estimates of μ and σ for the main (normal) process similarly affect the expected length of the code word stating an observed value y using a Huffman code based on the normal distribution.

If the Huffman code were based on the exact values μ', σ', we would obtain a code word length for observation y of

$$-\ln(\epsilon/\sigma'\sqrt{2\pi}) + (y-\mu')^2/2(\sigma')^2$$

The use of rounded values $\hat{\mu}$, $\hat{\sigma}$ gives an expected length (to second order) of

$$-\ln(\epsilon/\sigma'\sqrt{2\pi}) + \left(\frac{n}{n-1}\right)(y-\mu')^2/2(\sigma')^2 + 1/2n$$

where n is the number of "normal" process observations.

It is interesting that for each of the three parameters α, μ and σ, the effect of rounding on the expected total message length is to add 1/2 for each parameter.

Combining these results, we have that the expected length of the code word encoding an observed value y_i as an outlier (in the uniform error process model) is

$$-\ln \alpha' + (1-\alpha')/2N\alpha' + \ln R/\epsilon = -\ln(\alpha' g_i) \text{ (say)}$$

and the expected length of y_i is coded as a normal observation is

$$\begin{aligned} & -\ln(1-\alpha') + \alpha'/2N(1-\alpha') + \ln(\sigma'\sqrt{2\pi}/\epsilon) \\ & \quad + \left(\frac{n}{n-1}\right)(y-\mu')^2/2(\sigma')^2 + 1/2n \\ & = -\ln((1-\alpha')f_i) \text{ (say)} \end{aligned}$$

Both lengths are gross lengths, i.e., those obtained without use of the coding device described in section 3.2.

For each observation, define

$$\begin{aligned} t_i &= \alpha' g_i + (1-\alpha') f_i \\ w_i &= (1-\alpha') f_i / t_i \\ v_i &= \alpha' g_i / t_i \end{aligned}$$

4 The Uniform Error Process Model

4.1 Parameter estimation

For each observation, let t_i, f_i, g_i, v_i and w_i be defined as in section 3.5. Note that these quantities are functions of α', μ' and σ'. Then if α', μ' and σ' are minimum message length estimates, they satisfy

$$\begin{aligned} \mu' &= \sum_i w_i y_i / n \\ \sigma' &= \sqrt{\left(\sum_i w_i (y_i - \mu')^2 / (n-1)\right)} \\ n &= \sum_i w_i \\ 1-\alpha' &= (n + \frac{1}{2})/(N+1) \end{aligned}$$

These equations are easily solved by iteration. Initial guesses for α, μ and σ are used to calculate w_i $(i = 1, \ldots, N)$, then new values for α, μ and σ may be calculated. In the numerical examples discussed here, we use no-outlier estimates as the initial guesses for μ and σ, and 0.1 for α.

The iteration may collapse towards $\alpha = 0$ or $\alpha = 1$. However, this has happened only on data sets favouring an uncontaminated model or a uniform model respectively.

4.2 Message length

The total message length comprises the coded specifications of $\hat{\alpha}, \hat{\sigma}$ and $\hat{\mu}$, and the coded observations y_1, y_2,..., y_N. It is more convenient to regard the second parameter as $\lambda = \ln \sigma$, since λ has a uniform prior.

If the estimate θ' of a parameter θ is specified using a code in which the representable values of θ are separated by intervals of size δ, the rounded value $\hat{\theta}$ actually specified may differ from θ' by plus or minus $\delta/2$, and on average, we expect

$$E(\theta' - \hat{\theta})^2 = \delta^2/12$$

Further, the length of the Huffman code word for a value $\hat{\theta}$ will be approximately $-\ln(\delta h(\hat{\theta}))$ where $h(\theta)d\theta$ is the prior density of θ. In our model, we have assumed all parameters to have uniform priors, so the code word length for θ' can be written as $\ln(K/\delta)$ where K is the size of the prior range of θ. (Note that δ may be a function of θ' and/or estimates of other parameters).

Because our analysis of message length involves some approximations, it is possible that the interval δ may in some cases exceed the range K, which would yield a negative value for $\ln(K/\delta)$. We therefore calculate the code word length for specifying a parameter to precision δ in range K as $0.5 \ln(1 + K^2/\delta^2)$. The prior range size and values of $\delta^2/12$ for the three parameters are:

Parameter	Prior range size	$\delta^2/12$
α	0.5	$\alpha(1-\alpha)/N$
λ	$\ln(R/6\epsilon)$	$1/2(n-1)$
μ	R	σ^2/n

where N is the sample size, n is $(1-\alpha')N$, R is the known prior of the observations, and ϵ is the measurement or recording precision of the observations.

The length of the coded observations is $\sum_i \ln t_i$.

4.3 Revised precision estimate

The optimum precision with which a parameter estimate should be stated is related to the second derivative of the log likelihood function with respect to the estimate (Wallace 1984). In section 3.4, the optimum precisions are given on the assumption that each observation is unequivocally identified as an outlier or as a normal observation. The existence of some uncertainty about the identification has the effect of reducing the second derivative of the log likelihood function. Hence a small reduction in the total message length can be achieved by using slightly lower precisions than those stated in section 3.4. For instance, with the notation of section 3.5, the optimum precisions for stating α' is such that

$$E(\alpha' - \hat{\alpha})^2 \equiv 1/\sum_i (g_i - f_i)^2/t_i^2$$

rather than

$$E(\alpha' - \hat{\alpha})^2 = \alpha'(1 - \alpha')/N$$

For each data set, we calculate the reduction in message length due to reduced precision in α'. The calculation is more difficult for μ and σ, and has not been attempted. However, an approximate analysis shows that the effect for these parameters is expected to be smaller than for α.

References

[1] B. Abraham and G.E.P. Box. Linear models and spurious observations. *Journal of the Royal Statistical Society, Series C*, 27:131–138, 1978.

[2] G.E.P. Box and G.C. Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55:119–129, 1968.

[3] P.R. Freeman. On the number of outliers in data from a linear model. In J.M. Bernardo, M.H. DeGroot, A.F.M. Smith, and D.V. Lindley, editors, *Proceeding of the Valencia International Meeting on Bayesian Statistics, May 28- June 2, 1979, Valencia, Spain*, pages 349–365. Valencia University Press, 1980.

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

[5] C.S. Wallace. Estimation and inference by compact coding. Technical report TR 46, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, August 1984.