Least-Squares Fitting of a Hyperplane

Robert K. Moniot

October 20, 2002

Abstract

A method is developed for fitting a hyperplane to a set of data by least-squares, allowing for independent uncertainties in all coordinates of each data point, and including an error analysis.

Note: This paper is adapted from a technical report I wrote as a graduate student in the Department of Physics, University of California, Berkeley, in 1976.

1 Introduction

A recurring computational problem in the field of isotopic studies of terrestrial and extraterrestrial materials has been the interpretation of the observed mass spectra in terms of mixtures of various source components. Each source component is characterized by fixed, but often unknown, isotopic ratios, but it is present in variable amounts in different measured samples. One would like to verify the hypothesis that a given number of components is adequate to account for all observations, and, if possible, not only to determine the source component compositions, but also to resolve each measured sample into its original components, in order to separate different processes in its origin for study. This paper treats only the first steps in this sequence of analysis, i.e., the investigation into the number and compositions of possible source components.

The measured mass spectra may be denoted by $Y_i(\mu)$, where $i = 1, \ldots, n$ is the sample number, and μ is the mass number. Since μ takes on only discrete values $\mu_k,\ k = 1, \ldots, p$, each mass spectrum can be represented by a vector in a p-dimensional vector space, $Y_{ik} = Y_i(\mu_k)$. These are assumed to be made up of linear combinations of the component spectra,

$$Y_i(\mu) = \sum_{j=1}^{m} \alpha_{ij} g_j(\mu) \tag{1}$$

Where the α_{ij} are scalars between 0 and 1 subject to the normalization condition $\sum_{j=1}^{m} \alpha_{ij} = 1,\ i = 1, \ldots, n$ and the $g_j(\mu)$ are the m different component spectra.

The problem is analogous to that of curve resolution encountered, for example, in chromatography or spectrophotometry, where it has been treated with considerable success using the technique of principal component analysis. Lawton and Silvestre (1971), for example, have considered the case of two source components

and have developed a method for computing two bands of curves, each containing one of the source components. The method of principal component analysis, however, runs into difficulties if the data are characterized by widely different experimental uncertainties. This is often the case with mass spectroscopic data. Even if the relative uncertainties in isotopic ratios are similar, the ratios can vary by orders of magnitude. As Anderson (1963) pointed out, the method of principal component analysis is justified only if the ratio of the "uncertainty" variance to the "systematic," i.e. correlation, variance is the same for all components of the data. Nonconformity with this requirement may be remedied to some degree by rescaling the data according to their respective uncertainties. Here we abandon the method of principal component analysis for an alternative approach that is on a better statistical footing in that it takes full account of the estimated uncertainties of the data.

It is easily shown that data points consisting of linear combinations of components according to Equation (1) must lie in an $m-1$-dimensional subspace of the full p-dimensional vector space. This subspace is defined by the simplex whose vertices are the m distinct components. This paper deals with only the first step in component resolution, namely the determination of the parameters of this subspace. Furthermore, it considers only the simplest case, in which $p = m$, that is, the number of components is the same as the number of coordinates of the space (e.g. the number of isotopic ratios measured in each sample). Thus for 2-dimensional data we seek the equation of a straight line, for 3-dimensional data a plane, and in general a hyperplane of dimension one less than the space in which it is embedded. The general case of arbitrary $m \leq p$ is to be dealt with in a future paper.

2 Definition of problem

In accordance with the forgoing discussion, we assume that the measured data points should ideally lie on a hyperplane of $m-1$ dimensions in a space of m dimensions. If we let $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]$ denote a point on this hyperplane, the equation which the data point ideally should satisfy can be written

$$f(\boldsymbol{y}) = \sum_{k=1}^{m-1} a_k y_k + a_m - y_m = 0 \tag{2}$$

The measured data consist of a set of n vectors $\boldsymbol{Y_i}$, $i = 1 \ldots, n$. Each coordinate Y_{ik} of each data point has an associated experimental uncertainty σ_{ik}. (The σ_{ik} may or may not be known a priori. We assume that at least their relative magnitudes are known.) The experimental errors will cause the data points $\boldsymbol{Y_i}$ to lie scattered off the hyperplane of Equation (2). Therefore one seeks a "best fit" to the data.

The method of principal component analysis, referred to above, is equivalent to seeking the hyperplane that minimizes the sum of squares of perpendicular distances from the measured points to the hyperplane. As already mentioned, this method is unable to take proper account of the experimental uncertainties of the data, and is not invariant under a change of scale of one or more axes.

The usual method that one finds in the literature for obtaining a best fit of this kind is based on minimizing a sum of squares of the residuals $f(\boldsymbol{Y_i})$. This is called a regression of y_m against y_1 through y_{m-1}. The sum of squares of these residuals is

either unweighted or weighted by $1/\sigma_{im}^2$ (Bevington, 1991). This approach ignores the uncertainties in coordinates y_1 through y_{m-1}. It also gives different results depending on which coordinate is chosen as the "dependent" coordinate y_m.

The correct treatment that properly takes account of the experimental uncertainties was formulated by Deming (1943). Given the data Y_{ik} with associated uncertainties σ_{ik}, a set of corresponding "adjusted" values y_{ik} are sought which lie exactly on the hyperplane (2) and minimize the variance

$$S = \sum_i \sum_{k=1}^{m} \frac{1}{\sigma_{ik}^2} (Y_{ik} - y_{ik})^2 \tag{3}$$

The solution of this formulation of the problem is not straightforward. York (1966) first devised an approach, later improved by Williamson (1968), for the straight-line case. Here we extend Williamson's solution to arbitrary $m \geq 1$.

3 Derivation of solution

We begin with the constraint that the adjusted points $\boldsymbol{y_i}$ are required to satisfy the hyperplane equation, i.e. $f(\boldsymbol{y_i}) = 0,\ i = 1, \ldots, n$ where f is as defined in Equation (2). Defining residuals $V_{ik} = Y_{ik} - y_{ik}$ allows this constraint to be rewritten as

$$f(\boldsymbol{y_i}) = f(\boldsymbol{Y_i}) - \sum_{k=1}^{m-1} a_k V_{ik} + V_{im} = 0, \quad i = 1, \ldots, n \tag{4}$$

Now, given any values of the parameters a_k, we seek those values of y_{ik} that will minimize S for that choice of the a_k, subject to the constraint (4). Thus we require

$$\delta S = 0 = \sum_i \sum_{k=1}^{m} \frac{1}{\sigma_{ik}^2} V_{ik}\, \delta V_{ik} \tag{5}$$

From (4) we have

$$\delta f(\boldsymbol{y_i}) = 0 = -\sum_{k=1}^{m-1} a_k\, \delta V_{ik} + \delta V_{im}, \quad i = 1, \ldots, n \tag{6}$$

Multiplying each of the n equations (6) by its own undetermined multiplier λ_i and adding them all to Equation (5), we obtain

$$\sum_i \sum_{k=1}^{m-1} \left(\frac{V_{ik}}{\sigma_{ik}^2} - \lambda_i a_k\right) \delta V_{ik} + \sum_i \left(\frac{V_{im}}{\sigma_{im}^2} + \lambda_i\right) \delta V_{im} = 0 \tag{7}$$

Since the V_{ik} are independent, the coefficients of δV_{ik} in this equation must individually be zero, giving

$$\begin{array}{rcll} V_{ik} & = & \lambda_i a_k \sigma_{ik}^2, & i = 1, \ldots, n,\ k = 1, \ldots, m-1 \\ V_{im} & = & -\lambda_i \sigma_{im}^2, & i = 1, \ldots, n \end{array} \tag{8}$$

Substituting this result into Equation (4) and solving for λ_i yields

$$\lambda_i = W_i f(\boldsymbol{Y_i}) \tag{9}$$

where

$$W_i = \left[\sum_{k=1}^{m-1} a_k^2 \sigma_{ik}^2 + \sigma_{im}^2 \right]^{-1} \tag{10}$$

This allows S to be rewritten as

$$S = \textstyle\sum_i W_i f(\boldsymbol{Y_i})^2 \tag{11}$$

This expresses S in the form of a weighted sum of the residuals as defined for a conventional regression, but with weights that properly take account of the individual uncertainties σ_{ik}. Now S is to be minimized with respect to the parameters a_k. Setting $\partial S/\partial a_k = 0$ leads to the following set of equations analogous to the normal equations of the conventional regression:

$$\begin{aligned} &\textstyle\sum_i W_i f(\boldsymbol{Y_i})\, y_{ik} = 0, \quad k = 1, \ldots, m-1 \\ &\textstyle\sum_i W_i f(\boldsymbol{Y_i}) = 0 \end{aligned} \tag{12}$$

Since the parameters a_k occur in W_i, $f(\boldsymbol{Y_i})$, and $\boldsymbol{y_i}$, these equations are nonlinear and cannot be solved in closed form. However, recognizing that W_i and $\boldsymbol{y_i}$ are only weakly dependent on the parameters of the hyperplane, we can linearize Equations (12) by treating those quantities as constants. Writing out f in terms of the parameters, then, we obtain

$$\begin{aligned} \sum_{k=1}^{m-1} (\textstyle\sum_i W_i y_{ij} Y_{ik})\, a_k \quad &+ \quad (\textstyle\sum_i W_i y_{ij})\, a_m \quad = \quad \textstyle\sum_i W_i y_{ij} Y_{im}, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \ldots, m-1 \\ \sum_{k=1}^{m-1} (\textstyle\sum_i W_i Y_{ik})\, a_k \quad &+ \quad (\textstyle\sum_i W_i)\, a_m \quad = \quad \textstyle\sum_i W_i Y_{im} \end{aligned} \tag{13}$$

Identifying the parenthesized terms in Equation (13) as the elements of an $m \times m$ matrix M and the right-hand sides as elements of a vector $\boldsymbol{b}$ of length m, this set of equations is seen as a linear system of the form $M\boldsymbol{a} = \boldsymbol{b}$. Solution proceeds iteratively. Starting with an initial guess for the vector of coefficients $\boldsymbol{a}$, the matrix M and vector $\boldsymbol{b}$ are evaluated and the system $M\boldsymbol{a} = \boldsymbol{b}$ is solved for the new value of $\boldsymbol{a}$. This new value is used to re-evaluate M and $\boldsymbol{b}$, and the equation is solved again. The iteration is continued until convergence is obtained. In practice, it is not necessary to have a good starting guess for $\boldsymbol{a}$. From Equation (10) it can be seen that the initial choice $\boldsymbol{a} = 0$ gives, as the result of the first iteration, the same parameters as would be obtained if the σ_{ik} were zero for all k except m. This is the same result as would be given by the conventional weighted regression of y_m against the other coordinates. Incidentally, this means that weighted averages ($m = 1$) are computed correctly in one iteration. Experience has shown that the convergence is rapid for data sets where the fit is justified, and only a few iterations are necessary to obtain the coefficients to accuracies that are well within their uncertainties.

3.1 Refinement

It should be mentioned that for the sake of numerical stability, the measured points $\boldsymbol{Y_i}$ should be translated if necessary into a coordinate system whose origin is close to the mean of the points. This stability can be achieved automatically, and the

computation simplified somewhat, by reformulating the solution in the following way.

First, solve for a_m from the last of the equations (13):

$$a_m = \overline{Y}_m - \sum_{k=1}^{m-1} a_k \overline{Y}_k \tag{14}$$

where

$$\overline{Y}_k = \frac{\sum_i W_i Y_{ik}}{\sum_i W_i}, \quad k = 1, \ldots, m \tag{15}$$

This shows that the point $\overline{\boldsymbol{Y_i}}$ lies on the best-fit hyperplane. Now define $\boldsymbol{Y_i}' = \boldsymbol{Y_i} - \overline{\boldsymbol{Y}}$ and $\boldsymbol{z_i} = \boldsymbol{y_i} - \overline{\boldsymbol{Y}}$. From Equations (8) and (9) we have

$$z_{ij} = Y'_{ij} - W_i a_j \sigma_{ij}^2 f(\boldsymbol{Y_i}), \quad i = 1, \ldots, n, \; j = 1, \ldots m-1 \tag{16}$$

where we can express $f(\boldsymbol{Y_i})$ as

$$f(\boldsymbol{Y_i}) = \sum_{k=1}^{m-1} a_k Y'_{ik} - Y'_{im} \quad i = 1, \ldots, n \tag{17}$$

Then upon inserting (14) into the remaining equations (13) we find

$$\sum_{k=1}^{m-1} \left(\sum_i W_i z_{ij} Y'_{ik}\right) a_k \quad = \quad \sum_i W_i z_{ij} Y'_{im}, \quad j = 1, \ldots, m-1 \tag{18}$$

This reformulation has improved the numerical stability and reduced the order of the set of equations that needs to be solved on each iteration by 1.

4 Error analysis

The variances of the hyperplane parameters can be found by evaluating

$$\sigma^2(a_j) = \sum_i \sum_{k=1}^{m} \sigma_{ik}^2 \left(\frac{\partial a_j}{\partial Y_{ik}}\right)^2 \tag{19}$$

(This equation assumes the data Y_{ik} are uncorrelated.) Since the dependence of a_j on Y_{ik} is not linear as Equation (13) suggests, due to the dependence of W_i and y_{ik} on a_j, evaluation of this expression is very complicated. The original version of this paper contained an error in the result of this calculation, and a corrected calculation has not yet been done. To first order, however, ignoring the nonlinearity one obtains the approximation

$$\sigma^2(a_j) \approx M_{jj}^{-1} \tag{20}$$

that is, the variances of the parameters are given simply by the diagonal elements of the inverse of the normal matrix defined in Equation (13). (The off-diagonal elements of this matrix are the covariances of the parameters.) For well-behaved data such as those used for illustration by York (1966), this approximation is good to within a few percent.

If the experimenter does not have standard errors σ_{ik} for the measured quantities Y_{ik}, but only relative uncertainties, the resulting fit is the same using these relative uncertainties, but the variances in the fitted parameters are given by expression (20) multiplied by S/ν, where $\nu = n - m$ is the number of degrees of freedom of the problem. If the errors σ_{ik} are known a priori, then the goodness of fit can be inferred from the value of S/ν, which should be close to unity for normally distributed errors. This constitutes a test of the m-component hypothesis as set forth in the introduction.

5 Bibliography

Anderson, T. W. (1963). Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics,* **34,** 122–148.

Bevington, P. R. and Robinson, D. K. (1991). *Data Reduction and Error Analysis for The Physical Sciences,* McGraw-Hill, New York.

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

Lawton, W. H. and Sylvestre, E. A. (1971). Self modeling curve resolution. *Technometrics,* **13,** 617–633.

Williamson, J. H. (1968). Least-squares fitting of a straight line. *Canadian Journal of Physics,* **46,** 1845–1846.

York, D. (1966). Least-squares fitting of a straight line. *Canadian Journal of Physics,* **44,** 1079–1086.