Robust Estimation When More Than One Variable Per Equation of Condition Has Error

by

William H. Jefferys
Department of Astronomy
The University of Texas at Austin

Summary

In a least squares adjustment when more than one variable in an equation of condition has error, the results will be affected by unnecessary asymptotic bias unless the algorithm is properly formulated. Similar difficulties can be expected with robust estimation techniques that are based on extending least squares to a noneuclidean metric. This paper presents an algorithm for robust estimation in the case that each equation of condition contains several observations. The properties of the estimates produced by the algorithm are investigated, and a numerical example using real data on galaxies is given.

Some key words: Robust estimation, Maximum likelihood, Errors in variables.

Received___________________

Submitted to Biometrika

(Refereed version appeared in
Biometrika, **77**, 3, pp. 597-607 (1990))

1. Introduction

As Lybanon (1984) has pointed out, it is not widely appreciated outside of the statistical community that a special treatment of the least squares problem is required when there is more than one observation having error per equation of condition. Failure to formulate the problem correctly will result in an asymptotically biased estimator, even when fitting a straight line. Earlier work (Madansky, 1959; Celmins, 1973, 1984; Britt & Leucke, 1973; Golub & van Loan, 1980; Jefferys, 1980, 1981; see also Kendall & Stuart, 1979, §29, and references therein) has established a very general algorithm for performing least squares adjustments under a number of conditions: more than one observation per equation of condition, correlated observations, nonlinear equations of condition, and exact constraints among parameters. Since robust estimators of the type discussed by Huber (1981) can be thought of as generalizations of least squares to a noneuclidean metric, it is to be expected that they too will be subject to unnecessary asymptotic bias when applied to this case. In this paper I extend the results of Britt & Leucke, Celmins, and Jefferys to the robust estimation case.

2. Formulation of the Problem

Usually the equations of condition for an estimation problem are written explicitly, with each equation having been solved for the “dependent variable,” that is, the observation, on the left hand side as a function of “independent variables,” that is, parameters, on the right. Thus, for example, Huber (1981) writes for the nonlinear regression case

$$y_i = f_i(\vartheta),\ i = 1,\ldots,\ p \qquad (1)$$

where the y_i are the observations, $\vartheta=(\vartheta_1,\vartheta_2,\ldots,\vartheta_m)$ is the vector of the parameters, and the f_i are a set of p independent functions connecting the parameters with the observations.

If some of the equations of condition contain more than one observation, it is not possible to write them in the form of Eqs. (1). A simple example is a straight line fit where both x and y have error:

$$y_i = \alpha+\beta x_i,\ i = 1,\ldots,\ p. \qquad (2)$$

Given such an equation of condition, we cannot solve uniquely for the observations in terms of parameters only. The best we can do is to bring everything over to the left-hand side and express the equation of condition implicitly:

$$y_i-\alpha-\beta x_i = 0, \; i = 1,..., p. \tag{3}$$

It would be wrong to solve Eqs. (3) by minimizing

$$\sum_{i=1}^{p}(y_i-\alpha-\beta x_i)^2 \tag{4}$$

with respect to α and β, since this will result in an asymptotically biased estimator (in particular, the slope of the line will be underestimated). To treat this case, we introduce a new set of variables $y_{p+1}, y_{p+2},..., y_{2p}$ by the definition

$$y_{p+i} = x_i, \; i=1,...,p \tag{5}$$

and rewrite Eqs. (3) as

$$y_i-\alpha-\beta y_{p+i} = 0, \; i = 1,...,p. \tag{6}$$

We can simplify the notation even more by renaming the parameters. Using a superscript T to denote transpose, we set $a^T=(\alpha,\beta)=(a_1,a_2)$. We can then write Eqs. (6) in the form

$$f_i(y,a) = y_i-a_1-a_2 y_{p+i} = 0, \; i=1,...,p, \tag{7}$$

where we have collected the n=2p observations into a vector $y^T=(y_1,y_2,...,y_n)$. Finally we can collect the individual equations of condition f_i into a vector $f^T=(f_1,f_2,...,f_p)$ of equations of condition and write

$$f(y,a)=0. \tag{8}$$

Equations (8) are a general and convenient vector notation that can be extended to express the equations of condition for a wide class of estimation problems, both linear and nonlinear. In what follows, we shall assume that the equations of condition for the problem have been put into this form.

The next step is to identify the observation errors explicitly. Let us write

$$y = Y-v, \tag{9}$$

where the Y are the actual observations and $v^T=(v_1,v_2,...,v_n)$ is the error vector, and substitute Eq. (9) explicitly into Eq. (8) obtaining f(y,a)= f(Y–v,a). We will also introduce a loss function S(v), and our object will be to find a relative minimum of the loss function S subject to the constraints that the equations of condition (8) are satisfied. We require that S(v) vanish quadratically at the origin with a nonsingular, positive definite, diagonal Hessian matrix. The restriction to a diagonal Hessian is not neccessary and we discuss later how the method may be generalized to the nondiagonal case.

As an example, let us introduce one of the standard robust metric functions ρ(u). Then some possible loss functions for the robustified linear problem posed by Eqs. (7) (not the only, nor necessarily even the best ones) might be

$$S(v) = \sum_{i=1}^{p}\Big[\rho(v_i/s)+\rho(v_{p+i}/s)\Big], \text{ and} \tag{10a}$$

$$S(v) = \sum_{i=1}^{p}\rho(\sqrt{v_i^2+v_{p+i}^2/s}), \tag{10b}$$

where s is a scale factor that has to be estimated. The scales of the x and y variates can without loss of generality be assumed equal, since if they are unequal, but related by a constant, known factor κ, then rescaling the x data will make them equal. See Brown (1982) for further discussion of this point, and also Section 5 of this paper. For the linear errors-in-variables model, the first of these formulas corresponds to the choice advocated by Brown (1982); the second minimizes the robustified perpendicular distance between the data point and the fitted line, and has been advocated for the linear errors-in-variables model by Zamar (1985, 1987), who calls it the orthogonal regression M-estimator model. Both reduce to the ordinary least squares errors-in-variables model when the metric is Euclidean . In more complicated problems, for example when the data points involve more than two quantities measured with error, each of Eqs. (10) has an obvious generalization (see, e.g., Zamar 1985, 1987).

To solve the minimization problem, we follow Deming (1938) and Brown (1952) in recognizing that the residual vector v as well as the vector a of explanatory parameters are both free variables. In the statistical literature the v are known as incidental parameters. Since this is a problem in constrained minima, we introduce a p-vector λ of Lagrange multipliers and then find an extremum, relative to v and a, of

$$\tilde{S}(v,a,\lambda) = S(v)-\lambda^T f(Y-v,a). \tag{11}$$

Taking the variation, this means solving the equation

$$d\tilde{S} = \frac{\partial S}{\partial v}dv + \lambda^T\left(\frac{\partial f}{\partial y}dv-\frac{\partial f}{\partial a}da\right) = 0 \tag{12}$$

simultaneously with Eqs. (8). Since the variations dv and da are arbitrary, it follows that the coefficients of the infinitesimal variations dv_i and da_i in Eqs. (12) must each vanish separately. Using subscripts for brevity to denote the required matrices of partial derivatives

$$f_y = \frac{\partial f(y,a)}{\partial y} , \quad S_v = \frac{\partial S}{\partial v}, \ \ldots$$

we arrive at the conditions

$$S_v+\lambda^T f_y = 0, \tag{13a}$$

$$\lambda^T f_a = 0. \tag{13b}$$

Since Eqs. (8) and (13) are in general nonlinear, it is necessary to solve them by a method of successive approximations. To accomplish this let us first rewrite Eqs. (8), assuming that we want to find an improved solution $(\hat{y}_{new},\hat{a}_{new})=(y,a)$ close to a preliminary solution $(\hat{y},\hat{a})$. Using hats to indicate evaluation on $(\hat{y},\hat{a})$, and expanding in small quantities we find

$$\begin{aligned} f(y,a) &= f(Y-\hat{v}-(v-\hat{v}),\hat{a}+\hat{\delta}) \\ &= f(Y-\hat{v},\hat{a})-\hat{f}_y(v-\hat{v})+\hat{f}_a\hat{\delta} \qquad (14) \\ &= 0, \end{aligned}$$

where $\hat{\delta}$ is the required correction to the parameter vector. In Eq. (14) the partial derivative matrices are evaluated, as indicated, on $(\hat{y},\hat{a})$, which is the current best estimate of the solution. Quadratic terms have been neglected, but they vanish anyway when we converge on the solution. Finally let us define

$$\hat{f} = f(Y-\hat{v},\hat{a}), \qquad (15a)$$

$$\hat{\varphi} = \hat{f}+\hat{f}_y\hat{v}, \qquad (15b)$$

so that Eqs. (13-14) become (after evaluating the derivative matrices on the estimated vectors)

$$\hat{\varphi}-\hat{f}_y v+\hat{f}_a\hat{\delta} = 0. \qquad (16a)$$

$$S_v+\lambda^T\hat{f}_y = 0, \qquad (16b)$$

$$\lambda^T\hat{f}_a = 0. \qquad (16c)$$

Several methods are available to solve Eqs. (16), which are analogs of the methods found in the literature on robust estimation. One is Newton’s method; another is to generalize the method of iteratively reweighted least squares to this case. We shall discuss each of these two methods in turn. I have also developed a third iterative scheme to solve Eqs. (16) that is similar to the approach described by Huber (1975), but it does not appear to have any advantages in this case, and so I have not described it here.

3. Solution using iteratively reweighted least squares

Define a diagonal matrix D by

$$v^T D(v) = S_v(v), \qquad (17)$$

which exists for all v since S vanishes quadratically at 0. Then Eq. (16b) can be written

$$v^T D = -\lambda^T\hat{f}_y, \qquad (18)$$

so that

$$v^T = -\lambda^T \hat{f}_y D^{-1}. \tag{19}$$

At this point we note that for some metrics (for instance, hard redescenders like Tukey’s biweight) D is formally singular for those observations that are larger than the cutoff implied by the “tuning constant.” In such a case we alter D by replacing the zeros on the main diagonal by a small ε. Then at the very end we can take the limit as $\varepsilon \to 0$. It turns out that this prescription gives us a well-defined solution, which is equivalent to dropping from the solution those equations of condition that give rise to zeros on the main diagonal of D.

Substitute Eq. (19) into Eq. (16a) to obtain

$$\hat{\varphi} + \hat{f}_y D^{-1} \hat{f}_y^{\,T} \lambda + \hat{f}_a \hat{\delta} = 0 \tag{20}$$

Eq. (20) is solved for λ to give

$$\lambda = -\, W(\hat{\varphi} + \hat{f}_a \hat{\delta}), \tag{21}$$

where

$$W = (\hat{f}_y D^{-1} \hat{f}_y^{\,T})^{-1} \tag{22}$$

is a “weight matrix.” The terms of order $1/\varepsilon$ in Eq. (19) give rise to terms of order ε in Eq. (22) and vanish when the limit $\varepsilon \to 0$ is taken. This intuitively corresponds to the zero weighting that hard redescenders apply to equations of condition that involve residuals greater than the cutoff.

Substituting Eq. (21) into Eq. (16c), we arrive finally at the “normal equations”

$$(\hat{f}_a^T W \hat{f}_a) \hat{\delta} = -\, \hat{f}_a^T W \hat{\varphi}. \tag{23}$$

Finally, substituting Eq. (21) into Eq. (19), and dropping the term in $\hat{\delta}$ (which goes to zero anyway as the process is iterated) we obtain

$$\hat{v}_{new} = v = D^{-1} \hat{f}_y^T W \hat{\varphi}. \tag{24}$$

In Eq. (24), the terms in ε cancel those in $1/\varepsilon$, giving a finite result.

We now have all the equations we need for an iteration procedure. Given an approximate solution $(\hat{y},\hat{a})$ (we start with $\hat{y}=Y$, $\hat{v}=0$ unless a better value is known a priori), we plug into Eq. (15b) to obtain $\hat{\varphi}$. This is inserted into Eqs. (23) and (24) to obtain an updated $\hat{v}$ and a $\hat{\delta}$. Then the vector $\hat{a}$ is updated using $\hat{a}_{new} = a = \hat{a}+\hat{\delta}$. At the same time we update $\hat{y}$ using $\hat{y}_{new} = Y-\hat{v}_{new}$ Note that $\hat{v}_{new}$ is subtracted from the observation vector, not from the previous best-estimate vector $\hat{y}$. At this point the first iteration is complete, and subsequent iterations proceed in the same way until the process converges. In my software, convergence is assumed when the changes in $\hat{a}$ and $\hat{y}$ from one iteration to the next are smaller than a preset amount.

In solving Eqs. (16) as we have done, it is important that the matrices and vectors be evaluated using the most recently updated values of $\hat{y}$ and $\hat{a}$. Otherwise, an unnecessary source of bias will be introduced.

4. Solution using Newton’s Method

The solution using Newton’s method is less useful because it requires second derivatives of the metric function $\rho(\ u)$, which vanish with large u for many useful metrics. Unlike the iteratively reweighted least squares method, there does not appear to be a simple “limiting case” that can be applied to sidestep the singularities that arise as a result. Even for metrics that do not have a vanishing second derivative, numerical experience shows that Newton’s method sometimes diverges when iteratively reweighted least squares converges. On the other hand, it sometimes converges more rapidly than iteratively reweighted least squares. Therefore it may be useful despite its limitations, and the equations are derived below.

Let H be the Hessian matrix of S(v):

$$H_{ij} = \frac{\partial^2 S}{\partial v_i \partial v_j} \tag{25}$$

Then setting $v = \hat{v} + \Delta\hat{v}$, we can rewrite Eq. (16b) in the form

$$\hat{S}_v^T + \hat{H}\Delta\hat{v} + \hat{f}_y^T\lambda = 0, \tag{26}$$

where as usual the hats mean evaluation on hatted variables, so that

$$\Delta\hat{v} = -\hat{H}^{-1}(\hat{f}_y^T\lambda+\hat{S}_v^T). \tag{27}$$

Also, Eq. (16a) becomes

$$\hat{f}-\hat{f}_y\Delta\hat{v}+\hat{f}_a\hat{\delta} = 0. \tag{28}$$

Inserting Eq. (27) into Eq. (28), after some manipulation there results

$$\lambda = -W_n(\hat{\varphi}_N+\hat{f}_a\hat{\delta}), \tag{29}$$

where the “weight matrix” W_n is now

$$W_n = (\hat{f}_y\hat{H}^{-1}\hat{f}_y^T)^{-1} \tag{30}$$

and the right hand side vector $\hat{\varphi}_n$ is given by

$$\hat{\varphi}_n = \hat{f}+\hat{f}_y\hat{H}^{-1}\hat{S}_v^T. \tag{31}$$

Finally, this is inserted into Eq. (16c), giving the “normal equations”

$$(\hat{f}_a^T W_n \hat{f}_a)\hat{\delta} = -\hat{f}_a^T W_n \hat{\varphi}_n, \tag{32}$$

Finally, combining Eqs. (27) and (29), and dropping the term in $\hat{\delta}$ which goes to zero upon iteration, we obtain a formula for $\Delta\hat{v}$:

$$\Delta\hat{v} = \hat{H}^{-1}\hat{f}_y^T W_N \hat{\varphi}_n. \tag{33}$$

It is easily seen that Eqs. (31-33) bear a very strong resemblance to Eqs. (15b), (23) and (24), respectively. They can be set up and solved by appropriate modifications of the same software.

The iteration procedure for Newton’s method now proceeds as follows: Starting at an approximate solution $(\hat{y},\hat{a})$, we use Eq. (31) to determine $\hat{\varphi}_n$. Eq. (33) gives us $\Delta\hat{v}$, and Eq. (32) gives us $\hat{\delta}$. Finally as before we set $\hat{a}_{new} = a = \hat{a}+\hat{\delta}$, and (in a change from the iteratively reweighted least squares method) $\hat{v}_{new} = v = \hat{v}+\Delta\hat{v}$, $\hat{y}_{new} = y = \hat{y}-\Delta\hat{v}$. The iteration now complete, we repeat until convergence.

A comparison of Eqs. (30), (31) and (33) shows that this method has difficulties when the second derivative of the metric function vanishes. In this case, one of the terms in Eq. (33) is $O(\hat{H}^{-1})$, and since some eigenvalues of $\hat{H}$ will vanish if the second derivative of the metric function vanishes, the method will fail. It is true that for large u the vanishing of $\rho''(u)$ is equivalent to saying that the residual u has no influence on the solution; and it may be that simply excluding such observations from the solution will overcome this difficulty. However, this hypothesis has not been tested.

5. Consistency and the Choice of Loss Function

The choice of the loss function $S(v)$ is an important one. As a general principle one would want to demand that the loss function go over to the standard one for the errors-in-variables model when the metric function is Euclidean. Two possible choices that satisfy this criterion are given by Eqs. (10a-b). The first of these choices has been advocated by Brown (1982), but in a correction to his original paper he reported a counterexample to his consistency proof (Brown 1983). However, the counterexample he gave in fact only demonstrates a flaw in his consistency proof. Ironically, if Brown’s method is used on the counterexample, it nevertheless estimates the slope of the line consistently, a point that Brown appears to have missed. This turns out to be due to the special case that Brown considered—namely, fitting a line with unit slope and equal variances in the two coordinates. It turns out that Brown’s estimator is indeed inconsistent for an arbitrary slope (Wang, 1988). This is certainly a point against the use of this estimator.

Zamar (1985, 1987) has reported that the convergence properties of Brown’s estimator are not good, although I have had no difficulty with it. Zamar advocates the choice of Eq. (10b), which he has shown to be a consistent estimator in the linear errors-in-variables case when fitting a hyperplane in n-dimensional space under suitable conditions. In particular, Zamar assumes that the error distribution is spherically symmetric. Zamar says that this method has better convergence properties than does a method based on Eq. (10a).

It is interesting to notice that although both Eq. (10a) and Eq. (10b) go over to the standard least squares errors-in-variables model when the metric function $\rho(u)$ becomes Euclidean, they have very different properties when the metric is far from Euclidean. As an extreme example, consider the case when the metric function is the L_1 metric generated by $\rho(u)=|u|$. In this case, for a point lying away from the fitted line, Eq. (10a) possesses infinitely many solutions for the position of the fitted point when the slope is equal to 1 (Figure 1), and very different solutions for slopes that differ only slightly from this value, depending on whether the slope is less than 1 (Figure 2) or greater than 1 (Figure 3). On the other hand, Eq. (10b) always gives a unique solution for the fitted point that varies continuously with the slope (Figure 4), as does the least squares solution. In some sense, therefore, Eq. (10b) is closer to least squares than is Eq. (10a).

Another aspect of Eq. (10b) that makes it superior to Eq. (10a) is the fact that the metric of Eq. (10b) is invariant under a larger group of coordinate transformations than is that of Eq. (10a). Specifically, Eq. (10b) is invariant to rotations, whereas Eq. (10a) is not. This means that Eq. (10b) has a geometrically invariant meaning that is not shared by Eq. (10a).

If we adopt the principle that we should choose that metric that maintains as closely as possible the properties of the least squares metric, while still possessing the desired robust characteristics, then it would appear that grouping the data by observation point and measuring the error as a function of the orthogonal distance of the measured point to the fitted curve, as in Eq. (10b), is superior to treating each observed datum separately, as in Eq. (10a). The odd behavior of Eq. (10a) illustrated in Figures 1-3 may be related to the difficulties that have been reported as to its convergence properties, and the geometric invariance of Eq. (10a) under the rotation group is a decided advantage, since physically we would want our answers to be independent of the choice of coordinate system.

These remarks define the behavior of robust estimators in the errors-in-variables model when the surface being fitted is a hyperplane, but they do not address the problem of fitting a more general nonlinear function. Unfortunately, it is clear that in the latter case estimators of the type discussed in this paper cannot be consistent in general. This is true even if the metric function is Euclidean and the distribution of the errors is normal. This is easily be seen from a counterexample, and the counterexample can give us an idea of the asymptotic bias of the ordinary orthogonal regression least squares model in this case. This in turn can warn us when this method cannot be applied safely.

Let (ξ,η) be a set of measurements of points on a circle, the position of whose center is known. The “true” position of each point is (X,Y), and the errors $\xi-X$, $\eta-Y$ are $N(0,\sigma^2)$. An astronomical example of how such data might arise involves measurements of the relative positions of two members of a binary star system (assumed in a circular orbit with a face-on orientation). If no information were given to fix the time of each observation, we would still be able to estimate the radius of the orbit from these data. If R_0 is the “true” radius of the circle, then we will see that the estimated radius $b_n \rightarrow b > R_0$ in probability, and therefore the orthogonal regression least squares estimator of the radius of the circle is inconsistent.

Figure 5 shows the geometry of the problem. ρ is the distance from the observed point (ξ,η) to the “true” point (X,Y), and ϑ is the angle between the line from the center of the circle to the “true” point and the line from the “true” point to the observed point. The distance from the center of the circle to the “true” point is therefore r, where

$$r = R_0\sqrt{1+\alpha^2-2\alpha \cos \vartheta} = r(\rho,\vartheta) \tag{34}$$

independently of (X, Y), and

$$\alpha = \rho/R_0 \ . \tag{35}$$

The radius estimated from the orthogonal regression least squares model will converge in probability to the value of b that minimizes $E(b-r(\rho,\vartheta))^2$ (Wald 1949, Huber 1967). Since the expectation is a smooth function of b, the condition for a minimum is just

$$0 = \frac{\partial}{\partial b} E(b-r(\rho,\vartheta))^2 = 2\ E(b-r)\ , \tag{36}$$

or

$$b = E(r) . \tag{37}$$

Now if the errors in the (ξ,η) are normally distributed we can write:

$$\begin{aligned} E(r) &= \frac{1}{2\pi\sigma^2}\int_0^\infty d\rho\ \rho\ \exp\left(-\frac{\rho^2}{2\sigma^2}\right)\int_0^{2\pi} d\vartheta\ r(\rho,\vartheta) \\ &= \frac{1}{2\pi}\int_0^\infty du\ e^{-u}\int_0^{2\pi} d\vartheta\ r(\rho,\vartheta) \qquad (38) \\ &= \frac{1}{2\pi}\int_0^\infty du\ e^{-u}\int_0^{\pi} d\vartheta\ [r(\rho,\vartheta)+r(\rho,\vartheta+\pi)]. \end{aligned}$$

It is readily verified that for $\rho>0$ the integrand of the inner integral on the last line of Eqs. (38) satisfies

$$r(\rho,\vartheta)+r(\rho,\vartheta+\pi) \geq 2R_0 , \tag{39}$$

with equality possible only for $\vartheta=0,\pi$. Since the inequality is strict on all but a zero measure subset of the integrand's support, it follows that for $\rho>0$ we have:

$$\frac{1}{2\pi}\int_0^{2\pi} d\vartheta\ r(\rho,\vartheta) > R_0, \tag{40}$$

from which it follows immediately that whenever $\sigma>0$ we also have

$$b = E(\ r) > R_0 . \tag{41}$$

To gauge the degree to which the estimator fails to be consistent, we can obtain an asymptotic expression by assuming that $\alpha<<1$, and expanding the radical in Eq. (34) in powers of α, retaining terms through the second order. (This expansion fails for large α, i.e., large ρ, but the exponential dominates in this region and therefore the contribution to the integral from this region is asymptotically negligible.) Inserting this expression into Eq. (38), we can evaluate the resulting expression explicitly and finally arrive at

$$b \approx R_0 \left(1+\frac{1}{2}(\sigma/R_0)^2\right) . \tag{42}$$

This expression shows that the amount by which the orthogonal regression least squares estimator fails to be consistent in this case is quadratically small in the ratio (σ/R_0). Even moderately large scatter (say of the order of 10% of the radius of curvature R_0) may contribute only a negligible amount (in this case less than 1%) to the asymptotic bias of the estimator. Of course, it would be better if the estimator were consistent, but there does not seem to be a simple way to accomplish this in general. The bias correction is evidently a function of the actual distribution of the data, and if we replace the orthogonal regression least squares estimator by the corresponding orthogonal regression M-estimator estimator, it may also depend on the choice of metric function. It will depend in a complicated way on both of these, and since the true distribution is usually unknown, a general formula for the asymptotic bias would appear to be unattainable. Nevertheless, Eq. (42) provides a practical method for estimating the size of the bias, and hence of warning the user when the use of this method is likely to lead to trouble. This remark applies to both the robust and nonrobust case.

One approach to reducing the bias of these estimators may be to use methods of Fuller (1987; §3.2.4). Fuller suggests bias-correction terms that can be added to the equations of condition in the orthogonal regression least squares case. These terms depend on the second derivatives of the equations of condition, and so are messy to evaluate. However, it may be that the methods Fuller advocates can be adapted to the robust nonlinear orthogonal regression M-estimator estimation problem. This will be the subject of further research.

Another way to look at the asymptotic bias of the orthogonal regression least squares estimator when the function being fitted is nonlinear is to perform a suitable coordinate transformation that “straightens out” the fitted curve in the neighborhood of the data points. In the circle example this can only be done locally of course, but if we introduce a suitable coordinate patch about the point (X,Y), we find that the error distribution in these “straightened out” coordinates is not spherically symmetric with regard to the natural Euclidean metric of the patch coordinates. Since the ordinary orthogonal regression least squares estimator is asymptotically biased when fitting a straight line if the error distribution of the data is not spherically symmetric (Zamar, 1985) it is not surprising that the nonlinear example suffers similarly. In his paper Zamar presented numerical results demonstrating that for linear fits in the presence of asymmetric noise, the robustified orthogonal regression M-estimator estimator can dramatically reduce the asymptotic bias as compared to the corresponding orthogonal regression least squares estimator. Thus there is reason to believe that the same might be true for robustified nonlinear fits using the method of this paper. This will be the subject of future research.

6. Scale determination

Thus far in the discussion I have ignored the problem of scale determination. Now it is time to remedy this deficiency. The literature on robust estimation has many examples of robust scale estimators, and in my software the scale s has been estimated by solving the equation

$$\frac{1}{(p-k)I} S(v) \equiv 1 \qquad (43)$$

implicitly for s, which appears in the sum S(v) through Eqs. (10). In the software, the solution for s is made simultaneously with the other parameters. In Eq. (43), the factor I is a normalizing factor that depends on the desired asymptotic relative efficiency (ARE) and the particular metric function used. p is the number of equations of condition, i.e., the dimension of f, and k is the number of explanatory parameters, i.e., the dimension of a. Rey (1983) presents a selection of metrics and normalizing factors. This choice of s is not the only useful one. For example, Brown (1982) advocates using the median absolute deviation, whereas Huber (1975) discusses still other strategies.

7. Correlated Observations

The discussion so far has assumed that the observations y are uncorrelated and of unit weight so that the errors v satisfy $<vv^T>=I$, where I is the identity matrix, corresponding to the Hessian matrix of S(v) being a multiple of the unit matrix at v=0. This is in general not the situation, and so it is useful to consider the case of correlated observations, which includes unequal weights as a special case. To accomplish this, let us express our problem in terms of a correlated set of variables x=Qy so that the covariance matrix of the y’s is I and the covariance matrix of the x’s is $\sigma=QQ^T$. Similarly, we write for the correlated residuals u=Qv. Thus, if we assume a covariance matrix σ to be given a priori, then Q can be chosen to be any square root of σ. As σ is normally block-diagonal with only small matrices along its diagonal, its square root is easily obtained by any one of a number of methods, e.g., Cholesky decomposition of the blocks. If the observations are unequally weighted but uncorrelated, then σ is diagonal and Q can be chosen diagonal.

Since the new variables x are presumably closer to the actual problem than the normalized variables y, it is to be expected that the equations of condition will be expressed directly in terms of them, e.g., $g(x,a)=f(y,a)=f(Q^{-1}x,a)$. All that remains is to reexpress the problem explicitly in terms of g, x, and u, using Q to transform between the old and new variables. For the iteratively reweighted least squares solution, this results in the following equations:

$$\hat{f}_y = \hat{g}_x Q \tag{44a}$$

$$W = (\hat{g}_x Q D^{-1} Q^T \hat{g}_x^T)^{-1} \tag{44b}$$

$$\hat{\varphi} = \hat{g} + \hat{g}_x \hat{u} \tag{44c}$$

$$\hat{u} = Q D^{-1} Q^T \hat{g}_x^T W \hat{\varphi} \tag{44d}$$

$$(\hat{g}_a^T W \hat{g}_a)\hat{\delta} = -\hat{g}_a^T W \hat{\varphi} \tag{44e}$$

The solution by Newton’s method is left as an exercise for the reader.

8. Solution by Orthogonal Transformations

Although I have solved the minimization problem in a manner that mimics the use of normal equations in least squares problems, it is very easy to adapt the equations for solution using orthogonal transformations (Golub 1965; Golub & Reinsch 1970; Lawson & Hanson 1974) which may be advantageous because of the greater numerical stability of this method and its ability to overcome problems due to poor observability of parameters. The principle is very simple. Only Eq. (44e) or its equivalent has to be changed. First find a square root U of W:

$$W = U^T U \tag{45}$$

The matrix U has to be computed from W, but this is not difficult because in typical problems W is strongly block-diagonal and the square root operation only has to be performed on the individual blocks. In the practical problems I have solved where at most 4 observations appear in no more than 2 simultaneous equations of condition, the largest matrix whose square root is required has been of order 2.

Once we have U, it remains to set up the orthogonal decomposition problem that is equivalent to the least squares problem of Eq. (44e). That problem can be written

$$\|U\hat{g}_a\hat{\delta}+U\hat{\varphi}\| = \min. \tag{46}$$

where the vector $\hat{\delta}$ is the adjustable parameter. Except for replacing Eq. (44e) by the equivalent problem posed in Eq. (46), the iterative procedure remains unchanged. Thus, one bases the solution on a QR decomposition or a singular value decomposition of the matrix $U\hat{g}_a$.

9. Exact Constraints

In practical applications it frequently happens that the problem involves exact constraints among the parameters. This may come about because of physical considerations, or it may happen because the problem is most conveniently set up by using redundant parameters and then constraining them appropriately. In either case, it is useful to formulate the problem so as to handle this case.

Let the constraints be expressed as

$$h_j(a_1,a_2,...) = 0, \ j=1,2, \ ...,r \tag{47}$$

or equivalently as

$$h(a) = 0. \tag{48}$$

Then we introduce an additional set of Lagrange multipliers in the form of an r-vector μ, which after linearization results in the equations

$$(\hat{g}_a^T W\hat{g}_a)\hat{\delta}+\hat{h}_a^T\mu = -\hat{g}_a^T W\hat{\varphi} \tag{49a}$$

$$\hat{h}_a\hat{\delta} = -\hat{h} \ . \tag{49b}$$

which replace the normal equations (44e) and are solved for μ and $\hat{\delta}$ simultanously.

Similarly, exact constraints can also be handled easily when orthogonal decomposition is used to solve the “least squares” problem. As in the unconstrained case, the basic strategy is to reduce the unconstrained problem to a problem that looks like a problem in ordinary least squares, and then recast it in terms of orthogonal decompositions. Then the constraints are added to the problem. There are a number of approaches to adding the constraints, and the reader is referred to Lawson & Hanson (1974) for details.

Finally, note that when there are constraints, the number of degrees of freedom in the denominator of Eq. (43) is changed. In this case, the denominator should be written (p–m+r)I.

10. A Numerical Example

Instead of generating artificial data for the numerical example, I have chosen to present a problem with real data from the astronomical literature. Real data often have characteristics that are not well mimicked by artificial data. The data presented in Table 1 are from Dressler (1984). They consist of the integrated V magnitudes (V_{26}) and log of the central velocity dispersion (log σ) of a sample of 53 galaxies from two galaxy clusters, the Coma and Virgo clusters. According the the Faber-Jackson relation (Faber & Jackson, 1976), the relation between these two quantities is roughly linear, having the form

$$\log \sigma = a+bV_{26}, \tag{50}$$

where the parameter a depends on the distance to the cluster and b is a constant. Since there are two clusters, there are two distances and therefore two independent values of a. These data are presented graphically in Figure 6. In that figure the separation between the two clusters is clearly evident. Dressler identified four obvious outliers, two from each cluster, which are noted in the figure by their catalog numbers. From their position at the bottom of the figure, it is tempting to surmise that the outliers are actually a consequence of a physical deviation from the linear Faber-Jackson relation at the lower (fainter) end, but we have no theory to guide us on this point. Whatever the reason for the deviations of these points away from the general trend, they are outliers insofar as the linear Faber-Jackson relation is concerned. Dressler’s trend lines for the Faber-Jackson relation for these two clusters are also indicated in the figure. He obtained these lines by assuming a constant offset in magnitude between the two clusters and correcting the magnitudes of one of the clusters by this offset so as to superimpose the data points. He then estimated the slope of the relationship graphically (ignoring the four outliers), and finally drew a separate median line having that slope for each cluster.

The solutions reported here were all made with the program GaussFit (Jefferys, Fitzpatrick, & McArthur, 1988), which implements the algorithms described in this paper. In each case, a three-parameter reduction was performed that estimated the common slope b simultaneously with the values of a for each of the two clusters. The results are presented in Table 2. The first solution, which is taken to be the reference solution, was an orthogonal regression least squares fit to all the data points excepting the four outliers identified by Dressler. (In all solutions, the standard error in the V_{26} data was assumed to be ±0.125 magnitudes, and in log σ it was assumed to be ±0.02.) The parameters of this solution are nearly identical to the ones Dressler found. The second solution is an orthogonal regression least squares solution including all 53 data points. This solution is quite poor, with the parameters lying between two and three standard deviations away from the reference solution.

For comparison, a group of similar solutions was run, omitting in turn the ith data point from the data set, to produce a set of estimates $T_{n,-i}$ of each parameter. The rows marked “minimum” and “maximum” give the minimum and maximum of the $T_{n,-i}$ from this sequence of runs, and the row marked “range” gives the difference between the maximum and minimum values of $T_{n,-i}$. These rows allow one to judge the sensitivity of the solution to the individual data points. The largest deviations were found when the outliers were omitted. Next, pseudovalues P_i for each value of i were generated in the usual way using the formula $P_i=nT_n-(n-1)T_{n,-i}$, where T_n is the estimate from all the data points. Finally, Quenouille’s jackknife was applied to the pseudovalues and a jackknifed estimate of each parameter and its standard deviation was produced. In this way I obtained an estimate of the standard deviation of each parameter estimate using a different method than the usual one involving the inverse matrix of the normal equations. These estimates of the standard deviation turned out to be somewhat larger than the ones estimated from the least squares solution. The jackknifed estimates of the parameters and their standard deviations appear in the last two rows of the orthogonal regression least squares solution.

The standard jackknife is non-robust, and in order to obtain an impartial robust estimate of the parameters by a method different from the one advocated in this paper, the pseudovalues from the orthogonal regression least squares solution using all data points were subjected to the “trimmed jackknife” procedure of Hinkley & Wang (1980). See also Hinkley (1978). Two trimming parameters were used, 5% and 10% (see column 2). The standard deviations of the resulting parameter estimates were also calculated using Hinkley & Wang’s formulas. The resulting estimates were substantially closer to the reference solution than either the orthogonal regression least squares or jackknifed orthogonal regression least squares solutions. They are shown in the third group of rows in Table 2.

The last three groups of solutions are robustified orthogonal regression M-estimator solutions using the methods of this paper. Three different metrics have been used: Huber’s metric (Eq. 51), Tukey’s biweight (Eq. 52) and the metric “fair” (Rey, 1983) (Eq. 53). The adjustable parameter c of each metric was chosen so as to provide an ARE of 0.9 or 0.8, relative to normally distributed data, as shown in the second column of Table 2.

$$\rho(u) = \begin{cases} u^2, & \text{if } |u| \leq c \\ c(2|u|-c), & \text{if } |u| \geq c \end{cases} \tag{51}$$

$$\rho(u) = \begin{cases} (c^2/3)(1-[1-(u/c)^2]^3), & \text{if } |u| \leq c \\ (c^2/3), & \text{if } |u| \geq c \end{cases} \tag{52}$$

$$\rho(u) = 2\,c^2[|u|/c-\log(1+|u|/c)] \tag{53}$$

As before, pseudovalues were generated for each of the solutions, and the standard jackknife was used to estimate values and, more importantly, the standard deviations for each robustified solution. The jackknifed solutions are not much different from the robustified solutions on which they are based, except in the case of the Tukey metric, for which a further improvement was evident. In the case of an asymptotic relative efficiency of 0.8, the jackknifed solution is indistinguishable from the reference solution. This improvement is a serendipitous result not related to the purpose of this paper, since the principal reason for generating the jackknifed solutions was to estimate the standard deviations of the parameters, and not to make further improvements in the parameters themselves.

The Huber and “fair” metrics give results that are comparable to those of the trimmed jackknife. The Tukey metric gave even better results due to its stronger suppression of outliers. Indeed, the Tukey metric for an ARE of 0.8 gave a result that is nearly as good as the reference solution.

This example shows that the methods of this paper can substantially improve parameter estimates when the underlying data are contaminated by outliers. As it happens, the outliers in this example are particularly severe as all of them pull the solution in the same direction, and all are located so as to exert maximum leverage on the slope of the Faber-Jackson relation and therefore on all of the estimated quantities. Despite these difficulties, the algorithm made a considerable improvement in the result.

11. Conclusions

I have described an approach to robust estimation that can handle the situation when the equations of condition may contain more than one observation. Several methods to solve the resulting equations are given, which reduce to a simple modification of the classical least squares problem. By adapting existing least squares software, the algorithms have been implemented straightforwardly on a digital computer.

The methods described in this paper have been tested on a number of problems, both linear and nonlinear, using real as well as simulated data. The numerical example given in this paper, using real data on galaxies, shows that the method can substantially improve the estimates of the parameters of a problem when outliers are present.

I find that the iteratively reweighted least squares method of solution has the most dependable convergence, whereas the solution by Newton’s method sometimes diverges, probably due to the smallness of some of the eigenvalues of H. On the other hand, Newton’s method occasionally converges more rapidly than iteratively reweighted least squares. Therefore it is difficult to recommend one method over the other.

Acknowledgements

It is a pleasure to thank the many colleagues who have commented on this work, especially David Hinkley, who has given willingly and unstintingly of his time, and without whose interest this paper could not have been brought into its present form; Ray Carroll, for important critical comments on an early draft of the paper and for subsequent conversations; Sheng-Guang Shi and Suo-Jin Wang, who tutored me on the subject of consistent estimators; David Evans and Gerard deVaucouleurs, for suggesting sources of data for the numerical example, and Barbara McArthur and Michael Fitzpatrick, who programmed much of the GaussFit system that includes the algorithms described here. Finally, I am grateful to an anonymous referee, whose suggestions have substantially improved the final paper.

I am grateful to the National Aeronautics and Space Administration, which supported this research in connection with the Hubble Space Telescope Astrometry Team project under NASA Contract NAS8-32906.

References

Britt, H. I., & Luecke, R. H. (1973). The Estimation of Parameters in Nonlinear, Implicit Models. *Technometrics* **15**, 233-247.

Brown, D. (1955). A Matrix Treatment of the General Problem of Least Squares Considering Correlated Observations. *Ballistic Research Laboratories Report* No. 937. Aberdeen, Maryland: Aberdeen Proving Gound.

Brown, M. L. (1982). Robust Line Estimation With Errors in Both Variables. *J. Am. Statist. Assoc.* **77**, 71-79.

Brown, M. L. (1983). Robust Line Estimation With Errors in Both Variables. *J. Am. Statist. Assoc.* **78**, 1008.

Celmins, A. (1973). *Ballistic Research Laboratories Report* No. 1658. Aberdeen, Maryland: Aberdeen Proving Gound.

Celmins, A. (1984). Analysis of Residuals from Multidimensional Model Fitting. *Computers and Chemistry* **8**, 81-89.

Deming, W. E. (1938). *Statistical Adjustment of Data*. New York: John Wiley & Sons, Inc.

Dressler, A. (1984). Internal Kinematics of Galaxies in Clusters. I. Velocity Dispersions for Elliptical Galaxies in Coma and Virgo. *Astrophys. J.* **281**, 512-524.

Faber, S. M. & Jackson, R. (1976). Velocity Dispersions and Mass-to-Light Ratios for Elliptical Galaxies. *Astrophys. J.* **204**, 664-683.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.

Golub, G. H. (1965). Numerical Methods for Solving Linear Least Squares Problems. *Numer. Math.* **7**, 206-216.

Golub, G. H. & Reinsch, C. (1970). Singular Value Decomposition and the Least Squares Problem. *Numer. Math.* **14**, 403-420.

Golub, Gene H. & van Loan, Charles F. (1980). An Analysis of the Total Least Squares Problem. *SIAM J. Numer. Anal.* **17**, 883-893.

Hinkley, D. V. (1978). Improving the jackknife with special reference to correlation estimation. *Biometrika* **65**, 13-21.

Hinkley, D. V. & Wang, H. L. (1980). A Trimmed Jackknife. *J. R. Statist. Soc.* B, **42**, 347-356.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1. Berkeley: Univ. of California Press.
.

Huber, P. J. (1975). Robust Methods of Estimation of Regression Coefficients. Presented to the Second International Summer School of Problems of Model Choice and Regression Analysis at Rheinshardsbrum, G.D.R., November 8-18.

Huber, P. J. (1981). *Robust Statistics.* New York: John Wiley & Sons, Inc.

Jefferys, W. H. (1980). On the Method of Least Squares. *Astron. J.* **85**, 177-181.

Jefferys, W. H. (1981). On the Method of Least Squares. II. *Astron. J.* **86**, 149-155.

Jefferys, W. H., Fitzpatrick, M. J. & McArthur, B. E. (1988). GaussFit—A System for Least Squares and Robust Estimation. *Celestial Mechanics* **41**, 39-49.

Kendall, M. & Stuart, A. (1979). *The Advanced Theory of Statistics* Vol. 2. (Fourth Edition). New York: Macmillan.

Lawson, C. L. & Hanson, R. J. (1974). Solving Least Squares Problems. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Lybanon, M. (1984). A Better Least-Squares Method when Both Variables Have Uncertainties. Am. J. Physics **52**, 22-26.

Madansky, A. (1959). The Fitting of Straight Lines when Both Variables are Subject to Error. J. Am. Statist. Assoc. **54**, 173-205.

Rey, W. J. J. (1983). Introduction to Robust and Quasi-Robust Statistical Methods. Berlin: Springer-Verlag.

Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. Ann. Math. Statist. **20**, 595-601.

Wang, S. J. (1988). Private Communication.

Zamar, R. H. (1985). Orthogonal Regression M-estimators, Department of Statistics Tech. Report No.62. Seattle: University of Washington.

Zamar, R. H. (1987). Robust Estimation in the Errors in Variables Model. Submitted to Biometrika.

Coma Sample				Virgo Sample			
NGC or	Dressler			NGC or	Dressler		
IC No.	No.	V_{26}	log σ	IC No.	No.	V_{26}	log σ
N4839	31	12.60	2.449	N4168	…	11.39	2.242
N4926	49	13.12	2.394	N4239	…	12.53	1.716
I3959	69	14.23	2.285	N4365	…	9.98	2.412
I3957	70	14.86	2.166	N4374	…	9.37	2.480
	87	15.88	1.863	N4387	…	12.24	2.059
N4869	105	13.92	2.286	N4406	…	9.20	2.355
	107	15.45	1.761	N4434	…	12.17	2.009
N4906	118	14.36	2.209	N4458	…	12.01	1.949
N4898E	120	15.07	2.113	N4464	…	12.50	2.079
N4898W	121	14.07	2.301	N4472	…	8.56	2.474
N4876	124	14.53	2.243	N4473	…	10.28	2.268
	125	15.60	2.169	N4478	…	11.28	2.170
N4874	129	12.27	2.383	N4486	…	8.79	2.528
N4872	130	14.36	2.311	N4489	…	12.02	1.778
N4867	133	14.50	2.339	N4551	…	11.92	2.021
	136	15.52	2.251	N4552	…	9.95	2.391
I4051	143	13.46	2.361	N4564	…	11.30	2.185
N4889	148	11.85	2.584	N4621	…	9.88	2.338
I4011	150	15.31	2.007	N4636	…	9.82	2.303
N4886	151	13.98	2.180	N4649	…	8.90	2.514
	153	15.28	2.099	N4660	…	10.97	2.262
N4864	159	14.26	2.275	N4697	…	9.28	2.276
I4045	168	14.11	2.320	N4742	…	11.37	2.027
I4021	172	14.87	2.191				
I4012	174	14.82	2.247				
	193	15.37	2.059				
N4860	194	13.49	2.394				
	207	15.04	2.154				
N4881	217	13.67	2.274				
N4841B	240	12.88	2.383				

Table 1. Basic data for the numerical example (from Dressler, 1984).

Method	ARE	What	a_1	a_2	b
Reference Solution	1.0	all but 4	4.14	3.65	-0.132
		std.dev.	0.13	0.10	0.009
ORLS	1.0	all	4.65	4.01	-0.169
		std. dev	0.19	0.14	0.013
		minimum	4.51	3.91	-0.174
		maximum	4.73	4.07	-0.159
		range	0.21	0.16	0.015
		Jackknife	4.65	4.02	-0.169
		std. dev.	0.26	0.19	0.019
Trimmed Jackknife	5%	all	4.53	3.91	-0.159
		std. dev	0.19	0.14	0.013
	10%	all	4.45	3.87	-0.153
		std. dev	0.12	0.09	0.009
Huber	0.9	all	4.49	3.90	-0.158
		minimum	4.39	3.84	-0.162
		maximum	4.55	3.95	-0.151
		range	0.16	0.12	0.011
		Jackknife	4.46	3.88	-0.155
		std. dev.	0.23	0.17	0.017
	0.8	all	4.44	3.88	-0.154
		minimum	4.36	3.82	-0.159
		maximum	4.51	3.92	-0.148
		range	0.15	0.10	0.011
		Jackknife	4.42	3.87	-0.152
		std. dev.	0.23	0.16	0.016
Tukey	0.9	all	4.41	3.85	-0.152
		minimum	4.27	3.74	-0.158
		maximum	4.50	3.91	-0.141
		range	0.24	0.17	0.017
		Jackknife	4.32	3.78	-0.145
		std. dev.	0.34	0.24	0.024
	0.8	all	4.24	3.73	-0.140
		minimum	4.18	3.69	-0.145
		maximum	4.32	3.78	-0.135
		range	0.14	0.10	0.010
		Jackknife	4.12	3.65	-0.131
		std. dev.	0.19	0.14	0.014
“Fair”	0.9	all	4.51	3.92	-0.159
		minimum	4.40	3.85	-0.163
		maximum	4.57	3.96	-0.151
		range	0.17	0.11	0.012
		Jackknife	4.50	3.92	-0.158
		std. dev.	0.24	0.17	0.017
	0.8	all	4.47	3.89	-0.156
		minimum	4.38	3.83	-0.160
		maximum	4.54	3.94	-0.150
		range	0.16	0.10	0.011
		Jackknife	4.48	3.91	-0.157
		std. dev.	0.24	0.17	0.017

Table 2. Summary of Results.

Figures

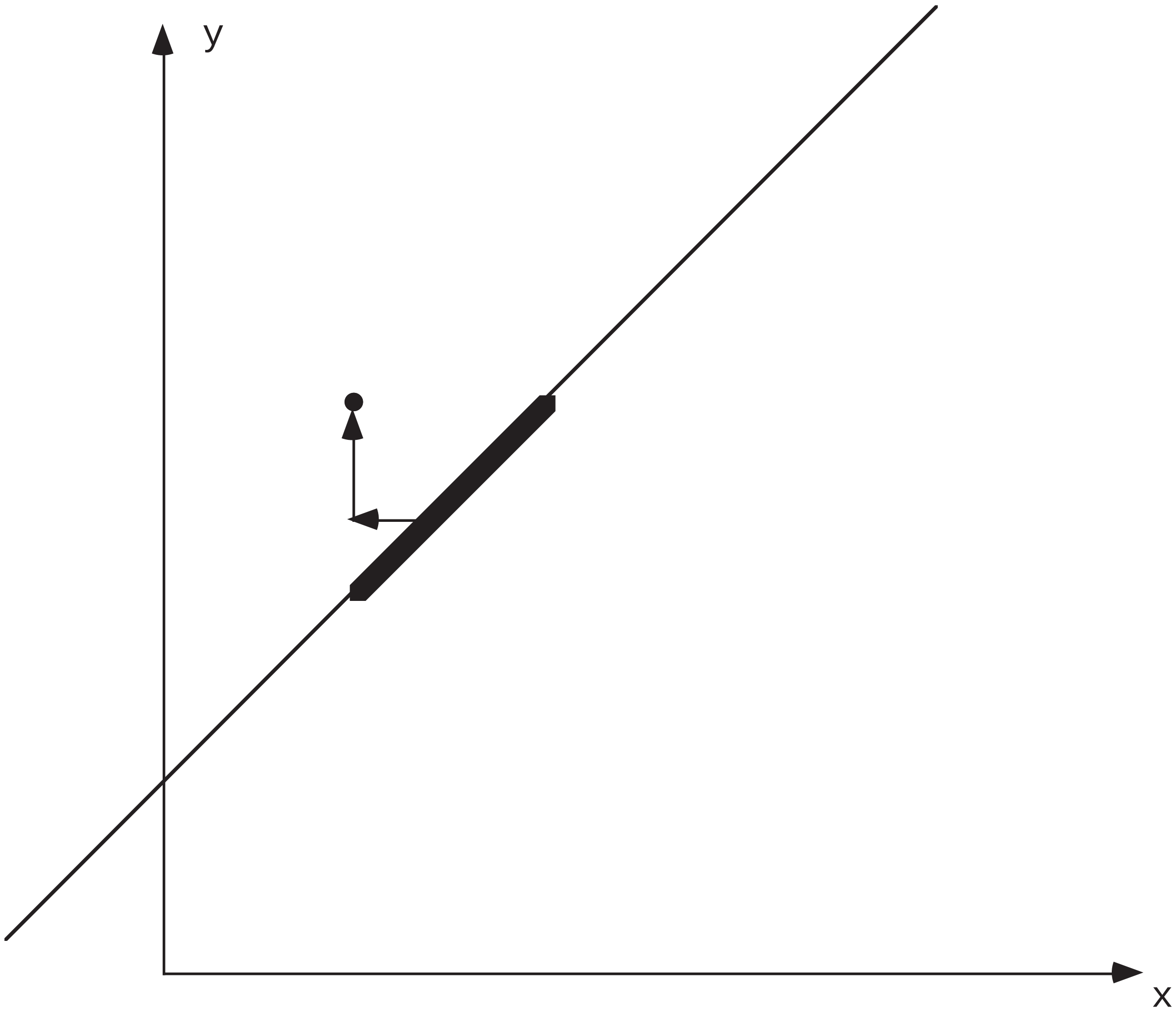

Figure 1

Figure 1: When ρ(u) = | u|, the loss function is given by Eq. 10a, and the slope of the line is 1, then there are infinitely many positions of the fitted point that correspond to a given observed point. All the points on the heavily marked segment of the fitted line have the same distance (in this metric) from the observed point.

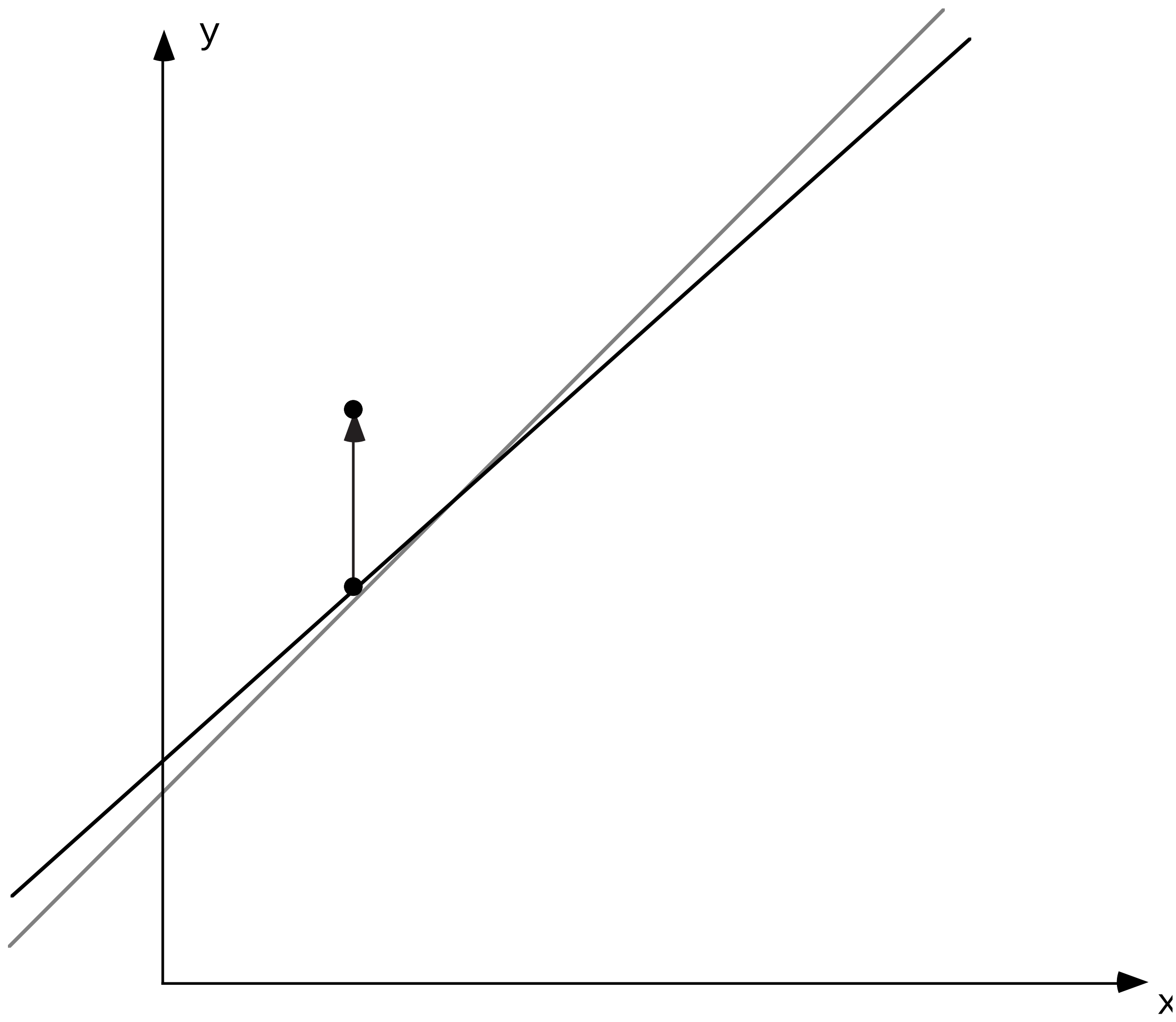

Figure 2

Figure 2: When the slope of the fitted line is less than 1, the fitted point jumps discontinuously from its arbitrary position in Figure 1 to the point on the fitted line that lies vertically above or beneath the observed point. The slope of the dashed line is unity, and the solid line is the fitted line.

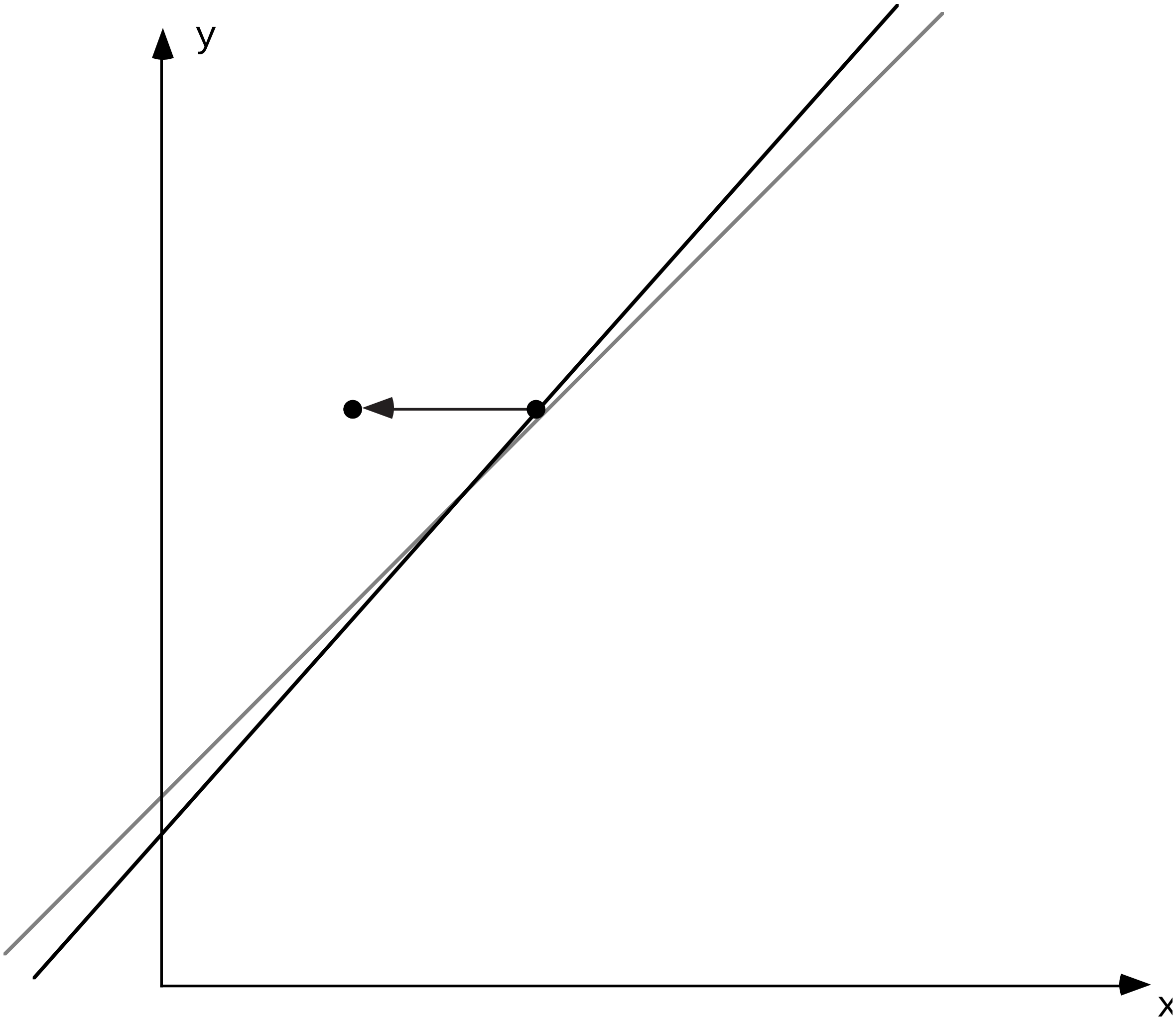

Figure 3

Figure 3: When the slope of the fitted line is greater than 1, the fitted point jumps discontinuously to the point that lies horizontally to the left or right of the observed point. The dashed and solid lines have the same meaning as in Figure 2.

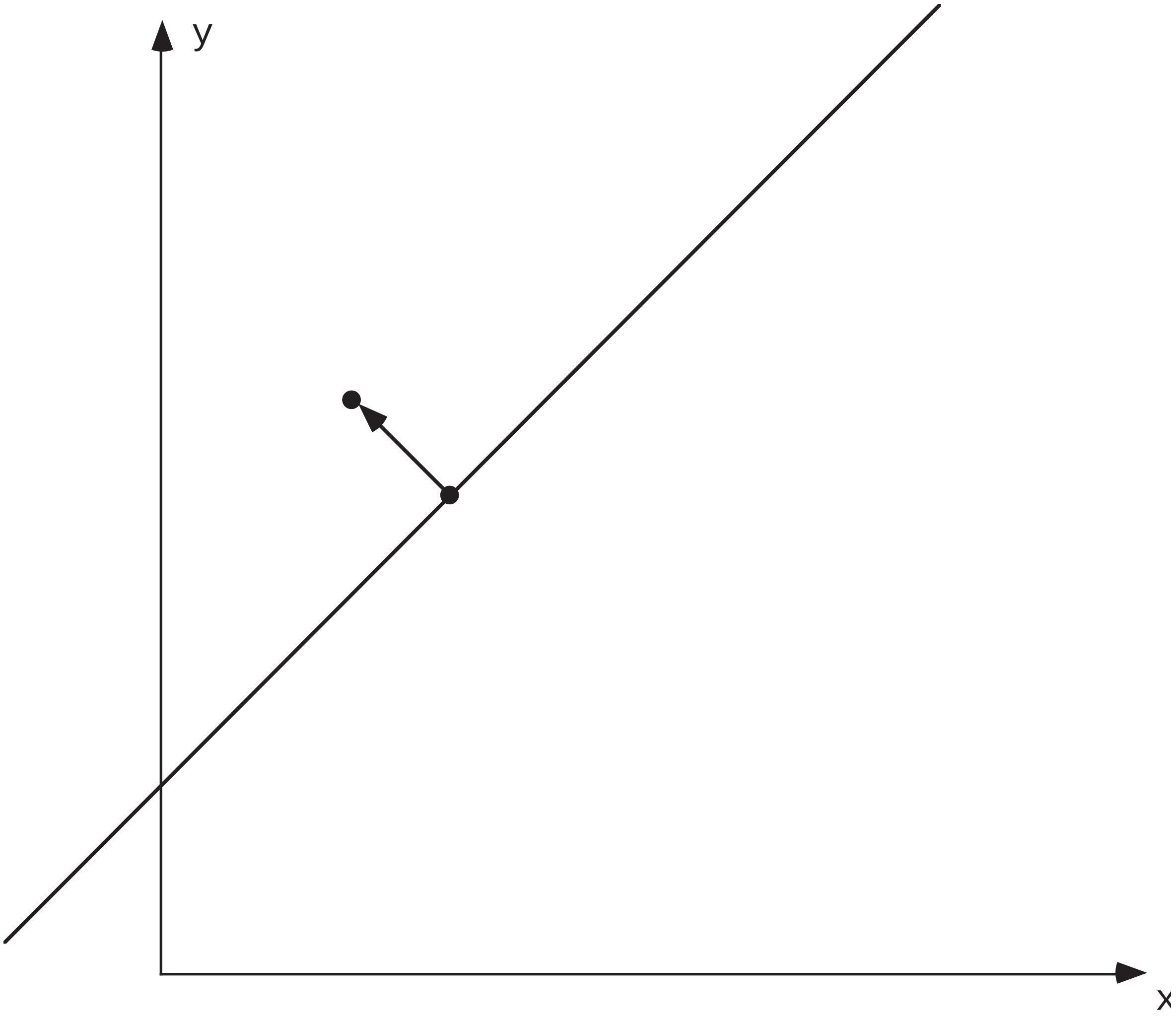

Figure 4

Figure 4: In the metric of Eq. 10b, there is no ambiguity in the position of the fitted point, which always lies at the foot of the perpendicular dropped from the observed point to the fitted line. As the slope of the fitted line varies, the position of the fitted point varies continuously.

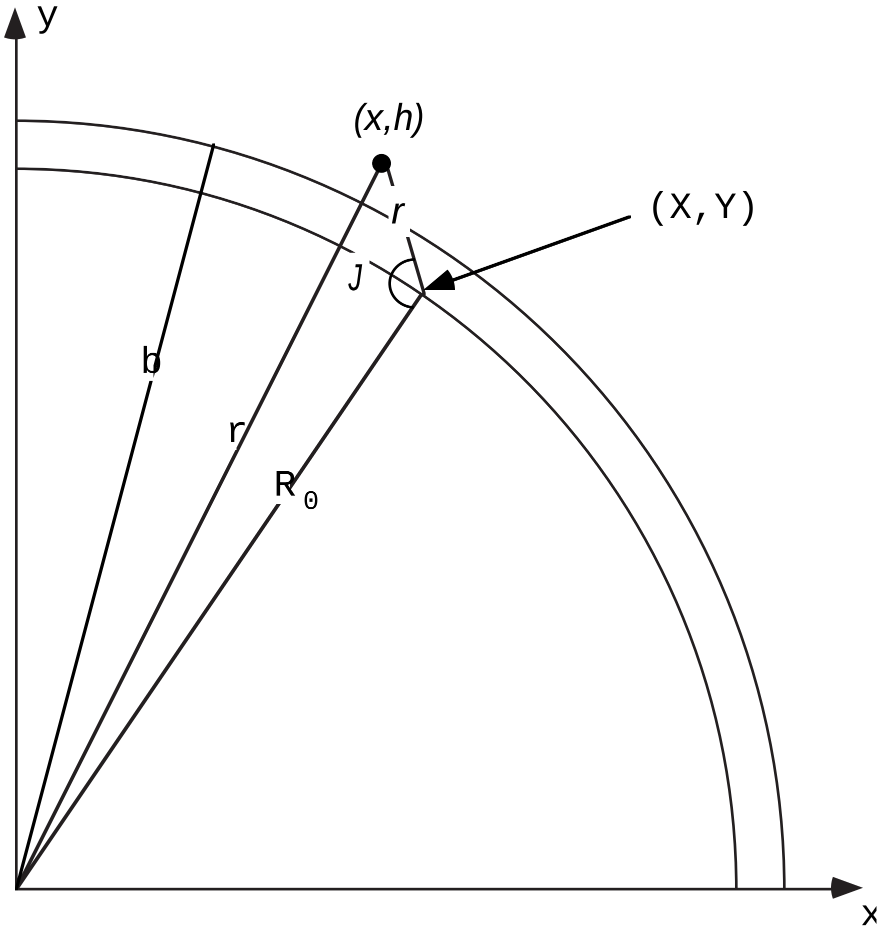

Figure 5

Figure 5: Geometry of the circle-fitting problem. The “true” point lies on the circle of radius R_0at (X, Y). The observed point at (ξ,η) is obtained by adding independent Gaussian random deviates to each of the (X, Y) coordinates of the “true” point. The observed point is distant from the “true” point by the amount ρ. The angle ϑ between the line from (X, Y) to the origin and the line from (X, Y) to the observed point is shown. Because of the curvature of the circle, the observed points tend to lie outside of the true circle, which leads to an overestimate b> R_0 of its radius.

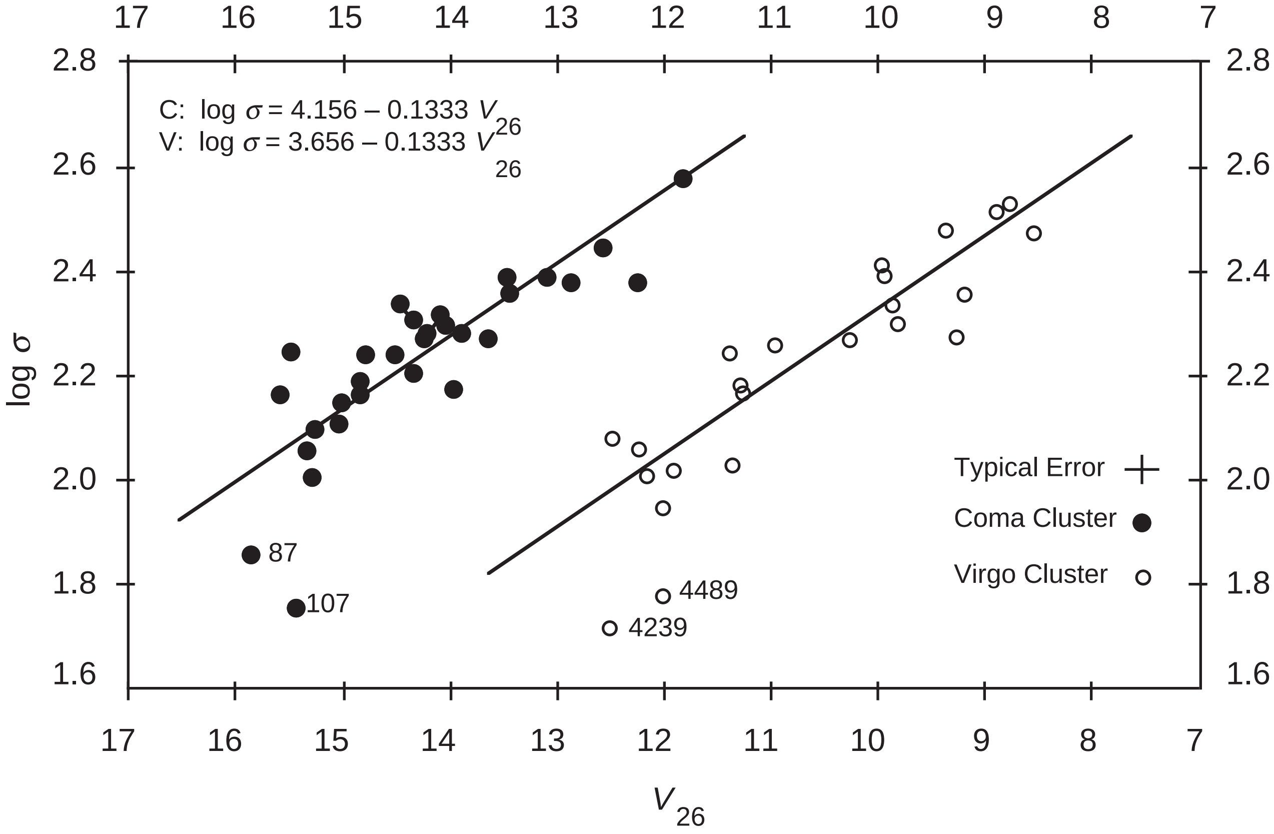

Figure 6: log of velocity dispersion (log σ) plotted against integrated V magnitude (V_{26}) for a sample of galaxies from two clusters. The four outliers noted by Dressler are identified by their catalog numbers. After Dressler (1984), Figure 6.