

Least Squares and matrices

Jonathan L.F. King
University of Florida, Gainesville FL 32611-2082, USA
squash@ufl.edu
Webpage <http://squash.1gainesville.com/>

The Problem

Suppose we have a collection \mathcal{K} of N points $Q_1, \dots, Q_j, \dots, Q_N$ in the plane.^{♥1} Consider now the line L with equation $y = \beta x + \alpha$. It has slope β and y -intercept α . At a given point $Q = (x, y)$, the vertical (signed) distance to L is

$$1: \quad v := [\alpha + \beta x] - y.$$

Letting v_j denote the vertical distance at Q_j , define the *least-square distance* from \mathcal{K} to L by

$$1': \quad g(\alpha, \beta) := \sum_{j=1}^N [v_j]^2.$$

Our goal is to find all pairs (α, β) which minimize g . It will turn out there is a *unique* minimum, *except* in the silly case that all the given points lie on one vertical line. That is, writing Q_j as (x_j, y_j) , except when $x_1 = \dots = x_N$.

The quantities that we will need are

$$X := \sum_{j=1}^N x_j, \quad Y := \sum_{j=1}^N y_j, \\ S := \sum_{j=1}^N x_j^2, \quad P := \sum_{j=1}^N y_j x_j.$$

(“ S ” is for Squares and “ P ” is for Product.)

Using Calculus

Evidently in computing the first-partials of g we will want to compute them for each v_j . From (1) we compute that

$$\frac{dv}{d\alpha} = 1, \quad \text{so} \quad \frac{dg}{d\alpha}(v^2) = 2v \cdot 1 \quad \text{and} \\ \frac{dv}{d\beta} = x, \quad \text{so} \quad \frac{dg}{d\beta}(v^2) = 2v \cdot x,$$

^{♥1}“Plane” is either $\mathbb{R} \times \mathbb{R}$ or $\mathbb{C} \times \mathbb{C}$.

by the Chain Rule. Consequently

$$\frac{dg}{d\alpha} = \sum_{j=1}^N 2v_j \quad \text{and} \quad \frac{dg}{d\beta} = \sum_{j=1}^N 2v_j x_j.$$

Thus, the pair (α, β) is a critical point of g IFF at (α, β) we have that

$$2: \quad 0 = \sum_{j=1}^N v_j \quad \text{and} \quad 0 = \sum_{j=1}^N v_j x_j.$$

Recall that v_j is $\alpha + x_j \beta - y_j$. So multiplying out and distributing the summations in (2) yields that

$$2': \quad 0 = N\alpha + X\beta - Y, \quad 0 = X\alpha + S\beta - P.$$

We can rewrite this to say that (α, β) is a critical point of g IFF

$$3: \quad Y = N\alpha + X\beta, \\ P = X\alpha + S\beta.$$

Matrices. Let \mathbf{M} denote the matrix $\begin{bmatrix} N & X \\ X & S \end{bmatrix}$ and let

$$D := \text{Det}(\mathbf{M}) \stackrel{\text{note}}{=} NS - X^2.$$

It follows from a standard^{♥2♥3} inequality that: *All the points x_1, \dots, x_N are equal IFF $D = 0$.* We henceforth assume that our scatterplot has at least two distinct x -values.

Bare-hands computation [or matrix algebra] shows that (3) has a unique solution, which is

$$4: \quad \alpha = \frac{1}{D} [SY - XP], \\ \beta = \frac{1}{D} [-XY + NP].$$

^{♥2}Jensen's Inequality implies that D is positive. For *that* assertion is equivalent to “ $D/[N^2] > 0$ ”, i.e. to

$$\frac{1}{N} \sum_{j=1}^N [x_j]^2 > \left[\frac{1}{N} \sum_{j=1}^N x_j \right]^2.$$

This has form $\frac{1}{N} \sum_1^N f(x_j) > f(\frac{1}{N} \sum_1^N x_j)$, where f is the squaring-map. Since f is strictly convex-up, Jensen's yields “ \geq ”, with equality IFF $x_1 = \dots = x_N$.

^{♥3}We use the Cauchy-Schwarz inequality [CS] with inner-product $\langle (p_1, \dots, p_N), (q_1, \dots, q_N) \rangle := \sum_1^N \overline{p_j} \cdot q_j$. For let $\mathbf{1} := (1, \dots, 1)$ and $\mathbf{w} := (x_1, \dots, x_N)$. CS gives

$$|\langle \mathbf{1}, \mathbf{w} \rangle|^2 \leq \langle \mathbf{1}, \mathbf{1} \rangle \cdot \langle \mathbf{w}, \mathbf{w} \rangle,$$

i.e. $X^2 \leq N \cdot S$. There is equality IFF \mathbf{w} is a multiple of $\mathbf{1}$, i.e. IFF all the x_j equal a common value.

Neat! (Exer. E1: Let $Q_j := (j, j^2)$. For $N = 2, 3, 4, 5$, find the best approximating line to scatterplot Q_1, \dots, Q_N . How do the slopes of the lines change as you increase N ? Taking two of the geometric points in the list, what happens to the fitting-line if you *repeat* each of them several times to make a new list?)

Using Linear Algebra

In matrix notation we can write (4) as

$$4': \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = M^{-1} \cdot \begin{bmatrix} Y \\ P \end{bmatrix},$$

suggesting that *least-squares* secretly contains linear algebra. We set the stage for a more general problem, then apply it to *least-squares*.

With \mathbf{F} either \mathbb{R} or \mathbb{C} , consider an N -dim'al \mathbf{F} -inner-product space $(\mathbf{H}, \langle \cdot, \cdot \rangle)$, a K -dimensional subspace $\mathbf{W} \subset \mathbf{H}$ and its *ortho-complement*

$$\mathbf{W}^\perp := \{ \mathbf{g} \in \mathbf{H} \mid \forall \mathbf{w} \in \mathbf{W}: \mathbf{g} \perp \mathbf{w} \}.$$

The *orthogonal projection* operator is the map $\text{Proj} : \mathbf{H} \rightarrow \mathbf{W}$ satisfying, for each $Q \in \mathbf{H}$, that

$$Q - \text{Proj}(Q) \in \mathbf{W}^\perp.$$

Point $P := \text{Proj}(Q)$ is the (unique) *closest-point* on \mathbf{W} to Q ; it minimizes $\langle \mathbf{w} - Q, \mathbf{w} - Q \rangle$ as \mathbf{w} ranges over \mathbf{W} .

Subspaces. One way to get a subspace is as the range of a linear map $A: \mathbf{F}^K \rightarrow \mathbf{H}$; so let

$$5: \quad \mathbf{W} := \text{Range}(A) \stackrel{\text{note}}{=} \{ AU \mid U \in \mathbf{F}^K \}.$$

Consider a point $P \in \mathbf{W}$. Then

$$6: \quad \begin{array}{l} \text{There is a \underline{unique} } U_0 \in \mathbf{F}^K \text{ with } AU_0 = P \\ \text{IFF } U \mapsto AU \text{ is 1-to-1, i.e, Rank}(A) = K. \end{array}$$

We want to state this rank-condition in terms of an adjoint operator, so equip \mathbf{F}^K with the [conjugate] dot-product.^{♥4} Thus we have a well-defined *adjoint map* $A^*: \mathbf{H} \rightarrow \mathbf{F}^K$, defined by

^{♥4}Actually, any inner-product on \mathbf{F}^K works in (8), but note that changing the IP will change what “ A^* ” means.

$$7: \quad \forall \mathbf{g} \in \mathbf{H} \text{ and } \forall U \in \mathbf{F}^K: \langle A^* \mathbf{g}, U \rangle = \langle \mathbf{g}, AU \rangle.$$

(Exer. E2: Show that $[A^*]^* = A$.) Hence we have linear maps $A^*A: \mathbf{F}^K \rightarrow \mathbf{F}^K$ and $AA^*: \mathbf{H} \rightarrow \mathbf{H}$. A standard result (Exer. E3) is that $\text{Ker}(A^*A) = \text{Ker}(A)$. A corollary of this is that $\text{Rank}(A^*A) = \text{Rank}(A)$.

So we can restate the above as

$$6': \quad \begin{array}{l} \text{There is a \underline{unique} } U_0 \in \mathbf{F}^K \text{ with } AU_0 = P \\ \text{IFF } U \mapsto AU \text{ is 1-to-1, i.e, Rank}(A) = K. \\ \text{IFF } A^*A \text{ is invertible.} \end{array}$$

The Problem. Fix a rank- K linear-map $A: \mathbf{F}^K \rightarrow \mathbf{H}$ and a point $Q \in \mathbf{H}$. We seek a formula for the unique point $U_0 \in \mathbf{F}^K$ so that $\|AU_0 - Q\|$ is the minimum of $\|AU - Q\|$ taken over all $U \in \mathbf{F}^K$.

The difference-vector $AU_0 - Q$ is orthogonal to every vector in (5). I.e, for each $U \in \mathbf{F}^K$, inner product $\langle AU_0 - Q, AU \rangle$ is zero. By (7), then,

$$\langle A^*AU_0 - A^*Q, U \rangle = 0.$$

But the only vector orthogonal to *all* $U \in \mathbf{F}^K$ is $\hat{0} \in \mathbf{F}^K$. Thus U_0 satisfies $A^*AU_0 = A^*Q$. Hence

$$8: \quad U_0 = [A^*A]^{-1}A^*Q,$$

courtesy (??').

Least squares. We can apply this to our line-fitting of (1). After all, RhS(??') is the square of the dot-product norm on $\mathbf{H} := \mathbf{F}^N$. We are minimizing the square-norm of column vector $[v_1, \dots, v_N]^t$. Our unknown vector is $U = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$; so $K = 2$. With

$$9: \quad A := \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad \text{and} \quad Q := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix},$$

we are minimizing the norm of $AU - Q$ over all U .

Applying (8), the minimum occurs at

$$8': \quad \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = [A^*A]^{-1}A^*Q$$

Of course, RhS(??') must equal RhS(??'). Indeed we find that $A^*A = M$ and $A^*Q = \begin{bmatrix} Y \\ P \end{bmatrix}$.

Finally, note that $\text{Rank}(A)$ equals K , i.e, equals 2, exactly when not all x_1, \dots, x_N are equal. This was precisely the “non-silly” condition we needed for the *Calculus* approach.

Fitting to a polynomial. To our N many data-points $Q_j = (x_j, y_j)$ in the $\mathbf{F} \times \mathbf{F}$ plane, we wish to least-squares fit the closest K -topped (i.e. $\text{Deg} < K$) polynomial

$$10: \quad \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_{K-1} x^{K-1}.$$

Copying what we did in (9), define our “unknown” col-vector $\mathbf{U} := [\alpha_0 \ \alpha_1 \ \dots \ \alpha_{K-1}]^t$, as well as

$$9': \quad \mathbf{A}_K := \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{K-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^{K-1} \end{bmatrix}, \quad \mathbf{Q} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

When $\text{Rank}(\mathbf{A}_K)$ equals K , then (8) applies, telling us that the closest-fit polynomial (10) has coefficients $\mathbf{U} = [\mathbf{A}_K^* \mathbf{A}_K]^{-1} \mathbf{A}_K^* \mathbf{Q}$.

Vandermonde matrices. The $N \times K$ matrix \mathbf{A}_K of (??') is called a **Vandermonde matrix**.^{♥5} When N and K equal a common value, L , then –it turns out–

$$11: \quad \text{Det}(\mathbf{A}_L) = \prod_{\substack{j,i \in [1..L] \\ \text{with } j > i}} [x_j - x_i].$$

Returning to the general $N \times K$ case, let L denote the number of *distinct* values in $\{x_1, \dots, x_N\}$, and suppose that $K \geq L$. Remove the duplicate rows, then only keep the first L many columns. We have thus produced an $L \times L$ Vandermonde matrix *inside* our original $N \times K$ matrix, and (11) implies that this $L \times L$ has non-zero determinant. We have thus proven:

Fix an arbitrary field \mathbf{F} , points $x_1, \dots, x_N \in \mathbf{F}$, and let L be the number of distinct points in this

11': *list. Then, for each $K \geq L$, the Vandermonde matrix $\mathbf{A}_K(x_1, x_2, \dots, x_N)$ has rank equaling L , the cardinality of set $\{x_1, x_2, \dots, x_N\}$.*

So we get a *unique* K -topped polynomial least-squares-closest to our N many data-points *exactly when* there are at least K distinct x -values among the points.

^{♥5}The Vandermonde-matrix Wikipedia article is nice.

Lagrange polynomials. Suppose points x_1, \dots, x_N are distinct. If K equals N , then *Lagrange Interpolation* tells us there is a *unique* K -topped polynomial whose graph passes through each of $Q_1, \dots, Q_j, \dots, Q_N$; the least-squares distance is zero.

When $K > N$, then there is a *family* of K -topped polynomials (a $[K-N]$ -dim'al family) which pass through the data-points; so no uniqueness in the least-squares fit.

Fitting to a family of functions. Fix an arbitrary set \mathbf{S} , functions $f_0, f_1, \dots, f_{K-1}: \mathbf{S} \rightarrow \mathbf{F}$, and let \mathcal{G} be the set of linear combinations $\sum_{j=0}^{K-1} c_j \cdot f_j()$.

A *scatterplot* is a multiset $\{Q_j\}_{j=1}^N$ of points

$$Q_j = (\mathbf{s}_j, \tau_j) \in \mathbf{S} \times \mathbf{F}.$$

Points $\{\mathbf{s}_j\}_{j=1}^N \subset \mathbf{S}$ are the *sample points*, and $\{\tau_j\}_{j=1}^N$ are the *target values*. [Previously we used “ x_j ” for a sample point, and “ y_j ” for a target value.]

We can use the preceding technique to find a function $g \in \mathcal{G}$ which minimizes the least-square distance to scatterplot $\{Q_j\}_{j=1}^N$. Namely, define this $N \times K$ matrix and column-vector

$$12: \quad \mathbf{A} := \begin{bmatrix} f_0(\mathbf{s}_1) & f_1(\mathbf{s}_1) & \dots & f_{K-1}(\mathbf{s}_1) \\ f_0(\mathbf{s}_2) & f_1(\mathbf{s}_2) & \dots & f_{K-1}(\mathbf{s}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(\mathbf{s}_N) & f_1(\mathbf{s}_N) & \dots & f_{K-1}(\mathbf{s}_N) \end{bmatrix}, \quad \mathbf{Q} := \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_N \end{bmatrix}.$$

When \mathbf{A} has rank K , then

$$8'': \quad \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{K-1} \end{bmatrix} := [\mathbf{A}^* \mathbf{A}]^{-1} \mathbf{A}^* \mathbf{Q}$$

is the coeff-vector giving this closest fnc $g()$.

13: *Appl: 1-variable polynomials.* The setup in (??') is a special case of (12), by setting

$$\mathbf{S} := \mathbf{F} \quad \text{and} \quad f_j := [x \mapsto x^j]. \quad \square$$

14: *Appl: Closest plane.* Suppose now you want to find the plane

$$(x, y) \mapsto a + bx + cy$$

least-square closest to $\{Q_j\}_1^N$, where $\mathbf{s}_j = (x_j, y_j)$, a point in $\mathbf{F} \times \mathbf{F}$. So apply (12) and (??"), where

$\mathbf{S} := \mathbf{F} \times \mathbf{F}$ and functions f_0, f_1, f_2 send $\mathbf{s} := (x, y)$ to, respectively: $1, x, y$.

Then $\begin{bmatrix} a \\ b \\ c \end{bmatrix} = [\mathbf{A}^* \mathbf{A}]^{-1} \mathbf{A}^* \mathbf{Q}$. □

Filename: Problems/Analysis/Calculus/least_squares.latex
As of: Monday 06Mar2006. Typeset: 17Nov2017 at 12:10.