leapm.tex, 4/21/1999

STRAIGHT LINE FITTING – A BAYESIAN SOLUTION

E. T. Jaynes
Wayman Crow Professor of Physics
Washington University
St. Louis, MO 63130, U.S.A.

Abstract. Fitting the "best" straight line to a scatter plot of data $D \equiv \{(x_1, y_1) \ldots (x_n, y_n)\}$ in which both variables x_i, y_i are subject to unknown error is undoubtedly the most common problem of inference faced by scientists, engineers, medical researchers, and economists. The problem is to estimate the parameters α, β in the straight line equation $y = \alpha + \beta x$, and assess the accuracy of the estimates. Whenever we try to discover or estimate a relationship between two factors we are almost sure to be in this situation. But from the viewpoint of orthodox statistics the problem turned out to be a horrendous can of worms; generations of efforts led only to a long line of false starts, and no satisfactory solution.

We give the Bayesian solution to the problem, which turns out to be eminently satisfactory and straightforward, although a little tricky in the derivation. However, not much of the final result is really new. Arnold Zellner (1971) gave a very similar solution long ago, but it went unnoticed by those who had the most need to know about it. We give a pedagogical introduction to the problem and add a few final touches, dealing with choice of priors and parameterizations.

In any event, whether or not the following solution has anything new in it, the currently great and universal importance of the problem would warrant bringing the result to the attention of the scientific community. Many workers, from astronomers to biologists, are still struggling with the problem, unaware that the solution is known.

Presented at the Tenth Annual MAXENT Workshop, University of Wyoming, July 1990. To appear in the Proceedings Volume, W. T. Grandy & L. Schick, Editors, Kluwer Academic Publishers, Holland.

1. INTRODUCTION

The following discussion is an instructive case history showing how nontrivial Bayesian results evolve. It illustrates three very important points:

(a) The difficulties are never mathematical; at no point do we encounter any mathematical problem that could not be dealt with by an undergraduate who had passed a first course in elementary algebra, and also had some sense of the proper order of carrying out limits. The long decades of error, in which the most famous figures in the field tried and failed to find a solution, were due to conceptual difficulties (inability to make the right connection between the real world problem and the mathematics of probability theory), lack of direction (failure to obey the rules of probability theory), or mathematical ineptness (throwing out the baby with the bath water by trying to pass to infinite limits too soon in the calculation).

(b) *After* one has the final correct solution, it all becomes intuitively obvious and we are chagrined at not having seen the answer immediately. This is typical of all nontrivial Bayesian solutions. In an ideal world, our intuition would be so powerful that we would have no need of probability theory. But in this world a little application of probability theory as logic can do wonders in educating our intuition.

(c) Finally, we realize that there are very simple rules of conduct about what the rules of probability theory are and about how to handle limits in potentially singular mathematics, that if followed would have bypassed all those difficulties and led us automatically, in a few lines to the final correct solution. But these rules are ignored – or worse, summarily rejected – by those whose mathematical training has concentrated on set theory rather than analysis. For those with this handicap, mere exhortations to follow the rules are not enough; the need to follow them can be appreciated only from studying case histories like the present one, demonstrating the specific consequences of not following them.

In the next three sections we survey the long and complicated background of the problem. The reader who wants to get on with the technical content of this work may turn immediately to Section 5 below.

2. HISTORY

This problem, seemingly straightforward from a Bayesian viewpoint, has a long history going back to Gauss (1809), who gave what are still the most used results (found, we note, by Bayesian methods; he interpreted his least – squares fitting as locating the most probable value of a parameter). The term "Gaussian distribution" derives from this work. Here we seek to generalize that solution to the case of unknown error in both variables. Orthodox statistical theory was helpless to do this, for three reasons.

Firstly, orthodox ideology requires one to specify which quantities are 'random' and which are not; yet nothing in the real problem tells one how to decide this. If some quantity X is declared to be 'random', then orthodoxy demands that we assign probabilibites to it, even though it may be a datum and therefore, in the context of our problem, a known constant. If a quantity α is declared to be 'nonrandom', then orthodoxy forbids us to assign probabilities to it at all, even though it may be the unknown quantity about which inference is being made. This view of things is already sufficient to prevent orthodoxy from solving the problem; the connection between the real world and the mathematics is seen backwards.

Secondly, the problem is swamped with nuisance parameters (with n data points we have $n+2$ nuisance parameters), which orthodox principles cannot cope with. Finally, because that ideology does not permit the use of the prior probabilities for unknown parameters (on the grounds that the parameters are not "random variables", only unknown constants), it denies itself any way to take prior information into account. As a result, orthodox statistics has never found a satisfactory way of estimating parameters θ in a model equation $y = f(x;\theta)$ when both the x and y measurements are subject to error; it has produced only contrived *adhockeries*, not derived from the principles of probability theory and with unacceptable performance.

Here Bayesians have the advantage because we see the connection between the real world problem and the mathematics in a very different way. We assign a probability to some quantity, not because it is 'random', but because it is *unknown*, and it is therefore a quantity about which probabilistic inference is needed and appropriate. This gives us immediately the technical power needed to solve the problem.

Because we admit the notion of probability of an hypothesis, the Bayesian mathematical apparatus is able to deal easily with nuisance parameters and prior information. So the Bayesian solution for the most general case might be on the level of a simple homework problem, but for some tricky points about limits that require careful explanation today (although earlier mathematicians, trained in analysis rather than set theory, considered them obvious).

Before 1960, the only person who could have solved this problem was Harold Jeffreys. His Bayesian treatment of regression (Jeffreys, 1939) finds easily the correct solutions to simpler problems, by methods that would have succeeded on our problem, had he tackled it. Then various writers attempted to deal with it by nonBayesian methods, as follows.

If the errors are known to be only in the y–measurements, the problem reverts to that of Gauss and the optimal solution found by everybody, Bayesian or nonBayesian, is the simple regression line that minimizes the sum of squares of the y–residuals. If the errors were known to be only in the x–measurements, one would instead minimize the sum of squares of the x–residuals. The fact that these two procedures yield different (often *very* different) estimated line slopes $\hat{\beta}$ is the origin of the difficulty that has troubled so many for so long.

The astronomer Strömberg (1940) advocated that we take the geometric mean of the two estimated slopes, and called this the 'impartial' regression line. It does indeed appear 'impartial' to interchange of x and y; but this suggests at once two other *ad hoc* devices that seem equally impartial: to take the bisector of the two simple regression lines, or to minimize the sum of squares of the perpendicular distances from the data points to the line.

All three of these devices have been advocated, all seem more or less reasonable intuitively, and all are impartial to interchange of x and y. Yet none was shown to follow from the principles of probability theory, none gives any indication of the accuracy of the estimates, and they yield three different results. We are left with no criterion of choice.

Of course, impartiality to interchange of x and y does not seem a desideratum if x and y are quantities of a different kind (such as the magnitude and redshift of a galaxy in a Hubble diagram). This indicates that we are really concerned with several different problems here, arising from different prior information of this type about the meaning of our variables; and shows another reason why methods which fail to use such prior information could not have succeeded.

Recently, some astronomers (Isobe, *et al*, 1990) have launched an ambitious effort to summarize the present situation; but they take no note of the existence of Bayesian methods, and so are able to give only a long, somewhat dreary, list of one *adhockery* after another, with no firm final conclusions. The same ambiguity that has puzzled astronomers for fifty years has given rise to equally long controversy among biologists (Sokal & Rohlf, 1981), as to how one is to choose the 'best' straight line.

Throughout those same fifty years, statisticians have also been trying to to deal with the problem. Abraham Wald (1940) made a stab at estimating β by orthodox methods when both x_i and y_i are subject to error. But his rule (separate the data points into two subsets, and take the slope of the line joining their centroids) gives a probable error many times that of the Bayesian result when the variances σ_x, σ_y are known, and its accuracy is indeterminate when they are unknown (in contrast, the Bayesian solution given below automatically estimates the noise level from the internal evidence in the data, and always gives accuracy estimates which correspond nicely with the indications of common sense). Unfortunately, Wald died in an airplane accident, so soon after his conversion to Bayesianity that he had no opportunity to correct this early attempt.

W. E. Deming (1943) concluded that the problem is fundamentally indeterminate and that it is necessary to know the ratio $\lambda = \sigma_y/\sigma_x$ in order to make any estimates of α, β. Then he estimated them by minimizing the sum of squares

$$\sum_i \left[\lambda^2(x_i - \hat{X}_i)^2 + (y_i - \hat{Y}_i)^2\right] \tag{1}$$

with respect to $\hat{X}, \hat{Y}$, subject to $\hat{Y} = \hat{\alpha} + \hat{\beta}\hat{X}$. The resulting estimates are not entirely unreasonable; they clearly do the right thing, reducing to the original simple solution, in the limits $\lambda \to 0$ and $\lambda \to \infty$, and interpolate somehow between these in intermediate cases. This method was advocated by Mandel (1964). But it is basically inapplicable; it is hard to imagine any real problem where we would know λ if we did not know σ_x and σ_y.

Cramér (1946) considers straight line fitting problems, but does not attempt to deal with this problem at all; like most orthodox writers, he is able to deal only with the case where the x_i are known without error (*i.e.*, $\sigma_x = 0$) and σ_y is known. Then the nuisance parameters go away and the maximum likelihood estimates of α, β succeed in reproducing the estimates of Jeffreys (but without his accuracy statements).

Kendall & Stuart (1961), in what was ostensibly the definitive account of the state of the art at the time, devote four Chapters, comprising 141 pages, to such problems. Yet their viewpoint is completely orthodox, and they take no note of the fact that the correct (Bayesian) solutions to most of their problems were already in print at that time. For example, in trying to estimate a correlation coefficient ρ, they offer several different *ad hockeries* which they acknowledge to be unsatisfactory; but do not mention the correct solution which had been given 22 years earlier by Jeffreys (1939).

In the case where both variables are subject to error, they have nothing to offer beyond repeating the method of Wald, noting its unsatisfactory nature, and closing the discussion at that. It is a sad performance; while they recognize that orthodox methods fail to give satisfactory solutions to most of their problems, still they do not recognize the Bayesian methods which had already demonstrated, in the hands of Jeffreys, their power to give the useful solutions that scientists needed.

This is particularly deplorable in view of the fact that Maurice Kendall was at St. John's College, Cambridge and presumably saw Jeffreys almost daily; he could hardly claim that he was unaware of what Jeffreys had to offer him, or that he had no opportunity to learn from Jeffreys. It was the orthodox ideology – promoted vigorously by R. A. Fisher – which led him to reject Jeffreys' approach without taking the trouble to examine its theoretical basis or its performance on real problems. No better illustration could be found of the devastating effects which that ideology has inflicted upon this field.

Zellner (1971, Chapter 5) finally gave the next Bayesian solutions beyond Jeffreys, and all but disposed of our present problem as far as theory is concerned. But he had so many useful solutions like this to present that the background discussion and applications are missing. It seems that few Bayesians and no orthodoxians took note of this important advance (the presence of the word 'Econometrics' in the title automatically prevented physicists and engineers from examining it, while the presence of the word 'Bayesian' prevented orthodoxians from examining it).

Kempthorne & Folks (1974) become so involved in the question: "Which quantities are random?" that they are led to formulate sixteen different problems corresponding to all possible answers; and then give up without finding any usable solution, and proceed to fall back on the aforementioned method of Wald still another time. Yet they reject any attempt to find a Bayesian solution, with charge (p. 439) that the likelihood function for this problem is "totally uninformative" (which shows only that they were uninformed about the work of Zellner). They proceed to deny the validity of the likelihood principle itself (which shows that they were uninformed about the work of dozens of other authors, from Jeffreys to Lindley).

This performance of Kempthorne & Folks is what the present writer was replying to in Jaynes (1976), by giving a bit of the general solution, to show just how much useful information is contained in that "completely uninformative" likelihood function. Unfortunately, at the time I too was ignorant of the work of Zellner for the reason just noted, and so gave only part of the full Bayesian solution. Let us hope that, for the next Edition of his book, Arnold can be persuaded to choose a different title which does not frighten away the very people who have the most need to know what is in it.

Steve Gull (1989) has just re-opened this discussion with an article which takes note of my old work (but unfortunately not the more complete work of Zellner which had appeared five years earlier) and improves on my solution. But he too did not find the full, final solution. The present work undertakes to complete this long discussion, now 180 years old. Whether it has finally succeeded will be judged, necessarily, by others.

3. TERMINOLOGY

As we warned in the Calgary tutorial (Jaynes, 1986), old terminology inherited from orthodox statistics can be totally inappropriate for Bayesian situations. Orthodoxy by its preoccupation with the question "Which quantities are random?" is obliged to draw several fine distinctions which a Bayesian finds not only irrelevant to the problem, but basically meaningless. But orthodoxy takes no note of such crucial things as prior information or the full shape of the likelihood function, and thus fails to draw the distinctions that *are* essential in real problems of inference.

We are glad to follow standard orthodox terminology whenever it is useful and has the same meaning in Bayesian and orthodox theory; but if Bayesians are ever to make ourselves understood to those with orthodox training, it is imperative that we jettison all orthodox terminology which is misleading in a Bayesian setting. We need to use different terms which are descriptive of their proper Bayesian meaning.

The terms "linear regression problem" or "linear model problem" are interpreted differently by different authors. One might expect this to mean fitting straight lines to the data: i.e., a model represented by a straight line model equation: $y_i = \alpha + \beta x_i + e_i$. However, in the standard literature [for example, Graybill (1961, p. 97)] a "linear model" is defined as one for which the model equation is linear in the parameters and the noise, not necessarily in the observations. Thus

$$\exp(y_i) = \alpha + \beta \cos(x_i^{3/2}) + e_i$$

is also a 'linear model'.

Both the Bayesian Steve Gull (1989) and the nearly-Bayesian Morris de Groot (1985) make a break with this terminology, and define the term "linear model" as referring to straight line fitting. So, while recognizing that our analysis applies equally well to fitting curves of any shape, we shall go along with de Groot and Gull by considering only straight line fitting here.

This still leaves us confronted with the term "regression". Kendall & Stuart (1961) draw a very great distinction between "regression" and "straight line fitting". They define the former as estimating y_i, given x_i, from the conditional sampling distribution $p(y|x) = p(xy)/p(x)$; the latter as estimating the parameters α, β in the model equation $y = \alpha + \beta x + e$, where e is "random error". Thus in regression our estimate of y_i is to depend only on x_i and the sampling distribution; it does not make any use of other observed data values $\{x_j, y_j\}$. In effect "regression" presupposes a perfectly factorized sampling distribution $p(\{x_j, y_j\}) = \prod_j p(x_j.y_j)$ without any variable parameters; it is only n repetitions of a trivial bivariate problem.

It seems to us that this would make the regression problem so specialized that it has virtually no applications. Almost always, our sampling distributions contain unknown parameters common to all observations, and for estimating y_i given x_i it is essential to take into account the other data values $\{x_j, y_j\}$ because they help us to estimate the common parameters. That is, if the model equation is a straight line $y = \alpha + \beta x$, we would always estimate y_i as $\hat{\alpha} + \hat{\beta} x_i$, where $\hat{\alpha}$, $\hat{\beta}$ are our estimates of the parameters.

Steve Gull also draws a distinction between "straight line fitting" and "regression" but one that that seems to us of a quite different nature. His distinction appears to refer to the conceptual meaning we attach to the variables rather than the form of the actual equations. We may consider $\{X_i, Y_i\}$ to be the unknown true values of some real physical quantities (such as load and deflection of a steel beam) connected by (1), and $\{e_i, f_i\}$ as the measurement errors in observing them. Then we would think of the scatter plot of the data points as a kind of cloudy image of the unknown true straight line (1), and estimating α and β amounts to fitting the best straight line to the data.

Alternatively, we could think of the observations $\{x_i, y_i\}$ as the true physical quantities (for example, the barometric pressures at New York and Boston, at noon on the i'th day) which are measured with negligible error. Then $\{e_i, f_i\}$ would represent, not mere measurement errors, but the variability of the physical phenomenon of interest, and we might think of $\{X_i, Y_i, \alpha, \beta\}$ as mental constructs invented by us to help us reason about it.

Perhaps α and β are thought of as 'propensities' for weather in the Eastern U. S. as a whole, representing the predictable component of the weather, while $\{e_i, f_i\}$ comprise the unpredictable component. We hasten to add that by "unpredictable" we do not mean the conventional Mind-Projection Fallacy meaning of "not determined by anything". We mean "determined by factors that are not in the data set of the weather forecaster." Steve calls this a 'regression problem'.

While we recognize the conceptual difference between Steve's 'regression' and 'straight line fitting' problems, we see very little mathematical difference. Of course, there may be great differences in the prior information in the two cases.

4. LEAPFROGGING ARTIFICIAL HORIZONS

During my year at St. John's College, Cambridge, I had to force myself not to get seriously involved in history, simply because it is addicting and there is too much of it there; you could easily find yourself trapped for a lifetime, and never again accomplish anything in contemporary science.

Perhaps one reason why some people hesitate to get seriously into Bayesian analysis is the same. A problem in sampling theory, which takes no note of prior information, is finite and having solved it, that is the end of it. But every Bayesian problem is open-ended; no matter how much analysis you have completed, this only suggests still other kinds of prior information that you might have had, and therefore still more interesting calculations that need to be done, to get still deeper insight into the problem.

A person who tries to present a Bayesian solution, being obliged to produce a finite sized manuscript in a finite time, must forego mentioning many other interesting things about the problem that he became aware of while writing it. My frustration at this came out particularly at the end of the "Bayesian Spectrum and Chirp Analysis" paper (Jaynes, 1987) which noted, helplessly, that we had examined only one almost trivial special case, and it would require several volumes to deal with all the interesting and important things that are to be found in that simple-looking model.

Since then, Larry Bretthorst (1988) has provided one of those volumes, and there is hope that he may produce a second.

In effect, anyone writing about a Bayesian solution must draw a kind of artificial horizon about the problem, beyond which he dare not tread however great the temptation. For most readers it does not matter just where that horizon is, because their horizon of expectations is well within it.

The aforementioned discussion of Bayesian linear regression in Jaynes (1976) was only a small part of that general reply to Oscar Kempthorne's antiBayesian charges. His provocation led me to note a few things about that allegedly impossible Bayesian solution, enough to show its vast superiority over the many different false starts of Kempthorne & Folks (1974). But because of space limitations I drew a rather tight horizon around the problem, just wide enough to answer Kempthorne.

Steve Gull is perhaps the only person who can be counted on to read one of my papers so deeply that he detects the horizon I chose, and gleefully takes a peek beyond it. In Gull (1989) he has leapfrogged that old horizon, and thereby goaded me into doing finally what I should have done 15 years ago: completing the discussion of Bayesian straight line fitting (and in the process leapfrogging his horizon). This leads to some results of current interest and importance, as well as a needed lesson in how to deal with potentially–singular mathematics.

5. FORMULATION OF THE PROBLEM

By hypothesis, there is an exact model equation

$$Y = \alpha + \beta X \tag{2}$$

in which α, β are the unknown parameters of interest. But the data D consist of n pairs of observed values:

$$D \equiv \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{3}$$

related to (2) by

$$x_i = X_i + e_i, \qquad y_i = Y_i + f_i, \qquad\qquad 1 \le i \le n \tag{4}$$

where e_i, f_i are measurement errors, and we consider the two problems of parameter estimation and prediction: given the data D and certain prior probability assignments to α, β, X_i, Y_i, e_i, f_i,

(A) *Straight Line Fitting:* What do we know about α and β?

(B) *Regression:* Given m more values $\{x_{n+1}, \ldots, x_{n+m}\}$, what do we know about the corresponding $\{y_{n+1}, \ldots, y_{n+m}\}$?

In estimating α, β, we have potentially $n+2$ nuisance parameters, $\{X_1 \ldots X_n, \sigma_x; \sigma_y\}$. Denoting prior information by I, the most general solution will then have a joint prior *pdf* for $n+4$ parameters:

$$p\,(\alpha, \beta, X_1 \ldots X_n, \sigma_x, \sigma_y | I) = p\,(\alpha, \beta | I)\, p\,(X_1 \ldots X_n, \sigma_x, \sigma_y | \alpha, \beta, I) \tag{5}$$

and an unending variety of different kinds of prior information I might be expressed by this function, which we understand is to be properly normalized to unit integral.

By the product rule, $p(\alpha, \beta|I)$ can always be factored as shown, although the possibility of a Borel–Kolmogorov paradox should be kept in mind. That is, by a probability $p(A|\alpha, \beta I)$ conditional on point values of α, β, we must understand the limit of the well-defined

$$P(A|d\alpha\, d\beta I) \equiv \frac{P(A, d\alpha\, d\beta|I)}{P(d\alpha\, d\beta|I)} \tag{6}$$

as $d\alpha \to 0$, $d\beta \to 0$. To avoid ambiguity it is necessary to prescribe the exact way in which the limit is to be approached.

For example, if we set $d\alpha = \epsilon\, g(\beta)$ and pass first to the limit $\epsilon \to 0$, our final results will in general depend on which function $g(\beta)$ we chose. But there is no 'right choice' or 'wrong choice' because it is for us to say which limit we want to take; *i.e.*, which problem we want to solve. Any choice presumably corresponds to a legitimate problem that we might want to reason about, and probability theory will then give us the correct solution to that problem. But having made one choice, we must stick to that choice throughout the calculation, otherwise we are switching problems in midstream and are pretty sure to generate contradictions.

This is Sermon #1 on mathematical limits; although it was given long ago by Kolmogorov, many who try to do probability calculations still fail to heed it and get themselves into trouble. The moral is that, unless they are defined in the statement of a problem, probabilities of the form $p(A|\alpha I)$ conditional on point values of a parameter, have no meaning until the specific limiting process is stated. More generally, probabilities conditional on any propositions of probability zero, are undefined.

In the following we use the abbreviations

$$x \equiv \{x_1 \dots x_n\}, \qquad Y \equiv \{Y_1 \dots Y_n\}, \qquad etc., \tag{7}$$

so that our data are denoted by $D = x, y$. Then the most general sampling *pdf* would have the functional form

$$p\,(x, y|\alpha, \beta, X, \sigma_x, \sigma_y, I) \tag{8}$$

and the most general solution we contemplate here would have the form

$$p\,(\alpha, \beta|x, y, I) = p\,(\alpha, \beta|I)\, \frac{p\,(x, y|\alpha, \beta, I)}{p\,(x, y|I)} \tag{9}$$

in which

$$p\,(x, y|\alpha, \beta, I) = \int d^n X\, d\sigma_x\, d\sigma_y\, p\,(x, y|\alpha, \beta, X, \sigma_x, \sigma_y, I)\, p\,(X, \sigma_x, \sigma_y|\alpha, \beta, I) \tag{10}$$

$$p\,(x, y|I) = \int d\alpha\, d\beta\, p\,(x, y|\alpha, \beta, I)\, p\,(\alpha, \beta|I) \tag{11}$$

and, writing $x^*, y^* \equiv \{(x_{m+1}, y_{m+1} \dots (x_{m+n}, y_{m+n})\}$, our most general predictive distribution is

$$p(y^*|x^*, x, y, I) = \int p(y^*|x^*, \alpha, \beta, I)\, p(\alpha, \beta|x, y, I)\, d\alpha d\beta \tag{12}$$

This defines our present horizon (but having found this solution, its extension to such details as more than two variables, correlated noise, noise known to vary with x, etc. is an easy homework problem, involving little more than promoting some of our symbols from numbers to matrices).

6. SPECIAL CASES

Note first how the standard solutions are contained in this as special cases. If, as is almost universally supposed, the prior *pdf* for the errors factors completely:

$$p\,(e_i f_i|I) = \prod_{i=1}^{n} p\,(e_i|I)\, p\,(f_i|I) \tag{13}$$

with common Gaussian distributions

$$p(e_i|I) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{e_i^2}{2\sigma_x^2}\right), \qquad p(f_i|I) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{f_i^2}{2\sigma_y^2}\right), \tag{14}$$

Now, dropping the prior information symbol I, which we suppose henceforth to be hidden in the right-hand side of all our probabilities, our sampling *pdf* is

$$\begin{aligned} p(x, y|\alpha, \beta, \sigma_x, \sigma_y, X) &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma_x\sigma_y} \exp\left\{ -\frac{(y_i - \alpha - \beta X_i)^2}{2\sigma_y^2} - \frac{(x_i - X_i)^2}{2\sigma_x^2} \right\} \\ &= p(y|\alpha, \beta, X, \sigma_y)\, p(x|X, \sigma_x) \end{aligned} \tag{15}$$

which we note factors as shown. By Bayes' theorem,

$$\begin{aligned} p(\alpha, \beta|x, y, \sigma_x, \sigma_y) &= \int d^n X\; p(\alpha, \beta, X|x, y, \sigma_x, \sigma_y) \\ &= \int d^n X\; p(\alpha, \beta, X|\sigma_x, \sigma_y)\, \frac{p(x, y|\alpha, \beta, \sigma_x, \sigma_y)}{p(x, y|\sigma_x, \sigma_y)} \end{aligned} \tag{16}$$

and whether σ_x, σ_y are known or unkwown, the solutions will depend on the data only through their first and second moments, which are sufficient statistics for α, β. Introducing the standard notations for the observed sample first and second moments,

$$\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \overline{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i\,, \tag{17}$$

$$\overline{x^2} \equiv \frac{1}{n}\sum x_i^2\,, \qquad \overline{xy} \equiv \frac{1}{n}\sum x_i\, y_i\,, \qquad \overline{y^2} \equiv \frac{1}{n}\sum y_i^2\,, \tag{18}$$

and the sample central moments and correlation coefficient

$$s_{xx} = s_x^2 \equiv \overline{x^2} - \overline{x}^2\,, \qquad s_{xy} \equiv \overline{xy} - \overline{x}\;\overline{y}\,, \qquad s_{yy} = s_y^2 \equiv \overline{y^2} - \overline{y}^2\,, \qquad r \equiv \frac{s_{xy}}{\sqrt{s_{xx}\, s_{yy}}}\,, \tag{19}$$

we note for later purposes that eventually all solutions involve the fundamental quadratic form determined by the data

$$\begin{aligned} Q(\alpha, \beta) &\equiv \frac{1}{n}\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \\ &= C_{11}(\alpha - \hat{\alpha})^2 + 2C_{12}(\alpha - \hat{\alpha})(\beta - \hat{\beta}) + C_{22}(\beta - \hat{\beta})^2 + C_0 \end{aligned} \tag{20}$$

where the matrix C is

$$C = \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \tag{21}$$

and the requirement that (20) be an identity in α, β uniquely determines the coefficients:

$$\begin{aligned} \hat{\beta} &= s_{xy}/s_{xx} \\ \hat{\alpha} &= \overline{y} - \hat{\beta}\,\overline{x} \\ C_0 &= s_{yy} - s_{xy}^2/s_{xx} = s_{yy}(1 - r^2)\,. \end{aligned} \tag{22}$$

Of these, $\hat{\alpha}$ and $\hat{\beta}$ are just the original least-squares estimates of α, β that one would make if the errors were only in the y_i–measurmenents. But before reaching the form $Q(\alpha, \beta)$, we have to integrate out the nuisance parameters $X_1 \ldots X_n$, and perhaps also σ_x, σ_y.

7. THE ONE–POINT PROBLEM

To orient our thinking about this, consider first the 'baby' problem of estimating X_1 given only the datum x_1.

If σ_x were known and we had only the data component x_1, from (15) we would have immediately for the posterior *pdf* for X_1:

$$p(X_1|x_1, I) = Ap(X_1|I) \exp\left[-\frac{(X_1 - x_1)^2}{2\sigma_x^2}\right] \tag{23}$$

where here and what follows, A always stands for a normalizing constant, not necessarily the same in all equations. Suppose our prior information had led us to estimate X_1 as about $x_0 \pm \delta$; we could indicate this by the prior *pdf*

$$p(X_1|I) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left[-\frac{(X_1 - x_0)^2}{2\delta^2}\right] \tag{24}$$

But we note that

$$\frac{(X_1 - x_0)^2}{\delta^2} + \frac{(X_1 - x_1)^2}{\sigma_x^2} = \frac{(X_1 - \hat{X}_1)^2}{\sigma^2} + (const.) \tag{25}$$

where the $(const.)$ is independent of X_1, and

$$\hat{X}_1 \equiv \frac{x_0/\delta^2 + x_1/\sigma_x^2}{1/\delta^2 + 1/\sigma_x^2}, \tag{26}$$

$$\frac{1}{\sigma^2} = \frac{1}{\delta^2} + \frac{1}{\sigma_x^2} \tag{27}$$

whereupon (23) becomes

$$p(X_1|x_1 I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(X_1 - \hat{X}_1)^2}{2\sigma^2}\right] \tag{28}$$

We would estimate X_1 as (mean $\pm$ standard deviation)

$$(X_1)_{est} = \hat{X}_1 \pm \sigma \tag{29}$$

a weighted average of the prior estimate x_0 and the datum x_1, weighted according to the respective variances.

If we had almost no prior information about the unknown true value X_1, then $\delta >> \sigma_x$ and this would reduce for all practical purposes to $(X_1)_{est} = x_1 \pm \sigma_x$. For example, if $\delta > 10\,\sigma$, then the exact solution is within one percent of this limiting value, and the prior information would hardly help at all. But if we had prior information fixing X_i to an uncertainty comparable to σ_x, this would evidently be cogent, enabling us to improve the accuracy of our estimates of α, β.

This discussion of the 'baby' problem is to condition us to the usual argument for passage to an improper prior, $\delta \to \infty$. Usually, the data will be highly informative compared to our prior information (indeed, data which tell us little that we did not know already, would be hardly worth gathering). But if as usual our prior information is vague compared to the accuracy of the data, then whether we keep the prior with finite δ or pass to the limit of an improper prior, $\delta \to \infty$, makes no difference in the results. This is the conventional argument, surely valid for the simple problem being considered.

8. THE REAL PROBLEM

But note that the above passage to the limit is, in principle, to be carried out only at the end of the calculation. In the real problem, the properly normalized posterior distribution (9) is a ratio of two integrals, (10) and (11), and if we want to approach an improper prior it is the limit of the ratio, not the ratio of the limits, that should be carried out according to the rules of probability theory. The former limit is the well-behaved solution that we want; the latter may or may not exist. Depending on how rapidly the likelihood factor cuts off away from its peak, the separate integrals (10), (11) may diverge in the limit.

In the integrations over X_i it does not matter, because the Gaussian factors guarantee convergence of the integrals. Then we may behave in a rather reckless way and still get the right answer; but this is a rather unfortunate accident, that encourages bad mathematical habits that will fail on other similar–looking problems. For example, had we used a Cauchy noise distribution $p(e_i) \propto (a^2 + e_i^2)^{-1}$ instead of the Gaussian (15), the limit of the ratio would still be a perfectly well-behaved quantity, but the ratio of the limits would not exist. Our present problem involves not only the safe integration over the X_i, but also integration over σ_x and σ_y, for which the limit of the ratio continues to be well-behaved, but attempting to calculate instead the ratio of the limits can get us into trouble.

Admittedly, the point we are making is quite trivial, since if one does not see the distinction between the limit of a ratio and a ratio of the limits, he cannot even grasp the concept of a derivative dy/dx. Nevertheless, the recent literature of probability theory has examples where the use of improper priors as limits of proper priors is rejected, because well-known authors failed to perceive this trivial point and tried to calculate the ratio of the limits instead of the limit of the ratio. See, for example, our exchange with DSZ over the Marginalization Paradox (Zellner, 1980).

Let us see how easy it is for the unwary to commit this error; but also how easy it is to avoid once we understand the point. If we assign uniform priors to the X_i on the above grounds, and the Jeffreys priors $d\sigma/\sigma$ to σ_x, σ_y, we may as noted integrate $\{X_1 \ldots X_n\}$ out of (15) without disaster, and this constructs for us the quadratic form $Q(\alpha, \beta)$:

$$p(\alpha, \beta, \sigma_x, \sigma_y | x, y, I) = \frac{A}{(\sigma_x \sigma_y)(\sigma_y^2 + \beta^2 \sigma_x^2)^{n/2}} \exp\left\{ -\frac{nQ(\alpha, \beta)}{2(\sigma_y^2 + \beta^2 \sigma_x^2)} \right\} \tag{30}$$

and now we face the mathematical subtlety that is the real point of all this discussion. If we try to get $p(\alpha, \beta | x, y, I)$ by integrating out σ_x, σ_y from this, the result diverges due to the factor $(\sigma_x \sigma_y)$, which expresses the Jeffreys prior indicating ignorance of σ_x, σ_y.

But in (30) we have violated the rules of probablity theory by passing to the limit of improper priors before doing the normalizing integral; we are in effect trying to calculate the ratio of the limits. We got away with this with the X_i, but not with the σ_x, σ_y. Had we calculated the normalizing integral first for proper priors there could have been no divergence; then passing to the limit of the improper priors afterward would be a perfectly safe, uneventful procedure leading to the useful result that we want.

In 1965, the writer did not yet perceive this and was carried along by the arguments of Deming and Mandel, that the problem was indeterminate; this led to a long comedy of errors which I have seen others repeating many times since. My notebook entry of that time says: "This is symptomatic of the fact that *the data of the problem do not provide any information at all about whether the errors are in x or in the y measurement.* Then supposing Deming's parameter $\lambda \equiv \sigma_y/\sigma_x$ known, we can see whether this enables us to get a Bayesian solution; instead of writing the joint prior proportional to $1/(\sigma_x\sigma_y)$, we use

$$p\,(\sigma_x, \sigma_y|I) \propto \frac{\delta(\sigma_y - \lambda\sigma_x)}{\sigma_x} \tag{31}$$

then integration over σ_y merely makes the substitution $\sigma_y = \lambda\sigma_x$, and it reduces to a convergent integral:

$$\begin{aligned} p(\alpha, \beta|x, y, I) &= \frac{A}{(\lambda^2 + \beta^2)^{n/2}} \int_0^\infty \frac{d\sigma_x}{\sigma_x^{n+1}} \exp\left\{-\frac{nQ(\alpha, \beta)}{2(\lambda^2 + \beta^2)}\right\} \\ &= \frac{A}{Q(\alpha, \beta)^{n/2}} \end{aligned} \tag{32}$$

which is just the bivariate t-distribution that we would have had for the simpler regression problem in which the errors are only in the y-measurement and σ_y is completely unknown. Then, for example, we can integrate out α to get the posterior marginal *pdf* for β:

$$p\,(\beta|x, y, I) \propto \left[\gamma^2 + (\beta - \hat{\beta})^2\right]^{-(n-1)/2} \tag{33}$$

where $\gamma^2 \equiv s_{yy}\,(1 - r^2)/s_{xx}$. As the initial pleasure at this nice result wore off, a little warning bell started ringing in my mind as it dawned on me that, unlike Deming's least squares result, (32) is independent of λ. How can it be that the problem is indeterminate if λ is unknown; yet the solution when λ is known does not depend on λ?

A few years later, the answer suddenly seemed intuitively obvious. Instead of supposing λ known, make the change of variables $(\sigma_x, \sigma_y) \to (\sigma, \lambda)$ in (30):

$$\sigma \equiv \sqrt{\sigma_y^2 + \beta^2\sigma_x^2}, \qquad\qquad \lambda \equiv \sigma_y/\sigma_x \tag{35}$$

The jacobian is

$$\frac{\partial(\sigma_x, \sigma_y)}{\partial(\sigma, \lambda)} = \frac{\sigma}{\lambda^2 + \beta^2} \tag{36}$$

from which we find that the element of prior probability transforms as

$$\frac{d\sigma_x d\sigma_y}{\sigma_x\sigma_y} = \frac{d\lambda}{\lambda}\frac{d\sigma}{\sigma} \tag{37}$$

and (30) becomes

$$p\,(\alpha, \beta, \sigma, \lambda|x, y, I) = A \cdot \frac{d\lambda}{\lambda} \cdot \frac{d\sigma}{\sigma^{n+1}} \exp\left\{-\frac{nQ(\alpha, \beta)}{2\sigma^2}\right\} \tag{38}$$

At the time, I drew the conclusion that λ is completely decoupled from the problem:

$$p\,(\alpha, \beta, \sigma, \lambda|x, y, I) = p\,(\lambda|x, y, I)\,p\,(\alpha, \beta, \sigma|x, y, I) \tag{39}$$

so whatever prior we had assigned to λ would just integrate out again into a normalization constant, and contribute nothing to the final result. The algebra now seems to tell us that, far from being essential to make a determinate problem, λ is completely irrelevant to our problem! At least, from (38) it is clear how it can be that integrating out λ leads to divergence; yet supposing λ known leads to a result independent of λ.

The 'solution' which I offered at the 1973 Waterloo, Ontario meeting (Jaynes, 1976) is then

$$p(\alpha, \beta | x, y, I) = \int_0^\infty p(\alpha, \beta, \sigma | I)\, d\sigma \propto Q(\alpha, \beta)^{-n/2}, \tag{40}$$

the same as (32). But as the pleasure at this nice result wore off for the second time, it dawned on me that λ *ought to be* relevant to the problem after all. The result (40) is identical with what everybody, from Gauss on, had found for the case that σ_x is known to be zero; the measurement errors are only in y. In the opposite extreme, where the errors are only in x, the roles of x and y ought to be interchanged; but the quadratic form $Q(\alpha, \beta)$ is not a symmetric function of x_i and y_i. So where did I go wrong?

Let us go back to (38) and the conclusion we drew from it. Seeing only (38), it appears that not only is λ irrelevant to α, β, the data x, y tell us nothing about λ, for $p(\lambda|x, y, I) = p(\lambda|I)$. That is what we meant by saying that λ is completely decoupled from the problem.

The error in this reasoning was that (38) was derived *only* in the case of the Jeffreys prior (37); it has been shown thus far only that λ is decoupled *for that prior.* It turns out that exactly the same error of interpretation generated the "marginalization paradox" that was about to burst upon us (Dawid, *et al*, 1973; Jaynes, 1980). But for a general prior $f(\sigma_x, \sigma_y)\, d\sigma_x d\sigma_y$ the transformation would be, in place of (36),

$$f(\sigma_x, \sigma_y)\, d\sigma_x d\sigma_y = f(\sigma_x, \sigma_y) \frac{\sigma}{\lambda^2 + \beta^2}\, d\lambda d\sigma \tag{41}$$

Now the joint prior for λ and σ is

$$g(\lambda, \sigma) = \frac{\sigma}{\lambda^2 + \beta^2}\, f\left(\frac{\sigma}{\sqrt{\lambda^2 + \beta^2}}, \frac{\sigma\lambda}{\sqrt{\lambda^2 + \beta^2}}\right) \tag{42}$$

which is in general very far from being decoupled!

But there is still another error in what we have done, which Steve Gull recognized. We integrated out the X_i with respect to independent uniform priors on the grounds that our prior uncertainty δ is large compared to σ_x so its exact value does not matter appreciably. Steve senses correctly that something is wrong here, but fixes his attention on the matter of independence of the X_i. It is true that if we have prior information making them logically related, we should take this into account by a correlated prior, and this may enable us to get better estimates of α, β; but we think this is a detail, not the crucial point.

The crucial point is that if we use any fixed prior for X_i, then as the estimated line rotates we are expressing different states of knowledge about the positions of the 'true' points (X_i, Y_i) on it; and this is true *even in the limit where the prior is uniform.* To see this most clearly, suppose that instead of integrating out the X_i we had integrated out the Y_i with respect to independent uniform priors. A short calculation shows the surprising fact that we get a different posterior *pdf* for α, β:

************* MORE COMING HERE!! *****************

But this is just the original Gauss solution for the case that $\sigma_y = 0$; the errors are only in x_i. Merely by changing from a prior uniform in X_i to one uniform in Y_i the roles of x_i, y_i have somehow become interchanged. This is far more than we bargained for; the conventional folklore which says that prior information does not matter as long as the prior uncertainty is large compared to the width of the likelihood function, is surely true in the conventional situations one had in mind before; but now we see that the principle needs to be stated more carefully in problems with many parameters.

The reasoning here is much like that for location parameters: our prior information need not be translationally invariant, yet the likelihood function for a location parameter will have an obvious translational invariance; so a translationally invariant prior will lead to the simplest final solution. This being the case, unless we have cogent prior information which is not translationally invariant, it would be foolish not to use a uniform prior (uniform, that is, over an interval wide enough to include all the high likelihood region). Indeed, this would be actively dishonest, in effect claiming to have more information than we possess.

In our present problem, it is of course true that our prior information need not be rotationally invariant. Indeed, even if x and y are quantities of exactly the same kind, the mixture $z = x\cos\theta + y\sin\theta$ may be without any meaning. Nevertheless, in solving this problem we shall find ourselves dealing, inevitably, with at least the *mathematics* of rotations in the $x-y$ plane.

You can see from the beginning that this must be so, because with Gaussian sampling distributions, only the second central moments of the data will appear in the sufficient statistic. But the collection of all those second moments forms a symmetric second rank tensor, and its minimum eigenvalue is found by reduction to diagonal form, always accomplished by a rotation of the coordinate axes. Therefore rotations will appear naturally in the likelihood function whether or not our prior information has rotational invariance. (In fact, we shall find a rotational invariance property exactly analogous to the translational invariance with a location parameter.)

This being the case, a rotationally invariant prior will lead to the (analytically) simplest solution, so unless we have cogent prior information which is not rotationally invariant, it will be prudent, both for pragmatic and philosophical reasons, to use an invariant prior.

************************** MORE HERE! ************************

REDUCTION OF THE RESULT

Examine the modified quadratic form that the mathematics led us to:

$$\begin{aligned} F(\alpha,\beta) &\equiv \frac{Q(\alpha,\beta)}{1+\beta^2} \\ &= \frac{(s_{yy}-\hat\beta^2 s_{xx}) + (\alpha-\hat\alpha)^2 + 2\overline{x}(\alpha-\hat\alpha)(\beta-\hat\beta) + \overline{x^2}(\beta-\hat\beta)^2}{1+\beta^2} \end{aligned} \tag{43}$$

It is evident by inspection that this has a single unique minimum with respect to α; by differentiation we find this is reached when $\alpha - \hat\alpha = -\overline{x}(\beta-\hat\beta)$. So, keeping α constantly fixed at this value, (43) reduces to

$$F(\beta) = \frac{(s_{xx}-\hat\beta^2 s_{xx}) + s_{xx}(\beta-\hat\beta)^2}{1+\beta^2} \tag{45}$$

which looks quite complicated and unsymmetrical in terms of the parameter β. But β is a very unsymmetrical parameter; the real sense of the modified quadratic form appears if we set $\beta = \tan\theta$ and rewrite this in terms of the parameter θ. All the complications cancel out, and it reduces magically to the standard form

$$F(\beta) = S(\theta) = s_{xx}\sin^2\theta - 2s_{xy}\sin\theta\cos\theta + s_{yy}\cos^2\theta \tag{46}$$

which we recognize as a second rank tensor element s'_{yy} in a coordinate system (x', y') rotated an angle θ from the original one. Separating off the rotationally invariant part, this becomes

$$S(\theta) = \frac{1}{2}(s_{xx} + s_{yy}) + \frac{1}{2}(s_{yy} - s_{xx}) \cos^2\theta - s_{xy} \sin^2\theta \tag{47}$$

To find the maxima and minima of this, define an angle ϕ by

$$\left\{ \begin{aligned} s_{xy} &= R \sin\phi \\ \frac{1}{2}(s_{xx} - s_{yy}) &= R\cos\phi \end{aligned} \right\}, \qquad (-\pi < \phi \le \pi) \tag{48}$$

or,

$$\begin{aligned} \tan\phi &= \frac{2s_{xy}}{s_{xx} - s_{yy}} \\ 2R &= \sqrt{(s_{xx} - s_{yy})^2 + 4s_{xy}^2} \ \ge \ 0 \end{aligned} \tag{49}$$

Note that this defines the branch of the function so that ϕ has the same sign as s_{xy}, and if $s_{xx} > s_{yy}$, then $|\phi| < \pi/2$, otherwise $\pi/2 \le |\phi| \le \pi$. Now we have simply

$$S(\theta) = \frac{1}{2}(s_{xx} + s_{yy}) - R\cos(2\theta - \phi) \tag{50}$$

so the minimum is reached at $\theta = \hat\theta \equiv \phi/2$, the maximum at right angles, $\theta = \hat\theta \pm \pi/2$. If the prior for α, β is rotationally invariant, we shall then estimate β as $\tan(\phi/2)$; now let us determine the accuracy of that estimate.

Whatever the prior, the quasi-likelihood function, which contains all the information the data have to give us about θ, is now

$$L(\theta) = \left[\frac{1}{S(\theta)}\right]^{n/2} \tag{51}$$

**************************** MORE COMING! ********************************

But although the algebra tells us this, can we understand it intuitively in a way that makes it obvious from the start? Yes, and in fact the solution to a more general problem than the one just discussed is equally obvious, if we look at this way. Each data point has some error which may be in either x, or y, or both. But if we make the estimates $\hat\alpha, \hat\beta$, then the component of error parallel to that line contributes nothing to the error in our estimate of either α or β. Only the component of error perpendicular to the estimated line matters.

Now an error vector (e_i, f_i) has components parallel and perpendicular to a regression line of slope $\beta = \tan\theta$ of

$$e_i\cos\theta + f_i\sin\theta = \frac{e_i + \beta f_i}{\sqrt{1+\beta^2}}, \qquad -e_i\sin\theta + f_i\cos\theta = \frac{-\beta e_i + f_i}{\sqrt{1+\beta^2}} \tag{52}$$

respectively. But these have mean square values of

$$\frac{\sigma_x^2 + \beta^2\sigma_y^2}{1+\beta^2}, \qquad \frac{\sigma_y^2 + \beta^2\sigma_x^2}{1+\beta^2} \tag{53}$$

respectively. Therefore the quantity $\sigma^2 = \sigma_y^2 + \beta^2\sigma_x^2$ generated by our integration over the X_i was, essentially, just the mean square value of the perpendicular component of error.

Recognizing this, it is clear that we need not have considered independent sampling probabilities for e_i, f_i; whether the errors are in x, in y, or in both; and whatever the shape or orientation of the concentration ellipse, only the perpendicular component of the error can matter and we shall be led to the same result. The algebra, once we learn how to state the problem correctly, gives us this result so fast it seems like magic.

PRIOR PROBABILITIES AND TRANSFORMATION GROUPS

The transformation group principle for assigning priors in the regression problem is quite simple. Given the straight line equation $y = \alpha + \beta x$ with a prior probability element

$$f(\alpha, \beta)\, d\alpha d\beta \tag{54}$$

we formulate **Problem P:** Given a data set $D \equiv \{(x_1, y_1)\ldots(x_n, y_n)\}$, estimate α and β.

Now consider a related problem: carry out a linear coordinate transformation $(x, y) \to (x', y')$ such that in the new variables (54) takes the form

$$y' = \alpha' + \beta' x' \tag{55}$$

with a prior probability element

$$g(\alpha', \beta')\, d\alpha' d\beta' \tag{56}$$

and consider **Problem P$'$:** Given a data set $D' \equiv \{(x_1', y_1')\ldots(x_n', y_n')\}$, estimate α' and β'.

Now if the two priors (54), (56) express the same prior information, it must be true that

$$f(\alpha, \beta)\, d\alpha d\beta = g(\alpha', \beta')\, d\alpha' d\beta' \tag{57}$$

or,

$$f(\alpha, \beta) = g(\alpha', \beta')\, \frac{\partial(\alpha', \beta')}{\partial(\alpha, \beta)} \tag{58}$$

This transformation equation tells how the two problems are related to each other, and it will hold whatever linear transformation we carry out, and whether or not we consider the problems P, P$'$ to be equivalent.

But now suppose that our prior information is invariant under the transformation. For example, if x is the distance from our present location to some origin of whose location we know nothing, then after the transformation $x' = x + a$, of walking a distance a, we are in the same state of ignorance; ignorance of one's location is a state of knowledge which is not changed by a small change in that location.

REFERENCES

Bretthorst, G. L. (1988), *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture Notes in Statistics, Vol. 48, Springer–Verlag, Berlin.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.

Dawid, A. P., Stone, M. & Zidek, J. V. (1973), "Marginalization Paradoxes in Bayesian and Structural Inference", J. Roy Stat. Soc. **B35**, pp. 189-233.

Deming, W. E. (1943), *Statistical Adjustment of Data*, J. Wiley, New York.

de Groot, M. H. (1985), *Probability and Statistics*, 2nd Edition, Addison-Wesley Pub. Co., Reading MA.

Graybill, F. A. (1961), *An Introduction to Linear Statistical Models*, Mc Graw-Hill Book Co., New York.

Gull, S. F. (1989), "Bayesian Data Analysis: Straight Line Fitting", in *Maximum Entropy and Bayesian Methods*, J. Skilling, Editor, Kluwer Academic Publishers, Holland.

Isobe, T., Feigelson, E. D., Akritas, M. G. & Babu G. J., (1990), "Linear Regression in Astronomy", Astrophys. Jour. Nov 20.

Jaynes, E. T. (1980), "Marginalization and Prior Probabilities", in *Bayesian Analysis in Econometrics and Statistics*, A. Zellner, Editor, North-Holland Publishing Co., Amsterdam. Reprinted in Jaynes (1983).

Jaynes, E. T. (1986), "Bayesian Methods: General Background", in Proceedings of the Workshop on Bayesian/Maximum Entropy methods in Geophysical inverse problems, Calgary, Canada, August 1984; J. H. Justice, Editor, Cambridge University Press.

Jaynes, E. T. (1987) "Bayesian Spectrum and Chirp Analysis", in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith & G. J. Erickson (editors), D. Reidel Publishing Company, Holland; pp. 1–37.

Kendall, M. G. & Stuart, A. (1961), *The Advanced Theory of Statistics*: Volume 2, Inference and Relationship, Hafner Publishing Co., New York.

Kempthorne, O. & Folks, L. (1974), *Probability, Statistics, and Data Analysis*, Iowa State University Press, Ames IA.

Mandel, J. (1964), *The Statistical Analysis of Experimental Data*, Interscience, New York. Straight orthodox *adhockeries*, one of which is analyzed in Jaynes (1976).

Sokal, R. R. & Rohlf, J. J. (1981), *Biometry: The Principles and Practice of Statistics in Biological Research*, 2nd edition, W. H. Freeman & Co., San Francisco.

Strömberg, G. (1940), Astrophys. J. **92**, 156.

Wald, A. (1940), "The fitting of straight lines if both variables are subject to error" Ann. Math. Stat. **11**, 284-300.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, J. Wiley & Sons, Inc., New York. Second printing (1987); R. E. Krieger Pub. Co., Malabar, Florida.

Zellner, A. (editor, 1980), *Bayesian Analysis in Econometrics and Statistics*, North-Holland Publishing Co.