PLOS | ONE

RESEARCH ARTICLE

Invariance properties for the error function used for multilinear regression

Mark H. Holmes[1]*, **Michael Caiola**[2]

[1] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York, United States of America, [2] Yerkes National Primate Laboratory/Udall Center, Emory University, Atlanta, Georgia, United States of America

* holmes@rpi.edu

OPEN ACCESS

Citation: Holmes MH, Caiola M (2018) Invariance properties for the error function used for multilinear regression. PLoS ONE 13(12): e0208793. https://doi.org/10.1371/journal.pone.0208793

Editor: J Malo, Universitat de Valencia, SPAIN

Received: February 27, 2018

Accepted: November 26, 2018

Published: December 26, 2018

Data Availability Statement: The data sets used are all publicly available: EMAP Surface Waters Lake Database: 1991-1994 Northeast Lakes Water Chemistry Data Summarized by Lake https://archive.epa.gov/emap/archive-emap/web/html/nelakes.html MAIA-Estuaries Summary Database: 1997 and 1998 Stations https://archive.epa.gov/emap/archive-emap/web/html/index-70.html National Data Buoy Center: Historical Meteorological Data - Station 46006 www.ndbc.noaa.gov/histsearch.php?station=46006 U.S. Census Bureau Statistical Abstract of the United States: 2012 Modeling wine preferences by data mining from physicochemical properties

Abstract

The connections between the error function used in multilinear regression and the expected, or assumed, properties of the data are investigated. It is shown that two of the most basic properties often required in data analysis, scale and rotational invariance, are incompatible. With this, it is established that multilinear regression using an error function derived from a geometric mean is both scale and reflectively invariant. The resulting error function is also shown to have the property that its minimizer, under certain conditions, is well approximated using the centroid of the error simplex. It is then applied to several multidimensional real world data sets, and compared to other regression methods.

Introduction

The problem considered here concerns the modeling assumptions made in multilinear regression, and their role in determining the error function. To provide a simple example, a principal component analysis (PCA) is used to find low dimensional subspace approximations of a data set. It has the property that if the data set is rotated, and a PCA is used, the rotated versions of the same low dimensional subspace approximations are obtained. This means that a PCA is rotationally invariant. This is one of the reasons that it is often used in face recognition [1, 2], visual tracking [3], and other pattern recognition problems.

In contrast, linear least squares is not rotationally invariant. However, unlike a PCA, it is scale invariant. What this means is that if you scale the variables in the data set, the resulting minimizer is the scaled version of what is obtained for the unscaled case (this is explained more precisely in the next section). Scale invariance is important, for example, if you want your minimizer to be independent of what units are used (e.g., inches versus centimeters). The fact that a PCA is scale dependent, and that it is possible to be fairly sensitive to the scaling, is well-known [4, 5].

A third type of invariance, which will play a central role in this paper, concerns the order the variables are listed or labeled. Specifically, if the minimizer is unaffected by the reordering of the variables, the result is said to be reflectively invariant. A simple example of where this is important is edge detection, where the minimizer should be unaffected by which axis is labeled x, y or z. It is not hard to show that a PCA is reflectively invariant, but that linear least squares

http://www.sciencedirect.com/science/article/pii/S0167923609001377.

Funding: This work was supported by National Science Foundation (MH and MC received funding) (https://www.nsf.gov/div/index.jsp?div=DMS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

is not. It needs to be pointed out that there are different forms of reflective invariance depending on the hyperplane used for the reflection. As an example, in computer vision it is desirable to be able to recognize an object regardless of whether you are looking at it, or at its horizontal reflection as seen when looking in a mirror [6]. In this case the reflection is through the vertical plane. There is also some variation in how to refer to reflective invariance as used here. As a case in point, it has been referred to as neutral data fitting because, for this form of invariance, the variables must be treated symmetrically [7, 8].

Obtaining a regression result with particular invariance properties has been considered earlier, although often framed in somewhat different language. One of the earliest studies concerned reflective invariance for a bivariate problem using area as a measure of error [9]. This is the E_A example illustrated in Fig 1. This was the impetus for Samuelson [10] to propose certain invariant properties one might expect, or require, of the error function. Since then numerous attempts have been made to find error functions that have one or more of these properties. An example of a straightforward approach for two variable linear regression is to concentrate on the slope of the regression line. It has been proposed that, after determining the lines for vertical and horizontal least squares, E_y and E_x in Fig 1, to simply use the line whose slope is determined using an averaging method involving the two slopes. A review and comparison of these, and similar methods, can be found in [11, 12].

A more fundamental approach is to concentrate on the error function. One possibility is to use true distance, E_D in Fig 1, and this is easily generalized to multilinear regression (a PCA is an example). Another possibility is to use area, E_A in Fig 1, which is what Woolley used, and this gives rise to what is called least product regression, or least area regression. Work has been done on how to generalize Woolley's idea, and use symmetry methods to obtain invariant error functions in two and three dimensions [7, 8]. However, this is not easily generalized to multilinear regression. An alternative, which is most relevant to the present study, is to use the geometric mean for each data point and then find the least squares value for this function [13]. This differs from the geometric mean considered here, which involves

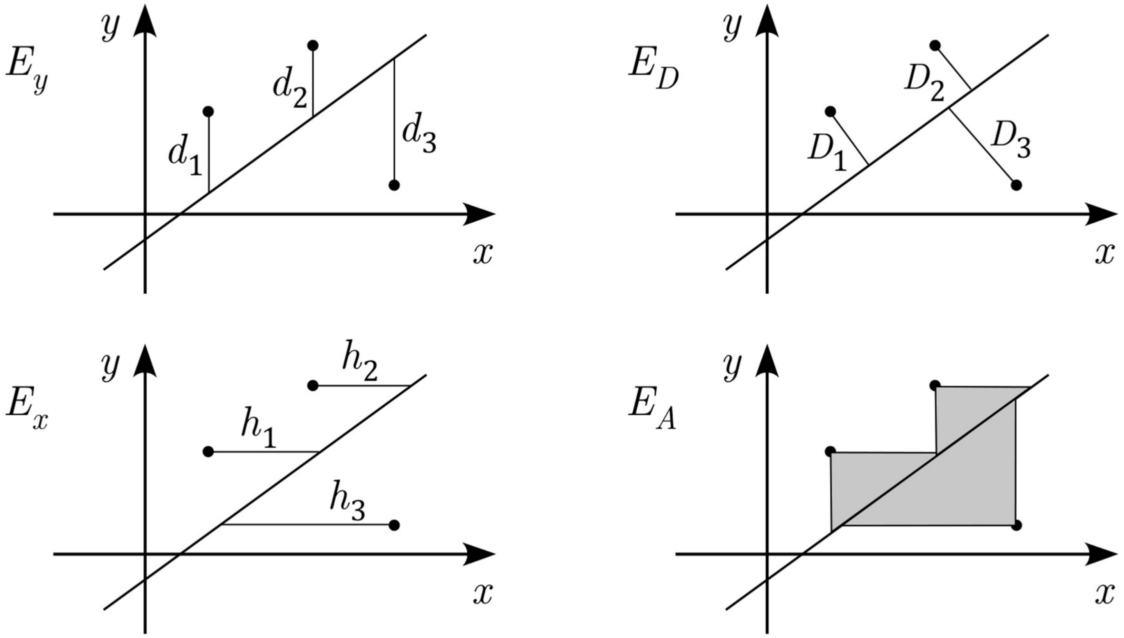

Fig 1. Error functions for bivariate regression. Shown are: $E_y = \sum d_i^2$, $E_D = \sum D_i^2$, $E_x = \sum h_i^2$, and $E_A = \frac{1}{2}\sum d_i h_i$.

https://doi.org/10.1371/journal.pone.0208793.g001

the geometric mean of the ordinary least squares error functions. The exception to this statement occurs when a hyperplane approximation is used, in which case the two formulations are equivalent. In the current study a hyperplane approximation is considered, but so are the other lower dimensional approximations that are possible (similar to what is done using a PCA).

In the next section it is shown that scale and rotational invariance are incompatible. This is done, for the case of two variables, by first characterizing mathematically what is needed to obtain particular invariant properties, and then to use similarity methods to demonstrate the incompatibility. In addition, in this two-dimensional setting, an error function that has several important invariance properties is formulated and discussed. In the subsequent two sections, the extension of this function to multilinear linear regression problems is considered, which includes showing that the minimizer is well-approximated using the centroid of an error simplex. With this, the error function is used to analyze real world data sets and compared to other regression methods.

Two variables

To begin, we start with the case of two variables. Assume that the data are (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), and they are centered. This means that $\sum x_i = \sum y_i = 0$. It is assumed in what follows that the data vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ are not orthogonal.

For the model function $y = \alpha x$, there are numerous ways to measure the error and four possibilities are shown in Fig 1. For a PCA the true distance is used, and this leads to the error function

$$\begin{aligned} E_D(\alpha) &= \sum_{i=1}^{n} D_i^2 \\ &= \frac{1}{1+\alpha^2}\sum_{i=1}^{n}(\alpha x_i - y_i)^2 \\ &= \frac{1}{1+\alpha^2}(\alpha^2\mathbf{x}\cdot\mathbf{x} - 2\alpha\mathbf{x}\cdot\mathbf{y} + \mathbf{y}\cdot\mathbf{y}). \end{aligned}$$

Because of the denominator, for this expression to be defined, α must be dimensionless. What this means is that, when x and y have different dimensions, it is first necessary to nondimensionalize the variables before writing down the formula for the error function. So, suppose the data are scaled as $X_i = x_i/S_x$ and $Y_i = y_i/S_y$, where S_x and S_y are positive. The model function is now $Y = \bar{\alpha}X$, and the corresponding error function is

$$\frac{1}{1+\bar{\alpha}^2}\sum_{i=1}^{n}(\bar{\alpha}X_i - Y_i)^2.$$

Minimizing this, and then transforming back to dimensional variables, one finds that

$$\alpha = \bar{\alpha}\,\frac{S_y}{S_x}\,, \tag{1}$$

where

$$\bar{\alpha} = \frac{1}{2}\left(-\lambda \pm \sqrt{\lambda^2 + 4}\right), \tag{2}$$

for

$$\lambda = \frac{\mathbf{X}\cdot\mathbf{X} - \mathbf{Y}\cdot\mathbf{Y}}{\mathbf{X}\cdot\mathbf{Y}}\,. \tag{3}$$

In the above expression, the + is used if $\mathbf{X}\cdot\mathbf{Y} > 0$ and the − is used if $\mathbf{X}\cdot\mathbf{Y} < 0$. What is evident from this calculation is that α depends on S_x and S_y. Typical choices for these scaling factors include $S_x = ||\mathbf{x}||_\infty$, $S_x = ||\mathbf{x}||_2/\sqrt{n}$, and $S_x = ||\mathbf{x}||_1/n$ (with similar expressions for S_y). Depending on the scatter in the data, the values of these quantities can be significantly different and this can result in rather dramatic differences in the corresponding value of α.

Scale invariance

There is a simple test for scale invariance that comes from the above analysis. To derive it, in the original x, y-coordinates and using the model equation $y = \alpha x$, whatever regression procedure is used will result in the slope α depending on the data. This is written as $\alpha = \alpha(\mathbf{x}, \mathbf{y})$. Using the scaling $X = x/S_x$ and $Y = y/S_y$, the model equation becomes $Y = \bar{\alpha}X$, where $\bar{\alpha} = \alpha S_x/S_y$. Now, scale invariance requires that $\bar{\alpha} = \alpha(\mathbf{X}, \mathbf{Y})$. Combining these two results, the requirement for scale invariance is that

$$\alpha(\mathbf{x}, \mathbf{y}) = \frac{S_x}{S_y}\alpha(\mathbf{x}/S_x, \mathbf{y}/S_y), \tag{4}$$

for any positive values of S_x and S_y. Using the infinitesimal generators $S_x = 1 + \varepsilon$ and $S_y = 1 + \delta$, then the above equation takes the form [14]

$$\begin{aligned} \alpha(\mathbf{x}, \mathbf{y}) &= \frac{1+\varepsilon}{1+\delta}\alpha(\mathbf{x}/(1+\varepsilon), \mathbf{y}/(1+\delta)) \\ &= \alpha(\mathbf{x}, \mathbf{y}) + \varepsilon[\alpha(\mathbf{x}, \mathbf{y}) - \nabla_x\alpha(\mathbf{x}, \mathbf{y})] - \delta[\alpha(\mathbf{x}, \mathbf{y}) + \nabla_y\alpha(\mathbf{x}, \mathbf{y})] + \cdots, \end{aligned}$$

where ∇_x is the gradient in the **x** variables (and similarly for ∇_y). The $O(\varepsilon)$ and $O(\delta)$ requirements are that

$$\mathbf{x}\cdot\nabla_x\alpha(\mathbf{x}, \mathbf{y}) = -\alpha(\mathbf{x}, \mathbf{y}), \qquad \text{and} \qquad \mathbf{y}\cdot\nabla_y\alpha(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y}).$$

These are easily solved using the n-dimensional version of spherical coordinates. Specifically, letting

$$\mathbf{x} = r\mathbf{X}(\phi_1, \phi_2, \cdots, \phi_{n-1}),$$

where $r = ||\mathbf{x}||_2$ and the ϕ_i's are the angular coordinates, then $\mathbf{x}\cdot\nabla_x\,\alpha = -\alpha$ reduces to $r\partial_r\,\alpha = -\alpha$. The general solution of this is $\alpha = c/r$, where c can depend on the ϕ_i's. Doing something similar for **y**, the conclusion is that, to be scale invariant, the minimizer must depend on the data as

$$\alpha = \frac{||\mathbf{y}||_2}{||\mathbf{x}||_2}F, \tag{5}$$

where F can depend on the values of the angular coordinates for **x** and **y**.

Rotational invariance

To determine the requirement for rotational invariance, assuming the angle of rotation is θ, then

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The model function now has the form $y' = \alpha' x'$, and invariance requires that

$$\alpha' = \frac{\alpha\cos\theta - \sin\theta}{\alpha\sin\theta + \cos\theta}.$$

Writing the minimizer of the error function as $\alpha = f(\mathbf{x}, \mathbf{y})$, then $\alpha' = f(\mathbf{x}', \mathbf{y}')$ results in the requirement that

$$\frac{f(\mathbf{x},\mathbf{y})\cos\theta - \sin\theta}{f(\mathbf{x},\mathbf{y})\sin\theta + \cos\theta} = f(\mathbf{x}\cos\theta + \mathbf{y}\sin\theta, -\mathbf{x}\sin\theta + \mathbf{y}\cos\theta).$$

Using the infinitesimal generator $\theta = \varepsilon$, then the $O(\varepsilon)$ requirement is

$$\mathbf{y}\cdot\nabla_x f - \mathbf{x}\cdot\nabla_y f + f^2 + 1 = 0, \qquad (6)$$

where $f = f(\mathbf{x}, \mathbf{y})$. As it should, the solution in Eqs (2) and (3) satisfies this nonlinear partial differential equation. What does not satisfy the equation is Eq (5). It is easiest to illustrate this assuming there are only two data points. In this case, the 2-dimensional spherical coordinates used in Eq (5) can be written as $x_1 = r\cos k$, $x_2 = r\sin k$, $y_1 = R\cos K$, and $y_2 = R\sin K$. Substituting Eq (5) into Eq (6), and reducing gives

$$\left(\frac{R^2}{r^2} + 1\right)\cos(k-K)F + \sin(k-K)\left(\frac{R^2}{r^2}\partial_k F + \partial_K F\right) = 1 + \frac{R^2}{r^2}F^2.$$

Given that F is independent of r and R, the above equation leads to the following two equations

$$\cos(k-K)F + \sin(k-K)\partial_k F = F^2,$$

and

$$\cos(k-K)F + \sin(k-K)\partial_K F = 1.$$

These equations are solvable by introducing the change of variables $s = k - K$, $t = k + K$, and from this one finds that it is not possible to find a function F that satisfies both equations. In other words, there is no function which satisfies both Eqs (5) and (6).

Scale and reflectively invariant error function

The conclusion from the above analysis is that it is not possible for the minimizer to be both scale and rotationally invariant. It is known that there are rotation and reflectively invariant error functions, and an example is one that uses true distance (E_D in Fig 1). So, the question considered here is whether there are error functions which satisfy all of the stated conditions except for rotational invariance.

It is relatively easy to find scale invariant error functions. For example, using the usual (vertical) least squares error

$$E_y(\alpha) = \sum_{i=1}^{n} (\alpha x_i - y_i)^2 = \alpha^2 \mathbf{x}\cdot\mathbf{x} - 2\alpha\mathbf{x}\cdot\mathbf{y} + \mathbf{y}\cdot\mathbf{y}, \tag{7}$$

the minimum occurs when $\alpha = \mathbf{x}\cdot\mathbf{y}/\mathbf{x}\cdot\mathbf{x}$. Similarly, using the (horizontal) least squares error

$$E_x(\alpha) = \sum_{i=1}^{n} (x_i - y_i/\alpha)^2 = \mathbf{x}\cdot\mathbf{x} - 2\mathbf{x}\cdot\mathbf{y}/\alpha + \mathbf{y}\cdot\mathbf{y}/\alpha^2, \tag{8}$$

the minimum occurs when $\alpha = \mathbf{y}\cdot\mathbf{y}/\mathbf{x}\cdot\mathbf{y}$. Both of these are scale invariant. What E_x and E_y are not, however, are reflectively invariant. To be reflectively invariant it is required that irrespective of which variables are considered independent or dependent, that an equivalent result is obtained. This means that the minimizer of E_x is the same as the one obtained for E_y. Mathematically, the requirement is that

$$\alpha(\mathbf{y},\mathbf{x}) = \frac{1}{\alpha(\mathbf{x},\mathbf{y})} . \tag{9}$$

The error function to be considered here is based on the geometric mean of the ordinary least squares error functions. In the case of two variables, the error function is

$$E(\alpha) = \sqrt{E_x(\alpha)E_y(\alpha)} , \tag{10}$$

where E_x and E_y are given in Eqs (7) and (8), respectively. Minimizing this one finds that

$$\alpha(\mathbf{x},\mathbf{y}) = \pm\sqrt{\frac{\mathbf{y}\cdot\mathbf{y}}{\mathbf{x}\cdot\mathbf{x}}}, \tag{11}$$

where the + is used if $\mathbf{x}\cdot\mathbf{y} > 0$ and the − is used if $\mathbf{x}\cdot\mathbf{y} < 0$. This satisfies the change of variables condition Eq (4), and this guarantees α is scale invariant. It is also reflectively invariant because it satisfies Eq (9). Another way to conclude that it is reflectively invariant is to note that the error function Eq (10) is a symmetric function of E_x and E_y.

The minimizer in Eq (11) is well-known and can be obtained in a number of ways. This includes using geometric mean regression [9], using the geometric means of the minimizers for Eqs (7) and (8), and by using simple symmetry arguments [10]. What is not known is a way to generalize it to multilinear regression to obtain scale and reflectively invariant low dimensional approximations of data. What is presented below is one way this might be possible.

For ordinary least squares, one of the standard measures on how well the linear model fits the data is the coefficient of determination R^2. For the multidimensional generalizations of Eq (10) considered later, a natural measure of fit involves the centroid of the error simplex. To explain what this is for this two dimensional problem, and connect it with R, let α_x and α_y be the minimizers of E_x and E_y, respectively. The centroid in this case is $\alpha_c = (\alpha_x + \alpha_y)/2$. The error measure to be introduced concerns how α, the minimizer of E, differs from the centroid, relative to the width of the simplex. The resulting formula is

$$C_F = \left|\frac{\alpha - \alpha_c}{\alpha_y - \alpha_x}\right|.$$

Now, using the law of cosines $\mathbf{x} \cdot \mathbf{y} = ||\mathbf{x}||_2 \cdot ||\mathbf{y}||_2 \cos\theta$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$, then $\alpha_x = \alpha/\cos\theta$ and $\alpha_y = \alpha\cos\theta$. Moreover, $R^2 = \cos^2\theta$, which gives a (signed) value of $R = \cos\theta$. Combining these formulas, the result is

$$C_F = \frac{1}{2} \cdot \frac{1-|R|}{1+|R|} .$$

Consequently, the better the fit (the closer R^2 is to one), the closer the minimizer of E is to the centroid of the error simplex formed from the minimizers of E_x and E_y.

Multilinear regression: Single component approximation

It is now assumed that there are m variables, so $\mathbf{p} = (p_1, p_2, \cdots, p_m)^T$. The centered data vectors for each variable are $\mathbf{p}_1 = (p_{11}, p_{12}, \cdots, p_{1n})^T$, $\mathbf{p}_2 = (p_{21}, p_{22}, \cdots, p_{2n})^T, \cdots, \mathbf{p}_m = (p_{m1}, p_{m2}, \cdots, p_{mn})^T$. It is assumed that no two of these vectors are orthogonal.

One of the central components of a PCA is the ability to find low dimensional approximations of the form $\mathbf{p} = \alpha_1\mathbf{v}_1 + \cdots + \alpha_k\mathbf{v}_k$, where $1 \leq k < m$. The question is whether something similar can be done for a scale and reflectively invariant approximation. One possibility, which is pursued here, is to use the geometric mean of the ordinary least squares functions.

We begin with a one dimensional subspace, which means that $\mathbf{p} = \alpha\mathbf{v}$. In what follows this is rewritten as $p_2 = \alpha_2 p_1, p_3 = \alpha_3 p_1, \cdots, p_m = \alpha_m p_1$. The corresponding individual error functions are then

$$\begin{aligned}
E_1 &= \sum_{i=1}^{n} [(p_{1i} - p_{2i}/\alpha_2)^2 + (p_{1i} - p_{3i}/\alpha_3)^2 + \cdots + (p_{1i} - p_{mi}/\alpha_m)^2], \\
E_2 &= \sum_{i=1}^{n} [(p_{2i} - \alpha_2 p_{1i})^2 + (p_{2i} - \alpha_2 p_{3i}/\alpha_3)^2 + \cdots + (p_{2i} - \alpha_2 p_{mi}/\alpha_m)^2], \\
\vdots &= \qquad\qquad \vdots \\
E_m &= \sum_{i=1}^{n} [(p_{mi} - \alpha_m p_{1i})^2 + (p_{mi} - \alpha_m p_{2i}/\alpha_2)^2 + \cdots + (p_{mi} - \alpha_m p_{m-1,i}/\alpha_{m-1})^2].
\end{aligned}$$

Letting $\alpha_1 = 1$, the general form of the above can be written as

$$\begin{aligned}
E_j &= \sum_{i=1}^{n} [(p_{ji} - \alpha_j p_{1i}/\alpha_1)^2 + (p_{ji} - \alpha_j p_{2i}/\alpha_2)^2 + \cdots + (p_{ji} - \alpha_j p_{mi}/\alpha_m)^2] \\
&= \mathbf{p}_j \cdot \mathbf{p}_j - 2\alpha_j \mathbf{p}_j \cdot \mathbf{p}_1 + \alpha_j^2 \mathbf{p}_1 \cdot \mathbf{p}_1/\alpha_1^2 + \cdots \\
&\qquad + \mathbf{p}_j \cdot \mathbf{p}_j - 2\alpha_j \mathbf{p}_j \cdot \mathbf{p}_m + \alpha_j^2 \mathbf{p}_m \cdot \mathbf{p}_m/\alpha_m^2.
\end{aligned} \tag{12}$$

After factoring, this can be written in the more compact form

$$E_j = \sum_{k=1}^{m} \left(\mathbf{p}_j - \frac{\alpha_j}{\alpha_k}\mathbf{p}_k\right) \cdot \left(\mathbf{p}_j - \frac{\alpha_j}{\alpha_k}\mathbf{p}_k\right). \tag{13}$$

It is worth pointing out that if one of the E_j's is zero, then they are all zero.

The resulting error function, which comes from the geometric mean of the E_j's, is $E = (E_1 E_2 \cdots E_m)^{1/m}$. As stated earlier, this is reflectively invariant because of the symmetric dependence of E on the individual error functions.

To make the connection between the minimizer of E and the error centroid, the minimizer for each E_j is needed. Finding them is straightforward. Namely, setting the first partials of E_j to zero, one finds that

$$\alpha_j = \frac{\mathbf{p}_j \cdot \mathbf{p}_1}{\mathbf{p}_1 \cdot \mathbf{p}_1}\,, \tag{14}$$

and, for $i \neq j$,

$$\alpha_i = \frac{\mathbf{p}_i \cdot \mathbf{p}_i}{\mathbf{p}_j \cdot \mathbf{p}_i}\,\alpha_j. \tag{15}$$

Note that these use the stated assumption that the data vectors are not orthogonal. Also, the above expressions hold in the case of when $j = 1$ because $\alpha_1 = 1$.

Assume now that the data are close to linear, which means that $\mathbf{p}_i = \alpha_{i0}\mathbf{p}_{10} + \varepsilon\mathbf{p}_{i1}$, for $i = 1, 2, \cdots, m$. Given a value for j in Eq (13), the corresponding asymptotic expansions of its minimizing coefficients, for small ε, have the form

$$\alpha_i \sim \alpha_{i0} + \varepsilon\alpha_{i1} + \varepsilon^2\alpha_{i2} + \cdots, \quad \text{for} \ \ i = 1, 2, \cdots, m.$$

Note that in the case $i = 1$, the coefficients are $\alpha_{10} = 1$ and $\alpha_{11} = \alpha_{12} = 0$. The α_{i0}'s are assumed to be known, and the other coefficients are determined by minimizing the error. In preparation for this, note that

$$\begin{aligned} \mathbf{p}_j - \frac{\alpha_j}{\alpha_k}\mathbf{p}_k \quad & \sim \alpha_{i0}\mathbf{p}_{10} + \varepsilon\mathbf{p}_{i1} - \frac{\alpha_{i0} + \varepsilon\alpha_{i1}}{\alpha_{k0} + \varepsilon\alpha_{k1}}(\alpha_{k0}\mathbf{p}_{10} + \varepsilon\mathbf{p}_{k1}) \\ & \sim \varepsilon\alpha_{j0}(\mathbf{q}_j - \mathbf{q}_k), \end{aligned}$$

where $(i = j, k)$

$$\mathbf{q}_i = \frac{1}{\alpha_{i0}}(\mathbf{p}_{i1} - \alpha_{i1}\mathbf{p}_{10}).$$

From Eq (13) it follows that

$$E_j \sim \frac{1}{\alpha_{j0}^2}\varepsilon^2\sum_{k=1}^{m}(\mathbf{q}_j - \mathbf{q}_k) \cdot (\mathbf{q}_j - \mathbf{q}_k).$$

Minimizing this determines the α_{i1}'s, and this yields

$$\alpha_i \sim \alpha_{i0} + \varepsilon\frac{\mathbf{p}_{10} \cdot (\mathbf{p}_{i1} - \alpha_{i0}\mathbf{p}_{11})}{\mathbf{p}_{10} \cdot \mathbf{p}_{10}}\,, \ \ \text{for} \ \ i = 2, 3, \cdots m. \tag{16}$$

What is significant about this is that the first two terms in the expansions for the α_i's do not depend on j. In other words, to $O(\varepsilon)$, the E_j's have the same minimizer. Because of this, it immediately follows that through terms of order ε, the minimizer of the error function E equals the centroid formed from the minimizers of the E_j's. Expressed mathematically, if $\boldsymbol{\alpha}$ is the minimizer of E, and $\boldsymbol{\alpha}_j$ the minimizer of E_j, then

$$\boldsymbol{\alpha} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\alpha}_j + O(\varepsilon^2). \tag{17}$$

It also follows from this analysis that, for small ε, the error function E is strictly convex in the neighborhood of the minimum.

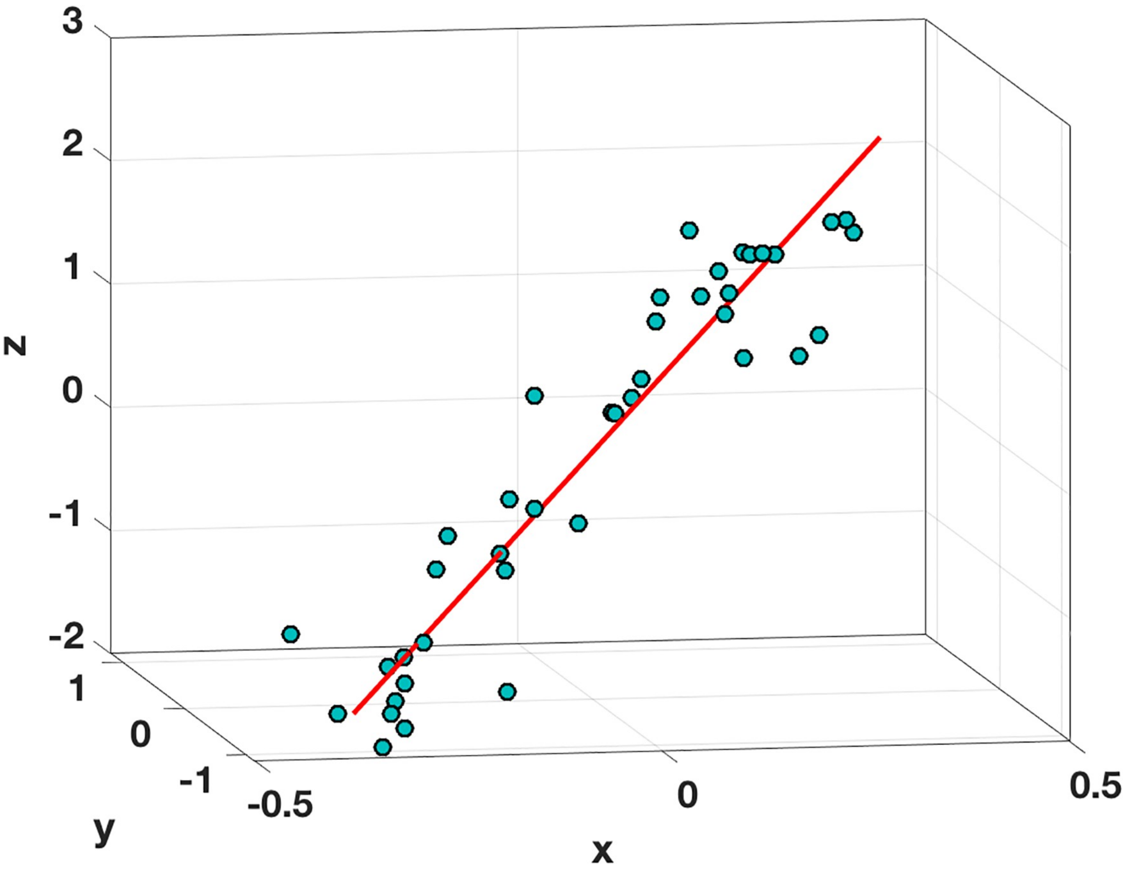

Fig 2. Data used for line fitting example in $\mathbb{R}^3$. The red line is the linear fit determined by minimizing $E = (E_x E_y E_z)^{1/3}$.

https://doi.org/10.1371/journal.pone.0208793.g002

Example in $\mathbb{R}^3$

As an example, in $\mathbb{R}^3$, for the line with $\alpha_2 = 3$ and $\alpha_3 = 5$, 40 randomized points within a distance of 0.2 of the line are used for the data (see Fig 2). For notational simplicity, let $p_1 = x$, $p_2 = y$, $p_3 = z$, $\alpha_2 = \alpha$, and $\alpha_3 = \beta$. The location of the minimizer of E was found using MATLAB's *fminsearch* command, and the location is shown in Fig 3. Also shown are the locations of the minimizers for E_x, E_y, and E_z, determined by Eqs (14) and (15), as well as the associated triangular region formed by these three points. For comparison, the location of the solution as determined using an unscaled PCA is shown. An important observation coming from this figure is that the minimizer for the scale (and reflectively) invariant error function $E = (E_x E_y E_z)^{1/3}$ is located near the centroid of the triangle. This is expected because of Eq (17).

It is worth having a way to characterize how close the minimizer and centroid are, to provide a measure for the goodness of fit. One possibility is to use the maximum distance relative to the width of the simplex, as given by the formula

$$C_F = \max\left\{\frac{|\alpha - \alpha_c|}{w_\alpha}, \frac{|\beta - \beta_c|}{w_\beta}\right\}, \tag{18}$$

where (α_c, β_c) is the centroid, $w_\alpha = \max\{\alpha_x, \alpha_y, \alpha_z\} - \min\{\alpha_x, \alpha_y, \alpha_z\}$, $w_\beta = \max\{\beta_x, \beta_y, \beta_z\} - \min\{\beta_x, \beta_y, \beta_z\}$. In these expressions, (α_x, β_x), (α_y, β_y), and (α_z, β_z) are the minimizers of E_x, E_y, and

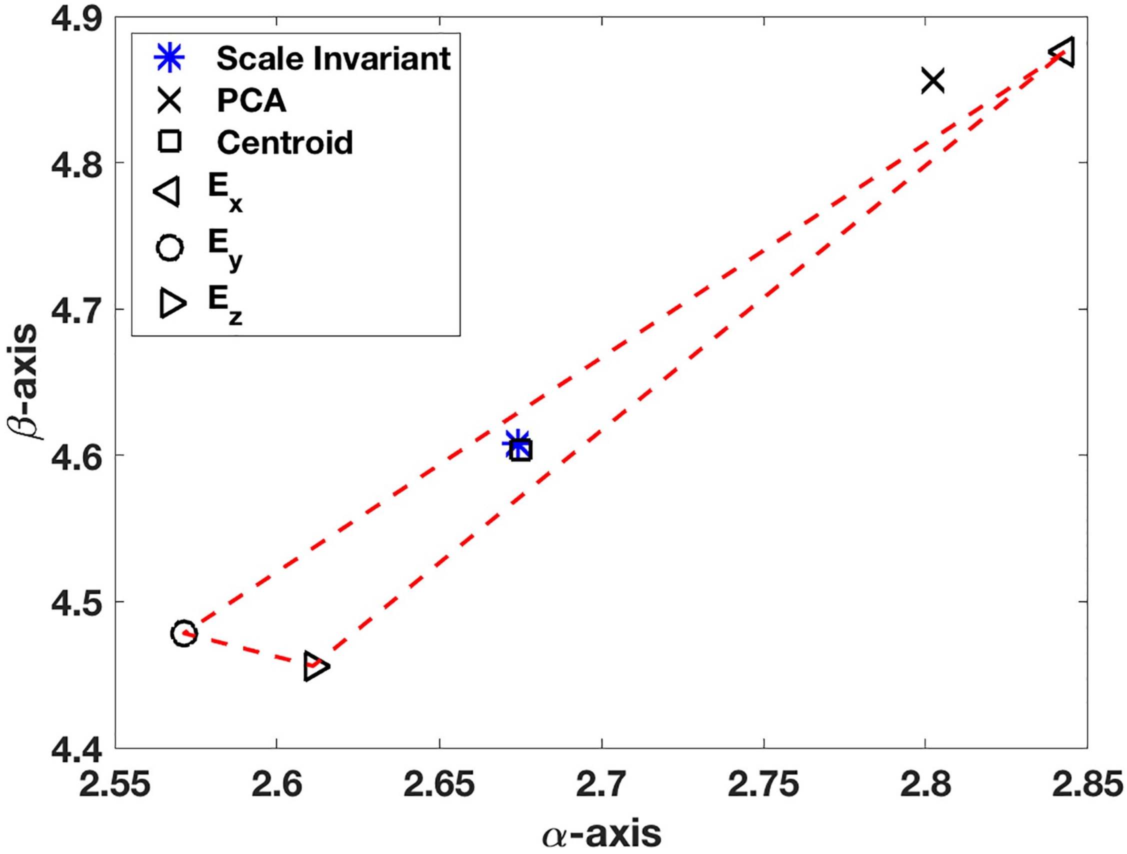

Fig 3. Location of minimizer for the error function $E = (E_x E_y E_z)^{1/3}$, as well as the location determined using an unscaled PCA. Vertices of the triangle (simplex) are the locations of the minimizers of E_x, E_y, and E_z.

https://doi.org/10.1371/journal.pone.0208793.g003

E_z, respectively, and they determine the error simplex in Fig 3. For the solution shown in Fig 3, $C_F \approx 0.01$.

Finally, in finding the minimizer the question of whether or not the error function is convex arises. To verify this, its contour and surface plot are shown in Fig 4. Note that, because of the assumed form of the model function in this example, the error functions E_x, E_y, and E_z, and E are undefined when either α or β are zero. The statement that E is convex refers to its dependence on α and β in the quadrant in which the minimizers are located.

Multilinear regression: (m − 1)-component approximation

The assumptions on the data are the same as for the one-dimensional approximation considered above. As for the model function, it is a hyperplane that is written as $p_m = \alpha_1 p_1 + \alpha_2 p_2 + \cdots + \alpha_{m-1} p_{m-1}$. The associated individual error functions are

$$E_j = \frac{1}{\alpha_j^2} \sum_{i=1}^{n} (p_{mi} - \alpha_1 p_{1i} - \alpha_2 p_{2i} - \cdots - \alpha_{m-1} p_{m-1,i})^2, \qquad \text{for } j = 1, 2, \cdots, m, \tag{19}$$

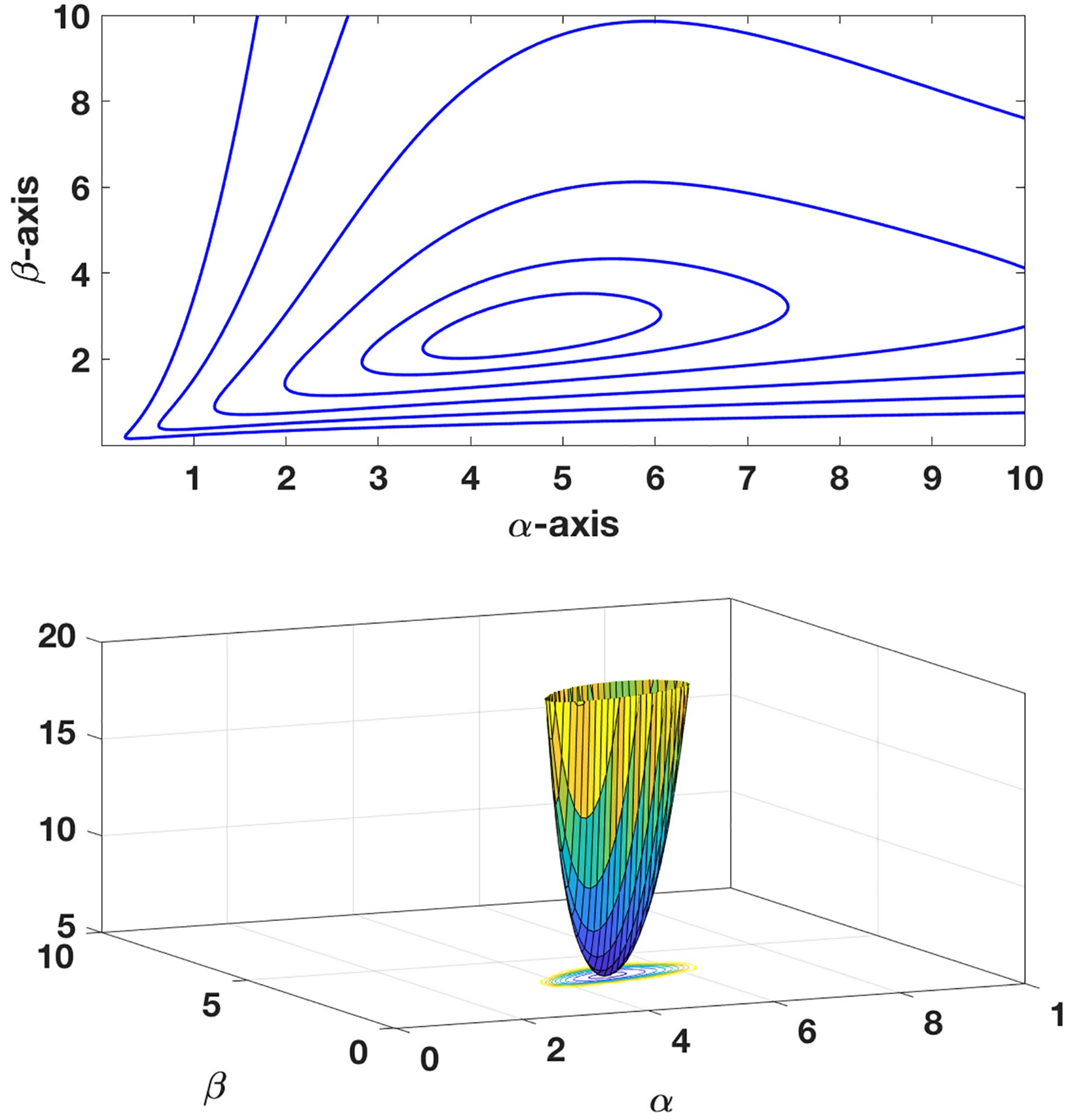

Fig 4. Contour and surface plots for error function $E = (E_xE_yE_z)^{1/3}$, using the the data in Fig 2.

https://doi.org/10.1371/journal.pone.0208793.g004

where $\alpha_m = 1$. With this, the composite error function is

$$\begin{aligned} E(\alpha_1, \quad \alpha_2, \cdots, \alpha_{m-1}) &= (E_1 E_2 \cdots E_m)^{1/m} \\ &= \frac{1}{(\alpha_1 \alpha_2 \cdots \alpha_{m-1})^{2/m}} \sum_{i=1}^{n} (p_{mi} - \alpha_1 p_{1i} - \alpha_2 p_{2i} - \cdots - \alpha_{m-1} p_{m-1,i})^2. \end{aligned} \tag{20}$$

Letting $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_{m-1}, -1)^T$ and $\mathbf{P} = (\mathbf{p}_1\ \mathbf{p}_2 \cdots \mathbf{p}_m)$, then this error function can be written as

$$E = \frac{1}{(\alpha_1 \alpha_2 \cdots \alpha_{m-1})^{2/m}} ||\mathbf{P}\boldsymbol{\alpha}||_2^2. \tag{21}$$

To obtain the connection between the minimizer of E and the centroid of the error simplex we need the minimizers for the E_j's. To determine them, note that for $k = 1, 2, \cdots, m - 1$,

$$\partial_{\alpha_k} \mathbf{P}\boldsymbol{\alpha} = \mathbf{p}_k.$$

Also, from Eq (19), $E_j = \mathbf{P}\boldsymbol{\alpha} \cdot \mathbf{P}\boldsymbol{\alpha}/\alpha_j^2$. Consequently,

$$\partial_{\alpha_k} E_j = \frac{2}{\alpha_j^2} \mathbf{P}\boldsymbol{\alpha} \cdot \left(\mathbf{p}_k - \frac{\delta_{kj}}{\alpha_j} \mathbf{P}\boldsymbol{\alpha} \right),$$

where δ_{kj} is the Kronecker delta. To determine the equation to solve to find the minimizer, consider the case for E_1. Setting $\nabla_\alpha E_1 = \mathbf{0}$, yields the $(m - 1) \times (m - 1)$ system of equations

$$\mathbf{A}\boldsymbol{\alpha} = \begin{pmatrix} \mathbf{P}\boldsymbol{\alpha} \cdot \mathbf{P}\boldsymbol{\alpha}/\alpha_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{22}$$

where $\mathbf{A}$ consists of the first $m - 1$ rows of $\mathbf{P}^T\mathbf{P}$. Since $\mathbf{P}\boldsymbol{\alpha} \cdot \mathbf{P}\boldsymbol{\alpha} = \boldsymbol{\alpha} \cdot (\mathbf{P}^T\mathbf{P})\boldsymbol{\alpha}$, then, letting $\mathbf{r}^T$ be the mth row of $\mathbf{P}^T\mathbf{P}$,

$$\mathbf{P}\boldsymbol{\alpha} \cdot \mathbf{P}\boldsymbol{\alpha} = \boldsymbol{\alpha} \cdot \begin{pmatrix} \mathbf{A} \\ \mathbf{r}^T \end{pmatrix} \boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\alpha} \cdot \mathbf{P}\boldsymbol{\alpha} - \mathbf{r} \cdot \boldsymbol{\alpha}.$$

This means that Eq (22) can be replaced with the equation $\mathbf{R}\boldsymbol{\alpha} = \mathbf{0}$, where $\mathbf{R}$ is the matrix obtained by removing the first row from $\mathbf{P}^T\mathbf{P}$. In a similar manner, the minimizer for E_j is found by solving the equation obtained by removing the jth row from $\mathbf{P}^T\mathbf{P}$. To be more explicit about what equation needs to be solved, for E_1, it is

$$\begin{pmatrix} \mathbf{p}_1 \cdot \mathbf{p}_2 & \mathbf{p}_2 \cdot \mathbf{p}_2 & \cdots & \mathbf{p}_{m-1} \cdot \mathbf{p}_2 \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{p}_1 \cdot \mathbf{p}_{m-1} & \mathbf{p}_2 \cdot \mathbf{p}_{m-1} & \cdots & \mathbf{p}_{m-1} \cdot \mathbf{p}_{m-1} \\ \mathbf{p}_1 \cdot \mathbf{p}_m & \mathbf{p}_2 \cdot \mathbf{p}_m & \cdots & \mathbf{p}_{m-1} \cdot \mathbf{p}_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{pmatrix} = \begin{pmatrix} \mathbf{p}_2 \cdot \mathbf{p}_m \\ \vdots \\ \mathbf{p}_{m-1} \cdot \mathbf{p}_m \\ \mathbf{p}_m \cdot \mathbf{p}_m \end{pmatrix}.$$

In general, the coefficient matrix for E_j is the $(m - 1) \times (m - 1)$ matrix that is obtained by removing the jth row and mth column from $\mathbf{P}^T\mathbf{P}$, and the right hand side is the $(m - 1)$-vector that is obtained by removing the jth entry in the mth column of $\mathbf{P}^T\mathbf{P}$.

To establish the connection between the centroid and the minimizer of E, assume that the data are close to planar. Specifically, letting $\mathbf{P} = \mathbf{P}_0 + \varepsilon\, \mathbf{P}_1$, then for small ε, $\boldsymbol{\alpha} \sim \boldsymbol{\alpha}_0 + \varepsilon\boldsymbol{\alpha}_1 + \varepsilon^2$

$\boldsymbol{\alpha}_2 + \cdots$, where $\mathbf{P}_0\boldsymbol{\alpha}_0 = \mathbf{0}$. Now, setting $\nabla_\alpha E = \mathbf{0}$, the problem to solve is

$$\mathbf{A}\boldsymbol{\alpha} = \frac{1}{m}\begin{pmatrix} 1/\alpha_1 \\ 1/\alpha_2 \\ \vdots \\ 1/\alpha_{m-1} \end{pmatrix} \mathbf{P}\boldsymbol{\alpha} \cdot \mathbf{P}\boldsymbol{\alpha}. \tag{23}$$

Also, $\mathbf{A} = \mathbf{A}_0 + \varepsilon\mathbf{A}_1 + \varepsilon^2\mathbf{A}_2 + \cdots$, where $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{A}_2$ are the first $(m-1)$-rows of $\mathbf{P}_0^T\mathbf{P}_0$, $\mathbf{P}_0^T\mathbf{P}_1 + \mathbf{P}_1^T\mathbf{P}_0$, and $\mathbf{P}_1^T\mathbf{P}_1$, respectively. With this, the $O(\varepsilon)$ problem that comes from Eq (23) is

$$\mathbf{A}_0\boldsymbol{\alpha}_1 + \mathbf{A}_1\boldsymbol{\alpha}_0 = \mathbf{0}, \tag{24}$$

and the $O(\varepsilon^2)$ problem is

$$\mathbf{A}_0\boldsymbol{\alpha}_2 + \mathbf{A}_1\boldsymbol{\alpha}_1 + \mathbf{A}_2\boldsymbol{\alpha}_0 = \frac{1}{m}\begin{pmatrix} 1/\alpha_{10} \\ 1/\alpha_{20} \\ \vdots \\ 1/\alpha_{m-1,0} \end{pmatrix} (\mathbf{P}_0\boldsymbol{\alpha}_1 + \mathbf{P}_1\boldsymbol{\alpha}_0) \cdot (\mathbf{P}_0\boldsymbol{\alpha}_1 + \mathbf{P}_1\boldsymbol{\alpha}_0). \tag{25}$$

In comparison, using the same form for the expansions for E_1, then one finds from Eq (22) that the $O(\varepsilon)$ equation is the same as the one given in Eq (24). This is also true for the other E_j's. Therefore, the $\boldsymbol{\alpha}_1$ term for the centroid and for minimizer of E are equal (as is the $\boldsymbol{\alpha}_0$ term). As for the $O(\varepsilon^2)$ term, for E_1, one finds from Eq (22) that the problem to solve is

$$\mathbf{A}_0\boldsymbol{\alpha}_2 + \mathbf{A}_1\boldsymbol{\alpha}_1 + \mathbf{A}_2\boldsymbol{\alpha}_0 = \begin{pmatrix} 1/\alpha_{10} \\ 0 \\ \vdots \\ 0 \end{pmatrix} (\mathbf{P}_0\boldsymbol{\alpha}_1 + \mathbf{P}_1\boldsymbol{\alpha}_0) \cdot (\mathbf{P}_0\boldsymbol{\alpha}_1 + \mathbf{P}_1\boldsymbol{\alpha}_0). \tag{26}$$

The equations for the other E_j's are the same except for the appropriate modification of the first vector on the right hand side of the equation. Given that $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are the same for E and the E_j's, it follows that the $O(\varepsilon^2)$ term in the centroid and the minimizer of E are equal. Therefore, the conclusion is that the minimizer of E and the centroid are equal through terms of order ε^2. Expressed mathematically, if $\boldsymbol{\alpha}$ is the minimizer of E, and $\boldsymbol{\alpha}_j$ the minimizer of E_j, then

$$\boldsymbol{\alpha} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\alpha}_j + O(\varepsilon^3). \tag{27}$$

Example in $\mathbb{R}^3$

As before, for notational simplicity, let $p_1 = x$, $p_2 = y$, $p_3 = z$, $\alpha_2 = \alpha$, and $\alpha_3 = \beta$. The model function can then be written as $z = \alpha x + \beta y$. In this case, from Eq (19), the individual error functions are:

$$\begin{aligned} E_x &= \sum_{i=1}^{n} (x_i - z_i/\alpha + \beta y_i/\alpha)^2 \\ &= \mathbf{x}\cdot\mathbf{x} - 2\mathbf{x}\cdot\mathbf{z}/\alpha + 2\beta\mathbf{x}\cdot\mathbf{y}/\alpha + \mathbf{z}\cdot\mathbf{z}/\alpha^2 - 2\beta\mathbf{y}\cdot\mathbf{z}/\alpha^2 + \beta^2\mathbf{y}\cdot\mathbf{y}/\alpha^2, \end{aligned} \tag{28}$$

$$\begin{aligned} E_y &= \sum_{i=1}^{n} (y_i - z_i/\beta + \alpha x_i/\beta)^2 \\ &= \mathbf{y}\cdot\mathbf{y} - 2\mathbf{y}\cdot\mathbf{z}/\beta + 2\alpha\mathbf{x}\cdot\mathbf{y}/\beta + \mathbf{z}\cdot\mathbf{z}/\beta^2 - 2\alpha\mathbf{x}\cdot\mathbf{z}/\beta^2 + \alpha^2\mathbf{x}\cdot\mathbf{x}/\beta^2, \end{aligned} \tag{29}$$

$$\begin{aligned} E_z &= \sum_{i=1}^{n} (z_i - \alpha x_i - \beta y_i)^2 \\ &= \mathbf{z}\cdot\mathbf{z} - 2\alpha\mathbf{x}\cdot\mathbf{z} - 2\beta\mathbf{y}\cdot\mathbf{z} + \alpha^2\mathbf{x}\cdot\mathbf{x} + 2\alpha\beta\mathbf{x}\cdot\mathbf{y} + \beta^2\mathbf{y}\cdot\mathbf{y}. \end{aligned} \tag{30}$$

The minimizer $\boldsymbol{\alpha}_x$ of E_x is

$$\boldsymbol{\alpha}_x = \frac{1}{(\mathbf{x}\cdot\mathbf{y})(\mathbf{y}\cdot\mathbf{z}) - (\mathbf{x}\cdot\mathbf{z})(\mathbf{y}\cdot\mathbf{y})} \begin{pmatrix} \mathbf{y}\cdot\mathbf{z} & -\mathbf{y}\cdot\mathbf{y} \\ -\mathbf{x}\cdot\mathbf{z} & \mathbf{x}\cdot\mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{y}\cdot\mathbf{z} \\ \mathbf{z}\cdot\mathbf{z} \end{pmatrix}, \tag{31}$$

the minimizer $\boldsymbol{\alpha}_y$ of E_y is

$$\boldsymbol{\alpha}_y = \frac{1}{(\mathbf{x}\cdot\mathbf{x})(\mathbf{y}\cdot\mathbf{z}) - (\mathbf{x}\cdot\mathbf{z})(\mathbf{x}\cdot\mathbf{y})} \begin{pmatrix} \mathbf{y}\cdot\mathbf{z} & -\mathbf{x}\cdot\mathbf{y} \\ -\mathbf{x}\cdot\mathbf{z} & \mathbf{x}\cdot\mathbf{x} \end{pmatrix} \begin{pmatrix} \mathbf{x}\cdot\mathbf{z} \\ \mathbf{z}\cdot\mathbf{z} \end{pmatrix}, \tag{32}$$

and the minimizer $\boldsymbol{\alpha}_z$ of E_z is

$$\boldsymbol{\alpha}_z = \frac{1}{(\mathbf{x}\cdot\mathbf{x})(\mathbf{y}\cdot\mathbf{y}) - (\mathbf{x}\cdot\mathbf{y})^2} \begin{pmatrix} \mathbf{y}\cdot\mathbf{y} & -\mathbf{x}\cdot\mathbf{y} \\ -\mathbf{x}\cdot\mathbf{y} & \mathbf{x}\cdot\mathbf{x} \end{pmatrix} \begin{pmatrix} \mathbf{x}\cdot\mathbf{z} \\ \mathbf{y}\cdot\mathbf{z} \end{pmatrix}. \tag{33}$$

The resulting error function based on the geometric mean is

$$\begin{aligned} E(\alpha,\beta) &= (E_x E_y E_z)^{1/3} \\ &= \frac{1}{\alpha^{2/3}\beta^{2/3}} \left(\mathbf{z}\cdot\mathbf{z} - 2\alpha\mathbf{x}\cdot\mathbf{z} - 2\beta\mathbf{y}\cdot\mathbf{z} + \alpha^2\mathbf{x}\cdot\mathbf{x} + 2\alpha\beta\mathbf{x}\cdot\mathbf{y} + \beta^2\mathbf{y}\cdot\mathbf{y}\right). \end{aligned} \tag{34}$$

Taking the first partials of this, and setting them to zero, one obtains the (nonlinear) system

$$\alpha^2\mathbf{x}\cdot\mathbf{x} + \alpha\beta\mathbf{x}\cdot\mathbf{y} + \beta\mathbf{y}\cdot\mathbf{z} = \mathbf{z}\cdot\mathbf{z} \tag{35}$$

$$\beta^2\mathbf{y}\cdot\mathbf{y} + \alpha\beta\mathbf{x}\cdot\mathbf{y} + \alpha\mathbf{x}\cdot\mathbf{z} = \mathbf{z}\cdot\mathbf{z}. \tag{36}$$

These equations can be written in somewhat simpler terms by letting

$$q = \alpha\sqrt{\frac{\mathbf{x}\cdot\mathbf{x}}{\mathbf{z}\cdot\mathbf{z}}} \quad \text{and} \quad p = \beta\sqrt{\frac{\mathbf{y}\cdot\mathbf{y}}{\mathbf{z}\cdot\mathbf{z}}}.$$

Using the law of cosines, then Eqs (35) and (36) become

$$q^2 + pq\cos\theta_{xy} + p\cos\theta_{yz} \quad = 1 \tag{37}$$

$$p^2 + pq\cos\theta_{xy} + q\cos\theta_{xz} \quad = 1, \tag{38}$$

where θ_{xy} is the angle between **x** and **y** (with similar definitions for the other angles). An analytical formula for the solution of these equations is not apparent, but it is a simple matter to compute the solution numerically.

As an example, for the plane with $\alpha = -10$ and $\beta = 6$, 40 randomized points within a distance of 0.15 of the plane were used for the data. The location of the minimizer of E was found by solving Eqs (35) and (36) using MATLAB's *fminsearch* command, and the location is shown in Fig 5. Also shown are the locations of the minimizers for E_x, E_y, and E_z, and the associated triangular region formed by these three points. For comparison, the location of the solution as determined using an unscaled PCA is shown. As expected from Eq (27), the of E minimizer is located near the centroid of the triangle. To quantify the difference, using the formula in Eq (18), $C_F \approx 0.02$. Finally, to demonstrate the convexity of the error function in the octant containing the minimizer, the contour and surface are shown in Fig 6.

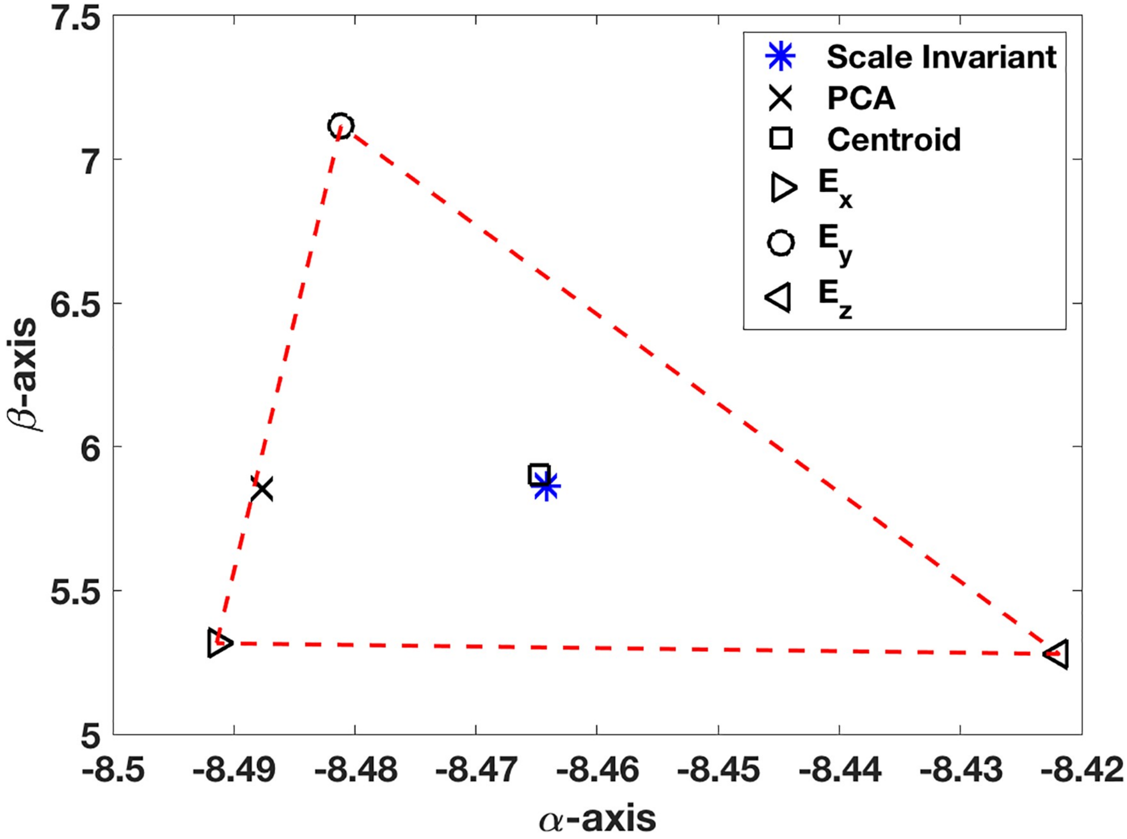

Fig 5. Typical locations for the various minimizers of the plane example. Vertices of the simplex are minimizers using individual coordinate projections E_x, E_y, and E_z.

https://doi.org/10.1371/journal.pone.0208793.g005

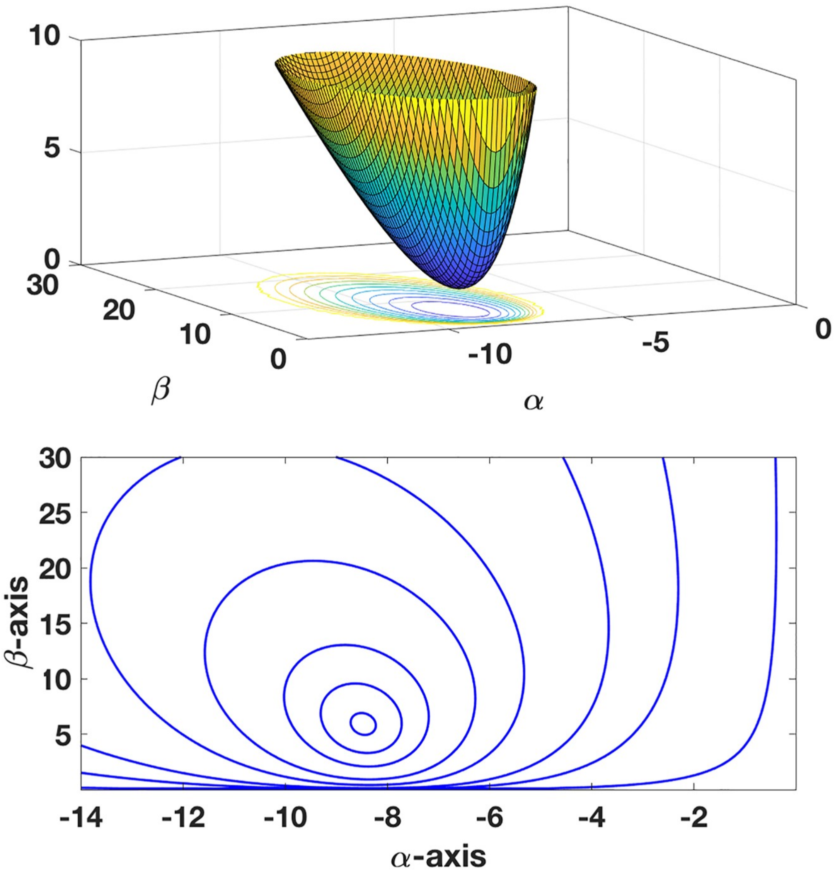

Fig 6. Contour and surface plots for error function $E = (E_xE_yE_z)^{1/3}$ for the planar fit data.

https://doi.org/10.1371/journal.pone.0208793.g006

Application to real world data sets

What follows are applications of the hyperplane function to various real world data sets. In the process, comparisons are made with a PCA. Also, three of the data sets have been used in other studies to compare multiple linear regression methods, and this is discussed in the respective application.

Crime data

As an illustration, and a comparison with a PCA, consider the data for the seven major crime rates and population for the larger cities in the U.S in 2009 [15]. Of the 105 cities reported, 79 were randomly chosen for the training set, and the testing dataset consists of those left out. With this, $n = 79$ and $m = 8$. The minimizer of Eq (20) was found using a modified Polak-Ribière descent procedure, with Armijo's method used to solve the line search problem. The modification is that if the search direction determined using Polak-Ribière is not a direction of descent, then the steepest descent direction is used instead. A description of the Polak-Ribière and Armijo's methods can be found in [5, 16]. The starting point for the descent procedure was a convex combination of the minimizers for the E_j's, which was computed using the formulas given earlier. Also, the normalization for the PCA uses $||\mathbf{p}_j||_2/\sqrt{n}$, for each column j of the centered training data matrix.

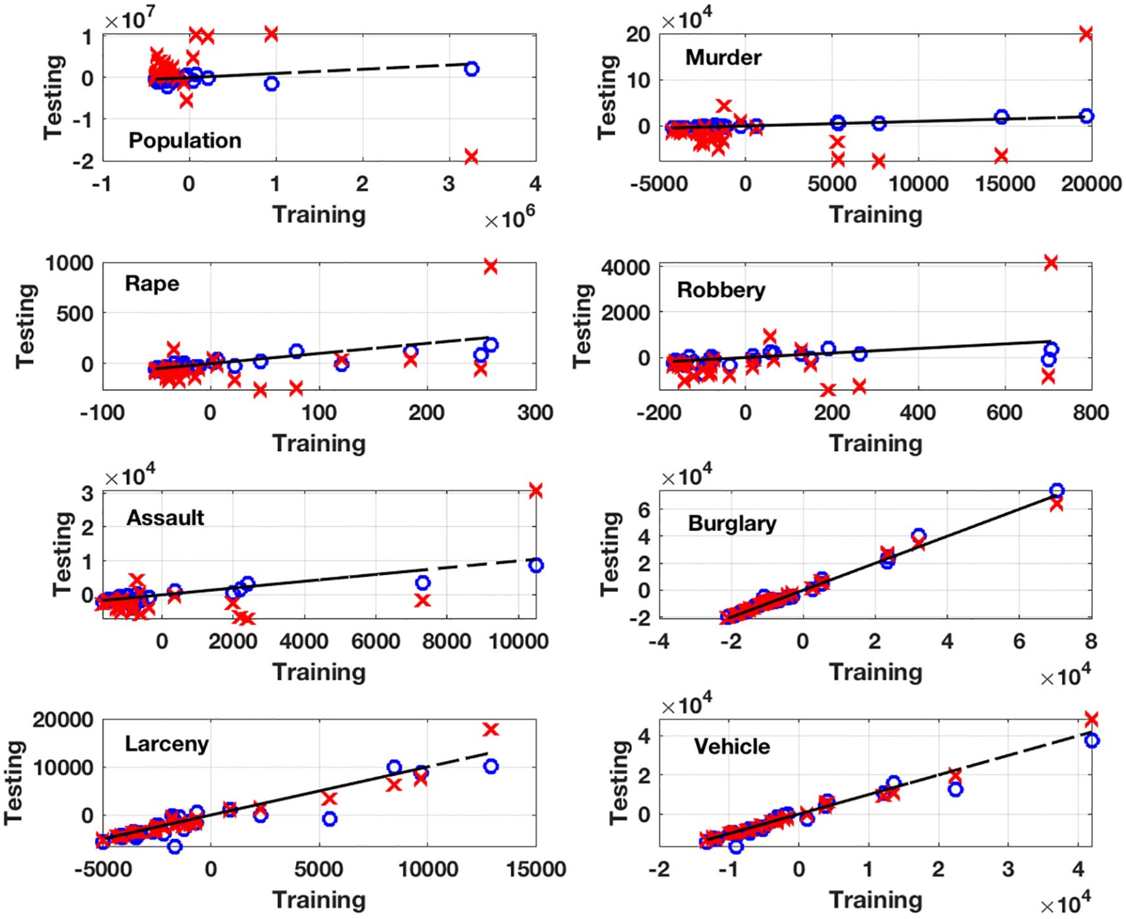

Fig 7. Comparison between fits using a hyperplane approximation to the crime and population data. Shown are the values obtained using the error function in Eq (20), indicated with the o's, and using a PCA (x). The dashed line in each graph corresponds when the training and testing data are equal.

https://doi.org/10.1371/journal.pone.0208793.g007

The resulting testing versus training comparison for each variable is shown in Fig 7. For Eq (20), the α_j's are determined by minimizing E and then using those same values for each graph. For example, for the graph associated with p_1, the model function is rewritten as

$$p_1 = -\frac{\alpha_2}{\alpha_1}p_2 - \frac{\alpha_3}{\alpha_1}p_3 - \cdots - \frac{\alpha_{m-1}}{\alpha_1}p_{m-1} + \frac{1}{\alpha_1}p_m, \tag{39}$$

and the resulting training-testing values are plotted. For the PCA, a seven component approximation is made. To make a more quantitative comparison between these two approaches, and because both methods are reflectively invariant, the normalized least squares errors for each of these graphs is given in Table 1. The normalization used is $N||\mathbf{p}_j||_1/n$, where n is the number of values in the training set and N the number in the testing set. It is seen, at least in this example, that the values using Eq (20) produce a uniformly better result than those obtained using a PCA. To make the point that the fit using a PCA varies with the scaling, the errors using two other scalings are also given in Table 1. The scaling used for PCA$_1$ is considerably worse than the PCA values, while the PCA$_2$ values in the last column are distinctly better.

Table 1. The E and PCA columns are the normalized least squares errors from Fig 7. The PCA$_1$ and PCA$_2$ columns are the errors for a PCA using two different column scalings.

	E	PCA	PCA$_1$	PCA$_2$
p_1	3.92e−01	2.98e+00	6.14e+02	1.24e−01
p_2	9.76e−02	2.97e+00	1.24e−01	8.65e−02
p_3	2.13e−01	8.23e−01	1.13e+00	6.70e−01
p_4	3.67e−01	1.41e+00	2.27e+00	1.94e+00
p_5	1.88e−01	7.78e−01	2.36e+00	7.44e−02
p_6	3.33e−02	2.23e−02	1.85e−01	2.19e−02
p_7	1.07e−01	6.92e−02	3.55e+01	5.80e−02
p_8	5.80e−02	3.58e−02	3.55e−01	3.64e−02

https://doi.org/10.1371/journal.pone.0208793.t001

Wine quality data

As a second example, data mining has been used to predict tasting preferences for wine [17]. The dataset for red wine consists of 1599 instances (vectors) containing values for 12 attributes, 11 being physicochemical properties of the wine and one the quality score for taste. The physicochemical properties in this case are: fixed acidity (α_1), volatile acidity (α_2), citric acid (α_3), residual sugar (α_4), chlorides (α_5), free sulfur dioxide (α_6), total sulfur dioxide (α_7), density (α_8), pH (α_9), sulphates (α_{10}), and alcohol (α_{11}). Following the protocol used in [17], the training set consists of 2/3 of the original (randomly chosen), and the testing set the remaining 1/3. They also used a regression error characteristic (REC) curve, which is defined in [18], to evaluate how well various regression procedures predict the taste score. To compare how Eq (20) does, using the training set, the minimizer is computed with the modified Polak-Ribière method described earlier, treating the data as defining a smooth function. The resulting REC curve determined using Eq (20) is shown in Fig 8. Also shown are the curves obtained using a

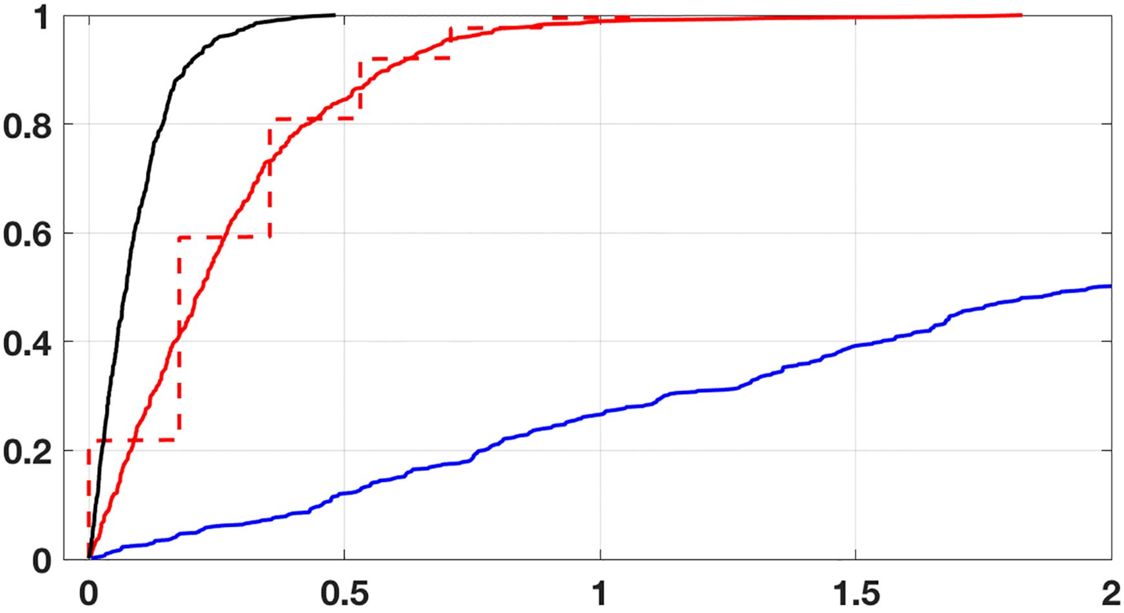

Fig 8. The regression error characteristic curve (REC) for the red wine testing data using the error function in Eq (20), solid red, using a PCA, blue, and least squares, black. The dashed red curve is the curve using Eq (20) after converting back to the original integer scale reported in [17].

https://doi.org/10.1371/journal.pone.0208793.g008

PCA as well as from using a standard multivariable least squares regression. The dashed curve is what is obtained using Eq (20) if the predicted values are converted back into integer scores as originally used in [17]. Finally, for the two and three dimensional examples considered earlier, it was found that the centroid of the error simplex furnished a fairly accurate approximation for the minimizer. A measure of how close they are is given in Eq (18). Generalizing this formula for the wine example, it is found that $C_F \approx 0.15$, which indicates they are reasonable close.

It is evident from Fig 8 that standard least squares provides better predictive values for the taste score than when using Eq (20). This can be quantified using the mean absolute deviation (MAD), which is defined as $||\mathbf{y} - \mathbf{y}^*||_1/N$, where $\mathbf{y}^*$ are the test values, $\mathbf{y}$ are the predicted values, and N is the number of observations in the testing set (note that the values are unscaled). For Eq (20) the MAD value is 1.5, while using least squares it is 0.5. Consequently, if given a particular bottle of red wine, least squares would provide a better predictor of how it tastes. What Eq (20) provides is a better model for how to modify the physicochemical properties to achieve a particular taste score, and the reason is reflective invariance. To illustrate, with the coefficients computed previously, the resulting MAD values for the other possible training-testing cases are given in Table 2. The fact that the least squares values are so poor is not surprising, and the reason was given earlier. Namely, the values obtained using p_m as the dependent variable are not equivalent to the values obtained if the model equation is solved for, say, p_1, and then considering it as the dependent variable.

Wave height data

Wave height at Buoy Station 46006, located in the northern Pacific Ocean, is measured hourly, along with the wind direction, wind speed, wind gust, dominant wave period, average wave period, barometric pressure, and water temperature [19]. The specific data considered here come from measurements over approximately 11 months. The dataset consists of 7960 instances containing values for the eight attributes. A reduced version of this dataset was examined in [20], in comparing how various regression procedures do on real ecological data. The reduction was to use the daily average values rather than the hourly data, but all of the values are considered here. Following the protocol used in [20], the training set consists of 3/4 of the original (randomly chosen), and the testing set the remaining 1/4. The comparison was made using the mean squared prediction error (MSPE), which is defined as $||\mathbf{y} - \mathbf{y}^*||_2^2/n$,

Table 2. The MAD values for red wine, as determined using Eq (20) and from conventional least squares. The last column is the ratio of the least squares value to the one obtained using Eq (20).

attribute	*E*	Least Squares	Ratio
fixed acidity	6.400e-01	7.455e+00	11.6
volatile acidity	2.299e-01	5.325e-01	2.3
citric acid	1.783e-01	2.481e+00	13.9
residual sugar	1.544e+00	8.459e+00	5.5
chlorides	8.297e-02	3.099e-01	3.7
free sulfur dioxide	1.241e+01	1.062e+02	8.6
total sulfur dioxide	3.848e+01	1.470e+02	3.8
density	9.008e-04	8.729e-03	9.7
pH	1.068e-01	3.286e+00	30.8
sulphates	3.350e-01	5.486e-01	1.6
alcohol	1.123e+00	2.075e+00	1.8
taste	1.493e+00	5.134e-01	0.3

https://doi.org/10.1371/journal.pone.0208793.t002

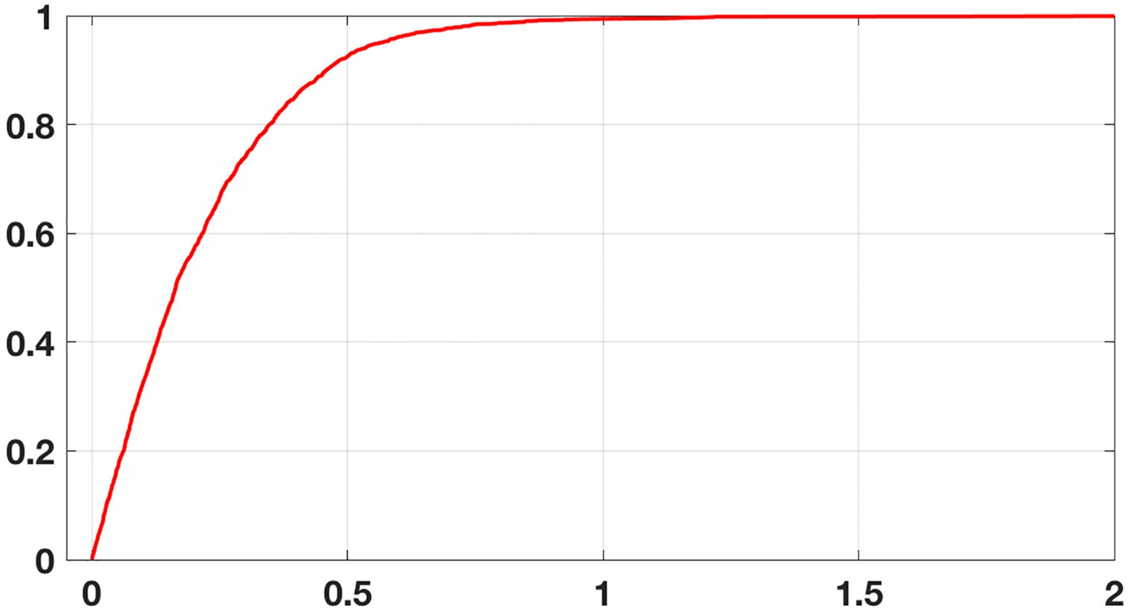

Fig 9. The regression error characteristic curve for the wave height data using the error function in Eq (20).

https://doi.org/10.1371/journal.pone.0208793.g009

where $\mathbf{y}^*$ are the test values for the wave height, $\mathbf{y}$ are the predicted values, and n is the number of observations in the testing set (note that the values are unscaled). Using Eq (20), the MSPE is about 1.2, which compares favorability with the mean MSPE of 12.3 found in [20]. The resulting MAD value is about 0.8, and the REC curve is shown in Fig 9. Finally, for this example, $C_F \approx 0.16$. This indicates that the centroid of the error simplex and the minimizer are close, although not as close as in the earlier examples.

Chemical reaction data

The chemical oscillator known as the Belousov-Zhabotinskii reaction can be described with the following five reactions [21]

$$\begin{aligned} A + Y &\rightarrow X + P, \\ X + Y &\rightarrow 2P, \\ A + X &\rightarrow 2X + 2Z, \\ 2X &\rightarrow A + P, \\ Z &\rightarrow \frac{1}{2} fY. \end{aligned}$$

The chemicals here are bromous acid (X), bromide (Y), cerium-4 (Z), bromate (A), and a product P. Given known initial concentrations for each species, measurements are made at later times to follow the evolution of the overall reaction. The complication is that one or more of these reactions are very fast, or occur at very low concentrations, so accurate measurements are difficult. What is demonstrated here is how the hyperplane fit can be used in the case of when four of the five concentrations are measured, and the fifth is determined using regression. What is important is that no matter which four are chosen, that the same (equivalent) value is obtained for the fifth species. To demonstrate, the reactions were run for 20 different initial concentrations, and the values for the five species were recorded at 60 second intervals up to 10 minutes. The resulting dataset consists of 180 instances, and these were randomly split into a training set (3/4) and a testing set (1/4). The values fitted are those for Z. Using Eq (20), the MAD value is 0.015 and the MSPE is 4.6×10^{-4}. Using a standard least squares fit the values are 0.012 and 3×10^{-4}, respectively. However, if you fit the values for, say, A, and then transform back to the equation for Z, then the MAD and MSPE values using Eq (20) remain unchanged while the least squares values are 0.024 and 10^{-3}. As a final comment, using a reflectively invariant error function with chemical density fits has the distinct advantage of being able to determine possible conservation laws inherent in the system.

Chlorophyll-*a* data

The link between phytoplankton and water chemistry has been the subject of several recent studies, although the connections with specific chemical species are incompletely understood. Several correlations have been made, and the physicochemical parameters most commonly considered are oxygen, pH, NH_4-N (ammonium nitrogen), NO_3-N (nitrate nitrogen), and PO_4-P (phosphate phosphorus) [20, 22–24]. The latter study used the values for the Chlorophyll-*a* density, along with various chemical properties, of lakes in the Northeast that were measured over a four year period [25]. Altogether, there are 20 useable variables in this study, and 500 observations. After removing incomplete entries and others that are not useable, the data set consists of 348 observations. The model requires specification of which variables to use, and after some analysis of the data it comes down to five: total dissolved aluminum, nitrate, total nitrogen, total phosphorous, total suspended solids, and turbidity. In comparison, in [20], 10 variables were initially used. Using Eq (20), the resulting MSPE is about 1.1, which compares favorability with the mean MSPE of 2.4 found in [20]. The resulting normalized MAD value is about 0.7, and the REC curve is shown in Fig 10. Finally, for this example, $C_F \approx 0.13$. This indicates that the centroid of the error simplex and the minimizer are close, similar to what was found for the wave height example.

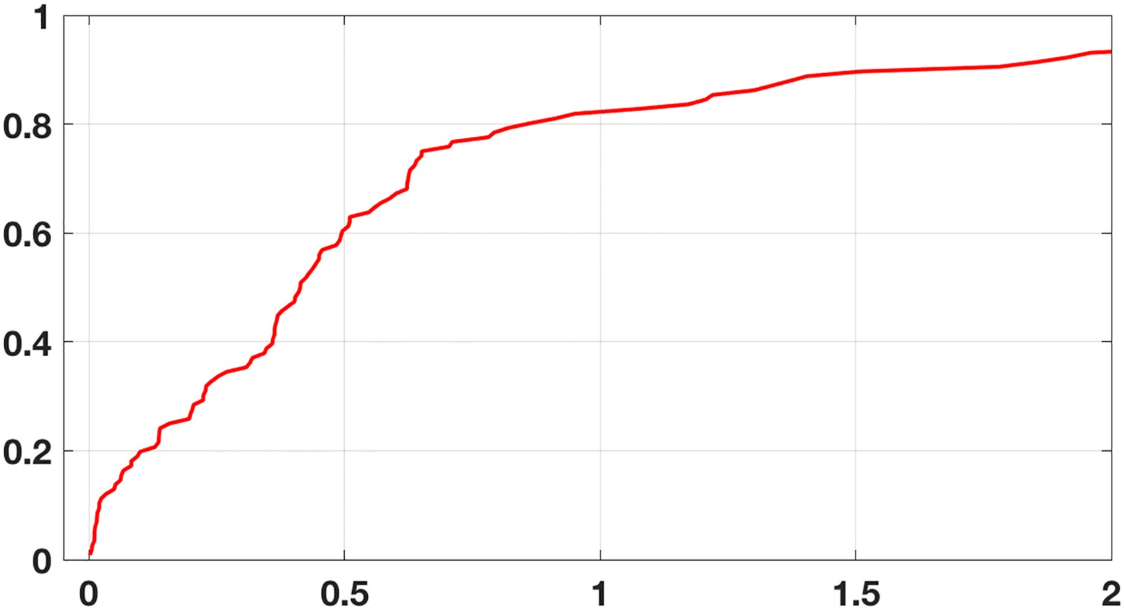

Fig 10. The regression error characteristic curve for the Chlorophyll-*a* data using the error function in Eq (20).

https://doi.org/10.1371/journal.pone.0208793.g010

Other dimensional fits

It is possible to write down the formulas for the other cases, but it is more informative to consider a particular case that illustrates how this is done. So, consider the case of when there are four variables, x, y, z, and w. There are three subspace approximations possible, corresponding to one, two and three dimensions. The one and three dimensional cases were discussed above, and so only the two dimensional case is considered. Writing the model functions as $z = \alpha_{11}x + \alpha_{12}y$, $w = \alpha_{21}x + \alpha_{22}y$. This can be written in matrix form as

$$\begin{pmatrix} z \\ w \end{pmatrix} = \mathbf{A}\begin{pmatrix} x \\ y \end{pmatrix}.$$

It is possible to rewrite the above equation in five different, but equivalent, ways. For example, if z and w are taken to be independent then

$$\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{B}\begin{pmatrix} z \\ w \end{pmatrix},$$

while if y and z are taken to be independent then

$$\begin{pmatrix} x \\ w \end{pmatrix} = \mathbf{C}\begin{pmatrix} y \\ z \end{pmatrix}.$$

The other three forms are

$$\begin{pmatrix} y \\ w \end{pmatrix} = \mathbf{D}\begin{pmatrix} x \\ z \end{pmatrix}, \quad \begin{pmatrix} y \\ z \end{pmatrix} = \mathbf{E}\begin{pmatrix} x \\ w \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} x \\ z \end{pmatrix} = \mathbf{F}\begin{pmatrix} y \\ w \end{pmatrix}.$$

The entrees in the above matrices are known expressions involving the original coefficients α_{11}, α_{12}, α_{21}, and α_{22}. For example,

$$\mathbf{B} = \mathbf{A}^{-1} = \frac{1}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}}\begin{pmatrix} \alpha_{22} & -\alpha_{21} \\ -\alpha_{12} & \alpha_{11} \end{pmatrix}.$$

The error function corresponding to the variable x is

$$E_x = \sum_{i=1}^{n}[(x_i - B_{11}z_i - B_{12}w_i)^2 + (x_i - C_{11}y_i - C_{12}z_i)^2 + (x_i - F_{11}y_i - F_{12}w_i)^2].$$

The three terms in the above sum correspond to the case when x is taken to be dependent. In a similar manner,

$$E_y = \sum_{i=1}^{n}[(y_i - B_{21}z_i - B_{22}w_i)^2 + (y_i - D_{11}x_i - D_{12}z_i)^2 + (y_i - E_{11}x_i - E_{12}w_i)^2].$$

The corresponding error to determine the α's is

$$E(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}) = (E_xE_yE_zE_w)^{1/4}.$$

Concluding remarks

The principal conclusions from this study are:

1. Scale and rotational invariance of the error function are incompatible.
2. Using the geometric mean of the ordinary least squares error functions, one obtains an error function which is scale and reflectively invariant, and which is easily extendable to low dimensional approximations for multilinear regression. For the two cases worked out, which correspond to a line and hyperplane approximation, the minimizer can be well approximated using the centroid of the error simplex obtained from the minimizers for the ordinary least squares error functions.

Because the error function used here is not quadratic, finding the minimizer requires using a nonlinear optimization procedure. The result is that more computational time is needed than for linear least squares. How much time depends on the size of the data matrix, and the number of variables involved. For the wine example considered earlier (12 variables and 1599 data vectors), linear least squares using MATLAB's *mldivide* routine takes about 1 msec, while the nonlinear optimization procedure takes about 100 msec (using a 2017 iMac). So, although the relative time is fairly large, the actual time is small. The proposed error function does have the advantage of having a warm start, which is the centroid approximation, and this helps reduce the computing time.

In terms of future work, it remains to determine if the centroid approximation applies to the other lower dimensional approximations. Also, there is the question as to the sensitivity of the minimizer to outliers in the training set. This is a problem for a PCA, and one approach to improve the robustness of a PCA is to switch from a ℓ_2-norm to a ℓ_1-norm [26, 27], a ℓ_p-norm [28], or a norm based on a generalized mean [29]. These are used, in part, because they preserve the rotational invariance of a PCA. It is straightforward to use these norms with the geometric mean function, and this will not affect its invariance properties. What has not been investigated is the sensitivity of the proposed error function to outliers, or whether switching norms might reduce any potential sensitivities.

Author Contributions

Conceptualization: Mark H. Holmes, Michael Caiola.

Data curation: Mark H. Holmes.

Formal analysis: Mark H. Holmes.

Funding acquisition: Mark H. Holmes.

Investigation: Mark H. Holmes, Michael Caiola.

Methodology: Mark H. Holmes.

Project administration: Mark H. Holmes.

Resources: Mark H. Holmes.

Software: Mark H. Holmes.

Supervision: Mark H. Holmes.

Validation: Mark H. Holmes.

Visualization: Mark H. Holmes.

Writing – original draft: Mark H. Holmes.

Writing – review & editing: Mark H. Holmes, Michael Caiola.

References

1. Belhumeur P, Hespanha J, Kriegman D. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell. 1997; 19(7):711–720. https://doi.org/10.1109/34.598228
2. Turk M, Pentland A. Eigenfaces for Recognition. J Cogn Neurosci. 1991; 3(1):71–86. https://doi.org/10.1162/jocn.1991.3.1.71 PMID: 23964806
3. Ross DA, Lim J, Lin RS, Yang MH. Incremental Learning for Robust Visual Tracking. Int J Comput Vis. 2008; 77(1):125–141. https://doi.org/10.1007/s11263-007-0075-7
4. Ramsey F, Schafer D. The Statistical Sleuth: A Course in Methods of Data Analysis. Cengage Learning; 2012.
5. Holmes MH. Introduction to Scientific Computing and Data Analysis. New York: Springer; 2016.
6. Henderson C, Izquierdo E. Multi-scale reflection invariance. In: 2016 SAI Computing Conference; 2016. p. 420–425.
7. Tofallis C. Multiple Neutral Data Fitting. Annals of Operations Research. 2003; 124(1):69–79. https://doi.org/10.1023/B:ANOR.0000004763.39347.bb
8. Goodman T, Tofallis C. Neutral Data Fitting by Lines and Planes. In: Iske A, Levesley J, editors. Algorithms for Approximation. Springer Berlin Heidelberg; 2007. p. 259–268.
9. Woolley EB. The Method of Minimized Areas as a Basis for Correlation Analysis. Econometrica. 1941; 9(1):38–62. https://doi.org/10.2307/1907173
10. Samuelson PA. A Note on Alternative Regressions. Econometrica. 1942; 10(1):80–83. https://doi.org/10.2307/1907024
11. Babu GJ, Feigelson ED. Analytical and Monte Carlo comparisons of six different linear least squares fits. Commun Stat Simul Comput. 1992; 21:533–549. https://doi.org/10.1080/03610919208813034
12. Ludbrook J. Linear regression analysis for comparing two measurers or methods of measurement: But which regression? Clin Exp Pharmacol Physiol. 2010; 37(7):692–699. https://doi.org/10.1111/j.1440-1681.2010.05376.x PMID: 20337658
13. Draper NR, Yang Y. Generalization of the geometric mean functional relationship. Comput Statist Data Anal. 1997; 23(3):355–372. http://dx.doi.org/10.1016/S0167-9473(96)00037-0
14. Bluman G, Cole J. Similarity Methods for Differential Equations. New York: Springer-Verlag; 1974.
15. U S Census Bureau. Crime Rates by Type: Selected Large Cities. In: Statistical Abstract of the United States: 2012. 131st ed. Washington, DC; 2012.
16. Nocedal J, Wright S. Numerical Optimization. Springer Series in Operations Research and Financial Engineering. Springer; 2006.
17. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. Modeling wine preferences by data mining from physicochemical properties. Decis Support Syst. 2009; 47(4):547–553. http://dx.doi.org/10.1016/j.dss.2009.05.016
18. Bi J, Bennett K. Regression error characteristic curves. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). Washington DC; 2003. p. 43–50.
19. NOAA. National Data Buoy Center: Historical Meteorological Data—Station 46006; 2006.
20. Murtaugh PA. Performance of several variable-selection methods applied to real ecological data. Ecol Lett. 2009; 12(10):1061–1068. https://doi.org/10.1111/j.1461-0248.2009.01361.x PMID: 19702634
21. Field RJ, Koros E, Noyes RM. Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the bromate-cerium-malonic acid system. J Amer Chem Soc. 1972; 94:8649–8664. https://doi.org/10.1021/ja00780a001
22. Heini A, Puustinen I, Tikka M, Jokiniemi A, Leppäranta M, Arvola L. Strong dependence between phytoplankton and water chemistry in a large temperate lake: spatial and temporal perspective. Hydrobiologia. 2014; 731(1):139–150. https://doi.org/10.1007/s10750-013-1777-1
23. George B, Kumar JIN, Kumar RN. Study on the influence of hydro-chemical parameters on phytoplankton distribution along Tapi estuarine area of Gulf of Khambhat, India. The Egyptian Journal of Aquatic Research. 2012; 38(3):157–170. http://dx.doi.org/10.1016/j.ejar.2012.12.010

24. Liu Y, Guo H, Yang P. Exploring the influence of lake water chemistry on chlorophyll a: A multivariate statistical model analysis. Ecol Modell. 2010; 221(4):681–688. http://dx.doi.org/10.1016/j.ecolmodel.2009.03.010

25. Stoddard J. EMAP Surface Waters Lake Database: 1991-1994 Northeast Lakes Water Chemistry Data Summarized by Lake. U.S. EPA NHEERL Western Ecology Division; 1996.

26. Ding CHQ, Zhou D, He X, Zha H. R_1-PCA: rotational invariant L_1-norm principal component analysis for robust subspace factorization. In: Proceedings of the 23rd International Conference on Machine Learning, ICML'06; 2006. p. 281–299.

27. Kwak N. Principal Component Analysis Based on L1-Norm Maximization. IEEE Trans Pattern Anal Mach Intell. 2008; 30(9):1672–1680. https://doi.org/10.1109/TPAMI.2008.114 PMID: 18617723

28. Kwak N. Principal Component Analysis by L_p-Norm Maximization. IEEE Trans Cybern. 2014; 44(5):594–609. https://doi.org/10.1109/TCYB.2013.2262936 PMID: 23807479

29. Oh J, Kwak N. Generalized mean for robust principal component analysis. Pattern Recognit. 2016; 54:116–127. https://doi.org/10.1016/j.patcog.2016.01.002