An alternative to R^2 to characterize the quality of fit for linear least squares using offsets of variable orientation related to uncertainties in the data.

Lewis Edward Wedgewood and Cody Wade Mischel

University of Illinois at Chicago

Abstract

It is not appropriate to use the determination coefficient, R^2, to characterize the quality of fit for a least squares fitted line. In this paper, the maximum of R^2 is found as a function of the rotation angle of the data and gives the quality of fit for the line found by linear least squares with perpendicular offsets. The same rotation method is used to derive the perpendicular offset fit to the data, which yields two possible solutions where the correct root can be identified by a simple discriminant. These results are then generalized for any arbitrarily oriented offset, bringing about a new measure for the quality fit of a line, Q^2. Unlike the determination coefficient, R^2, this quality of fit measure is invariant to rotational transformations of the data and is specific to the offset's orientation, which is directly related to the uncertainties in x- or y-data. Finally, this paper provides a method to determine the slope and intercept of a fitted line, as well as its quality of fit, given any estimate of the uncertainty ratio.

Key words: quality of fit, linear least squares, perpendicular offsets, variable offsets, uncertainty

Introduction

A common practice in mathematics, engineering, and science is to fit a straight line to data, thus leading to the question, "What is the quality of the fit?" Although standard deviations for the fit, slope, and intercept are well-known and can be easily calculated, this question is very often answered by giving the square of the correlation coefficient R^2, which we will refer to as the determination coefficient. The determination coefficient is seemingly easy to interpret. It varies between zero and one where a value of zero indicates that the data has no preferred direction and a value of one indicates a perfect fit with all data points falling on a straight line. However, this interpretation is naïve. The determination coefficient depends not only on the degree of scatter of the data points, but also on the slope or average inclination of the data with respect to the axes. This can be seen from the definition of R^2. If we consider a data set (x_i, y_i) of n points, then R^2 for such a data set has several equivalent definitions. For the purpose of this paper, we use the following definition of the determination coefficient:

$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \tag{1}$$

where the x-variance, y -variance, and covariance are defined as $SS_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x_i \rangle)^2$, $SS_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y_i \rangle)^2$, and $SS_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x_i \rangle)(y_i - \langle y_i \rangle)$ respectively, with the

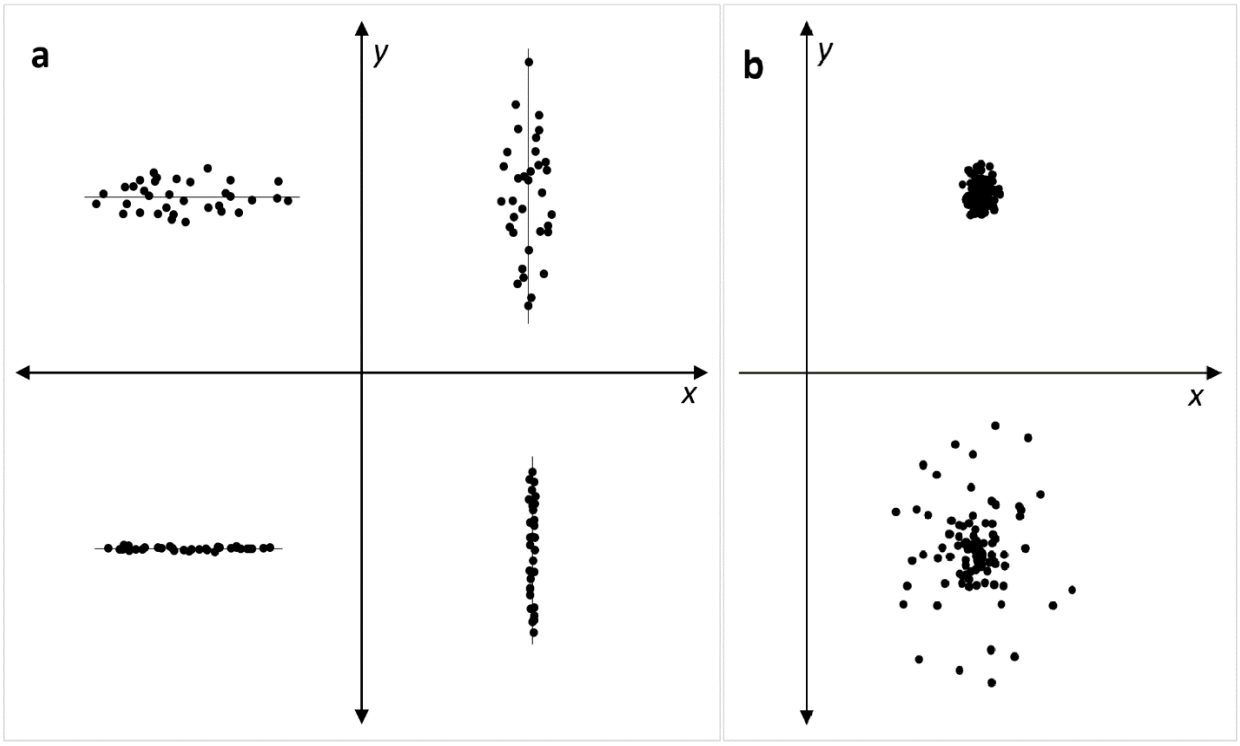

Figure 1 *Examples of data sets with determination coefficient, R^2, equal to zero. (a) R^2 equal to zero due to data being perfectly oriented in the direction of either axis. (b) R^2 equal to zero due to random scatter of points that have no definitive orientation or covariance. In all these examples, the degree of scatter does not impact the value of R^2.*

averages of x_i and y_i represented as $\langle x_i \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\langle y_i \rangle = \frac{1}{n}\sum_{i=1}^{n} y_i$. To form R^2, the covariance squared is made dimensionless by dividing by the product of the two variances. The covariance compares the tendency of x_i to deviate from $\langle x_i \rangle$ to how y_i deviates from $\langle y_i \rangle$. That is, when x_i values with a tendency to be above $\langle x_i \rangle$ occur together with y_i values with a tendency to be above $\langle y_i \rangle$, which implies that x_i values below $\langle x_i \rangle$ tend to accompany y_i below $\langle y_i \rangle$, then $SS_{xy} > 0$. If $SS_{xy} < 0$ then there is a negative correlation between the x_i and y_i deviations from their averages. Clearly, R^2 is a meaningful quantity, but it is not an appropriate measure of the quality of fit. In this paper, we propose a measure of quality of fit Q^2 which is appropriate for characterizing linear least squares fit of data.

To illustrate how the naïve interpretation of R^2 can fail, consider situations where $SS_{xy} = 0$ as shown in Figs. 1a and 1b. In Fig. 1a, four sets of data points are plotted that have a preferred direction that lies parallel to either of the axes. It is demonstrated in these figures that no matter how closely the points lie along the line of preferred direction, the determination coefficient $R^2 = 0$. Had these data sets been rotated so as not to lie along one of the axes, a transformation that does not affect the scatter of the data, then the determination coefficient would have been nonzero. By comparison in Fig. 1b, two examples show data that is randomly scattered, which also leads to $R^2 = 0$. These two data sets in Fig. 1b

can also be rotated, but the determination coefficient will remain zero. It is only the degree of scatter that should be associated with quality of fit. In §1, we derive the maximum of the determination coefficient, R^2_{max} that only reflects the degree to which the data line along a straight line and is invariant to the inclination of the data. We then develop a geometric interpretation of this result which makes it possible to determine that R^2_{max} is the quality of fit for a line determined using linear least squares with perpendicular offsets.

In addition, the rotational method lends itself to an alternative derivation of the linear least squares with perpendicular offsets. This is sometimes referred to as total least squares because it gives the minimum sum of the square residuals or error. The traditional derivation for this problem yields a quadratic equation for the slope, and hence, two possible solutions. It is a nuisance to visually choose the correct solution and mathematically unsatisfying. The rotational method used in this paper yields the same two solutions for the slope and intercept; however, it also yields a very simple discriminant to distinguish which is the correct solution. Sampaio[1] has used a rotation combined with an iterative scheme to estimate this result numerically. In §2, we derive the discriminant for the traditional solution. We then use the rotational method to derive a purely analytical solution, which yields the same numerical result but has a simpler discriminant.

Textbooks[2] often state that least squares using vertical offsets assumes that all the uncertainty is in the y-variable, while least squares using horizontal offsets assumes that all the uncertainty is in the x-variable. In addition, least squares with perpendicular offsets assumes that the uncertainty in x and y is equal. This observation can be generalized using the rotational method. In §3 we demonstrate how to fit a straight line to data using offsets of arbitrary angle to the axes. The resulting slope and intercept depend on this angle. What is most interesting is that we can relate the slope of the offset to the ratio of the relative uncertainty of the x-data to that of the y-data. This gives a direct relation between the relative uncertainty in the x-data versus y-data set and how it affects the slope and intercept for the fitted line. This leads to a simple relation between the uncertainty ratio and the slope and intercept of a line fit with arbitrary offsets.

Finally, in §4 we generalize R^2_{max} the quality of fit for perpendicular offsets to the quantity Q^2 the quality of fit specific to any offset. An illustration of the main results of the paper are given in the appendix for readers who wish to apply the method.

The problem of fitting a straight line to data has been addressed by many authors. Notably, York[3,4] published a method of fitting straight lines with correlated error in both variables. Several others[5,6,7,8,9] have subsequently published articles extending and improving York's method. The method we present in this paper requires a constant error or uncertainty ratio σ_x^2/σ_y^2 for all data points. This leads to simpler results that can easily be implemented and for which Q^2 can be expressed.

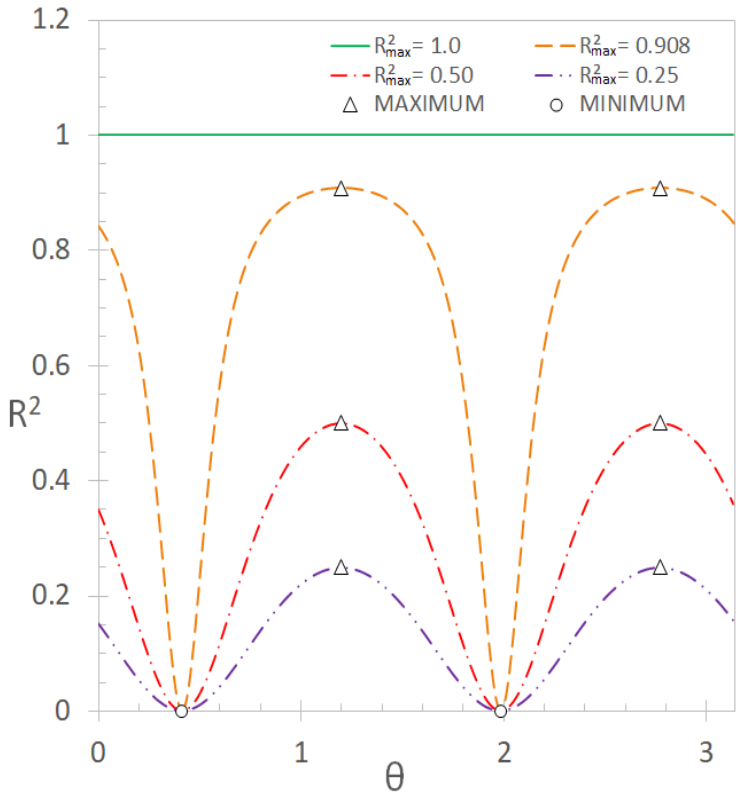

Figure 2 *Periodicity of R^2 for various data sets depicted by R^2_{max} as they are rotated about the origin at an angle θ (radians) from 0 to π. An extremum can be found every π/4, with the maxima and minima represented as squares and circles, respectively. For R^2_{max}=1, the data set is represented perfectly by a straight line thus making R^2=1 independent of θ.*

§1 The Determination Coefficient

Our first goal is to separate the effect of the scatter of the data from the inclination of the data with the axes. We do this by considering the data in an arbitrarily rotated coordinate system. A pure rotation can be defined by the transformation:

$$\hat{x}_i = x_i \cos\theta - y_i \sin\theta \quad \text{and} \quad \hat{y}_i = x_i \sin\theta + y_i \cos\theta \tag{2}$$

where θ is the angle between the fixed and the rotated axes. Here we take (x_i, y_i) as the data with reference to fixed axes and $(\hat{x}_i, \hat{y}_i)$ as the data points with reference to rotated axes. Using Eq. (2), the determination coefficient can be expressed in the rotated frame as:

$$R^2(\theta) = \frac{SS^2_{\hat{x}\hat{y}}}{SS_{\hat{x}\hat{x}} SS_{\hat{y}\hat{y}}} = \frac{\left(SS_{xy}\cos 2\theta + \frac{1}{2}[SS_{xx} - SS_{yy}]\sin 2\theta\right)^2}{\left(\frac{1}{2}[SS_{xx}+SS_{yy}] + \frac{1}{2}[SS_{xx}-SS_{yy}]\cos 2\theta - SS_{xy}\sin 2\theta\right)\left(\frac{1}{2}[SS_{xx}+SS_{yy}] - \frac{1}{2}[SS_{xx}-SS_{yy}]\cos 2\theta + SS_{xy}\sin 2\theta\right)} \tag{3}$$

The determination coefficient has been expressed in terms of trigonometric functions of double angles to emphasize the periodicity over a rotation of 180˚. Plots for typical data sets are shown

in Fig. 2. The plots have two extrema. The maximum value corresponds to rotating the data such that the line fit using perpendicular offsets will have a slope of ± 1, and therefore, occurs every 90˚. There is a minimum value which is always zero and corresponds to the angle which rotates the data to give a zero or infinite slope to a fitted line and again occurs every 90˚, that is, the data lies along a preferred direction along either axis.

The precise positions of these extrema can be found by taking the derivative of Eq. (3) and setting it equal to zero. After simplification, this yields the result:

$$\frac{dR^2}{d\theta} = \frac{\left((SS_{xy})^2 - SS_{xx}SS_{yy}\right)\left(2SS_{xy}\sin 2\theta - (SS_{xx}-SS_{yy})\cos 2\theta\right)\left(2SS_{xy}\cos 2\theta + (SS_{xx}-SS_{yy})\sin 2\theta\right)}{\left(\frac{1}{2}[SS_{xx}+SS_{yy}] + \frac{1}{2}[SS_{xx}-SS_{yy}]\cos 2\theta - SS_{xy}\sin 2\theta\right)^2\left(\frac{1}{2}[SS_{xx}+SS_{yy}] - \frac{1}{2}[SS_{xx}-SS_{yy}]\cos 2\theta + SS_{xy}\sin 2\theta\right)^2} = 0 \quad (4)$$

The positions of the extrema are given by:

$$\theta_{max} = \frac{1}{2}\tan^{-1}\left(\frac{SS_{xx}-SS_{yy}}{2SS_{xy}}\right) + \frac{k\,\pi}{2} \quad \text{and} \quad \theta_{min} = \frac{1}{2}\tan^{-1}\left(-\frac{2SS_{xy}}{SS_{xx}-SS_{yy}}\right) + \frac{k\,\pi}{2} \quad (5a,b)$$

where k is an arbitrary integer. In fact, k can be set equal to zero above without loss of generality. Equations (5a) and (5b) can be re-inserted into Eq. (3) to give analytical expressions for the two unique extrema Eqs. (6a) and (6b), respectively:

$$R^2_{max} \equiv R^2(\theta_{max}) = \frac{(SS_{xx}-SS_{yy})^2 + 4(SS_{xy})^2}{(SS_{xx}+SS_{yy})^2} \quad \text{and} \quad R^2_{min} \equiv R^2(\theta_{min}) = 0 \quad (6a,b)$$

Equation (6a) is a remarkably simple expression that is only affected by the degree of scatter in a data set. The quantity R^2_{max} is invariant to a pure rotation of the coordinates as given by Eq. (2). In fact, R^2_{max} is invariant to any rotation/translation of the data. In addition, the numerator and denominator of the right side of Eq. (6a) are both invariant to rotation. In Fig. 3, a single data set is rotated to several different angles and R^2 is compared with R^2_{max} at each angle. In this example, the value of $R^2_{max} = 0.9082$ remains constant for all orientations while R^2 varies from zero to 0.9082. The figure emphasizes that R^2 has orientation dependence while R^2_{max} does not. Because $R^2(\theta)$ ranges from zero to one, it may be obvious that the expressions in Eqs. (6a) and (6b) correspond to the maximum and the minimum. For the sake of completeness, the second derivative of $R^2(\theta)$ can be found and evaluated for the roots θ_m given in Eqs. (5a) and (5b). The result of this operation is

$$\left.\frac{d^2R^2}{d\theta^2}\right|_{\theta_{max}} = -32\frac{\left[(SS_{xx}-SS_{yy})^2 + 4(SS_{xy})^2\right]\left[SS_{xx}SS_{yy} - (SS_{xy})^2\right]}{(SS_{xx}+SS_{yy})^4} \leq 0 \quad \text{and}$$

$$\left.\frac{d^2R^2}{d\theta^2}\right|_{\theta_{min}} = 2\frac{(SS_{xx}-SS_{yy})^2 + 4(SS_{xy})^2}{SS_{xx}SS_{yy} - (SS_{xy})^2} \geq 0 \quad (7a,b)$$

where Eq. (7a) is the second derivative evaluated with Eq. (5a) and Eq. (7b) is the second derivative evaluated with Eq. (5b). This confirms that the maximum and minimum of $R^2(\theta)$ have been correctly identified in Eqs. (6a) and (6b).

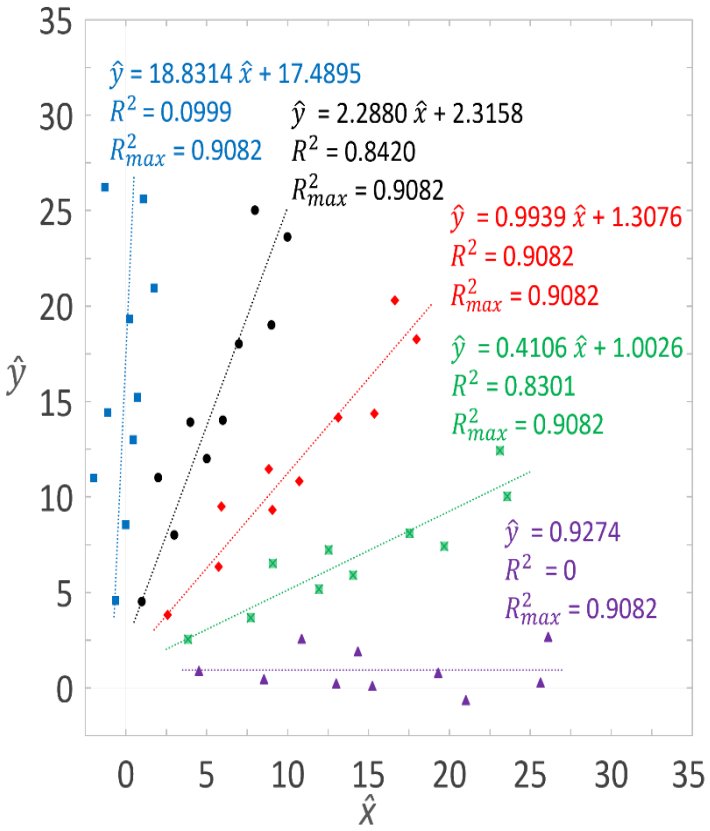

Figure 3 *Comparison of the rotational dependence of R^2 and independent R^2_{max} for a single data set as it is rotated about the origin at an angle θ (radians). The best fit line equations are formed using perpendicular offsets.*

We now wish to understand the degree of scatter R^2_{max} more deeply. The scatter of points can be interpreted in geometric terms by considering the covariance matrix:

$$\Sigma = \begin{pmatrix} SS_{xx} & SS_{xy} \\ SS_{xy} & SS_{yy} \end{pmatrix} \tag{8}$$

The eigenvalue problem $|\Sigma - \lambda I| = 0$, where I is the identity matrix and λ is the eigenvalue, leads to two eigenvalues:

$$\lambda_1 = \frac{1}{2}\left(SS_{xx} + SS_{yy}\right) + \frac{1}{2}\sqrt{\left(SS_{xx} - SS_{yy}\right)^2 + 4\left(SS_{xy}\right)^2} \tag{9}$$

$$\lambda_2 = \frac{1}{2}\left(SS_{xx} + SS_{yy}\right) - \frac{1}{2}\sqrt{\left(SS_{xx} - SS_{yy}\right)^2 + 4\left(SS_{xy}\right)^2} \tag{10}$$

The eigenvalues give a measure of the length and breadth of the data points independent of the orientation of the data with respect to the axes. This can be illustrated as an ellipse defined by the eigenvalues and eigenvectors. The degree of scatter can now be interpreted as:

$$R^2_{max} = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2 \tag{11}$$

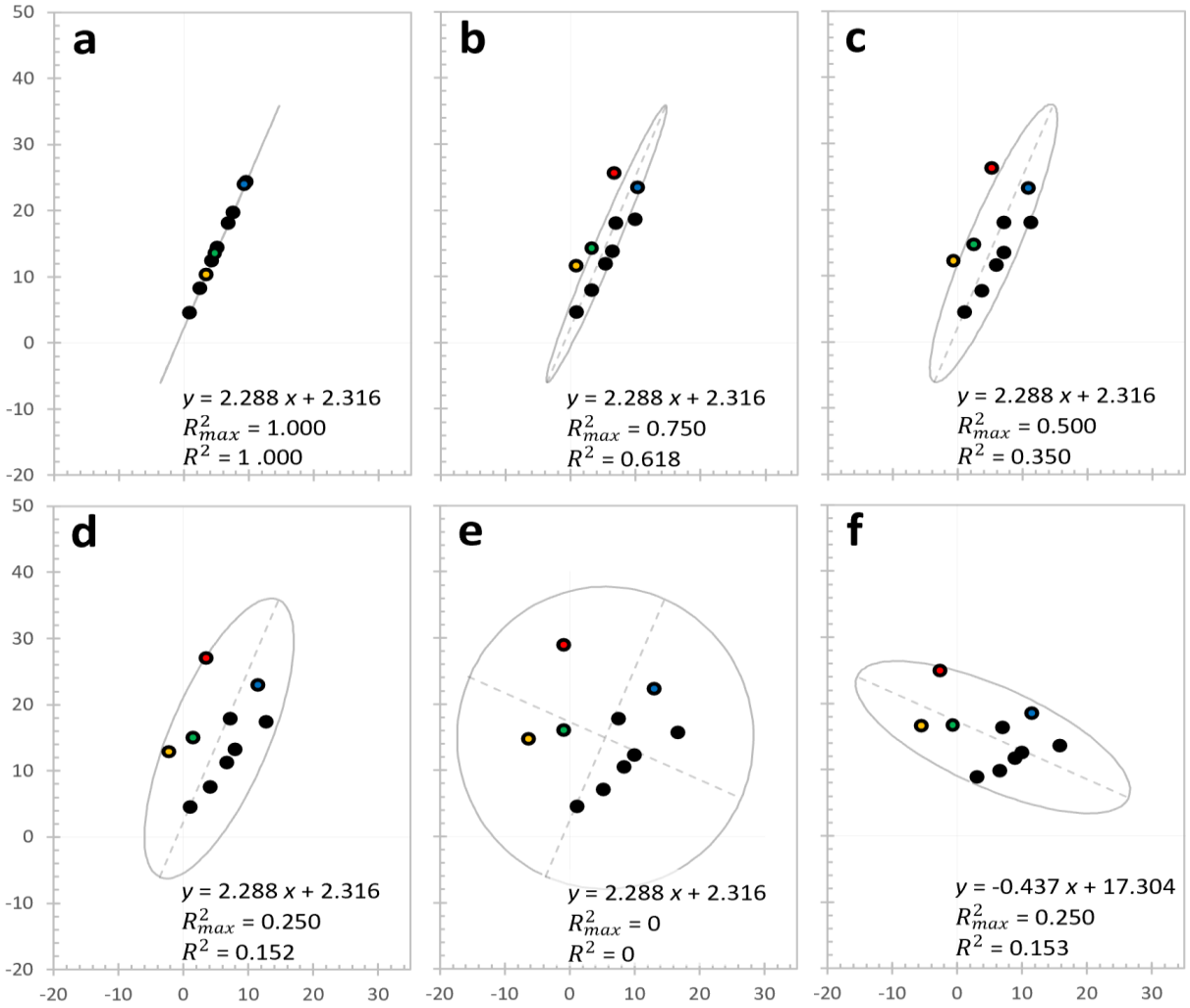

Figure 4 *Compares the effects of scatter on R^2_{max} as a data set is systematically spread out further from a line. Ellipses have been drawn around the data sets to indicate the principle axes given by λ_1 and λ_2. a) The data points lie perfectly on a straight line with R^2_{max} and R^2 equal to one. b,c,d) Data points stretched, thus reducing R^2_{max} to 0.750, 0.500, and 0.250 respectively. e) Data set oriented such that the principle axes forms a perfect circle meaning there is no discernable orientation; R^2_{max} and R^2 are nearly 0. f) Data has been further stretched such that the minor and major principle axes have switched; the line of best fit is now perpendicular to the earlier cases. R^2_{max} can be seen to increase again to 0.250, but now in relation to the new, perpendicular best-fit line.*

which is equivalent to Eq. (6a). It is now clear that when the principal minor axis is zero, then R^2_{max} is unity. In this case, the data defines an infinitesimally narrow ellipse and all the data points must fall on a straight line. In contrast, when the principal axes are equal $\lambda_1 = \lambda_2$ the

degree of scatter is zero. In this case, the ellipse is a circle and the data has no preferred direction.

We can also consider the eigenvectors given by the solution of $(\Sigma - \lambda_i I)v_i = 0$ where v_i represents two vectors, one associated with each of the two eigenvalues. Solving this problem gives the slope of the two vectors:

$$\frac{v_{1,y}}{v_{1,x}} = \frac{-\frac{1}{2}(SS_{xx}+SS_{yy})+\frac{1}{2}\sqrt{(SS_{xx}-SS_{yy})^2+4(SS_{xy})^2}}{SS_{xy}} = -B + \sqrt{B^2+1} \tag{12}$$

$$\frac{v_{2,y}}{v_{2,x}} = \frac{-\frac{1}{2}(SS_{xx}+SS_{yy})-\frac{1}{2}\sqrt{(SS_{xx}-SS_{yy})^2+4(SS_{xy})^2}}{SS_{xy}} = -B - \sqrt{B^2+1} \tag{13}$$

where $B = \left(SS_{xx} - SS_{yy}\right)/\left(2\,SS_{xy}\right)$. The x and y components of the vectors are represented by $v_{1,x}$ and $v_{1,y}$, respectively. The slopes of the two eigenvectors are perpendicular to one another. We show in the next section that R^2_{max} describes the scatter of the data about a line found by least squares using perpendicular offsets. In Fig. 4, a data set is originated on a given line, and then systematically spread out from the line in a sequence of plots. Since the line does not lie along the 45˚ line, the values of R^2 and R^2_{max} are not in general equal, but are only equal when either the data points have no scatter and both values are unity, or when there is no preferred direction and both values are zero. In Fig. (4f), the points are now scattered about a line that is perpendicular to the initial line. We have colored four of the points so that the relative configuration can be discerned as the points are moved.

§2 Fitting linear data with perpendicular offsets

The problem of fitting a line to a data set has been extensively studied and has broad application. For a typical set of data (x_i, y_i), this problem can be solved using vertical offsets. It is appealing to use perpendicular offsets because this leads to the absolute minimum residual error. For perpendicular offsets, the sum of the square residuals for n data points is given by Weisstein[10]

$$\phi_\perp = \sum_{i=1}^{n} \frac{[y_i-(a\,x_i+b)]^2}{1+a^2} \tag{14}$$

where a is the slope and b is the intercept of the fitted line. To find the extrema for $\phi_\perp$, the derivatives with respect to a and b yield

$$a = -B \pm \sqrt{B^2+1} \text{ where } B = \frac{SS_{xx}-SS_{yy}}{2SS_{xy}} \tag{15}$$

$$b = \langle y_i \rangle - a\langle x_i \rangle \tag{16}$$

This is a well-known solution, so we do not repeat the derivation here. Notice that the two slopes given by Eq. (15) are identical to the slopes of the eigenvectors in Eqs. (12) and (13). The problem remains that the sign in front of the radical that leads to the appropriate value for the

minimum error $\phi_\perp$ cannot be determined *a priori*. To determine the sign in Eq. (15), we use the second partial derivative test.[11] This requires evaluating the four second partials of Eq. (14) with respect to the slope and intercept. These are given below:

$$\frac{\partial^2\phi_\perp}{\partial a^2} = 2\frac{\left(b^2-2b\langle y_i\rangle+\langle y_i^2\rangle-\langle x_i^2\rangle\right)\left(3a^2-1\right)+2a\left(a^2-3\right)\left(b\langle x_i\rangle-\langle x_i y_i\rangle\right)}{\left(a^2+1\right)^3} \tag{17}$$

$$\frac{\partial^2\phi_\perp}{\partial b^2} = \frac{2}{a^2+1} \tag{18}$$

$$\frac{\partial^2\phi_\perp}{\partial a\partial b} = \frac{\partial^2\phi_\perp}{\partial b\partial a} = 2\frac{\langle x_i\rangle\left(1-a^2\right)+2a\left(\langle y_i\rangle-b\right)}{\left(a^2+1\right)^2} \tag{19}$$

Because Eq. (18) is always positive, the test depends only on the discriminant given by:

$$D(a,b) = \frac{\partial^2\phi_\perp}{\partial a^2}\frac{\partial^2\phi_\perp}{\partial b^2} - \frac{\partial^2\phi_\perp}{\partial a\partial b}\frac{\partial^2\phi_\perp}{\partial b\partial a} \tag{20}$$

where the slope and intercept are evaluated according to Eqs. (15) and (16). After considerable simplification, this gives the discriminant for both possible solutions as:

$$D_\pm = 2\left\{\left(SS_{xx}-SS_{yy}\right)\left(\frac{\left(SS_{xx}-SS_{yy}\right)^2+3\left(SS_{xy}\right)^2}{\left(SS_{xx}-SS_{yy}\right)^2+4\left(SS_{xy}\right)^2}\right) \pm \frac{SS_{xy}}{|SS_{xy}|}\frac{\left(\left(SS_{xx}-SS_{yy}\right)^2+\left(SS_{xy}\right)^2\right)}{\sqrt{\left(SS_{xx}-SS_{yy}\right)^2+4\left(SS_{xy}\right)^2}}\right\} \tag{21}$$

The plus/minus sign in Eq. (21) corresponds to the plus/minus sign in Eq. (15) and represents two discriminants for the two possible solutions. To demonstrate that D_+ and D_- always have opposite signs, we take the product of the two discriminants.

$$(D_+)(D_-) = -16\frac{\left(SS_{xy}\right)^6}{\left(\left(SS_{xx}-SS_{yy}\right)+4\left(SS_{xy}\right)^2\right)^2} \le 0 \tag{22}$$

This confirms that one solution from Eqs. (15) and (16) corresponds to a minimum and the other to a saddle point according to the second partial test.

The discriminant in Eq. (21) is very useful for fitting data numerically. It allows the selection of the correct solution automatically without having to visually refer to a plot of the data. Unfortunately, the discriminant in Eq. (21) is somewhat unwieldy. We now demonstrate that a simpler method can be used to derive the results above for fitting a line to data using perpendicular offsets and leads to a simpler discriminant.

In this method, we fit the data with vertical offsets using a horizontal line such that $\hat{y} = \hat{b}$ where $\hat{b}$ is a constant for a given value of θ according to the rotation transformation given in Eq. (2). As in §1, the circumflexes on the symbols signify that we are in a frame that has been rotated about the origin by an angle of θ. The sum of the square residuals for n data points in the rotated frame is given by:

$$\phi_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{b}-\hat{y}_i\right)^2 \tag{23}$$

where $\hat{b} = \langle \hat{y}_i \rangle = \langle x_i \rangle \sin\theta + \langle y_i \rangle \cos\theta$ is the well-known least-squares fit for a zero-order polynomial using vertical offsets. The subscript zero on ϕ_0 indicates the order of the polynomial being fit. Also, we replace $\hat{y}_i = x_i \sin\theta + y_i \cos\theta$ according to Eq. (2). Making these substitutions into Eq. (23) and performing the squaring operation leads to

$$\phi_0 = \frac{1}{n}\sum_{i=1}^{n}(\langle y_i \rangle^2 + y_i^2 - 2\langle y_i \rangle\, y_i)\cos^2\theta + 2(\langle x_i \rangle\langle y_i \rangle - \langle y_i \rangle x_i - \langle x_i \rangle\, y + x_i\, y_i)\sin\theta\cos\theta + (\langle x_i \rangle^2 + x_i^2 - 2\langle x_i \rangle\, x_i)\sin^2\theta \quad (24)$$

After performing the summation and converting to double-angle trigonometric functions, Eq. (24) simplifies to:

$$\phi_0 = -\frac{1}{2}\left[SS_{xx} - SS_{yy}\right]\cos 2\theta + SS_{xy}\,\sin 2\theta + \frac{1}{2}\left[SS_{xx} + SS_{yy}\right] \quad (25)$$

Eq. (25) gives the sum of the squared residuals for a horizontal line fit to the data using vertical offsets as a function of the angle of rotation. This means that the offsets are also perpendicular to the line. We wish to determine the angle θ that gives the minimum error $\phi_0(\theta)$. The derivative with respect to θ can easily be found and set to zero.

$$\frac{d\phi_0}{d\theta} = \left[SS_{xx} - SS_{yy}\right]\sin 2\theta + 2\,SS_{xy}\,\cos 2\theta = 0 \quad (26)$$

Solving Eq. (26) for the angles that give extrema yields:

$$\theta_k = \frac{1}{2}\tan^{-1}\left(-\frac{2SS_{xy}}{SS_{xx} - SS_{yy}}\right) + \frac{k\,\pi}{2} = \frac{1}{2}\tan^{-1}\left(-\frac{1}{B}\right) + \frac{k\,\pi}{2} \quad (27)$$

where k is again an arbitrary integer and B is defined in Eq. (15). We now reverse the rotation for the horizontal line $\hat{y} = \hat{b}$ to the original frame so that $x\sin\theta + y\cos\theta = \langle x_i \rangle\ \sin\theta + \langle y_i \rangle\ \cos\theta$ or

$$y = -(\tan\theta)x + \ (\tan\theta)\langle x_i \rangle\ + \langle y_i \rangle \quad (28)$$

so that the slope is $a = -\tan\theta$ and the intercept is $b = (\tan\theta)\langle x_i \rangle\ + \langle y_i \rangle$. This is a very simple result and yields the same pair of solutions as given in Eqs. (15) and (16).

To determine the minimum error, we take the second derivative of ϕ_0 with respect to θ

$$\frac{d^2\phi_0}{d\theta^2} = 2\left[SS_{xx} - SS_{yy}\right]\cos 2\theta - 4\,SS_{xy}\,\sin 2\theta \quad (29)$$

After substitution of Eq. (27) into Eq. (29), the second derivative simplifies to

$$\frac{d^2\phi_0}{d\theta^2} = 2(-1)^k\left[SS_{xx} - SS_{yy}\right]\sqrt{\frac{4\,SS_{xy}{}^2}{\left(SS_{xx} - SS_{yy}\right)^2} + 1} \quad (30)$$

Because the radical and the factor of two are always positive, the discriminant for the desired minimum of ϕ_0 can be expressed as

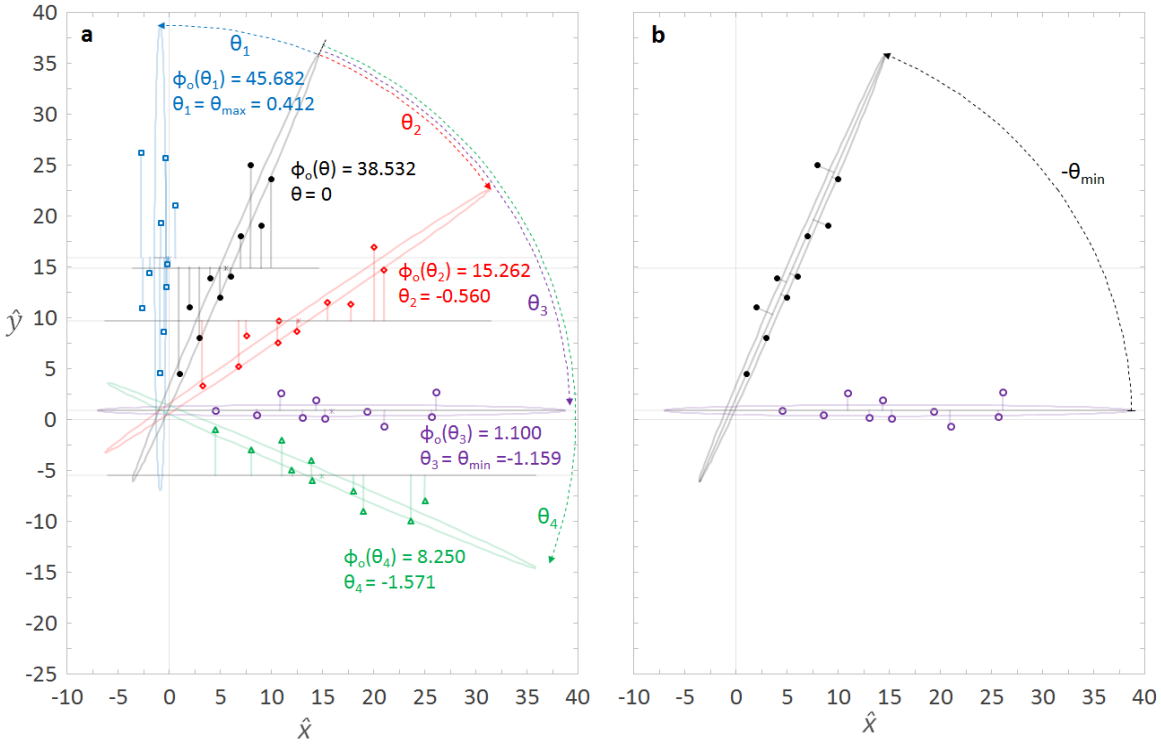

Figure 5 *Rotational method to obtain line of best fit with perpendicular offsets by minimizing ϕ_0. The horizontal, zero-order line $\hat{y} = \hat{b}$ is shown for each angle. a) Original data set, shown as solid circles, can be rotated about the origin by any given angle θ; several are displayed as hollowed data points. ϕ_0 can be calculated for each data set by taking the vertical offsets with respect to a horizontal line passing through the centroid. ϕ_0 is found to be minimized when the data set is rotated horizontally, depicted by hollow-circle data set, thus giving θ_{min}. b) The minimized data set and its horizontal fit can be rotated back $-\theta_{min}$ to give a fit now based off perpendicular offsets for the original data set.*

$$\text{If } \left[SS_{xx} - SS_{yy}\right] > 0 \text{ Then } k \text{ is even, for example, } k = 0 \qquad (31a)$$

$$\text{If } \left[SS_{xx} - SS_{yy}\right] < 0 \text{ Then } k \text{ is odd, for example, } k = 1 \qquad (31b)$$

where other values of k are redundant. Comparing this result with Eq. (21), the discriminant for the traditional method, the inequalities in Eqs. (31a,b) are much simpler to apply. The method is illustrated in Fig. 5a where a set of data points in the fixed coordinates is represented by the solid circles. Four possible rotations of the data are shown including the angles that give the minimum and maximum values of ϕ_0. In Fig. 5b, the rigid rotation returning the points from the optimal angle to the original position is shown. Notice that the perpendicular offsets are preserved when the data is again in its original position.

The slope of the fitted line can be expressed in terms of B by substitution of Eq. (27) into the expression for the slope:

$$a_{\perp} = -(\tan\theta) = -\left(\tan\left[\frac{1}{2}\tan^{-1}\left(-\frac{1}{B}\right) + \frac{k\,\pi}{2}\right]\right) = -B \pm B\sqrt{1+\frac{1}{B^2}} \qquad (32)$$

where the last expression on the right was derived using the following two trigonometric identities:

$$\tan(\alpha+\beta) = \frac{\tan\alpha+\tan\beta}{1-\tan\alpha\tan\beta} \text{ and } \tan\left[\frac{1}{2}\tan^{-1}(x)\right] = \frac{-1\pm\sqrt{1+x^2}}{x} \qquad (33)$$

We have added the subscript $\perp$ to the slope $a_{\perp}$ to indicate the offsets are perpendicular. This expression for the slope in Eq. (32) is very similar to the expression for the slope in Eq. (15) with the simplification that the correct sign in front of the radical is easily determined by the discriminant $[SS_{xx} - SS_{yy}]$.

$$a_{\perp} = -B + B\sqrt{1+\frac{1}{B^2}} \quad \text{if} \quad [SS_{xx} - SS_{yy}] > 0 \qquad (34a)$$

$$a_{\perp} = -B - B\sqrt{1+\frac{1}{B^2}} \quad \text{if} \quad [SS_{xx} - SS_{yy}] < 0 \qquad (34b)$$

$$b_{\perp} = \langle y_i\rangle - a_{\perp}\langle x_i\rangle \qquad (35)$$

Notice that the expression above preserves the sign of B in front of the radical. This factor of B could be absorbed into the radical to give the same expression as in Eq. (15), but the sign information would be lost. It is this additional information that leads to the simpler form of the discriminant. Clearly, this new formulation is easier to apply. Equation (35) is known as the centroid equation.

Finally, it can be proven that the result of Eq. (5a) rotates the data set such that the perpendicular offset fitted line lies at a 45˚ angle to the axes. If we recognize that Eq. (5a) can be written without loss of generality as $\theta = \frac{1}{2}\tan^{-1}(B)$, we then add this pure rotation to the angle of the fitted line to get

$$\tan^{-1}\left(-B \pm \sqrt{B^2+1}\right) + \frac{1}{2}\tan^{-1}(B) = \pm\frac{\pi}{4} \qquad (36)$$

which is a trigonometric identity. This proves the above statement.

§3 Fitting linear data with arbitrary offsets

We now turn our attention to fitting data with offsets of arbitrary orientation to the axes. To achieve this goal, we follow the method used in the previous section starting with Eq. (23). However, rather than fitting a horizontal line to the data as it is rotated, we fit a straight line $\hat{y} = \hat{a}\hat{x} + \hat{b}$ where $\hat{a}$ is the slope and $\hat{b}$ is the intercept for the line fitted to data rotated by an angle θ (see Fig. 6). The sum of the square residuals for n data points in the rotated frame is given by:

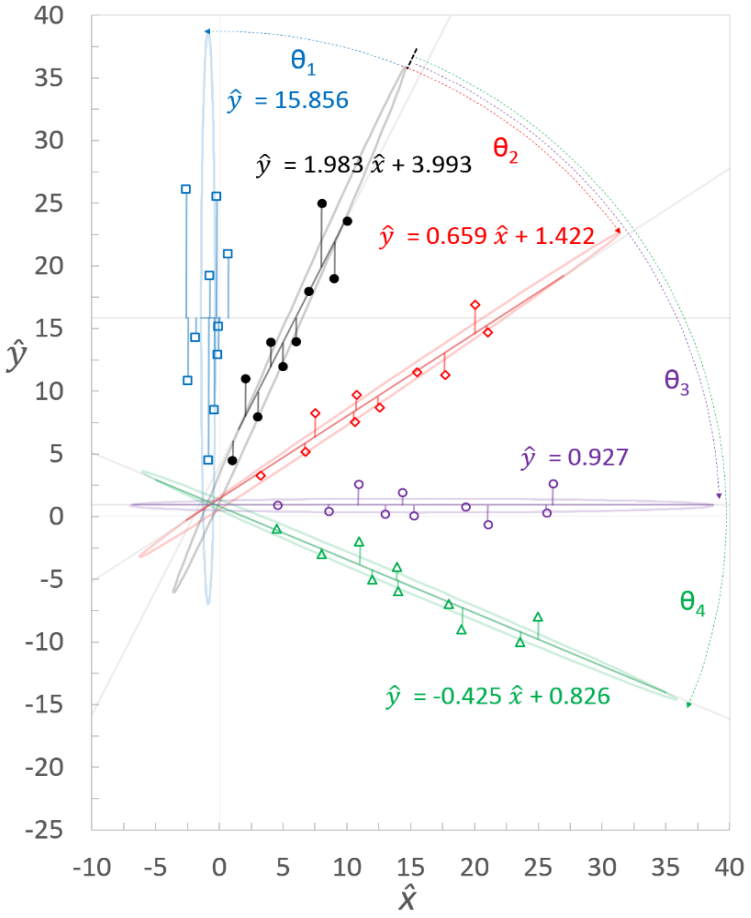

Figure 6 *A fitted line $\hat{y} = \hat{a}\hat{x} + \hat{b}$ using vertical offsets for a data set that has been rotated about the origin by angle θ. The original data set is shown as solid circles. The hollowed data sets have been rotated by the same angles found in Figure 5.*

$$\phi_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{a}\hat{x}_i + \hat{b} - \hat{y}_i\right)^2 \tag{37}$$

where the values of $\hat{a}$ and $\hat{b}$ are calculated in the rotated frame using offsets that are vertical with respect to the rotated axes.

$$\hat{a} = \frac{SS_{\hat{x}\hat{y}}}{SS_{\hat{x}\hat{x}}} = \frac{SS_{xy}\cos 2\theta + \frac{1}{2}[SS_{xx} - SS_{yy}]\sin 2\theta}{\frac{1}{2}[SS_{xx} - SS_{yy}]cos2\theta - SS_{xy}\,sin2\theta + \frac{1}{2}[SS_{xx} + SS_{yy}]} \tag{38}$$

$$\hat{b} = \frac{\langle\hat{y}_i\rangle SS_{\hat{x}\hat{x}} - \langle\hat{x}_i\rangle SS_{\hat{x}\hat{y}}}{SS_{\hat{x}\hat{x}}} = \frac{\left(\langle x_i^2\rangle\langle y_i\rangle - \langle x_i y_i\rangle\langle x_i\rangle\right)\cos\theta + \left(\langle y_i^2\rangle\langle x_i\rangle - \langle x_i y_i\rangle\langle y_i\rangle\right)\sin\theta}{\frac{1}{2}[SS_{xx} - SS_{yy}]cos2\theta - SS_{xy}\,sin2\theta + \frac{1}{2}[SS_{xx} + SS_{yy}]} \tag{39}$$

Equations (38) and (39) represent the slope, $\hat{a}$, and intercept, $\hat{b}$, in the rotated frame using vertical offsets with respect to the rotated axes. This slope and intercept have also been expressed in terms of the averages and variances in the fixed frame and the angle of rotation θ. The result of this rotation is shown in Fig. 6 for the identical four rotations as in Fig. 5a from the

previous section. Notice that the rotation through θ_3 leads to perpendicular offsets and the same fit as a rotation through θ_{min} in Fig. 5a. Also, in Fig. 6, the rotation θ_1 brings the data to a vertical inclination; however, the use of vertical offset leads to a horizontal fitted line. This is a peculiarity of the least squares method but does indeed represent the minimum error for that angle. Other rotations give lines with offsets that are not perpendicular. If we substitute $\hat{x}_i = x_i \cos\theta - y_i \sin\theta$ and $\hat{y}_i = x_i \sin\theta + y_i \cos\theta$ along with Eqs. (38) and (39) into Eq. (37) and we perform the summation, the result can be simplified to:

$$\phi_1 = \frac{2\langle x_i y_i\rangle\langle x_i\rangle\langle y_i\rangle - \langle x_i y_i\rangle^2 - \langle x_i^2\rangle\langle y_i\rangle^2 - \langle y_i^2\rangle\langle x_i\rangle^2 + \langle x_i^2\rangle\langle y_i^2\rangle}{\frac{1}{2}[SS_{xx} - SS_{yy}]cos2\theta - SS_{xy}\, sin2\theta + \frac{1}{2}[SS_{xx} + SS_{yy}]} \tag{40}$$

This result can be differentiated with respect to θ and set equal to zero to give the same result as Eq. (27) from the previous section using the horizontal line to fit the data. This is not a surprising result. The quantity $SS_{\hat{x}\hat{y}}$ is zero when evaluated at the extrema which means the slope $\hat{a}$ is zero and the intercept $\hat{b} = \langle \hat{y}_i \rangle$ as in the previous section. See Eqs. (38) and (39). Instead, we are interested in developing expressions for the slope a and intercept b for any arbitrary rotation, and hence, an arbitrary slope of the offset. This can be achieved by substituting Eqs. (38) and (39) along with $\hat{x} = x\cos\theta - y\sin\theta$ and $\hat{y} = x\sin\theta + y\cos\theta$ into the equation $\hat{y} = \hat{a}\hat{x} + \hat{b}$ and simplifying to:

$$y = \left(\frac{SS_{xy}\cos\theta - SS_{yy}\sin\theta}{SS_{xx}\cos\theta - SS_{xy}\sin\theta}\right)x + \langle y_i\rangle - \left(\frac{SS_{xy}\cos\theta - SS_{yy}\sin\theta}{SS_{xx}\cos\theta - SS_{xy}\sin\theta}\right)\langle x_i\rangle \tag{41}$$

so that

$$a = \frac{SS_{xy}\cos\theta - SS_{yy}\sin\theta}{SS_{xx}\cos\theta - SS_{xy}\sin\theta} \qquad \text{and} \qquad b = \langle y_i\rangle - a\langle x_i\rangle \tag{42a,b}$$

Notice that Eq. (42b) is the centroid equation for the intercept.

The least squares slope and intercept for offsets of any arbitrary slope are given by the expressions above. Notice that for $\theta = 0$ the vertical offset expressions for a and b are recovered, and for $\theta = \pm\frac{\pi}{2}$ the expressions for horizontal offsets are recovered. Also, by inserting Eq. (27) into Eqs. (42a) and (42b), the results for perpendicular offsets can be recovered. This rotation back to the original data position is shown in Fig. 7 for each of the four rotations in Fig. 6. In Fig. 7c, the perpendicular offset result from §2 is reproduced. The plot in Fig. 7b illustrates the fit for offsets that have been rotated from the vertical by $-\theta_3$. In Fig. 7d, a 90° rotation yields a fit using horizontal offsets. Finally, in Fig. 7a we have parallel offsets. This leads to a very poor fit of the data and can be avoided. To explain this situation, Fig. 8 shows the slope $\hat{a}$ and intercept $\hat{b}$ in the rotated frame as a function of the angle of rotation. Also shown is the error ϕ_1. Starting at $\theta = 0$ and rotating in the positive (counterclockwise) direction, the slope increases as the data becomes more vertical. This also increases the values

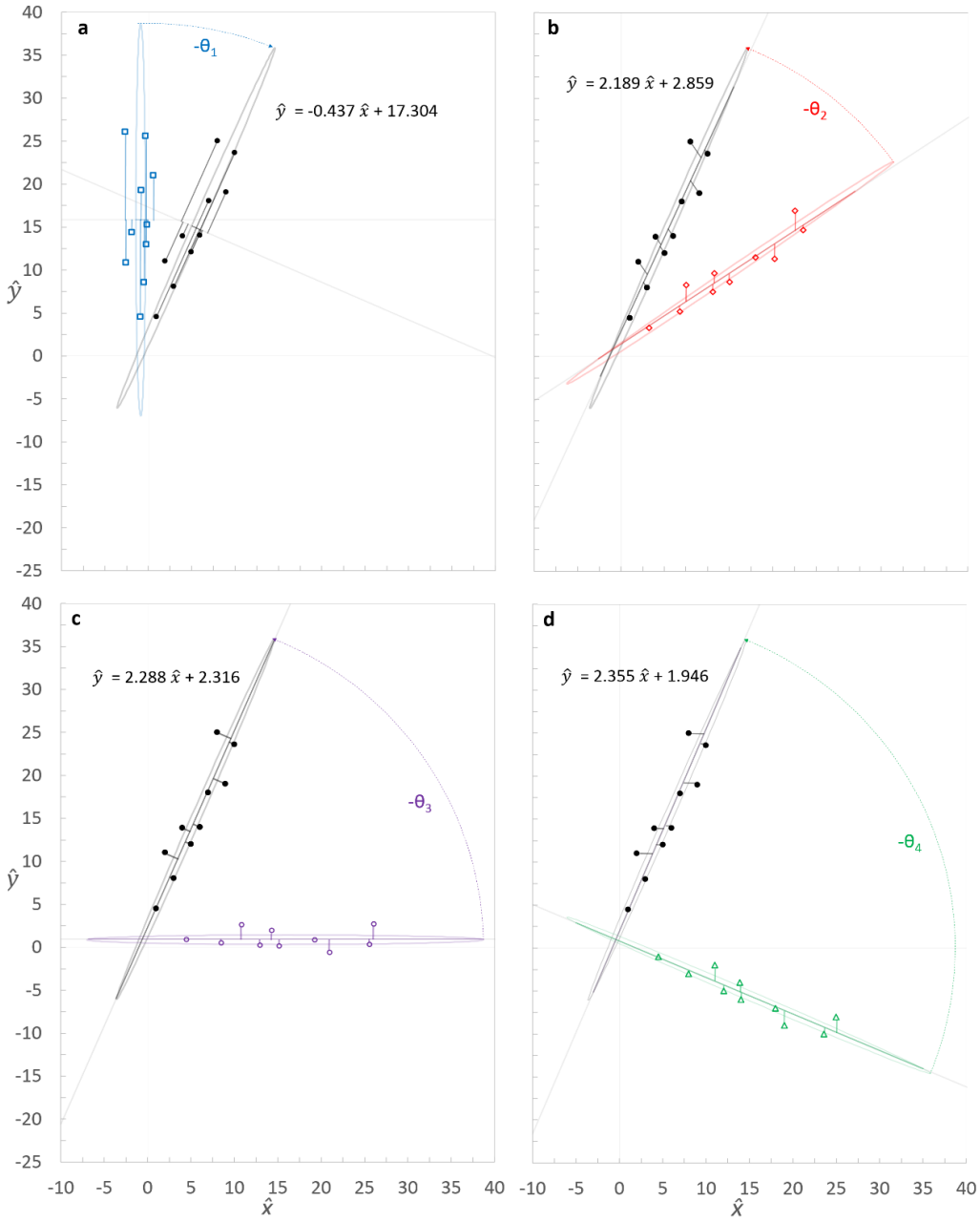

Figure 7 *Transformation of rotated data set and line of best fit with vertical offsets back to non-rotated coordinate system. a) Shows worse-case scenario where vertical offsets are taken for a vertically oriented data set. b) Intermediate case of offset results that are between vertical and perpendicular fits. c) Perpendicular offsets acquired after transforming back from horizontal oriented data. d) Horizontal offsets result when rotation transformation of π/2 is used.*

of ϕ_1. Eventually, the vertical offsets become more and more parallel to the fitted line causing the error to dramatically increase. The minimization problem leads to a fitted line that becomes horizontal which gives the minimum error.

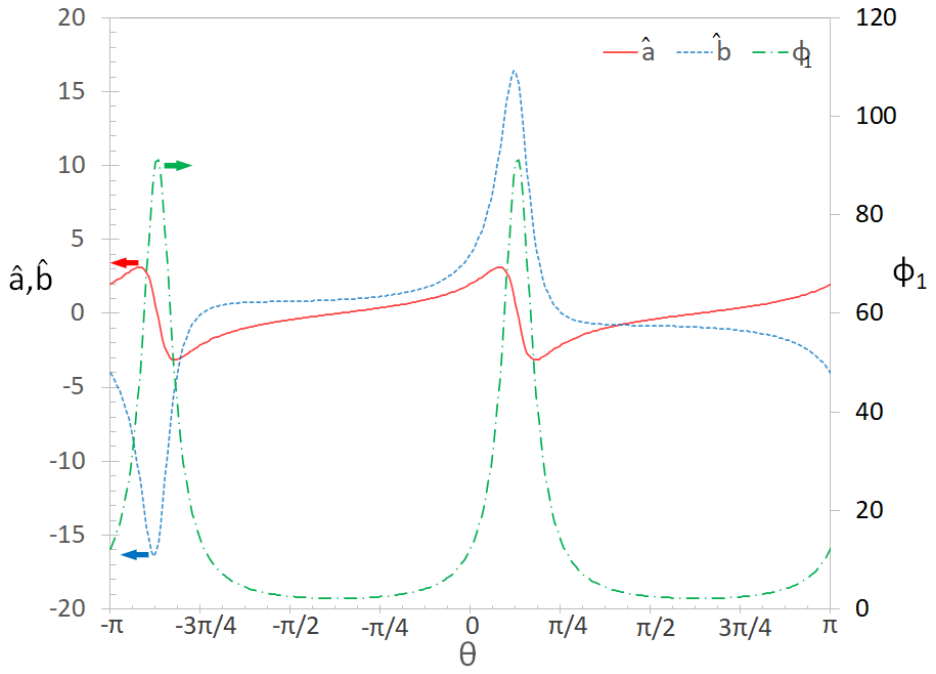

Figure 8 *Relationship of the line of best fit (using vertical offsets) slope, $\hat{a}$; line of best fit intercept, $\hat{b}$; and the sum of the square residuals with respect to the angle of rotation θ.*

We now relate the offsets to the relative confidence levels of the data. We use the results of Neri[12] in which Eq. (14) is modified by a weighting factor W_i derived using the propagation of error law such that

$$\phi_{\sigma_i} = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \frac{[y_i-(a\,x_i+b\,)]^2}{1+a^2} W_i = \sum_{i=1}^{n} \frac{[y_i-(a\,x_i+b\,)]^2}{\sigma_{y,i}^2+\sigma_{x,i}^2\,a^2} \quad \text{where } W_i = \frac{1+a^2}{\sigma_{y,i}^2+\sigma_{x,i}^2\,a^2} \qquad (43)$$

For the purpose of this paper, the uncertainties in $x_i$and y_i, independent from one another, are assumed equal for x_i to give $\sigma_{x,i}^2 = \sigma_x^2$ where σ_x^2 a constant and equal for y_i to give $\sigma_{y,i}^2 = \sigma_y^2$ where σ_y^2 a constant. This assumption means that all offsets will have the same slope. Under these conditions, Eq. (43) simplifies to:

$$\phi_\sigma = \sum_{i=1}^{n} \frac{[y_i-(a\,x_i+b\,)]^2}{\sigma_y^2+\sigma_x^2\,a^2} \qquad (44)$$

These uncertainties can be related to an offset of some orientation to the axes and a weighting factor which can be determined. At zero uncertainty in x-data, when $\sigma_x^2 \to 0$, the least squares vertical offset is recovered; conversely, at zero uncertainty in y-data, when $\sigma_y^2 \to 0$, the least squares horizontal offset is recovered. Lastly, when both the x and y uncertainties are equal, $\sigma_x^2 = \sigma_y^2$, perpendicular offsets are obtained.

We now wish to derive a similar expression to Eq. (44) in terms of slope and intercept of the fitted line and the slope of the offset. The sum of the square residuals for an arbitrary offset can be derived by first determining the square distance from the data point (x_i, y_i) and the line

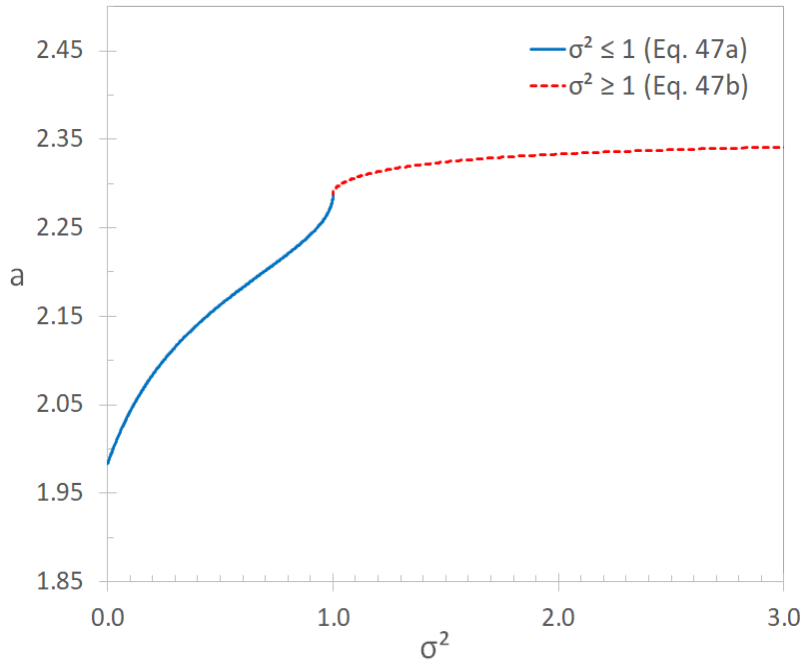

Figure 9 *Relationship of the slope of the line of best fit (using vertical offsets) to the ratio of uncertainties, σ^2. The solid line represents the solution given from Eq. 47a for $\sigma^2 \leq 1$; the dashed line represents the solution for Eq. 47b for $\sigma^2 \geq 1$. The transition between the two solutions at $\sigma^2 = 1$ is continuous and smooth. Additionally, when the lines meet at $\sigma^2 = 1$, perpendicular offsets with equal uncertainties is recovered.*

$y = ax + b$ using an offset with slope $\gamma = (y - y_i)/(x - x_i)$. By combining these relations with the distance formula, $d_i^2 = (x - x_i)^2 + (y - y_i)^2$ and after simplification yields:

$$\phi_v = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \frac{[y_i - (a\,x_i + b)]^2}{\frac{(a-\gamma)^2}{1+\gamma^2}} w \tag{45}$$

where we have added a weighting factor of w to account for the fact that the uncertainties in Eq. (43) change both the slope of the offset and multiply in a weighting factor. Equating the expressions ϕ_σ and ϕ_v in Eqs. (44) and (45) can be done in at least two ways:

$$\frac{(a-\gamma)^2}{1+\gamma^2} = 1 + \sigma^2 a^2 \quad \text{for } \sigma^2 \leq 1 \text{ where } \quad w = \frac{1}{\sigma_y^2} \tag{46a}$$

$$\frac{(a-\gamma)^2}{1+\gamma^2} = \frac{1}{\sigma^2} + a^2 \quad \text{for } \sigma^2 \geq 1 \quad \text{where } \quad w = \frac{1}{\sigma_x^2} \tag{46b}$$

where $\sigma^2 = \sigma_x^2/\sigma_y^2$ is the ratio of the uncertainties. The equations above lead to two expressions for σ^2. We recognize that the rotational method gives the slope of the offsets as $\gamma = \cot\theta$. This relation can be used in Eq. (42a) for the slope of the line a to eliminate θ and combine the result with Eqs. (46a) and (46b) to eliminate γ. Hence, expressions for σ^2 can be derived.

$$\sigma^2 = \frac{[a^2-1]-2a(SS_{yy}-aSS_{xy})/(SS_{xy}-aSS_{xx})}{a^2\left(1+\left[(SS_{yy}-aSS_{xy})/(SS_{xy}-aSS_{xx})\right]^2\right)} \quad \text{for } \sigma^2 \leq 1 \text{ where } a_v \leq a \leq a_\perp \qquad (47a)$$

$$\sigma^2 = \frac{1+\left[(SS_{xy}-aSS_{xx})/(SS_{yy}-aSS_{xy})\right]^2}{[1-a^2]-2a(SS_{xy}-aSS_{xx})/(SS_{yy}-aSS_{xy})} \quad \text{for } \sigma^2 \geq 1 \text{ where } a_\perp \leq a \leq a_h \qquad (47b)$$

Here $a_v = SS_{xy}/SS_{xx}$ is the slope for vertical offsets while $a_h = SS_{yy}/SS_{xy}$ is the slope for horizontal offsets. The results are shown in Fig. 9. The solid line shows how the slope changes with relative uncertainty. For $\sigma^2 = 0$ the vertical offset slope is recovered and as $\sigma^2 \to \infty$ the slope for horizontal offsets is recovered. The curves for Eq. (47a) meets that of Eq. (47b) at $\sigma^2 = 1$ when the offsets are perpendicular. These equations give the complete relation between σ^2 and the slope of the fitted line. The method illustrated above demonstrates how relative uncertainty in data effects the slope of the fitted line. The intercept can be calculated using the centroid equations as usual.

§4 Quality of fit for arbitrary offsets

In §1, we establish that R^2_{max} is a measure of how well a data sets conforms to a straight line independent of the orientation of the data. We were able to interpret R^2_{max} in geometric terms using the eigenvalues of the covariance matrix. In §2, we established that the fit using perpendicular offsets was the line for which R^2_{max} represented the quality of fit. In §3, we generalized our method to fit data with lines of arbitrary offsets. We now wish to generalize Eq. (11) for R^2_{max} to represent the quality of fit for these fits with arbitrary offsets. As shown in Fig. 10, these lines will pass through the centroid point of the data. In general, they will have a slope different from the slope given by a perpendicular offset fit. The angle ψ between the line fit with perpendicular offsets and the line fit with arbitrary offsets is given by $m \equiv \tan\psi = \frac{a-a_\perp}{1+aa_\perp}$ where $a_\perp$ is the slope for of the line fit with perpendicular offsets and a is the slope of the line fit with arbitrary offsets. Using the eigenvalues λ_1 and λ_2 as the major and minor axes of the ellipse, respectively, the distance traced across the ellipse by the line fit with arbitrary offsets can be found as:

$$\Lambda_1 = \sqrt{\frac{(1+m^2)\lambda_1^2\lambda_2^2}{m^2\lambda_1^2+\lambda_2^2}} \quad \text{where } m^2 = \left(\frac{a-a_\perp}{1+aa_\perp}\right)^2 \qquad (48)$$

We now define a generalized quality of fit Q^2 based on Eq. (11) where λ_1 is replaced by Λ_1:

$$Q^2 = \left(\frac{\Lambda_1-\lambda_2}{\Lambda_1+\lambda_2}\right)^2 = \left(\frac{\lambda_1\sqrt{1+m^2}-\sqrt{m^2\lambda_1^2+\lambda_2^2}}{\lambda_1\sqrt{1+m^2}+\sqrt{m^2\lambda_1^2+\lambda_2^2}}\right)^2 \qquad (49)$$

In Fig. 11 we show a plot of Q^2 as a function of θ. Rotating through 90˚ towards the perpendicular fit gives a reasonable set of linear fits for a range of offsets from vertical to horizontal. The perpendicular offset gives the optimal fit of R^2_{max} to the data and is recovered when m = 0 or $a = a_\perp$. Rotating away from the perpendicular fit moves towards suboptimal fit.

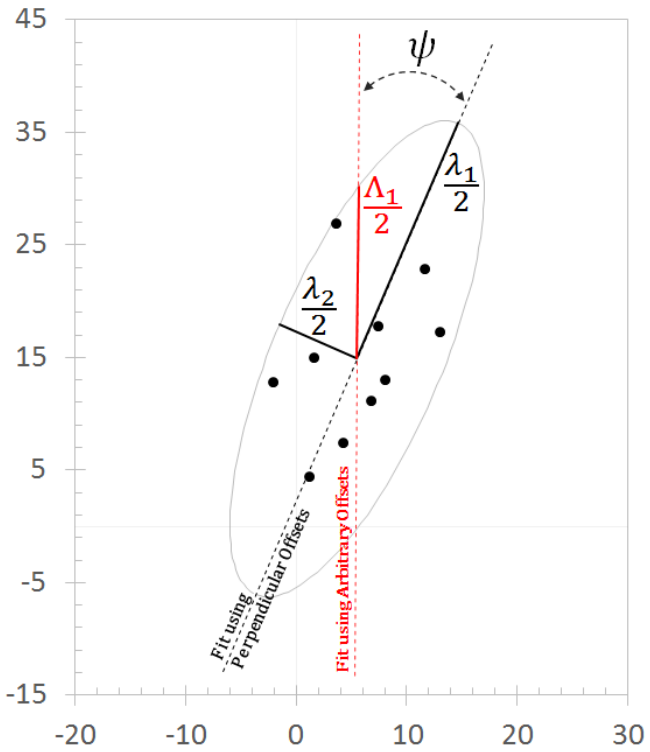

Figure 10 *Comparison of arbitrary fits using non-perpendicular offsets to a fit using perpendicular offsets. All fits pass through a common centroid point. The angle between the perpendicular offset fit and the arbitrary non-perpendicular offset fit is represented by ψ. λ_1 and λ_2 are the principle axes of the ellipse given by the perpendicular offset fit; Λ_1 is the first principle axis of the ellipse using any arbitrary offset.*

This includes offsets that lie parallel to the preferred direction of the data and creates fitted line that is at a right angle to the perpendicular offset line. This leads to a Q^2 of zero because $\Lambda_1 = \lambda_2$. The quantity Q^2 gives a rationale measure of quality of fit for lines of arbitrary slope passing through the data through the centroid point. This quantity can be used in place of R^2 which is not appropriate as a measure of quality of fit.

Conclusions

The common practice of fitting data with linear least squares using vertical offsets and describing the quality of the fit with the determination coefficient R^2 can be misleading. We

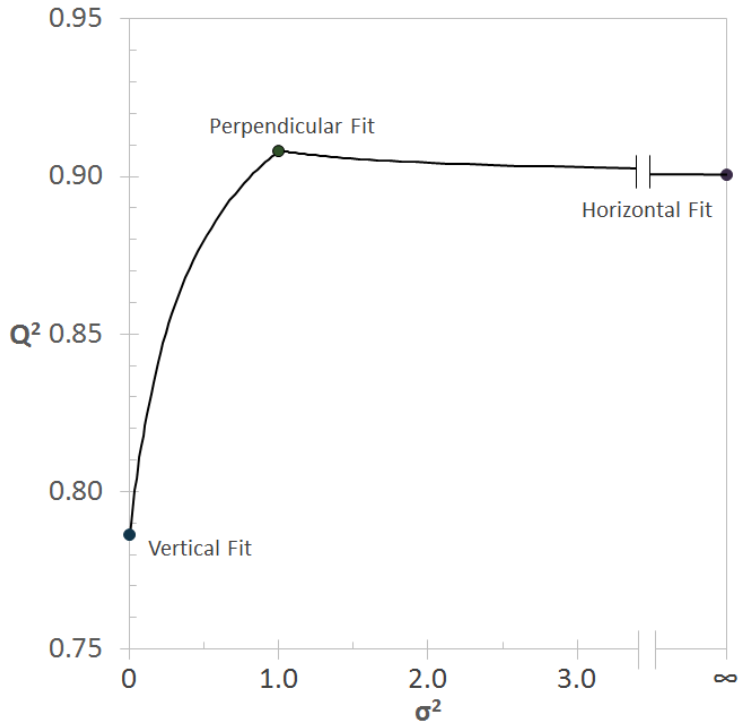

Figure 11 *Correlation of the quality of fit, Q^2, with respect to the uncertainty ratio, σ^2. When the uncertainty ratio is exactly equal to one, in which the confidence levels in x and y are equal, the value of Q^2 is obtained that pertains to a perpendicular fit. When the uncertainty ratio is zero or approaches infinity, the Q^2 values obtained match that of a vertical or horizontal fit, respectively.*

have shown that the determination coefficient can severely under-estimate the quality of the fit of a given line to the data. When no information about confidence levels is known, it would be preferable to use perpendicular offsets which assume equal confidence levels and gives the highest quality of fit. When the confidence levels are available, or at least a good estimate of them, then more appropriate offsets could be used allowing for more adjustability in the lower confidence level variable.

In this paper, we have presented Q^2, a more meaningful measure for the degree of scatter in a data set which was derived using a pure rotation to a data set. This quantity is invariant to the slope or inclination of the data with the axes and describes the scatter of the data about the best fit line given offsets. The quantity Q^2 varies from zero to one and can be interpreted much like the coefficient of determination; however, the interpretation is not clouded by the effects of the slope of the data. This leads to more meaningful and reliable comparisons of the quality of fit of data. We do recognize that the degree of scatter is particular to the scatter about the line found using perpendicular offsets. Although there are ways to account for the discrepancy causes by using offsets other than perpendicular, the changes are small for typical data sets.

An extremely simple discriminant has been developed for choosing the correct solution of least squares using perpendicular offsets. The method of rotating the data to optimize the fit of a horizontal line gives an equivalent result to the traditional method of varying the slope and intercept. The rotational method is simpler in that only the angle of rotation is varied which leads to a simpler optimization problem. Recasting the problem in this rotation method reveals that only the sign of the differences in the variances, namely, $SS_{xx} - SS_{yy}$, is needed to determine the correct solution.

The rotation method has been used to find the fits of straight lines using variable offsets. The method makes it easy to use offsets with any inclination to the axes including horizontal, vertical, perpendicular, or any inclination in between. Moreover, these variable offsets can be related to the uncertainties in the data if we assume that all x-values have equal uncertainty and likewise all y-values have equal uncertainties. Although this is a case of limited application, it does allow fitting data and estimating the relative effect of error in one variable compared to the other.

Appendix

In this appendix, we illustrate the method with an example. We use the same data set used to produce Figs. 3, 5, 6, and 7 which consists of the following ten points (x_i, y_i): (1, 4.5), (2, 11), (3, 8), (4, 13.9), (5, 12), (6,14), (7, 18), (8, 25), (9, 19), and (10, 23.6). Using the definitions in the introduction, the averages are: $\langle x_i \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i = 5.5$ and $\langle y_i \rangle = \frac{1}{n}\sum_{i=1}^{n} y_i = 14.9$. The variances and covariance are: $SS_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x_i \rangle)^2 = 8.25$, $SS_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y_i \rangle)^2 = 38.532$, and $SS_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x_i \rangle)(y_i - \langle y_i \rangle) = 16.36$. The least squares fit using perpendicular offsets can now be found. The quantity $B = \frac{SS_{xx} - SS_{yy}}{2SS_{xy}} = -0.9255$ and the discriminant $\left[SS_{xx} - SS_{yy}\right] = -30.282 < 0$. Because the discriminate is negative, we use Eq. (34b) to calculate the slope, $a_{\perp} = -B - B\sqrt{1 + \frac{1}{B^2}} = 2.288$. Had the discriminate been positive, Eq. (34a) would have been applied. Now use the centroid equation (Eq. (35)) to calculate the intercept, $b_{\perp} = \langle y_i \rangle - a_{\perp}\langle x_i \rangle = 2.316$.

To calculate the quality of fit when using perpendicular offsets, we need to calculate the eigenvalues using Eqs. (9) and (10) respectively to give $\lambda_1 = 45.682$ and $\lambda_2 = 1.100$ which leads to $Q^2 = R^2_{max} = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2 = 0.9082$. If there is no information about the relative uncertainties for the x-values versus the y-values, then it is reasonable to use the perpendicular offset slope and intercept along with the $Q^2 = R^2_{max}$ value to describe the quality of the fit.

However, if the relative uncertainty is fixed, then the effect of the uncertainty on the slope and intercept can be determined using Eqs. (47a) and (47b). For the example above, the uncertainty ratio is $\sigma^2 = 1$ for the case of perpendicular offsets. We can examine other

uncertainty ratios, for example, if $\sigma^2 = 0.3$, then Eq. (47a) is used to solve for the slope. This can be done numerically to give $a = 2.116$ and again using the centroid equation (Eq. 42b) gives $b = 3.262$. In contrast, if $\sigma^2 = 3.0$, then Eq. (47b) is used to determine a slope of $a = 2.341$ which again using the centroid equation gives $b = 2.025$. Notice that varying the uncertainty has a significant impact on the fitted line.

Finally, the quality of the fits must be evaluated for the two uncertainties above. For $\sigma^2 = 0.3$, the slopes already calculated above can be used to find $m^2 = 0.0008738$ which gives $\Lambda_1 = 28.861$ according to Eq. (48). Using Eq. (49) the quality for this case is $Q^2 = 0.8586$ which is lower than the optimal value for perpendicular offsets. For $\sigma^2 = 3.0$, a similar calculation yields $Q^2 = 0.9030$ which is again reduced from the optimal value. Using non-perpendicular offsets will always reduce the quality of the fit.

References

[1] Sampaio, Jorge H.B. "An iterative procedure for perpendicular offsets linear least squares fitting with extension to multiple linear regression." *Applied Mathematics and Computation* **176** (1), 91-98. (2006).

[2] Morrison, S. J. *Statistics for Engineers an Introduction.* (John Wiley & Sons, 2009).

[3] York, D. "Least-Squares Fitting Of A Straight Line." *Canadian Journal of Physics* **44**, 1079-1086. (1966).

[4] York, D. "Least-Squares Fitting Of A Straight Line With Correlated Errors." *Earth And Planetary Science Letters* **5**, 320-324. (1969).

[5] Reed, Cameron B. "Linear least-squares fit with errors in both coordinates." *American Journal of Physics* **57** (7), 642-646. (1989).

[6] Reed, Cameron B. "Linear least-squares fits with errors in both coordinates. II: Comments on parameter variances." *American Journal of Physics* **60** (1), 59-62. (1992).

[7] Moreno, C., and H. Bruzzone. "Parameter's variances of a least-squares determined straight line with errors in both coordinates." *Meas. Sci. Technol.* **4**, 635-636. (1993).

[8] Borcherds, P. H., and C. V. Sheth. "Least Squares fitting of a straight line to a set of data points." *Eur. J. Phys.* **16,** 204-210. (1995).

[9] Sheth, C. V., A. Ngwengwe, and P. H. Borcherds. "Least Squares fitting of a straight line to a set of data points: II. Parameter variances." *Eur. J. Phys.* **17**, 322-326. (1996).

[10] Weisstein, Eric W. "Least Squares Fitting." *MathWorld--A Wolfram Web Resource.* <http://mathworld.wolfram.com/LeastSquaresFitting.html>. (2019).

[11] Stewart, James. *Calculus.* Seventh Edition. (Brooks/Cole, 2012).

[12] Neri, F., G. Saitta, and S. Chiofalo. "An accurate and straightforward approach to line regression analysis of error-affected experimental data." *Journal of Physics E: Scientific Instruments* **22** (4), 215-217. (1989).

[13] Spiess, Andej-Nikolai, and Natalie Neumeyer. "An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach." *BMC Pharmacology* **10** (6). (2010).