Least Squares Optimization

The following is a brief review of least squares optimization and constrained optimization techniques. I assume the reader is familiar with basic linear algebra, including the Singular Value decomposition (as reviewed in my handout *Geometric Review of Linear Algebra*).

Least squares (LS) problems are those in which the objective function may be expressed as a sum of squares. Such problems have a natural relationship to distances in Euclidean geometry, and the solutions may be computed analytically using the tools of linear algebra.

1 Regression

Least Squares regression is the most basic form of LS optimization problem. Suppose you have a set of measurements, y_n gathered for different parameter values, x_n. The LS regression problem is to find:

$$\min_p \sum_{n=1}^{N} (y_n - px_n)^2$$

We rewrite the expression in terms of column N-vectors as:

$$\min_p \|\vec{y} - p\vec{x}\|^2$$

Now we describe three ways of obtaining the solution. The traditional (non-linear-algebra) approach is to use calculus. If we set the derivative of the expression with respect to p equal to zero and solve for p, we get:

$$p_{\text{opt}} = \frac{\vec{y}^T\vec{x}}{\vec{x}^T\vec{x}}.$$

Technically, one should verify that this is a minimum (and not a maximum or saddle point) of the expression. But since the expression is a sum of squares, we know the solution must be a

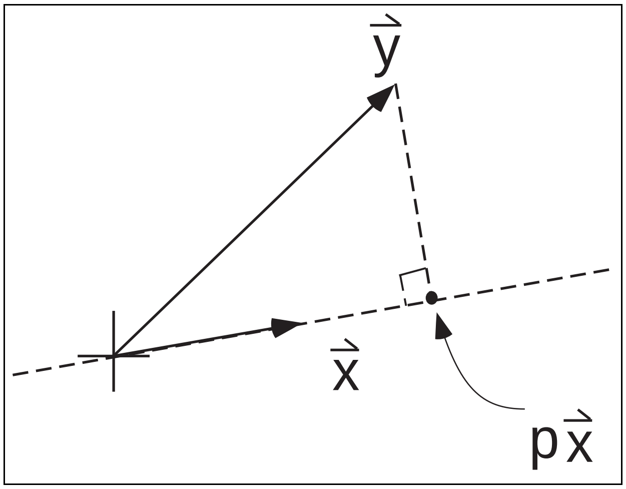
• Author: Eero Simoncelli, Center for Neural Science, and Courant Institute of Mathematical Sciences.
• Created: 15 February 1999. Last revised: 21 July 2003.
• Send corrections or comments to eero.simoncelli@nyu.edu

minimum.

A second method of obtaining the solution comes from considering the geometry of the problem in the N-dimensional space of the data vector. We seek a scale factor, p, such that the scaled vector $p\vec{x}$ is as close as possible (in a Euclidean-distance sense) to $\vec{y}$. Geometrically, we know that the scaled vector should be the projection of $\vec{y}$ onto the line in the direction of $\vec{x}$:

$$p\vec{x} = (\vec{y}\cdot\hat{x})\hat{x} = \frac{(\vec{y}\cdot\vec{x})}{\|\vec{x}\|^2}\vec{x}$$

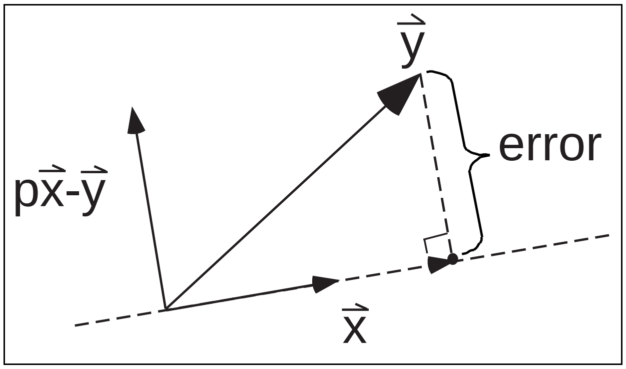
Thus, the solution for p is the same as above.



A third method of obtaining the solution comes from the so-called **orthogonality principle**. The concept is that the error vector for the optimal p should be perpendicular to $\vec{x}$:

$$\vec{x}\cdot(p\vec{x}-\vec{y}) = 0.$$

Solving for p gives the same result as above.



Generalization: Fitting with a basis set

The basic regression problem generalizes to fitting the data with a sum of basis functions, f_{mn}:

$$\min_{\{p_m\}} \sum_{n=1}^{N}(y_n - \sum_m p_m f_{mn})^2$$

or in matrix form:

$$\min_{\vec{p}} \|\vec{y} - F\vec{p}\|^2$$

where F is a matrix whose columns contain the basis functions. For example, if we wanted to include an additive constant in the fitting done in the previous section, F would contain a column with the x_n's, and another column of all ones.

As before there are three ways to obtain the solution: using (vector) calculus, using the geometry of projection, or using the orthogonality principle. The geometric solution can be greatly simplified by first computing the SVD of matrix F [verify]. The orthogonality method is the simplest to obtain, so we show it here. The generalization of the orthogonality principle to a multi-dimensional basis is quite simple: The error vector should be perpendicular to *all* of the basis vectors. This may be expressed directly in terms of the matrix F:

$$F^T * (\vec{y} - F\vec{p}) = 0$$

Solving for $\vec{p}$ gives:

$$\vec{p}_{\text{opt}} = (F^T F)^{-1} F^T \vec{y}$$

The square matrix $(F^T F)$ will be invertible if (and only if) F is full-rank, which is equivalent to saying that the basis vectors are linearly independent. If they are not, one can use the pseudo-inverse of the matrix, which gives a solution that is optimal but not unique.

Generalization: Weighting

Sometimes, the data come with additional information about which points are more reliable. For example, different data points may correspond to averages of different numbers of experimental trials. The regression formulation is easily augmented to include weighting of the data points. Form an $N \times N$ diagonal matrix W with the appropriate error weights in the diagonal entries. Then the problem becomes:

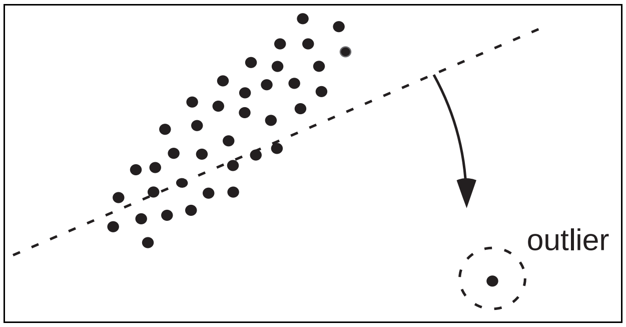
$$\min_{\vec{p}} \|W(\vec{y} - F\vec{p})\|^2$$

and, using the same methods as described above, the solution is

$$\vec{p}_{\text{opt}} = (F^T W^T W F)^{-1} F^T W^T W \, \vec{y}$$

Generalization: Robustness

The most serious problem with LS regression is non-robustness to outliers. In particular, if you have one extremely bad data point, it will have a strong influence on the solution. A simple remedy is to iteratively discard the worst-fitting data point, and re-compute the LS fit to the remaining data.



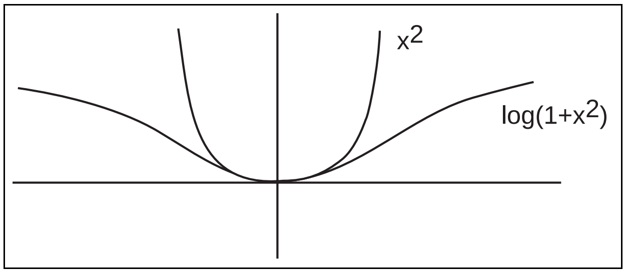
Alternatively one can consider the use of a so-called “robust error metric” $d(\cdot)$ in place of the squared error:

$$min_{\vec{p}} \sum_n d(y_n - F_n \vec{p}).$$

For example, a common choice is the “Lorentzian” function:

$$d(e_n) = log(1 + (e_n/\sigma)^2),$$

plotted at the right along with the squared error function. Note that this function gives less penalty to large errors.



Use of such a function will, in general, mean that we can no longer get an analytic solution to the problem. In most cases, this means that we’ll have to use a numerical algorithm (e.g., gradient descent) to search the parameter space for a minimum. We may not find a minimum, or we may get stuck in a local minimum.

2 Total Least Squares (Orthogonal) Regression

In classical least-squares regression, errors are defined as the squared distance from the data points to the fitted function, as measured along a particular axis direction. But if there is not a clear assignment of “dependent” and “independent” variables, then it makes more sense to measure errors as the squared *perpendicular* distance to the fitted function. The drawback of this formulation is that the fitted surfaces must be subspaces (lines, planes, hyperplanes).

Suppose one wants to fit the N-dimensional data with a subspace (line / plane / hyperplane) of dimensionality $N - 1$. The space is conveniently defined as containing all vectors perpendicular to a unit vector $\hat{u}$, and the optimization problem may thus be expressed as:

$$\min_{\vec{u}} \|M\vec{u}\|^2, \qquad \text{s.t.} \quad \|\vec{u}\|^2 = 1,$$

where M is a matrix containing the data vectors in its rows.

Performing a Singular Value Decomposition (SVD) on the matrix M allows us to find the solution more easily. In particular, let $M = USV^T$, with U and V orthogonal, and S diagonal with positive decreasing elements. Then

$$\begin{aligned} \|M\vec{u}\|^2 &= \vec{u}^T M^T M \vec{u} \\ &= \vec{u}^T V S^T U^T U S V^T \vec{u} \\ &= \vec{u}^T V S^T S V^T \vec{u} \end{aligned}$$

Since V is an orthogonal matrix, we can modify the minimization problem by substituting the vector $\vec{v} = V^T\vec{u}$, which has the same length as $\vec{u}$:

$$\min_{\vec{v}} \vec{v}^T S^T S \vec{v}, \qquad \text{s.t.} \quad \|\vec{v}\| = 1.$$

The matrix $S^T S$ is square and diagonal, with diagonal entries s_n^2. Because of this, the expression being minimized is a weighted sum of the components of $\vec{v}$ which must be greater than the square of the smallest (last) singular value, s_N:

$$\begin{aligned} \vec{v}^T S^T S \vec{v} &= \sum_n s_n^2 v_n^2 \\ &\geq \sum_n s_N^2 v_n^2 \\ &= s_N^2 \sum_n v_n^2 \\ &= s_N^2 \|\vec{v}\|^2 \\ &= s_N^2. \end{aligned}$$

where we have used the constraint that $\vec{v}$ is a unit vector in the last step. Furthermore, the expression becomes an equality when $\vec{v}_{\text{opt}} = \hat{e}_N = [0\ 0\ \cdots\ 0\ 1]^T$, the standard basis vector associated with the Nth axis [verify].

We can transform this solution back to the original coordinate system to get a solution for $\vec{u}$:

$$\begin{aligned} \vec{u}_{\text{opt}} &= V\vec{v}_{\text{opt}} \\ &= V\hat{e}_N \\ &= \vec{v}_N, \end{aligned}$$

which is the Nth column of the matrix V. In summary, the minimum value of the expression occurs when we set $\vec{v}$ equal to the column of V associated with the minimal singular value.

The formulation can easily be augmented to include a shift of origin. That is, suppose we wish to fit the data with a line / plane / hyperplane that does not necessarily pass through the origin:

$$\min_{\vec{u},u_0} \|M\vec{u} - u_0\vec{1}\|^2, \qquad \|\vec{u}\| = 1.$$

where $\vec{1}$ is a column vector of ones. For a given $\vec{u}$, the optimal solution for u_0 is easily found to be $u_0 = \bar{m}^T\hat{u}$, where $\bar{m}$ is a vector whose components are the average of each column of M [verify].

Suppose we wanted to fit the data with a line/plane/hyperplane of dimension $N-2$? We could first find the direction along which the data vary least, project the data into the remaining $(N-1)$-dimensional space, and then repeat the process. Because V is an orthogonal matrix, the secondary solution will be the second column of V (i.e., the column associated with the second-largest singular value). In general, the columns of V provide a basis for the data space, in which the axes are ordered according to variability. We can solve for a vector subspace of any desired dimensionality in which the data are closest to lying.

The total least squares problem may also be formulated as a pure (unconstrained) optimization problem using a form known as the **Rayleigh Quotient**:

$$\min_{\vec{u}} \frac{\vec{u}^T M^T M \vec{u}}{\vec{u}^T \vec{u}}.$$

The length of the vector doesn't change the value of the fraction, so one typically solves for a unit vector. As above, this fraction takes on values in the range $[s_N^2, s_1^2]$, and is equal to the minimum value when $\vec{u} = \vec{v}_N$, the first column of the matrix V.

Relationship to Principal Component Analysis

Often, one has a data set in some large-dimensional space, but the actual data are close to lying within a much smaller-dimensional subspace. In such cases, one would like to project the data into the small-dimensional space in order to summarize or analyze them. Specifically, the least squares formulation of the problem is: find a subspace that captures most of the summed squared vector lengths of the data. This problem is just a variant of the TLS problem discussed above. The axes (basis) for this low-dimensional space are known as the **Principal Components** of the data, and correspond to the columns of V (the second orthogonal matrix in the SVD) associated with the largest singular values. In a typical problem setting, one looks at the decreasing sequence of singular values and decides how many dimensions are necessary to adequately represent the data. One then transforms the data into this subspace (by projecting onto these axes).

Relationship to Eigenvector Analysis

The Total Least Squares and Principal Components problems are often stated in terms of **eigenvectors**. The eigenvectors of a square matrix are a set of vectors that the matrix re-scales:

$$S\vec{v} = \lambda\vec{v}.$$

The scalar λ is known as the **eigenvalue** associated with $\vec{v}$.

The problems we've been considering can be restated in terms of eigenvectors by noting a simple relationship between the SVD and eigenvector decompositions. The total least squares problems all involve minimizing expressions

$$\|M\vec{v}\|^2 = \vec{v}^T M^T M \vec{v}$$

Substituting the SVD $(M = USV^T)$ gives:

$$\vec{v}^T V S^T U^T U S V^T \vec{v} = \vec{v}(V S^T S V^T \vec{v})$$

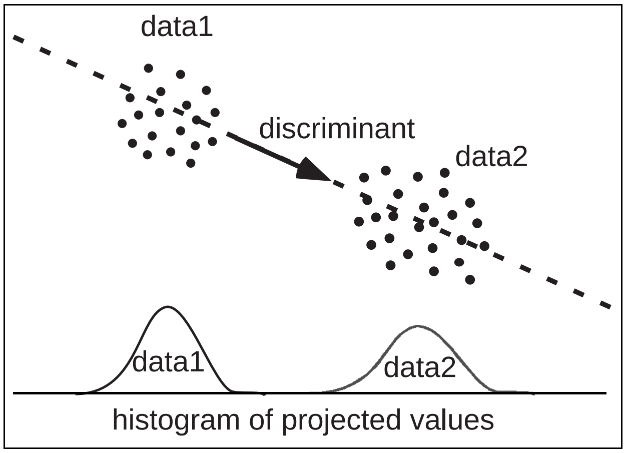
Consider the parenthesized expression. When $\vec{v} = \vec{v}_n$, the nth column of V, this becomes

$$M^T M \, \vec{v}_n = (VS^TSV^T) \, \vec{v}_n = Vs_n^2\vec{e}_n = s_n^2\vec{v}_n,$$

where $\vec{e}_n$ is the nth standard basis vector. That is, the $\vec{v}_n$ are eigenvectors of (M^TM), with associated eigenvalues $\lambda_n = s_n^2$. Thus, we can solve total least squares problems by seeking the eigenvectors of the symmetric matrix M^TM.

3 Fisher's Linear Discriminant

Suppose we have two sets of data gathered under different conditions, and we want to find a line/plane/hyperplane that best separates the two sets of data. This problem may be expressed as a LS optimization problem (the formulation is due to Fisher (1936)).



We seek a vector $\vec{u}$ such that the projection of the data sets maximizes the discriminability of the two sets. Intuitively, we'd like to maximize the distance between the two data sets. But a moment's thought should convince you that the distance should be considered relative to the variability within the two data sets. Thus, an appropriate expression to maximize is the ratio of the squared distance between the means of the classes and the sum of the within-class squared distances:

$$\max_{\vec{u}} \frac{[\vec{u}^T(\bar{a} - \bar{b})]^2}{\frac{1}{M}\sum_m[\vec{u}^T\vec{a}'_m]^2 + \frac{1}{N}\sum_n[\vec{u}^T\vec{b}'_n]^2}$$

where $\{\vec{a}_m, 1 \leq m \leq M\}$ and $\{\vec{b}_n, 1 \leq n \leq N\}$ are the two data sets, $\bar{a}, \bar{b}$ represent the averages (centroids) of each data set, and $\vec{a}'_m = \vec{a}_m - \bar{a}$ and $\vec{b}'_n = \vec{b}_n - \bar{b}$.

Rewriting in matrix form gives:

$$\max_{\vec{u}} \frac{\vec{u}^T[(\bar{a} - \bar{b})(\bar{a} - \bar{b})^T]\vec{u}}{\vec{u}^T[\frac{A^TA}{M} + \frac{B^TB}{N}]\vec{u}}$$

where A and B are matrices containing the $\vec{a}'_m$ and $\vec{b}'_n$ as their rows. This is now a quotient of quadratic forms, and we transform to a standard Rayleigh Quotient by finding the eigenvector matrix associated with the denominator[1]. In particular, since the denominator matrix

[1] It can also be solved directly as a generalized eigenvector problem.

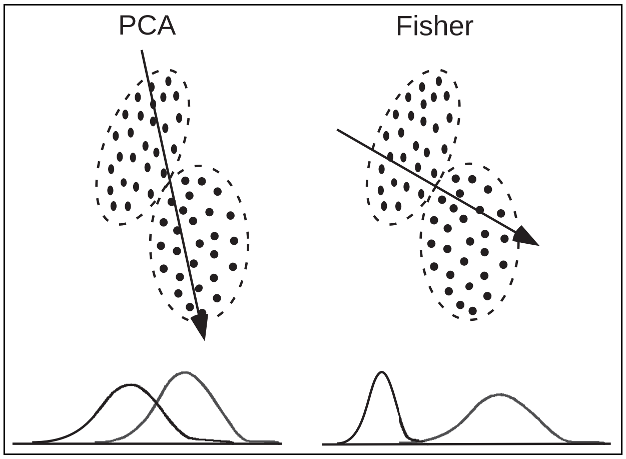
is symmetric, it may be factorized as follows

$$[\frac{A^T A}{M} + \frac{B^T B}{N}] = V D^2 V^T$$

where V is orthogonal and contains the eigenvectors of the matrix on the left hand side, and D is diagonal and contains the square roots of the associated eigenvalues. Assuming the eigenvalues are nonzero, we define a new vector relate to $\vec{u}$ by an invertible transformation: $\vec{v} = D V^T \vec{u}$. Then the optimization problem becomes:

$$\max_{\vec{v}} \frac{\vec{v}^T [D^{-1} V^T (\bar{a} - \bar{b})(\bar{a} - \bar{b})^T V D^{-1}] \vec{v}}{\vec{v}^T \vec{v}}$$

The optimal solution for $\vec{v}$ is simply the eigenvector of the numerator matrix with the largest associated eigenvalue.[2] This may then be transformed back to obtain a solution for the optimal $\vec{u}$.



To emphasize the power of this approach, consider the example shown to the right. On the left are the two data sets, along with the Principal Components of the full data set. Below this are the histograms for the two data sets, as projected onto the first component. On the right are the same two plots with Fisher's Linear Discriminant. It is clear the latter provides a much better separation of the two data sets.

[2] In fact, the rank of the numerator matrix is 1 and a solution can be seen by inspection to be $\vec{v} = D^{-1} V^T (\bar{a} - \bar{b})$.