MATH 167: APPLIED LINEAR ALGEBRA
Least-Squares

Jesús De Loera, UC Davis

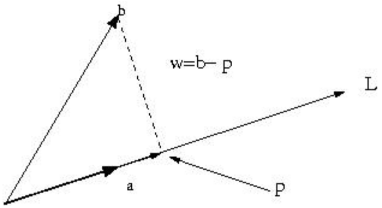
October 30, 2014

Least Squares

- We do a series of experiments, collecting data. We wish to see patterns!!
- We expect the output b to be a linear function of the input t $b = \alpha + t\beta$, but we need to **determine** α, β.
- At different times t_i we measure a quantity b_i.
- **EXAMPLE:** A police man is interested on clocking the speed of a vehicle by using measurements of its relative distance. Assuming the vehicle is traveling at constant speed, so we know linear formula, but errors exist.
- At $t = t_i$, the error between the measured value b_i and the value predicted by the function is $e_i = b_i - (\alpha + \beta t_i)$.
- We can write it as $e = b - Ax$ where $x = (\alpha, \beta)$. e is the **error vector**, b is the **data vector**. A is an $m \times 2$ matrix.
- We seek the line that minimizes the total squared error or Euclidean norm $\|e\| = \sqrt{\sum_{i=1}^{m} e_i^2}$.
- **GOAL:** Given $m \times n$ matrix A and m-vector b, Find x that minimizes $\|b - Ax\|$.
- We assume $m \geq n$.

Distance and projection are closely related to each other!!!

- Fundamental question: If we have a subspace S, is there a formula for the projection p of a vector b into that subspace?
- Imaging b as **data from experiments**, b is not in S, due to error of measurement, its projection p is the best choice to replace b. Key idea of **LEAST SQUARES** for **regression analysis**
- Let us learn how to do this projection for a line! b is projected into the line L given by the vector a. (PICTURE!).



- The projection of vector b onto the line in the direction a is $p = \frac{a^T b}{a^T a} a$.

- Note: $\|b - Ax\|$ is the distance from b to the point Ax which is element of the column space!
- Key point: The optimal solution is x that minimizes that distance!
- **Theorem** The smallest error vector $e = b - Ax$ is must be perpendicular to the column space (picture!).
- Thus for each column a_i we have $a_i^T(b - Ax) = 0$. Thus in matrix notation: $A^T(b - Ax) = 0$, This gives the normal equations $A^T Ax = A^T b$.
- **Theorem** The best estimate is given by $x = (A^T A)^{-1} A^T b$. and its projection is $p = A((A^T A)^{-1} A^T)b$.
- **Lemma** $A^T A$ is a symmetric matrix. $A^T A$ has the same Nullspace as A.
  Why? if $x \in N(A)$, then clearly $A^T Ax = 0$. Conversely, if $A^T Ax = 0$ then $x^T A^T Ax = \|Ax\| = 0$, thus $Ax = 0$.
- **Corollary** If A has independent columns, then $A^T A$ is square, symmetric and invertible.

- **Example** Consider the problem $Ax = b$ with

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & -1 & -2 \\ 2 & 1 & -1 \end{bmatrix} \qquad b^T = (1, 0, -1, 2, 2).$$

- We can see that there is no EXACT solution to $Ax = b$, use NORMAL EQUATION!

$$A^T A = \begin{bmatrix} 16 & -2 & -2 \\ -2 & 11 & 2 \\ -2 & 2 & 7 \end{bmatrix} \qquad A^T b = \begin{bmatrix} 8 \\ 0 \\ -7 \end{bmatrix}$$

- Solving $A^T Ax = A^T b$ we get the least square solution $x^* \approx (0.4119, 0.2482, -0.9532)^T$ with error $\|b - Ax^*\| \approx 0.1799$.

- **Example** A sample of lead-210 measured the following radioactivity data at the given times (time in days). Can YOU predict how long will it take until one percent of the original amount remains?

time in days	0	4	8	10	14	18
mg	10	8.8	7.8	7.3	6.4	6.4

- A linear model does not work. There is an exponential decay on the material $m(t) = m_0 e^{\beta t}$, where m_0 is the initial radioactive material and β the decay rate. By taking logarithms

$$y(t) = log(m(t)) = log(m_0) + \beta t$$

- Thus we can now use **linear** least squares to fit on the logarithms $y_i = log(m_i)$ of the radioactive mass data. In this case we have

$$A^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 4 & 8 & 10 & 14 & 18 \end{bmatrix} \quad b^T = [2.302585093, 2.17475172$$

- Thus $A^TA = \begin{bmatrix} 6 & 54 \\ 54 & 700 \end{bmatrix}$. Solving the NORMAL form system we get $log(m_0) = 2.277327661$ and $\beta = -0.0265191683$
  Thus the original amount was 10 mg. After 173 days it will below one percent of the radioactive material.
- There is nothing special about polynomials or exponential functions in the application. We can deal with approximating function is al linear combination of some prescribed functions $h_1(t), h_2(t), \ldots, h_n(t)$. Then we receive data y_i at time t_i and the matrix A has entry $A_{ij} = h_i(t_j)$.
- The least squares method can be applied when the measurement of error is not can be applied to situations when not all observations are trusted the same way!
- Now the error is $\sqrt{(b-Ax)^T C(b-Ax)}$. Then the weighted least square error is given by the new equations

$$A^TCAx = A^TCb, \quad \text{and} \quad x = (A^TCA)^{-1}A^TCb.$$

Review of Orthogonal Vectors and Subspaces .

- In real life vector spaces come with additional **METRIC properties!! We have notions of distance and angles!!** You are familiar with the **Euclidean vector space** $\mathbb{R}^n$:
- Since kindergarden you know that the distance between two vectors $x = (x_1, \ldots, x_n)$ $y = (y_1, \ldots, y_n)$ is given by

$$dist(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

- We say vectors x, y are **perpendicular** when they make a 90 degree angle. When that happens the triangle they define is right triangle! (WHY?)
- **Lemma** Two vectors x, y in $\mathbb{R}^n$ are perpendicular if and only if

$$x_1y_1 + \cdots + x_ny_n = xy^T = 0$$

When this last equation holds we say x, y are **orthogonal**.
- **Orthogonal Bases:** A basis $u_1, \ldots, u_n$ of V is orthogonal if $\langle u_i, u_j \rangle = 0$ for all $i \neq j$.
- **Lemma** If $v_1, v_2, \ldots, v_k$ are orthogonal then they are linearly independent.

The Orthogonality of the Subspaces

- **Definition** We say two say two subspaces V, W of $\mathbb{R}^n$ are **orthogonal** if for $u \in V$ and $w \in W$ we have $uw^T = 0$.
- Can you see a way to detect when two subspaces are orthogonal?? **Through their bases!**
- **Theorem:** The row space and the nullspace are orthogonal. Similarly the column space is orthogonal to the left nullspace.
- **proof:** The dot product between the rows of A^T and the respective entries in the vector y is zero.
- Therefore the rows of A^T are perpendicular to any $y \in N(A^T)$

$$A^T y = \begin{bmatrix} Column\,1\,of\;A \\ \vdots \\ Column\,n\,of\;A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

where $y \in N(A^T)$.

- There is a stronger relation, for a subspace V of $\mathbb{R}^n$ the set of all vectors orthogonal to V is the **orthogonal complement** of V, denoted $V^\perp$.
- Warning Spaces can be orthogonal without being complements!
- **Exercise** Let W be a subspace, its orthogonal complement is a subspace, and $W \cap W^\perp = 0$.
- **Exercise** If $V \subset W$ subspaces, then $W^\perp \subset V^\perp$.
- **Theorem** (Fundamental theorem part II) $C(A^T)^\perp = N(A)$ and $N(A)^\perp = C(A^T)$. Why?
- **proof:** First equation is easy because x is orthogonal to all vectors of row space $\leftrightarrow$ x is orthogonal to each of the rows $\leftrightarrow$ $x \in N(A)$. The other equality follows from exercises.
- **Corollary** Given an $m \times n$ matrix A, the nullspace is the orthogonal complement of the row space in $\mathbb{R}^n$. Similarly, the left nullspace is the orthogonal complement of the column space inside $\mathbb{R}^m$
- WHY is this such a big deal?

- **Theorem** Given an $m \times n$ matrix A, every vector x in $\mathbb{R}^n$ can be written in a unique way as $x_n + x_r$ where x_n is in the nullspace and x_r is in the row space of A.
- **proof** Pick x_n to be the orthogonal projection of x into $N(A)$ and x_r to be the orthogonal projection into $C(A^T)$. Clearly x is a sum of both, but why are they unique?
- If $x_n + x_r = x'_n + x'_r$, then $x_n - x'_n = x_r - x'_r$ Thus the must be the zero vector because $N(A)$ is orthogonal to to $C(A^T)$.
- This has a beautiful consequence: Every matrix A, when we think of it as a linear map, transforms the row space into its column space!!!

An important picture

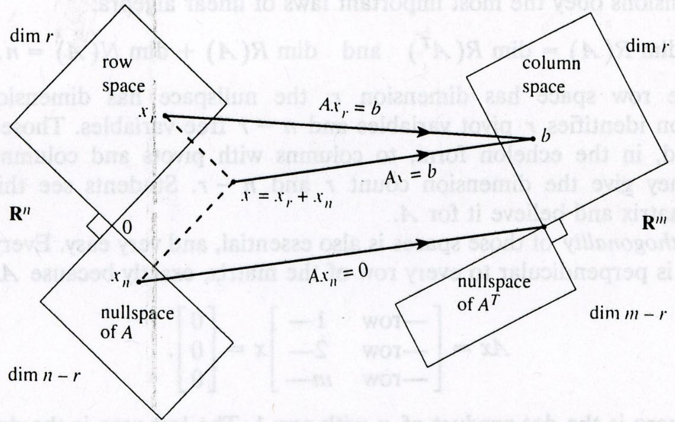


Figure 1. The action of A: Row space to column space, nullspace to zero.

Orthogonal Bases and Gram-Schmidt

- A basis $u_1, \ldots, u_n$ of a vector space V is **orthonormal** if it is orthogonal and each vector has unit length.
- **Observation** If the vectors $u_1, \ldots, u_n$ are orthogonal basis, their normalizations $\frac{u_i}{\|u_i\|}$ form an orthonormal basis.
- **Examples** Of course the standard unit vectors are orthonormal.

  Consider the vector space of all quadratic polynomials $p(x) = a + bx + cx^2$, using the L^2 inner product of integration:

$$\langle p, q \rangle = \int_0^1 p(x)q(x)dx$$

  The standard monomials $1, x, x^2$ form a basis, but do **not** form an orthogonal basis!

$$\langle 1, x \rangle = 1/2, \quad \langle 1, x^2 \rangle = 1/3, \quad \langle x, x^2 \rangle = 1/4$$

- An orthonormal basis is given by

$$u_1(x) = 1, \quad u_2(x) = \sqrt{3}(2x - 1), \quad u_3(x) = \sqrt{5}(6x^2 - 6x + 1).$$

Why do we care about orthonormal bases?

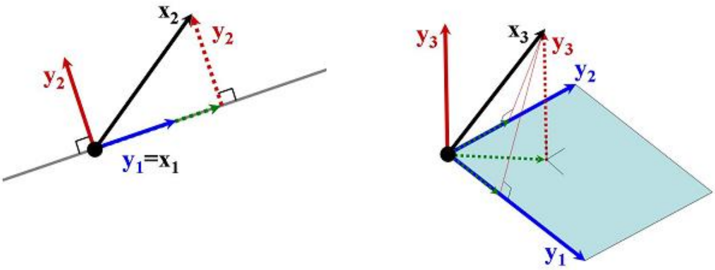
- **Theorem** Let $u_1, \ldots, u_n$ be an orthonormal bases for a vector space with inner product V. The one can write any element $v \in V$ as a linear combination $v = c_1 u_1 + \cdots + c_n u_n$ where $c_i = \langle v, u_i \rangle$, for $i = 1, \ldots, n$. Moreover the norm $\|v\| = \sqrt{\sum c_i^2}$.
- **Example** Let us rewrite the vector $v = (1,1,1)^T$ in terms of the orthonormal basis
  $u_1 = (\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}})^T, u_2 = (0, \frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}), u_3 = (\frac{5}{\sqrt{30}}, \frac{-2}{\sqrt{30}}, \frac{1}{\sqrt{30}})$
  Computing the dot products $v^T u_1 = \frac{2}{\sqrt{6}}, v^T u_2 = \frac{3}{\sqrt{5}}$, and $v^T u_3 = \frac{4}{\sqrt{30}}$. Thus
  $$v = \frac{2}{\sqrt{6}} u_1 + \frac{3}{\sqrt{5}} u_2 + \frac{4}{\sqrt{30}} u_3$$
- **Challenge:** Figure out the same kind of formulas if the vectors are just orthogonal!!!

- A key reason to like matrices that have orthonormal vectors: The least-squares equations are even nicer!!!
- **Lemma** If Q is a rectangular matrix with orthonormal columns, then the normal equations simplify because $Q^TQ = I$:
  - $Q^TQx = Q^Tb$ simplifies to $x = Q^Tb$
  - Projection matrix simplifies $Q(Q^TQ)^{-1}Q^T = QIQ^T = QQ^T$.
  - Thus the projection point is $p = QQ^Tb$, thus

$$p = (q_1^Tb)q_1 + (q_2^Tb)q_2 + \cdots + (q_n^Tb)q_n$$

- So how do we compute orthogonal/orthonormal bases for a space?? We use the GRAM-SCHMIDT ALGORITHM.
- **Input** Starting with a linear independent vectors $a_1, \ldots, a_n$, **Output:** orthonormal vectors $q_1, \ldots, q_n$.

- So how do we compute orthogonal/orthonormal bases for a space?? We use the GRAM-SCHMIDT ALGORITHM.
- Here is the geometric idea:



- **input** Starting with a linear independent vectors $a_1, \ldots, a_n$, **output:** orthogonal vectors $q_1, \ldots, q_n$.
  - Step 1: $q_1 = a_1$
  - Step 2: $q_2 = a_2 - (\frac{a_2^T q_1}{q_1^T q_1})q_1$
  - Step 3: $q_3 = a_3 - (\frac{a_3^T q_1}{q_1^T q_1})q_1 - (\frac{a_3^T q_2}{q_2^T q_2})q_2$
  - Step 4: $q_4 = a_4 - (\frac{a_4^T q_1}{q_1^T q_1})q_1 - (\frac{a_4^T q_2}{q_2^T q_2})q_2 - (\frac{a_4^T q_3}{q_3^T q_3})q_3$

    $\vdots \quad \vdots \quad \vdots \quad \vdots$

  - Step j: $q_4 = a_j - (\frac{a_j^T q_1}{q_1^T q_1})q_1 - (\frac{a_j^T q_2}{q_2^T q_2})q_2 - \ldots (\frac{a_j^T q_{j-1}}{q_{j-1}^T q_{j-1}})q_{j-1}$
- At the end NORMALIZE all vectors if you wish to have unit vectors!! (DIVIDE BY LENGTH).

EXAMPLE

Consider the subspace W spanned by $(1, -2, 0, 1)$, $(-1, 0, 0, -1)$ and $(1, 1, 0, 0)$. Find an orthonormal basis for the space W.
ANSWER:

$$(\frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}}, 0, \frac{1}{6}), (\frac{-1}{\sqrt{3}}, \frac{-1}{\sqrt{3}}, 0, \frac{-1}{\sqrt{3}}), (\frac{1}{\sqrt{2}}, 0, 0, \frac{-1}{\sqrt{2}})$$

- In this way, the original basis vectors $a_1, \ldots, a_n$ can be written in a "triangular" way!
  If $q_1, q_2, \ldots, q_n$ are orthogonal Just think of $r_{ij} = a_j^T q_i$

$$a_1 = r_{11}(q_1 / q_1^T q_1), \tag{1}$$

$$a_2 = r_{12}(q_1 / q_1^T q_1) + r_{22}(q_2 / q_1^T q_1) \tag{2}$$

$$a_3 = r_{13}(q_1 / q_1^T q_1) + r_{23}(q_2 / q_2^T q_2) + r_{33}(q_3 / q_3^T q_3) \tag{3}$$

$$\vdots \tag{4}$$

$$a_n = r_{1n}(q_1 / q_1^T q_1) + r_{2n}(q_2 / q_2^T q_2) + \cdots + r_{nn}(q_n / q_n^T q_n). \tag{5}$$

  Where $r_{ij} = a_j^T q_i$.
- Write this equations in matrix form! we obtain $A = QR$ where $A = (a_1 \quad \ldots \quad a_n)$ and $Q = (q_1 \quad q_2 \ldots q_n)$ and $R = (r_{ij})$.

- **Theorem (QR decomposition)** An $m \times n$ matrix A with independent columns can be factor as $A = QR$ where the columns of Q are orthonormal and R is upper triangular and invertible.
- NOTE: A and Q have the same column space. R is an invertible and upper triangular
- The simplest way to compute the QR decomposition:
  1. Use Gram-Schmidt to get the q_i orthonormal vectors.
  2. Matrix Q has columns $q_1, \ldots, q_n$
  3. The matrix R is filled with the dot products $r_{ij} = a_j^T q_i$.
- **Key Point**: Every matrix has two decompositions LU and QR.
- They are both useful for different reasons!! One is for solving equations, the other good for least-squares.