Bayesian Data Analysis: Straight-line fitting

Stephen F. Gull
Cavendish Laboratory,
Madingley Road,
Cambridge CB3 0HE, U.K.

Abstract

A Bayesian solution is presented to the problem of straight-line fitting when both variables x and y are subject to error. The solution, which is fully symmetric with respect to x and y, contains a very surprising feature: it requires a informative prior for the distribution of sample positions. An uninformative prior leads to a bias in the estimated slope.

1. Introduction

An apparently simple data analysis problem that often arises is that of fitting a straight line to measurements of two quantities (x,y). Suppose that we have N such measurements $\{x_i, y_i\}$ and that they are each subject to independent Gaussian errors (σ_x, σ_y) (for the moment assumed known). Our task is to find a relationship of the form:

$$\hat{y} = a\hat{x} + b\ , \quad \text{where} \quad x_i = \hat{x}_i \pm \sigma_x\ ; \quad y_i = \hat{y}_i \pm \sigma_y\ .$$

Note that we are considering a problem in which there is an underlying exact relation for the (unknown) quantities $(\hat{x},\hat{y})$ and that the measurements $\{x_i, y_i\}$ are subject to error. A related, but different problem is the case where there is very little experimental error, but the measurements refer to different objects with an intrinsic spread of (a,b) values. An example of this type would be the height and weight distributions of a set of students. Problems of this latter type are known as regression and, although they are clearly interesting, they are not the type of problem considered here.

2. The joint distribution

We now begin a careful Bayesian analysis of the straight-line-fitting problem, and will derive the joint probability distribution of the data and the parameters. For the case where both variables are subject to error we cannot avoid introducing the "hidden variables" $\{\hat{x}_i\}$ (and $\hat{y}_i = a\hat{x}_i + b$), which are a set of N nuisance parameters. We need these before we can even write down the likelihood function:

$$\mathrm{pr}(x,y\,|\,\hat{x},a,b,\sigma_x,\sigma_y) = (4\pi^2\sigma_x^2\sigma_y^2)^{-N/2} \exp -\Big(\sum_i (x_i-\hat{x}_i)^2/\sigma_x^2 + \sum_i (y_i-\hat{y}_i)^2/\sigma_y^2\Big)/2\ .$$

When we have completed the assignment of the joint p.d.f., we will integrate the nuisance parameters out of the posterior distribution.

To make further progress we need to refine our thinking about the nature of the problem. The variables x and y may not be of the same type, but it is

J. Skilling (ed.), Maximum Entropy and Bayesian Methods, 511–518.

usually as natural to plot x against y as y against x. We must therefore treat x and y in a symmetrical fashion. Recognising this, a sensible way to treat the problem is to suppose that there are separate scalings and offsets of the x and y variables that map them both into a given interval, for example (-1,+1). We define new scaled variables $\hat{X}$ and $\hat{Y}$, which are related as follows:

$$\hat{X} = (x-x_0) / R_x , \qquad \hat{Y} = (y-y_0) / R_y ,$$
$$a = R_y / R_x , \qquad b = y_0 - a\, x_0 .$$

This procedure closely follows what we do in practice when plotting points on graph-paper or on a display screen - we ascertain the range of both variables and plot accordingly. In this way our relationship takes the simple form:

$$\hat{X} = \pm\hat{Y} .$$

In what follows we will derive formulae appropriate for the positive sign. In order to cope with this ambiguity of the sign of the slope, we should, strictly, always compute both cases, and compare their posterior probabilities. In many cases it will be obvious which case is better. Two other extreme cases that might also be worth considering separately are the degenerate cases $\hat{X} = 0$ and $\hat{Y} = 0$.

At this point the reader may be forgiven for thinking that we have gone backwards; we started with two variables (a,b) and have replaced them by four (x_0,y_0,R_x,R_y). However, we will find that there are great advantages to be had from this more symmetrical formulation of the problem.

We start our development with the prior for $pr(x_0,y_0,R_x,R_y)$. Because the units of x and y are related to R_x and R_y, we can reasonably take R_x and R_y to be scale parameters, and the offsets x_0 and y_0 to be location parameters. We therefore take the prior as uniform in $\log R_x$, $\log R_y$, x_0 and y_0:

$$pr(x_0,y_0,R_x,R_y)\, dx_0\, dy_0\, dR_x\, dR_y \;\propto\; dx_0\, dy_0\, d(\log R_x)\, d(\log R_y) ,$$
$$\propto\; d(\log a)\, d(ba^{-1/2})\, d(\log R)\, d(x_0 a^{1/2}) ,$$

where $R = (R_x\, R_y)^{1/2}$ is a symmetric range parameter. We should also, for completeness, specify some sensible ranges for these parameters. In fact, the posterior distribution is normalisable over infinite ranges of x_0 and y_0 when there are more than two samples, and we shall return to the question of what (a_{min}, a_{max}) and (R_{min}, R_{max}) should be later.

The final expression for the prior in terms of our original variables a and b (and the range and offset of $\hat{X}$) is very instructive. In particular, the $(da\; db\; a^{-3/2})$ part of this prior can be compared to that obtained by Jaynes (1967) for an allegedly similar problem: he finds $(da\; db\; (a^2 + 1)^{-3/2})$, using a transformation group argument. Whilst I am always very wary of disagreeing with Ed, I note the following points.

1) the functional relationship derived by Jaynes:

$$a^3 f(a,b) = f(1/a, -b/a) \quad \text{(in the present notation)},$$

is satisfied by both candidate priors... and many others - this functional relationship is too weak to determine the prior uniquely.

2) The $(da\, db\, (a^2 + 1)^{-3/2})$ prior is the correct answer to a different problem. Suppose that (as in the Bertrand problem (Jaynes 1973)) a straw is thrown at random onto a piece of square graph paper. Imagine then that this straw defines a regression line. The rotational symmetry inherent in this second problem is now sufficient to determine the prior uniquely, but is not relevant to the straight-line fitting problem, even when the variables are of the same type. Indeed, for the particular example given by Jaynes, that of the daily temperature variations at New York and Boston (which is actually a regression problem, rather than line-fitting), it is rather difficult to understand why we should want to consider rotation of one axis onto the other.

For these reasons I believe that the hypothesis space defined here by (x_0,y_0,R_x,R_y) is more useful for the line-fitting problem than that implied by Jaynes' prior, but it was his prior (and the obvious non-uniqueness of the *functional equation*) that stimulated my interest in this problem.

We now arrive at a very interesting stage. The joint p.d.f. can be written as:

$$\mathrm{pr}(x,y,\hat{x},x_0,y_0,R_x,R_y) = \mathrm{pr}(x_0,y_0,R_x,R_y)\ \mathrm{pr}(\hat{x}|x_0,y_0,R_x,R_y)\ \mathrm{pr}(x,y|\hat{x},\sigma_x,\sigma_y),$$

where irrelevant conditionals have been dropped. Our remaining problem is the prior $\mathrm{pr}(\hat{x}|x_0,y_0,R_x,R_y)$. At first sight it may seem peculiar that our answers are going to depend on our prior knowledge of the *distribution of the "true"* $\hat{x}$, and I imagine that strong objections will be voiced from some directions. However, my intuition about this matter has now been educated a little, and it is from this part of the prior that the most unexpected (and pleasing) feature of the Bayesian solution emerges. Let us take for this prior the independent Gaussian form:

$$\mathrm{pr}(\hat{x}|x_0,y_0,R_x,R_y) \;=\; (2\pi R_x^2)^{-N/2} \exp -\sum_i((\hat{x}_i-x_0)^2/R_x^2)/2 \ .$$

This form can be derived by invoking the principle of Maximum Entropy, using constraints on $\langle\Sigma(\hat{x}-x_0)^2\rangle = N R_x^2$ and $\langle\Sigma\hat{x}\rangle = N x_0$. Note also that, because of the definition of the parameters, this prior is fully symmetric with respect to x and y. Perhaps the choice of a Gaussian prior for $\hat{x}$ and $\hat{y}$ does not really correspond to our best intuition for this problem; we might prefer to consider the points spread evenly over the graph paper. However, we shall continue to use a Gaussian prior, because it makes the algebra tractable, if not actually pleasant.

We now write down the full symmetric joint p.d.f.:

$$\mathrm{pr}(x,y,\hat{x},x_0,y_0,\log R_x,\log R_y|\sigma_x,\sigma_y) \;=$$

$$(8\pi^3R_x^2\sigma_x^2\sigma_y^2)^{-N/2} \exp - \sum_i ((\hat{x}_i-x_0)^2/R_x^2+(x_i-\hat{x}_i)^2/\sigma_x^2+(y_i-\hat{y}_i)^2/\sigma_y^2)/2 \ ,$$

which, using Bayes' theorem, is then proportional to the posterior distribution $\mathrm{pr}(\hat{x},x_0,y_0,\log R_x,\log R_y|x,y,\sigma_x,\sigma_y)$.

3. Estimation of parameters

At this point we draw a polite veil over the algebra as we integrate the nuisance parameters $\hat{X}$ out of the problem. We note that the $\hat{X}$ have independent

Gaussian distributions which lead in turn to Gaussian distributions for x_0 and y_0:

$$\mathrm{pr}(x_0,y_0,\log R_x,\log R_y\,|\,x,y,\sigma_x,\sigma_y) = \int d^N\hat{x}\ \mathrm{pr}(\hat{x},x_0,y_0,\log R_x,\log R_y\,|\,x,y,\sigma_x,\sigma_y).$$

This yields estimators for x_0 and y_0, together with their covariance matrix:

$$<x_0> = \sum_i x_i/N = \overline{x}; \quad <y_0> = \sum_i y_i/N = \overline{y}; \quad \text{or} \quad <b> = \overline{y} - a\,\overline{x}\ .$$

$$<\delta x_0^2> = (\sigma_x^2 + aR^2)/N; \qquad <\delta x_0 \delta y_0> = R^2/N; \qquad <\delta y_0^2> = (\sigma_y^2 + a^{-1}R^2)/N,$$

and

$$<\delta b^2> = (\sigma_y^2 + a^2\sigma_x^2)/N.$$

Note that the error estimates for x_0 and y_0 depend on the range parameter R, but that the error in the intercept value b depends only on the measurement errors σ_x and σ_y. We take this opportunity to integrate x_0 and y_0 out of the problem also, and to express the answer in terms of a and R. Finally, we find:

$$\log \mathrm{pr}(\log a,\log R\,|\,x,y) = \text{constant} - (N-1)/2\ \log(a\sigma_x^2R^2+\sigma_x^2\sigma_y^2+a^{-1}\sigma_y^2R^2)$$
$$- \frac{(V_{xx}(aR^2+\sigma_y^2)-2V_{xy}R^2+V_{yy}(a^{-1}R^2+\sigma_x^2))}{2\ (a\sigma_x^2R^2+\sigma_x^2\sigma_y^2+a^{-1}\sigma_y^2R^2)}\ ,$$

where the sample sum-squares are defined:

$$V_{xx} = \sum_i (x_i-\overline{x})^2; \quad V_{xy} = \sum_i (x_i-\overline{x})(y_i-\overline{y}); \quad V_{yy} = \sum_i (y_i-\overline{y})^2.$$

There is little insight to be gained in developing this formula further analytically, but it is interesting to note its behaviour in certain limits. The estimated slope $\hat{a}$ is close to either V_{xy}/V_{xx} or V_{yy}/V_{xy}, depending on the relative sizes of σ_x and σ_y; its error is determined by the measurement errors, not the range parameter. The range parameter R is similarly determined by either $R_x^2 \sim V_{xx}/N$ or $R_y^2 \sim V_{yy}/N$, and its error $\delta \log R \sim N^{-1/2}$.

4. Discussion

We now illustrate this formula with a computer example. Figure 1 shows a dataset of 100 samples together with the best-fitted line. This looks to be a sensible fit, though we claim little credit for this in itself, because an equally good job can be done by eye. Figure 2 shows the posterior distribution of the interesting parameters R_x and R_y, confirming the presence of a single, well-defined maximum in the posterior p.d.f. We see also that there are certain problems of normalisation of the posterior distribution, because the p.d.f. tends to a constant value as $R_x \dashrightarrow 0$ and $R_y \dashrightarrow \infty$. As R_x or $R_y \dashrightarrow \infty$ the distribution falls off sufficiently fast to be integrable over an infinite range. This therefore raises again the question of a "sensible" cut-off for R_{xmin} and R_{ymin}. We can answer the question of what a sensible cut-off means by investigating just what these cut-offs would have to be so that the contribution from the quadrant $(R_x,R_y) \dashrightarrow 0$ made a 50 per cent contribution to the posterior probability integral. For our dataset we find R_{xmin} and $R_{ymin} < \exp(\exp(-1000))$. This is clearly a crazy number, and indicate that we are solving an essentially well-posed problem.

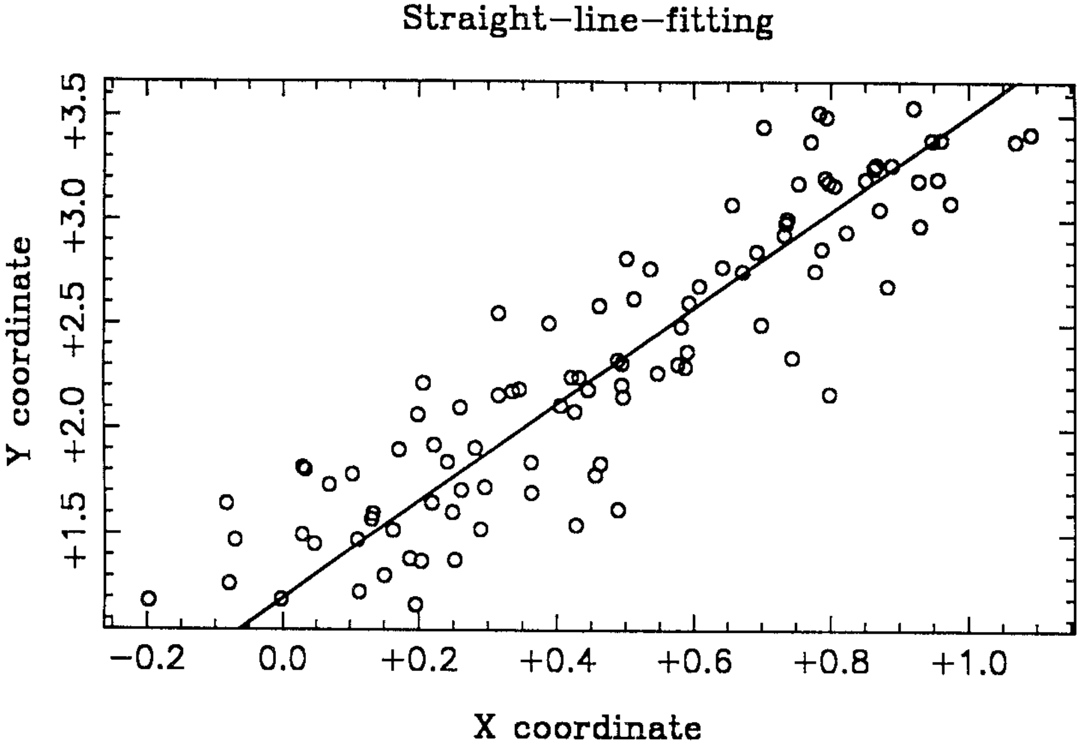

Figure 1. The dataset used: there are 100 uniformly-spaced samples.

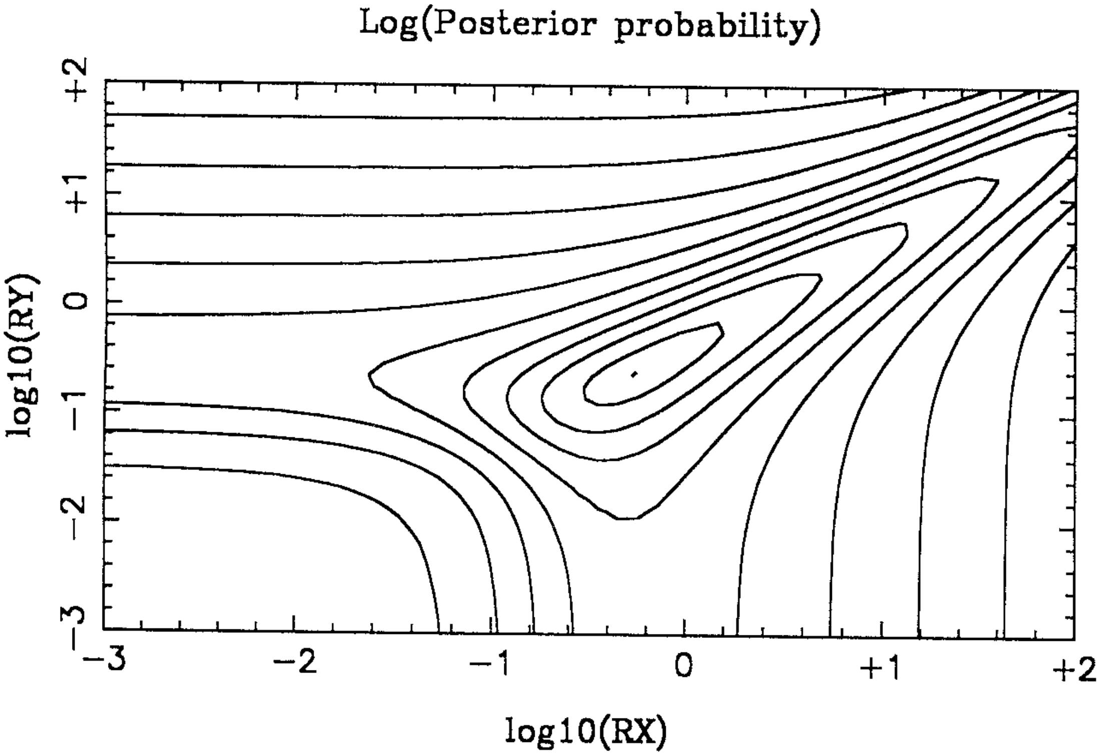

Figure 2. Posterior distribution of the range parameters R_x and R_y. The contour intervals are logarithmic, each level representing a probability difference of exp(100).

In other cases, though, we could well imagine that these numbers would not be so crazy, but instead give us insight into very real dilemmas. For example, suppose that the range of the data in one variable, say y, is very nearly covered by its error σ_y. This could easily happen, and implies $R_x \sim 0$. In this case are we really so sure that there is any real variation of y present in the data? Only our prior probability of the range of R_{ymin} can help us here - R_{ymin} *does* matter. Of course, a change of prior for a, such as that suggested by Jaynes, can make this p.d.f. integrable, but at the cost of disguising what is an essential part of the problem. When the x variation is similar to σ_x, then R_{xmax} is also important.

Whilst the assignment of priors for a and b leads to interesting discussions, it does not actually affect the numerical estimates greatly. It is a bit like arguing about whether to use V/N or V/(N-1) when calculating standard deviations; the prior information becomes swamped as we gather more data. There are, however, much more important matters that are raised by our formula. The prior $pr(\hat{x}|x_0,y_0,R_x,R_y)$ is the most contentious part of the analysis, for the reason that there are so many nuisance parameters. We cannot swamp the $\hat{x}$ by gathering more data: we introduce a new $\hat{x}$ for each sample. We therefore have to be rather more careful about this prior.

Our first instinct, perhaps, would be to say "$\hat{x}$ is a location parameter" and assign to it a uniform prior over an infinite interval. There is no mathematical difficulty in this, indeed the analysis is far easier, and corresponds to our case R --> ∞. I freely admit that this was the first case that I tried, and I only abandoned it because it doesn't work. Indeed, if it had worked, then this analysis would have stayed in my research notebook as a trivial application of Bayesian methods. To see that the formula goes wrong, look at it in the limits $\sigma_x = 0$, $R = \infty$:

$$\log pr(a|x,y,\sigma_y) = \text{constant} - ((N-1)/2) \log a - ((a^2V_{xx} + aV_{xy})/\sigma_y^2)/2 .$$

The last term is fine, but the first term biasses the answer, increasing a by one-half of a standard deviation. But this term cannot just be dropped! We could get rid of it by re-formulating our hypothesis space in a different way, by dropping the symmetry with respect to x and y. But that in turn would exacerbate the problem for the complementary case $\sigma_y = 0$, where the present one-half standard deviation bias would be doubled.

All my Bayesian friends have objected at this point that "there's no such concept as bias in Bayesian analysis". It is true that there is no meaningful, exact definition of bias except in a frequentist sense. What I mean here is that the answer given by the R --> ∞ estimator is usually wrong, and in a given direction. The dictionary calls this "bias".

When a Bayesian calculation gives the wrong answer, it simply means that the hypothesis space contains wrong information. Here , we assembled the joint p.d.f. in a systematic way that I recommend be used in *all* Bayesian calculations (Gull 1988), so it is easy to see what went wrong. It was clear at the time that we needed the prior $pr(\hat{x}|x_0,y_0,R_x,R_y)$ for *all the N samples simultaneously*. We might swallow the "location parameter" argument for the first sample, but for N all at once it looks very strange. Suppose that the first (N-1) samples all lie in $-3 < x < 2.5$. Do we really believe that the next sample can be anywhere in $(-\infty, \infty)$? Our original Gaussian prior amounts to the reasonable suggestion that we learn about the mean and variance of the $\hat{x}$ distribution from the sample. We can of course do this, so that R is in general well-determined by the sample. Seen this way, we might it think it advisable to learn some other parameters of the shape of the $\hat{x}$ distribution as well. This will improve our results, but probably not very much, and at a terrible price: we would then be unable to perform the required integrals analytically.

We can see now why the range parameter R corrects the bias of the estimated slope. Suppose, again, that $\sigma_x = 0$. The $((N-1)/2 \log a)$ term from the determinant increases the slope by one-half of a standard deviation, but, as R is reduced, the $\hat{y}$ are gently squeezed in range, reducing the slope. When R reaches its most likely value the bias in the slope is exactly corrected. Seen the other way around: the range parameter biasses the slope against the weaker direction of the error bars; the determinant term corrects this. The formula given earlier seems to work for all combinations of σ_x and σ_y.

Can such a simple problem really require so complicated a solution? If all you want is the answer I can recommend an estimator for a:

$$\min \frac{(a\, V_{xx} - 2\, V_{xy} + a^{-1}\, V_{yy})}{(a\, \sigma_x^2 + a^{-1}\, \sigma_y^2)} \,.$$

This is our answer with R --> ∞ and the determinant term dropped, so it will probably work. In can be derived by an ingenious argument (Brian Ripley, private communication, see also Ripley 1987 and Sprent 1969). The problem is scaled, not on the range of the data, but on the size of the errors σ_x and σ_y. The range itself is then allowed to go to infinity. If you scale on σ_x and σ_y, then there is no longer a 'weak' direction to be biassed, so no problems appear with the R --> ∞ solution. However, we note that a finite range $R^2 \sim V/N$ is still more likely than $R = \infty$, and because the problem is no longer scaled symmetrically on the range of the data, bias would return if R were reduced. In any case, scaling on the size of the errors looks a bit peculiar if either σ_x or σ_y is zero. Again, this modification of the hypothesis space seems to be an attempt to disguise what is a real problem. One is even led to speculate that scaling on σ_x and σ_y is a subconscious admission that some statisticians are more interested in the errors than they are in the data themselves!

Finally, we examine the problem of determining the level of the errors σ_x and σ_y if they are unknown. This does not involve any more analysis, because we have already been careful to retain all factors of σ_x and σ_y from the likelihood. We assign an uninformative prior for these variables, uniform in $\log \sigma_x$ and $\log \sigma_y$, so that our previous formula will also be the posterior: $\mathrm{pr}(\log\sigma_x, \log\sigma_y, \log a, \log R \,|\, x, y)$, which is illustrated in Figure 3. As we would expect, if only one of σ_x or σ_y is unknown, then the data determine the other extremely well, but it is too much to expect that <u>both</u> can be determined simultaneously. Rather, it is the combination $(a\sigma_x^2 + a^{-1}\sigma_y^2)$ that is accurately determined, but the error cannot be very reliably assigned to x or y individually. However, Figure 3 does show that there is just a little information about the ratio σ_y/σ_x contained in the dataset, presumably reflecting the fact that the $\hat{x}$ were uniformly sampled, rather than taken from a Gaussian distribution. Looking at the data by eye confirms this feeling; for a uniform distribution one can guess the relative contributions to the error. This indicates that there might be some real merit in using a more complicated hypothesis space, despite the difficulties of the computations involved.

5. Conclusions

The apparently simple Bayesian problem of straight-line fitting with both variables subject to error contains a subtle twist. The ranges of the variables

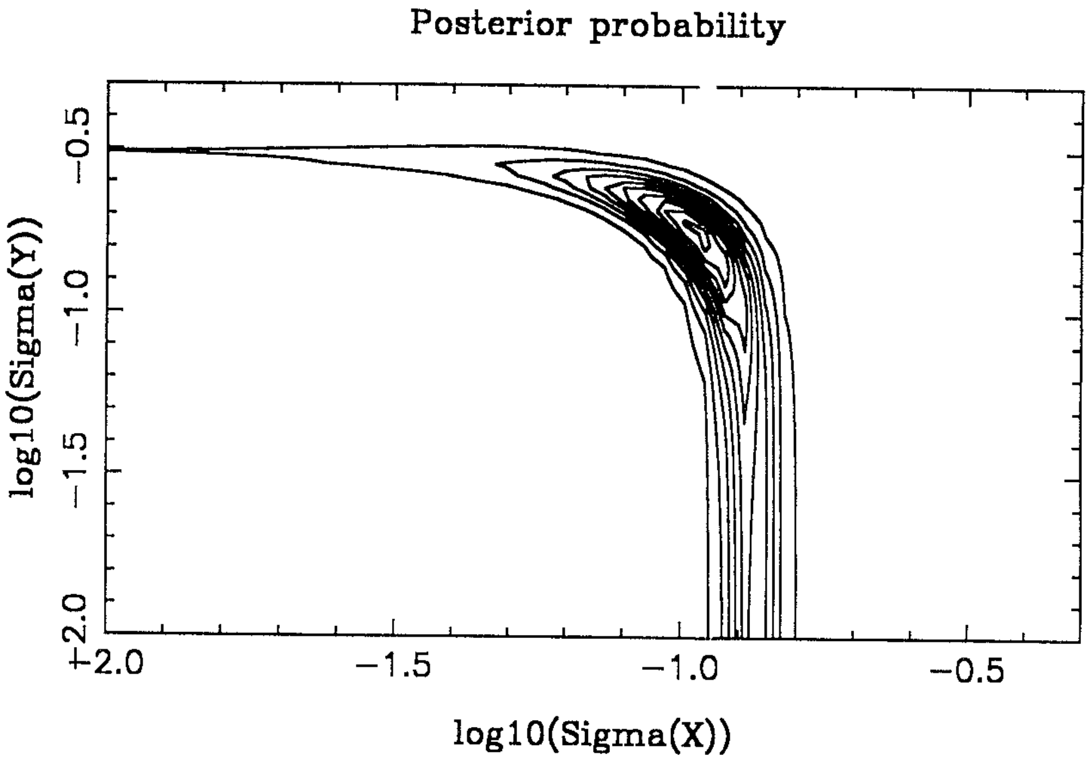

Figure 3. Posterior distribution of σ_x and σ_y. The contours are linear.

are usually well-determined by the dataset, equivalent to an "informative" prior for the sample positions. The use of a uniform, uninformative prior would lead to a bias in the estimated slope. The use of informative priors containing range parameters is a common feature of Bayesian analyses of this type: the "Classic" Maximum Entropy presented here by Skilling (1988) is another example.

Acknowledgments

I am grateful to John Skilling, Geoff Daniell and Brian Ripley for discussions about Bayesian regression.

References

Gull, S.F. (1988). Bayesian inductive inference and maximum entropy. In Maximum Entropy and Bayesian Methods in Science and Engineering, Vol. 1, ed. G.J. Erickson & C.R. Smith, pp 53-74. Kluwer, Dordrecht.

Jaynes, E.T. (1967). Reply to comments by Oscar Kempthorne, following Bayesian Intervals vs. Confidence Intervals. Reprinted in E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics, ed. R. Rosenkrantz (1983), pp190-209. Reidel, Dordrecht.

Jaynes, E.T. (1973). The well-posed problem. Reprinted in E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics, ed. R. Rosenkrantz (1983), pp133-148. Reidel, Dordrecht.

Ripley, B.D. (1987). Regressian techniques for the detection of analytical bias, Analyst, 112, 377-383.

Skilling, J. (1988). Classic Maximum Entropy. In these Preceedings.

Sprent, P. (1969). Models in Regression and Related Topics. Methuen, London.