DEALING WITH DUFF DATA

D.S. SIVIA
Rutherford Appleton Laboratory
Chilton, Oxon, OX11 0QX, United Kingdom

Abstract. The classical problem of statistical outliers does not really exist in the Bayesian context because its detection is indicative of an inadequate analysis model. Nevertheless, we may sometimes want to deal with data where quirky things occasionally happen. We describe two elementary probability formulations for handling this case and show, with the aid of a simple example, how they lead to robust estimation.

Key words: Statistical Outliers, Robust Bayesian Estimation

1. Introduction

Fitting a straight line to a pertinent set of data, such as those shown in Fig. 1(a), is one of the most common problems in data analysis. By and large, it reduces to just an easy exercise in the use of least-squares. Sometimes, however, there are unexpected glitches in the measurements! While these may be due to real physical phenomena, and therefore indicate that our simple linear model is in need of revision, they may also be due to nothing more than an intermittent fault in a detector. In this paper, we are concerned only with the second case. Although the so-called outliers are often readily picked out, and ignored, by our eyes and brain, they can severely skew the results of a least-squares fit; this is illustrated in Fig. 1(b). How can we guard ourselves against this situation in an automatic way?

Two elementary probability formulations for handling outliers are described separately in Sections 2 and 3. Their use is demonstrated with the aid of a simple example; namely, the estimation of the mean of data subject to Gaussian noise. The behaviour of the alternative models presented is contrasted and discussed further in the conclusions of Section 4.

2. A Conservative Formulation

Suppose we are given estimates of a certain quantity by various laboratories; for example, the age of a scroll found on an archaeological dig, based on radioactive dating. Since the different institutions will have equipment of varying sophistication, we will ask that they also provide an error-bar (1σ) to indicate the measure of the reliability of their results. Our ultimate task will be to use this information from the N laboratories, which we will call data $\{D_k, \sigma_k\}$, for $k = 1, 2, \ldots, N$, and obtain a best estimate and error-bar for the quantity of interest μ.

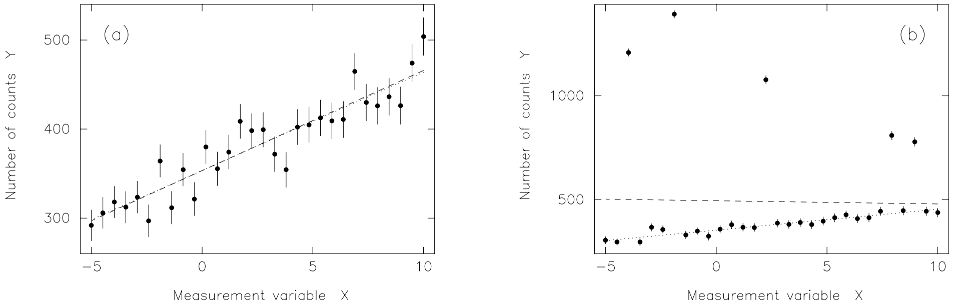

Figure 1: Fitting a straight line to pertinent set of data, which may occasionally be corrupted by a large rogue signal. The dashed line shows the traditional least-squares solution, whereas the dotted line is the result of a simple Bayesian analysis that allows for the possibility of outliers.

2.1. GAUSSIAN DATUM WITH UNCERTAIN ERROR-BAR

The delivery of (just) a datum D and an error-bar σ is really a shorthand way of saying that the appropriate likelihood function is, or can be approximated by, a Gaussian *pdf*

$$\Pr(D|\mu,\sigma) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\,\exp\left(-\frac{|\mu-D|^2}{2\sigma^2}\right) \quad . \tag{1}$$

The quoted error-bar will invariably be based upon ideal conditions, where everything is under control. Given that quirky things can sometimes happen, we could adopt a very conservative attitude and say that the experimentalist has only provided us with a lower-bound on the uncertainty σ_{min}. With a suitably pessimistic upper-bound σ_{max}, a Jeffreys' *pdf* can be assigned for the (unknown) error-bar

$$\Pr(\sigma|\sigma_{min},\sigma_{max}) \;=\; \frac{1}{\ln(\sigma_{max}/\sigma_{min})}\times\frac{1}{\sigma} \quad , \tag{2}$$

for $\sigma_{min} \le \sigma \le \sigma_{max}$, and zero otherwise. The marginal likelihood for the data can then be expressed in terms of the two *pdfs* above as follows:

$$\Pr(D|\mu,\sigma_{min},\sigma_{max}) \;=\; \int_0^\infty \Pr(D|\mu,\sigma)\;\Pr(\sigma|\sigma_{min},\sigma_{max})\,d\sigma \quad , \tag{3}$$

where we have used the product rule to expand the joint *pdf* for D and σ on the right-hand side, and dropped some unnecessary condition statements. Substituting from Eqns. (1) and (2) into (3), and carrying out integral, we find that $\Pr(D|\mu,\sigma_{min},\sigma_{max})$ is given by

$$\frac{1}{2\,|\mu-D|\,\ln(\sigma_{max}/\sigma_{min})}\left[erf\left(\frac{|\mu-D|}{\sigma_{min}\sqrt{2}}\right) - erf\left(\frac{|\mu-D|}{\sigma_{max}\sqrt{2}}\right)\right] \quad . \tag{4}$$

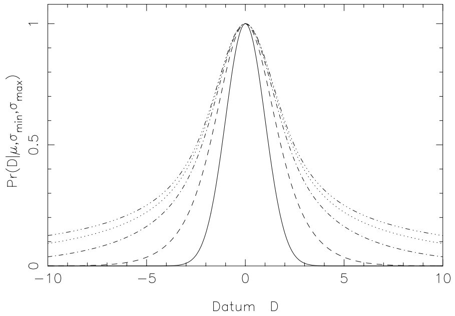

Figure 2: The marginal likelihood function of Eqn. (4), with $\mu = 0$ and $\sigma_{min} = 1$, vertically scaled to unit amplitude. The solid line is for $\sigma_{max} = 1$ and the others, in order of increasing widths and tails, are for $\sigma_{max} = 3$, 9, 27 and $\rightarrow \infty$.

The marginal likelihood function of Eqn. (4) is shown in Fig. 2. It has the pleasing property that the width of the central bump is principally controlled by σ_{min}, whereas the decay of the tails is determined by σ_{max}. When the upper and lower bounds become coincident, Eqn. (4) reduces to Eqn. (1) with $\sigma = \sigma_{min}$. At the opposite extreme, as $\sigma_{max} \rightarrow \infty$, Eqn. (4) simplifies to

$$\Pr(D|\mu,\sigma \geq \sigma_{min}) \;\propto\; \frac{1}{|\mu - D|}\, erf\left(\frac{|\mu - D|}{\sigma_{min}\sqrt{2}}\right) \quad . \tag{5}$$

In this limit, we find that the central bump is about twice as wide as the equivalent Gaussian of Eqn. (1) with $\sigma = \sigma_{min}$ and that the tails of the *pdf* decay like a Jeffreys' prior.

2.2. THE INFERENCE

If we assume that our N data $\{D_k\}$ from the various laboratories are independent, and we treat the quoted error-bars as being the most optimistic estimates of their reliability $\{\sigma_{0k}\}$, then Bayes' theorem tells us that our inference about the quantity of interest μ is given by

$$\Pr(\mu|\{D_k,\sigma_k \geq \sigma_{0k}\}) \;\propto\; \Pr(\mu) \times \prod_{k=1}^{N} \Pr(D_k|\mu,\sigma_k \geq \sigma_{0k}) \quad . \tag{6}$$

Computationally, of course, it is much better to work with the logarithm of the posterior *pdf* . Thus, with a uniform prior for μ, we have

$$\ln\left[\Pr(\mu|\{D_k,\sigma_k \geq \sigma_{0k}\})\right] \;=\; Const + \sum_{k=1}^{N} \ln\left[\frac{1}{|\mu - D_k|}\, erf\left(\frac{|\mu - D_k|}{\sigma_{0k}\sqrt{2}}\right)\right] \;. \tag{7}$$

This can be contrasted with the equivalent form for a traditional least-squares analysis, where we believe all the quoted error-bars without question:

$$\ln\left[\Pr(\mu|\{D_k,\sigma_k = \sigma_{0k}\})\right] \;=\; Const - \frac{1}{2}\sum_{k=1}^{N} \frac{|\mu - D_k|^2}{\sigma_{0k}^2} \quad . \tag{8}$$

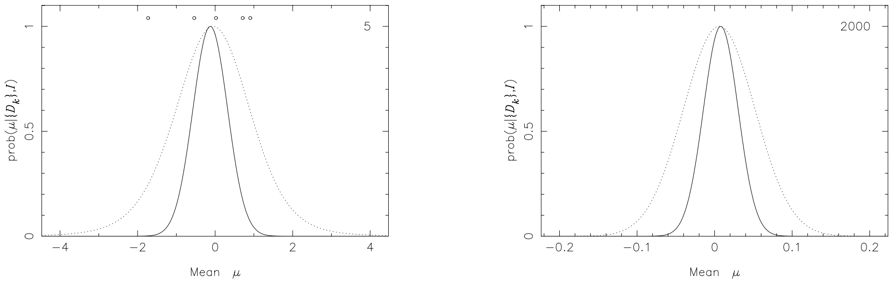

Figure 3: The posterior *pdf* for the mean μ, after 5 and 2000 data, when there are no outliers. The solid line is for the least-squares assumptions of Eqn. (8) and the dotted line is for the conservative formulation of Eqn. (7).

Let us illustrate the use of Eqns. (7) and (8) with the aid of a few computer simulations. First of all, consider the case when there are no outliers. Figure 3 shows the resulting posterior *pdfs* for μ from the analysis of just 5 data, and 2000 data, where the quoted (and actual) error-bar on each point is of unit magnitude. We see that the wide tails from the conservative likelihood function of Eqn. (5) are lost very quickly, and (reassuringly) that the best estimates of μ are almost the same. We do pay a penalty for our pessimism, however, in that the error-bar for the inferred parameter is about twice as large as that from the ordinary least-squares likelihood function.

Now let's repeat the exercise when there is one (20σ) outlier; the results are shown in Fig. (4). After just two data, the posterior *pdf* from Eqn. (7) is bimodal: the points are sufficiently discordant that one of them is probably quirky, but at this stage we can't say which. By contrast, the traditional analysis of Eqn. (8) puts the best estimate of μ half-way between the measurements, and implies that both data are highly unlikely! With the measurement of a third datum, we start to become fairly confident that it is the point on the far right that is duff; and Eqn. (7) starts to ignore it rapidly (or so it appears). As we would expect, the skewing effect of the outlier on the ordinary least-squares analysis becomes less pronounced as it is diluted with an increasing number of *bone fide* data.

3. The "Good-and-Bad Data" Formulation

Although the automatic protection against quirky measurements afforded by the conservative formulation above is very desirable, we have seen that the resulting inference can be unduly pessimistic when nothing has actually gone wrong. Rather than allowing for the possibility of a whole continuum of bad scenarios, it might be better to just consider the occasional small chance of a serious error. As such, a suitable alternative for the assignment of Eqn. (2) would be

$$\Pr(\sigma|\sigma_0,\beta,\gamma) \;=\; \beta\,\delta(\sigma-\gamma\sigma_0) \,+\, (1-\beta)\,\delta(\sigma-\sigma_0) \quad , \tag{9}$$

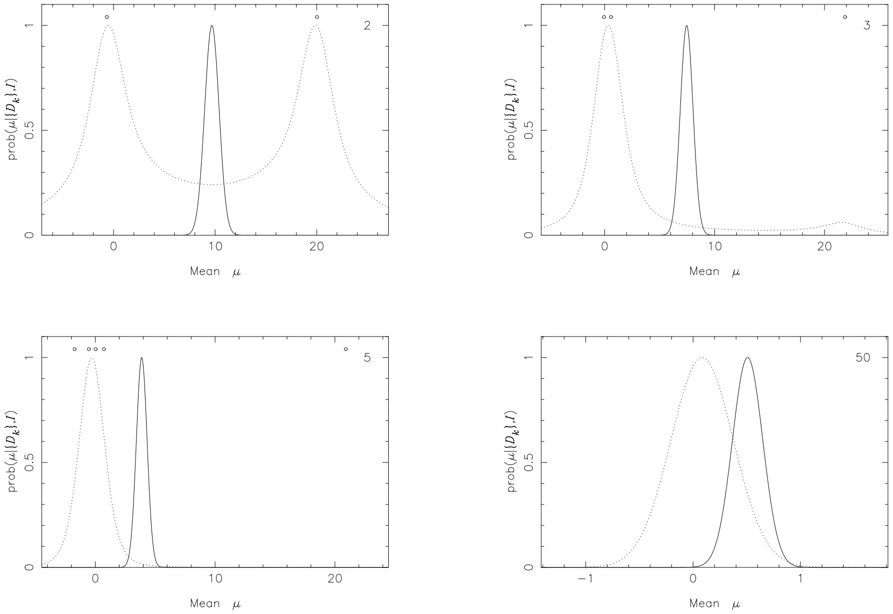

Figure 4: The posterior *pdf* for the mean μ, after 2, 3, 5 and 50 data, when there is one outlier. The solid line is for the least-squares assumptions of Eqn. (8) and the dotted line is for the conservative formulation of Eqn. (7).

where $0 \leq \beta << 1$ and $\gamma >> 1$. Substituting from Eqns. (1) and (9) into the appropriate analogue of Eqn. (3), we find that the marginal likelihood for a datum $\Pr(D|\sigma_0, \beta, \gamma)$ is now given by

$$\frac{1}{\sigma_0\sqrt{2\pi}}\left[\frac{\beta}{\gamma}\,\exp\left(-\frac{|\mu-D|^2}{2\,\gamma^2\sigma_0^2}\right)\;+\;(1-\beta)\exp\left(-\frac{|\mu-D|^2}{2\sigma_0^2}\right)\right]\;. \qquad (10)$$

Thus, marginalising out β and γ as nuisance parameters, and using Bayes' theorem, the posterior *pdf* for μ becomes

$$\Pr(\mu|\{D_k,\sigma_{0k}\})\;\propto\;\Pr(\mu)\int\!\!\int\Pr(\beta,\gamma)\;\Pr(\{D_k\}|\mu,\{\sigma_{0k}\},\beta,\gamma)\;d\beta\,d\gamma\;, \qquad (11)$$

where the likelihood of the data is given by the product of N terms like Eqn. (10).

Figure 5 shows the results of carrying out such an analysis for the situation corresponding to the case of Fig. 4. The double integral of Eqn. (10) was not carried out explicitly, but was approximated by expanding the logarithm of the joint posterior *pdf* for μ, β and γ in a quadratic Taylor series about its maximum (with uniform priors). This is obviously an inadequate procedure for the bimodal case of two discordant data but, as the comparison with the posterior *pdf* of μ

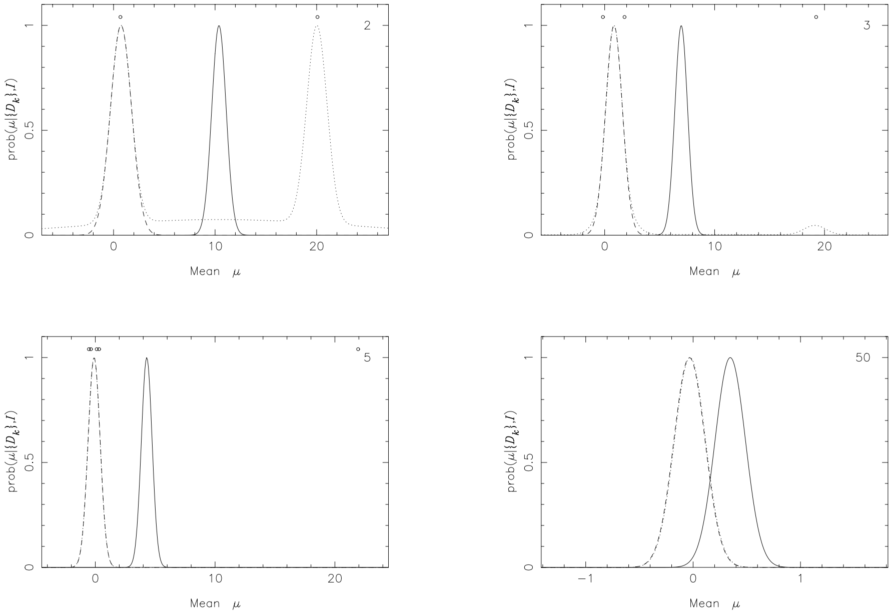

Figure 5: The posterior *pdf* for the mean μ, after 2, 3, 5 and 50 data, when there is one outlier. The solid line is for the least-squares assumptions of Eqn. (8) and the dashed line is for the "good-and-bad data" formulation, where the marginalisation of Eqn. (11) has been carried out in the quadratic approximation; the dotted line is the posterior *pdf* conditional on the best-fit values of β and γ.

conditional on the best-fit values of β and γ suggests, it soon becomes quite a good one. Thus, it would seem that we have achieved our goal of robust estimation in the presence of an outlier without paying the penalty of over-pessimism incurred in the conservative formulation of the previous section.

The "good-and-bad" data formulation above was, in fact, put forward almost thirty years ago by Box and Tiao (1968). They weren't so keen on marginalising out β and σ, however, preferring instead to just asses the sensitivity of the results with respect to these nuisance parameters. This seems a little odd given that they had no hesitation in marginalising out an unknown σ_0, assuming that the good measurement error-bar was the same for all the data.

4. Discussion and Conclusions

We have considered two elementary probability formulations for dealing with the occasional quirky data. Their principal assumptions can be stated, respectively, as follows: (1) the error-bars could (only) be worse than those quoted; and (2) the measurements can either be good or bad. Both models treat every datum on an equal footing, so that we are not "throwing out" any points. They are computationally more expensive than traditional least-squares, because they involve the evaluation of logarithms, exponentials and error functions, but not unduly so given the speed of modern computers. For linear problems, such as the estimation of the mean or the fitting of a straight line, they are also algorithmically more demanding because the likelihood functions are no longer unimodal. These numerical drawbacks are more than offset, however, in terms of the great robustness they bring to the analysis. The flip side of this strength can also be their weakness, of course, since we may be tempted to be lazy and not think about the possible causes (and cures) of the unexpected data.

The conservative formulation of Section 2 is more general than the "good-and-bad" model of Section 3, since every datum is associated with a continuum of uncertainty rather than all being tied to the same degree of potential error; the former also has the advantage of having fewer (if any) nuisance parameters, particularly in the $\sigma_{max} \to \infty$ limit of Eqn. (5). There is a price to be paid for the pessimism of the first model in that the error-bars on the inferred parameters are usually larger than they need be (by a factor of about two). For example, the inferred gradient and intercept of the straight line $Y = mX + c$ from the data of Fig. 1(b) is $m = 10.0 \pm 1.7$ and $c = 353.5 \pm 8.6$ with the likelihood function of Eqn. (5) (where the expected datum μ is now equal to Y), whereas $m = 9.7 \pm 0.9$ and $c = 355.1 \pm 4.3$ with the likelihood function of Eqn. (10) (with β and γ marginalised out in the quadratic approximation). By way of comparison, the least-squares formulation of Eqn. (8) yields $m = -1.5 \pm 0.8$ and $c = 495.1 \pm 3.8$; the (true) values used for the simulation were $m = 10$ and $c = 350$.

There is a practical footnote that should be added to the above example. In an attempt to circumvent the problems associated with the multimodal likelihood functions for the robust formulations, we used the solution of the (linear) least-squares analysis for the gradient and intercept as an initial guess for the optimisation of the *pdfs* in Eqns. (7) and (11) with a *simplex* algorithm (Nelder and Mead, 1965). While this worked well for the conservative formulation of Section 2, even under relatively extreme conditions, the approach was far less successful for the "good-and-bad" data model.

In conclusion, we see that the problem of outliers can easily be dealt with by probability theory. Both of the elementary formulations presented here provide for much more robust estimation than does a naive least-squares analysis. The conservative approach seems to offer some practical benefits over the two-valued alternative, but the latter avoids the problem of over-pessimism.

References

Box, G.E.P. and Tiao, G.C.: 1968, 'A Bayesian approach to some outlier problems', *Biometrika* **55**, 119-129.

Nelder, J.A. and Mead, R.: 1965, 'A simplex method for function minimization', *Compuer J.* **7**, 308-313.