# 21

# Outliers and robustness

Probably everybody who has been involved in quantitative measurements has found himself in the following situation. You are trying to measure some quantity $\theta$ (which might be, for example, the right ascension of Sirius, the mass of a $\pi$-meson, the velocity of seismic waves at a depth of 100 km, the melting point of a new organic compound, the elasticity of consumer demand for apples, etc.). But the apparatus or the data taking procedure is always imperfect and so, having made $n$ independent measurements of $\theta$, you have $n$ different results $(x_1, \ldots, x_n)$. How are you to report what you now know about $\theta$? More specifically, what 'best' estimate should you announce, and what accuracy are you entitled to claim?

If these $n$ data values were closely clustered together making a reasonably smooth, single-peaked histogram, you would accept the solutions given in the previous chapters, and might feel that the problem of drawing conclusions from good data is not very difficult, even without any probability theory. But your data are not nicely clustered: one value, $x_j$, lies far away from the nice cluster made by the other $(n-1)$ values. How are you to deal with this outlier? What effect does it have on the conclusions that you entitled to draw about $\theta$?

We have seen in Chapters 4 and 5 how the appearance of astonishing, unexpected data may cause the resurrection of dead hypotheses; it appears that something like that may be at work here. In fact, any surprisingly ugly looking data with unexpected features might raise the question in your mind. Here we consider only the special case of outliers, leaving it as an exercise for the reader to work out the corresponding theory for other kinds of unexpected structure.

## 21.1 The experimenter's dilemma

The problem of outliers in data has been a topic of lively discussion since the 18th century, when it arose in astronomy, geodesy, calorimetry, and doubtless many other measurements. Let us interpret 'apparatus' broadly as meaning any method for acquiring data. On the philosophical side, two opposite views have been expressed repeatedly.

(I) Something must have gone wrong with the apparatus; the outlier is not part of the good data and we must throw it out to avoid getting erroneous conclusions.

(II) No! It is dishonest to throw away any part of your data merely because it was unexpected. That outlier may well be the most significant datum you have, and it must be taken into account in your data analysis, otherwise you are 'fudging' the data arbitrarily and you can make no pretense of scientific objectivity.

From these statements we can understand why the issue can arouse controversy that is very hard to resolve. Not only has an element of righteous ethical fervor crept in; it is also clear that both positions do contain elements of truth. How can they be reconciled?

On the pragmatic side, several arbitrary *ad hoc* recipes were invented (such as the Chauvenet criterion found in the astronomy textbooks of a century ago) to decide when to reject an outlier. It is curious that the arbitrary criteria for rejection (two standard deviations, etc.) seem to have taken no note of the following, which we think is essential for any rational approach to the problem.

Pondering the two statements above, we see that they reflect different *prior information* about the apparatus. This is the crucial factor – ignored in all the aforementioned criteria. To take it into account properly requires, not still more *ad hockeries*, but straightforward probability analysis.

Position (I) seems reasonable if we know that the means of gathering data is unreliable, and it is indeed likely to break down without warning and give erroneous data. If we already expect this, then the appearance of a wild outlier seems far more likely to be due to 'apparatus failure' than to a real effect.[1]

Position (II), on the other hand, is the reasonable stance for one who has absolute confidence in his apparatus: he is sure that his voltmeter always gives readings reliable to $\pm 0.5\%$, and could not be in error by 5%; or that his telescope was aimed within 10 arc seconds of the recorded direction, and cannot be off by 1 degree. Then the appearance of an outlier must be accepted as a significant event, however unexpected; to ignore it could be to miss an important discovery.

But (I) and (II) are extreme positions, and the real experimenter is almost always in some intermediate situation. Presumably, if he knew that his apparatus was very unreliable, he would prefer not to take data with it at all; but in a field like biology or economics one may be obliged to use whatever 'apparatus' Nature has provided. On the other hand, few scientists – even in the best laboratories of the National Bureau of Standards – are ever so confident of their apparatus that they will affirm dogmatically that it *cannot* go awry.

One would like to see the estimate in the form of an unequivocal statement like $(\theta)_{\text{est}} = A \pm B$, where $A, B$ are two definite numbers, presumably two functions of the data $D \equiv (x_1, \ldots, x_n)$. But what two functions? When the data are closely clustered together, it is surely a reasonable guess to take $A = \overline{x} \equiv n^{-1} \sum x_i$, the sample mean, as the estimated value. The observed scatter of the data values $x_i$ indicates the *reproducibility* of the measurements, and one might suppose that this indicates also their *accuracy*. If so, it might seem reasonable to calculate the mean-square deviation from the mean, or sample

[1] We saw another example of this phenomenon in Chapter 6, in the discussion following Eq. (6.97). There probability theory told us that, if large fluctuations in counting rate are to be expected as an artifact of the apparatus, then the observed fluctuations become less cogent for estimating changes in beam intensity.

variance: $s^2 \equiv n^{-1}\sum(x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2$ and choose $B = s$, the sample standard deviation. A more educated intuition familiar with elementary results of probability theory can improve on this by taking $B = s/\sqrt{n}$, and even if it is not shown to be optimal by any clearly stated criterion, the conclusion

$$(\theta)_{\text{est}} = \overline{x} \pm \frac{s}{\sqrt{n}} \tag{21.1}$$

would not be criticized as wildly unreasonable in location or accuracy.

**Exercise 21.1.** We have seen in Chapter 7 that, under rather general conditions, a Gaussian sampling distribution, $p(x|\theta, \sigma) \propto \exp\{-(x-\theta)^2/2\sigma^2\}$, leads us to take our point estimate of $\theta$ as the data mean $\overline{x}$. Show that any sampling distribution with a rounded top (that is, $p(x|\theta) = a_0 - a_1(x-\theta)^2 + \cdots$) will lead us to the same mean-value estimate in the limit where the data are closely clustered.

If the data are not closely clustered, the above discussion seems to consider only two possible actions: keep the outlier and give it full credence; or throw it out altogether. Is there a more defensible intermediate position?

## 21.2 Robustness

Another viewpoint toward such problems has arisen recently, represented by Huber (1981) and noted briefly in Chapter 6. It still seeks to deal with them by intuitive *ad hoc* procedures that do not take explicit note of prior information or probability theory; but it does look for an intermediate position. One seeks data analysis methods that are *robust*, which means insensitive to the exact sampling distribution of errors or, as it is often stated, insensitive to the model, or are *resistant*, meaning that large errors in a small proportion of the data do not greatly affect the conclusions.

The general idea, stated vaguely, is that theoretical 'optimality', in the sense we have used it in previous chapters, is not always a good criterion in practice. Often we are unsure of the correct model; then a method which is useful for a variety of different models, even though not optimal for any, may be preferable to one that is exactly suited to one specific model, but misleading for others.

Evidently, there could be some merit in this view; but the 'robustnik/exploratory' school of thought, represented by Tukey and Mosteller (for example, Tukey, 1977), carries this to the point of deprecating all considerations of optimality. However, attempts to define this position less vaguely become troublesome. Given data $D$ and any two estimators $f(D)$, $g(D)$ of some parameter, is there any explicit definition of the term 'robust' or 'resistant' which would make it meaningful to say that one is 'more robust' or 'more resistant' than the other? If so, then within a given set $S$ of feasible estimators there is necessarily

an 'optimally robust' one $a(D)$ and an 'optimally resistant' one $b(D)$, not necessarily unique.[2]

The point we make here is that if any intuitive property, such as robustness, is held to be desirable, then as soon as this property is defined with enough precision to allow transitive comparisons, an optimality principle follows inexorably. So one cannot consistently advocate any well-defined inference property and at the same time reject optimality principles.

Equally troublesome is the fact that robust/resistant qualities – however defined – must be bought at a price: the price of poorer performance when the model is correct. Indeed, this performance can be very much poorer; for it is clear that the most robust procedure of all – the 'optimal' procedure if one asks only for robustness – is the one that ignores the model, the data, and the prior information altogether, and considers all parameters zero, all hypotheses false! There must be, inevitably, some trade-off between the conflicting requirements of robustness and accuracy.[3] Advocates of robust/resistance have an obligation to show us just what trade-off, i.e. how much deterioration of performance, they are asking us to accept.

In estimating a location parameter, for example, the sample median $M$ is often cited as a more robust estimator than the sample mean. But here it is obvious that this 'robustness' is bought at the price of insensitivity to much of the relevant information in the data. Many different data sets all have the same median; the values above or below the sample median may be moved about arbitrarily without affecting the estimate. Yet those data values surely contain information highly relevant to the question being asked, and all this is lost. We would have thought that the whole purpose of data analysis is to extract all the information we can from the data.

Thus, while we agree that robust/resistant properties may be desirable in some cases, we think it important to emphasize their cost in performance. In the literature, *ad hoc* procedures have been advocated on no more grounds than that they are 'robust' or 'resistant', with no mention of the quality of the inference they deliver, much less any comparison of performance with alternative methods; yet alternative methods such as Bayesian ones are criticized on grounds of lack of robustness, without any supporting factual evidence.

Those who criticize Bayesian methods on such grounds are simply revealing that they do not understand how to use Bayesian methods. We wish to show that Bayesian data analysis, properly applied, automatically delivers robustness and resistance whenever those qualities are desirable; in fact, it does so in a way that agrees qualitatively with what previous *ad hoc* intuitive procedures have done, but improves on them quantitatively, because Bayesian methods never throw away relevant information. In other words, present robust methods are, like the other orthodox methods, only intuitive approximations to what a full Bayesian analysis gives automatically.

Indeed, this situation is very much like that noted in Section 5.6, where we discussed horse racing and weather forecasting. The new information – there called the data – was not known

[2] The term 'robust/resistant estimator' was coined by John W. Tukey; the present writer suggested to him that this must mean, literally, 'an estimator which resists being robust', but he denied it.

[3] In parameter estimation, the orthodox criterion of admissibility suffers from just the same defect; a procedure which ignores the data and always estimates $\theta^* = 5$ is admissible if the point $\theta = 5$ is in the parameter space; yet it is clear that almost any 'inadmissible' estimation rule would be superior to this 'admissible' one.

with certainty to be true, and we saw how Bayesian analysis takes that into account automatically. Here it is the model – part of the prior information – that is in doubt, but that makes no difference in principle because the 'data' and the 'prior information' are just two components of our total evidence which enter into probability theory in the same way. In the present case, a detailed Bayesian analysis reveals some very interesting and unexpected insight.

Reasoning that is unresponsive to changes in the model must be also in some way unresponsive to changes in the data. Is this what we really want? We think the answer is: Sometimes: that is, in problems where we are unsure of the model *but nevertheless sure of the meaning of the parameters in it.* But if we are sure of the model, then robust/resistance is the last thing we want in our data analysis procedure; it would waste data by throwing away cogent information.

Again, we must take explicit note of the prior information before the issue can be judged. As demonstrated below, if we choose our sampling distribution to represent properly our prior knowledge of the phenomenon that is generating the data, Bayesian analysis gives us automatically both robustness/resistant qualities when we are unsure of the model, and optimal performance when we are sure of it.

We may, however, make one concession. Intuitive devices of the Tukey genre can take into account, after a fashion, all kinds of special, one-time transitory contingencies that would be difficult – and not even desirable – to build into a model. A formal probability model ought to describe nontransitory situations that deserve more careful treatment and recording for future use. As a mathematician once put it: 'A *method* is a *device* that you use twice.'

But this one-time intuition is necessarily also a one-man operation, because it offers no rationale or criterion of optimality for what it does so that others could judge its suitability. If your one-time intuition differs from mine, then, without a normative theory of rational inference, we are at an impasse that cannot be resolved. But a logically consistent 'normative theory of rational inference' means necessarily (because of the theorems of Cox) a Bayesian theory.

Let us examine first the Bayesian treatment of the most common situation, in which the data are classified into only two categories: good and bad.

## 21.3 The two-model model

We have a 'good' sampling distribution

$$G(x|\theta) \tag{21.2}$$

with a parameter $\theta$ that we want to estimate. Data drawn urn-wise from $G(x|\theta)$ are called 'good' data. But there is also a 'bad' sampling distribution

$$B(x|\eta), \tag{21.3}$$

possibly containing an uninteresting parameter $\eta$. Data from $B(x|\eta)$ are called 'bad' data; they appear to be useless or worse for estimating $\theta$, since their probability of occurring has

nothing to do with $\theta$. Our data set consists of $n$ observations

$$D = (x_1, \ldots, x_n). \tag{21.4}$$

But the trouble is that some of these data are good and some are bad, and we do not know which is which (however, we may be able to make guesses: an obvious outlier – far out in the tails of $G(x|\theta)$ – or any datum in a region of $x$ where $G(x|\theta) \ll B(x|\eta)$ comes under suspicion of being bad).

In various real problems we may, however, have some prior information about the process that determines whether a given datum will be good or bad. Various probability assignments for the good/bad selection process may express that information. For example, we may define

$$q_i \equiv \begin{cases} 1 \text{ if the ith datum is good} \\ 0 \text{ if it is bad,} \end{cases} \tag{21.5}$$

and then assign joint prior probabilities

$$p(q_1 \cdots q_n|I) \tag{21.6}$$

to the $2^n$ conceivable sequences of good and bad.

## 21.4 Exchangeable selection

Consider the most common case, where our information about the good/bad selection process can be represented by assigning an exchangeable prior. That is, the probability of any sequence of $n$ good/bad observations depends only on the numbers $r$, $(n - r)$ of good and bad ones, respectively, and not on the particular trials at which they occur. Then the distribution (21.6) is invariant under permutations of the $q_i$, and by the de Finetti representation theorem (Chapter 18), it is determined by a single generating function $g(u)$:

$$p(q_1 \cdots q_n|I) = \int_0^1 \mathrm{d}u\, u^r\, (1-u)^{n-r}\, g(u). \tag{21.7}$$

It is much like flipping a coin with unknown bias where, instead of 'good' and 'bad', we say 'heads' and 'tails'. There is a parameter $u$ such that if $u$ were known we would say that any given datum $x$ may, with probability $u$, have come from the good distribution; or with probability $(1 - u)$ from the bad one. Thus, $u$ measures the 'purity' of our data; the closer to unity the better. But $u$ is unknown, and $g(u)$ may, for present purposes, be thought of as its prior probability density (as was, indeed, done already by Laplace; further technical details about this representation are given in Chapter 18). Thus, our sampling distribution may be written as a probability mixture of the good and bad distributions:

$$p(x|\theta, \eta, u) = uG(x|\theta) + (1-u)B(x|\eta)\,, \qquad 0 \le u \le 1. \tag{21.8}$$

This is just a particular form of the general parameter estimation model, in which $\theta$ is the parameter of interest, while $(\eta, u)$ are nuisance parameters; it requires no new principles beyond those expounded in Chapter 6.

Indeed, the model (21.8) contains the usual binary hypothesis testing problem as a special case, where it is known initially that all the observations are coming from $G$, or they are all coming from $B$, but we do not know which. That is, the prior density for $u$ is concentrated on the points $u = 0, u = 1$:

$$p(u|I) = p_0\,\delta(1-u) + p_1\,\delta(u), \tag{21.9}$$

where $p_0 = p(H_0|I)$, $p_1 = 1 - p_0 = p(H_1|I)$ are the prior probabilities for the two hypotheses:

$$\begin{aligned} H_0 &\equiv \text{all the data come from the distribution } G(x|\theta), \\ H_1 &\equiv \text{all the data come from the distribution } B(x|\eta). \end{aligned} \tag{21.10}$$

Because of their internal parameters, they are composite hypotheses; the Bayesian analysis of this case was noted briefly in Chapter 4. Of course, the logic of what we are doing here does not depend on value judgments like 'good' and 'bad'.

Now consider $u$ unknown and the problem to be that of estimating $\theta$. A full nontrivial Bayesian solution tends to become intricate, since Bayes' theorem relentlessly seeks out and exposes every factor that has the slightest relevance to the question being asked. But often much of that detail contributes little to the final conclusions sought (which might be only the first few moments, or percentiles, of a posterior distribution). Then we are in a position to seek useful approximate algorithms that are 'good enough' without losing essential information or wasting computation on nonessentials. Such rules might conceivably be ones that intuition had already suggested, but, because they are good mathematical approximations to the full optimal solution, they may also be far superior to any of the intuitive devices that were invented without taking any note of probability theory; it depends on how good that intuition was.

Our problem of outliers is a good example of these remarks. If the good sampling density $G(x|\theta)$ is very small for $|x| > 1$, while the bad one $B(x|\eta)$ has long tails extending to $|x| \gg 1$, then any datum $y$ for which $|y| > 1$ comes under suspicion of coming from the bad distribution, and intuitively one feels that we ought to 'hedge our bets' a little by giving it, in some sense, a little less credence in our estimate of $\theta$. Put more specifically, if the validity of a datum is suspect, then intuition suggests that our conclusions ought to be less sensitive to its exact value. But then we have just about stated the condition of robustness (only now, this reasoning gives it a rationale that was previously missing). As $|x| \to \infty$ it is practically certain to be bad, and intuition probably tells us that we ought to disregard it altogether.

Such intuitive judgments were long since noted by Tukey and others, leading to such devices as the 'redescending psi function', which achieve robust/resistant performance by modifying the data analysis algorithms in this way. These works typically either do not deign to note even the existence of Bayesian methods, or contain harsh criticism of

Bayesian methods, expressing a belief that they are not robust/resistant and that the intuitive algorithms are correcting this defect – but never offering any factual evidence in support of this position.

In the following we break decades of precedent *actually examining* a Bayesian calculation of outlier effects, so that one can see – perhaps for the first time – what Bayesianity has to say about the issue, and thus give that missing factual evidence.

## 21.5 The general Bayesian solution

Firstly, we give the Bayesian solution based on the model (21.8) in full generality, then we study some special cases. Let $p(\theta\eta u|I)$ be the joint prior density for the parameters. Their joint posterior density given the data $D$ is

$$f(\theta, \eta, u|DI) = Af(\theta, \eta, u|I)\, L(\theta, \eta, u), \tag{21.11}$$

where $A$ is a normalizing constant, and, from (21.8),

$$L(\theta, \eta, u) = \prod_{i=1}^{n}\Big[uG(x_i|\theta) + (1-u)B(x_i|\eta)\Big] \tag{21.12}$$

is their joint likelihood. The marginal posterior density for $\theta$ is

$$p(\theta|DI) = \int\int \mathrm{d}\eta\mathrm{d}u f(\theta, \eta, u|DI)\,. \tag{21.13}$$

To write (21.12) more explicitly, factor the prior density:

$$f(\theta, \eta, u|I) = h(\eta, u|\theta, I)f(\theta|I), \tag{21.14}$$

where $f(\theta|I)$ is the prior density for $\theta$, and $h(\eta, u|\theta, I)$ is the joint prior for $(\eta, u)$, given $\theta$. Then the marginal posterior density for $\theta$, which contains all the information that the data and the prior information have to give us about $\theta$, is

$$f(\theta|D, I) = \frac{f(\theta|I)\overline{L}(\theta)}{\int \mathrm{d}\theta f(\theta|I)\overline{L}(\theta)}, \tag{21.15}$$

where we have introduced the quasi-likelihood

$$\overline{L}(\theta) \equiv \int\int \mathrm{d}\eta\mathrm{du}\; L(\theta, \eta, u)h(\eta, u|\theta, I). \tag{21.16}$$

Inserting (21.12) into (21.16) and expanding, we have

$$\begin{aligned}\overline{L}(\theta) = \int\int \mathrm{d}\eta\mathrm{d}u\; h(\eta, u|\theta, I)\Big[&u^n L(\theta) + u^{n-1}(1-u)\sum_{j=1}^{n} B(x_j|\eta)L_j(\theta)\\ &+ u^{n-2}(1-u)^2\sum_{j<k} B(x_j|\eta)B(x_k|\eta)L_{jk}(\theta) + \cdots\\ &+ (1-u)^n B(x_1|\eta)\cdots B(x_n|\eta)\Big],\end{aligned} \tag{21.17}$$

in which

$$\begin{aligned} L(\theta) &\equiv \prod_{i=1}^{n} G(x_i|\theta) \\ L_j(\theta) &\equiv \prod_{i\neq j} G(x_i|\theta) \\ L_{jk}(\theta) &\equiv \prod_{i\neq j,k} G(x_i|\theta)\ldots \quad \text{etc.} \end{aligned} \tag{21.18}$$

are a sequence of likelihood functions for the good distribution in which we use all the data, all except the datum $x_j$, all except $x_j$ and $x_k$, and so on. To interpret the lengthy expression (21.17), note that the coefficient of $L(\theta)$,

$$\int_0^1 \mathrm{d}u \int \mathrm{d}\eta\, h(\eta, u|\theta, I)\, u^n = \int \mathrm{d}u\, u^n h(u|\theta, I), \tag{21.19}$$

is the probability, conditional on $\theta$ and the prior information, that all the data $\{x_1, \ldots, x_n\}$ are good. This is represented in the Laplace–de Finetti form (21.7) in which the generating function $g(u)$ is the prior density $h(u|\theta, I)$ for $u$, conditional on $\theta$. Of course, in most real problems this would be independent of $\theta$ (which is presumably some parameter referring to an entirely different context than $u$); but preserving generality for the time being will help to bring out some interesting points later.

Likewise, the coefficient of $L_j(\theta)$ in (21.17) is

$$\int \mathrm{d}u\, u^{n-1}(1-u) \int \mathrm{d}\eta B(x_j|\eta) h(\eta, u|\theta, I). \tag{21.20}$$

Now the factor

$$\mathrm{d}\eta \int \mathrm{d}u\, u^{n-1}(1-u) h(\eta, u|\theta I) \tag{21.21}$$

is the joint probability density, given $I$ and $\theta$, that any specified datum $x_j$ is bad, that the $(n-1)$ others are good, and that $\eta$ lies in $(\eta, \eta + \mathrm{d}\eta)$. Therefore the coefficient (21.20) is the probability, given $I$ and $\theta$, that the $j$th datum would be bad and would have the value $x_j$, and the other data would be good. Continuing in this way, we see that, to put in it words, our quasi-likelihood is:

$$\begin{aligned} \overline{L}(\theta) = {} & \text{prob(all the data are good)} \times \text{(likelihood using all the data)} \\ & + \sum_j \text{prob(only } x_j \text{ bad)} \times \text{(likelihood using all data except } x_j) \\ & + \sum_{j,k} \text{prob(only } x_j, x_k \text{ bad)} \times \text{(likelihood using all except } x_j, x_k) \\ & + \cdots \\ & + \sum_j \text{prob(only } x_j \text{ good)} \times \text{(likelihood using only the datum } x_j) \\ & + \text{prob(all the data are bad).} \end{aligned} \tag{21.22}$$

In shorter words: the quasi-likelihood $\overline{L}(\theta)$ is a weighted average of the likelihoods for the good distribution $G(x|\theta)$ resulting from every possible assumption about which data are good, and which are bad, weighted according to the prior probabilities of those assumptions. We see how every detail of our prior knowledge about how the data are being generated is captured in the Bayesian solution.

This result has such wide scope that it would require a large volume to examine all its implications and useful special cases. But let us note how the simplest ones compare with our intuition.

## 21.6 Pure outliers

Suppose the good distribution is concentrated in a finite interval

$$G(x|\theta) = 0, \qquad |x| > 1\,, \tag{21.23}$$

while the bad distribution is positive in a wider interval which includes this. Then any datum $x$ for which $|x| > 1$ is known with certainty to be an outlier, i.e. to be bad. If $|x| < 1$, we cannot tell with certainty whether it is good or bad. In this situation our intuition tells us quite forcefully: Any datum that is known to be bad is just not part of the data relevant to estimation of $\theta$ and we shouldn't be considering it at all. So just throw it out and base our estimate on the remaining data.

According to Bayes' theorem this is almost right. Suppose we find $x_j = 1.432$, $x_k = 2.176$, and all the other $x$'s less than unity. Then, scanning (21.24) it is seen that only one term will survive:

$$\overline{L}(\theta) = \int \mathrm{d}u \int \mathrm{d}\eta\; h(\eta, u|\theta, I) B(x_j|\eta) B(x_k|\eta) L_{jk}(\theta) = C_{jk}(\theta) L_{jk}(\theta). \tag{21.24}$$

As discussed above, the factor $C_{jk}$ is almost always independent of $\theta$, and since constant factors are irrelevant in a likelihood, our quasi-likelihood in (21.15) reduces to just the one obtained by throwing away the outliers, in agreement with that intuition.

But it is conceivable that in rare cases $C_{jk}(\theta)$ might, after all, depend on $\theta$; and Bayes' theorem tells us that such a circumstance would make a difference. Pondering this, we see that the result was to be expected if only we had thought more deeply. For if the probability of obtaining two outliers with values $x_j$, $x_k$ depends on $\theta$, then the fact that we got those particular outliers is in itself evidence relevant to inference about $\theta$.

Thus, even in this trivial case Bayes' theorem tells us something that unaided intuition did not see: even when some data are known to be outliers, their values might still, in principle, be relevant to estimation of $\theta$. This is an example of what we meant in saying that Bayes' theorem relentlessly seeks out and exposes every factor that has any relevance at all to the question being asked.

In the more usual situations, Bayes' theorem tells us that whenever any datum is known to be an outlier, then we should simply throw it out, if the probability of getting that particular

outlier is independent of $\theta$. For, quite generally, a datum $x_i$ can be known with certainty to be an outlier only if $G(x_i|\theta) = 0$ for all $\theta$; but in that case every likelihood in (21.24) that contains $x_i$ will be zero, and our posterior distribution for $\theta$ will be the same as if the datum $x_i$ had never been observed.

## 21.7 One receding datum

Now suppose the parameter of interest is a location parameter, and we have a sample of ten observations. But one datum $x_j$ moves away from the cluster of the others, eventually receding out 100 standard deviations of the good distribution $G$. How will our estimate of $\theta$ follow it? The answer depends on which model we specify.

Consider the usual model in which the sampling distribution is taken to be simply $G(x|\theta)$ with no mention of any other 'bad' distribution. If $G$ is Gaussian, $x \sim N(\theta, \sigma)$, and our prior for $\theta$ is wide (say $> 1000\,\sigma$), then the Bayesian estimate for quadratic loss function will remain equal to the sample average, and our far-out datum will pull the estimate about ten standard deviations away from the average indicated by the other nine data values. This is presumably the reason why Bayesian methods are sometimes charged with failure to be robust/resistant.

However, that is the result only for the assumed model, which in effect proclaims dogmatically: I know in advance that $u = 1$; all the data will come from $G$, and I am so certain of this that no evidence from the data could change my mind. If one actually had this much prior knowledge, then that far-out datum would be highly significant; to reject it as an 'outlier' would be to ignore cogent evidence, perhaps the most cogent piece of evidence that the data provide. Indeed, it is a platitude that important scientific discoveries have resulted from an experimenter having that much confidence in his apparatus, so that surprising new data were believed; and not merely rejected as 'accidental' outliers.

If, nevertheless, our intuition tells us with overwhelming force that the deviant datum *should* be thrown out, then it must be that we do not really believe that $u = 1$ strongly enough to adhere to it in the face of the evidence of the surprising datum. A Bayesian may correct this by use of the more realistic model (21.8). Then the proper criticism of the first procedure is not of Bayesian *methods*, but rather of the saddling of Bayesian methodology with an inflexible, dogmatic model which denies the possibility of outliers. We saw in Section 4.4 on multiple hypothesis testing just how much difference it can make when we permit the robot to become skeptical about an overly simple model.

Bayesian methods have inherent in them all the desirable robust/resistant qualities, and they will exhibit these qualities automatically, whenever they *are* desirable – if a sufficiently flexible model permits them to do so. But neither Bayesian nor any other methods can give sensible results if we put absurd restrictions on them. There is a moral in this, extending to all of probability theory. In other areas of applied mathematics, failure to notice some feature (like the possibility of the bad distribution $B$) means only that it will not be taken into account. In probability theory, failure to notice some feature may be tantamount to making irrational assumptions about it.

Then why is it that Bayesian methods have been criticized more than orthodox ones on this issue? For the same reason that city B may appear in the statistics to have a higher crime rate than city A, when the fact is that city B has a lower crime rate, but more efficient means for detecting crime. Errors undetected are errors uncriticized.

Like any other problem in this field, this can be further generalized and extended endlessly, to a three-model model, putting parameters in (21.6), etc. But our model is already general enough to include both the problem of outliers and conventional hypothesis testing theory; and a great deal can be learned from a few of its simple special cases.