Lecture 7: QR factorization and least square regression

Francesco Preta

July 2020

Orthogonal projection

Let U be a subspace of $\mathbb{R}^n$. For every $\mathbf{y}$ in $\mathbb{R}^n$ we want to find a vector $\hat{\mathbf{y}}$ in U that minimizes the distance from U. In particular, we have the following:

Definition 1. Let U be a subspace of $\mathbb{R}^n$. For $\mathbf{y} \in \mathbb{R}^n$, we define the orthogonal projection $\text{proj}_U(\mathbf{y})$ as

$$\text{proj}_U(\mathbf{y}) = \text{argmin}_{\hat{\mathbf{y}} \in U} ||\mathbf{y} - \hat{\mathbf{y}}||^2$$

The orthogonal projection has the usual meaning in $\mathbb{R}^2$ and $\mathbb{R}^3$, as shown in the following examples:

Example: let $U = \text{span}(\begin{bmatrix} 1 \\ 1 \end{bmatrix})$ and consider $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$. Minimizing the square distance from U is equivalent to finding the α such that the following function is minimized

$$f(\alpha) = ||\alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 3 \end{bmatrix} ||^2 = (\alpha - 2)^2 + (\alpha - 3)^2$$

some calculus shows us that the minimum for this function is given by $\alpha = \frac{5}{2}$. Therefore

$$\hat{\mathbf{y}} = \begin{bmatrix} \frac{5}{2} \\ \frac{5}{2} \end{bmatrix}$$

Notice that the segment from $\mathbf{y}$ to $\hat{\mathbf{y}}$, is perpendicular to the line representing the subspace U, which justifies the name orthogonal projection.

Example: consider

$$U = \text{span}(\mathbf{e}_1, \mathbf{e}_2) = \left\{ \begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix} \mid \alpha, \beta \in \mathbb{R} \right\}$$

and $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. In order to find the orthogonal projection, we need to minimize the function

$$g(\alpha, \beta) = (1 - \alpha)^2 + (2 - \beta)^2 + 3^2$$

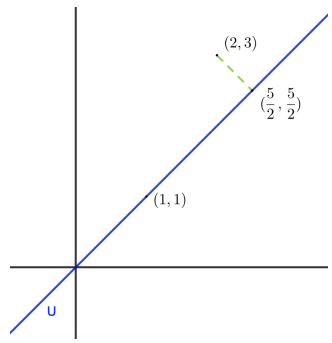

Figure 1: Orthogonal projection of $\mathbf{y} = \begin{bmatrix} 2 & 3 \end{bmatrix}$ in $\mathbb{R}^2$ to the subspace U generated by $\begin{bmatrix} 1 & 1 \end{bmatrix}$

which has a minimum for $\alpha = 1$ and $\beta = 2$. Then

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

again, the segment from $\mathbf{y}$ to $\hat{\mathbf{y}}$, is orthogonal to the subspace U represented by the plane $z = 0$, which justifies the name.

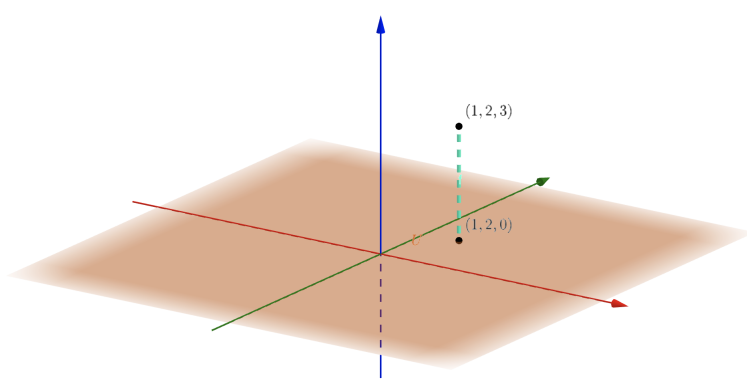

Figure 2: Orthogonal projection of $\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$ in $\mathbb{R}^3$ to the subspace U generated by $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$

In general we have the following:

Theorem 1. *Let U be a subspace of $\mathbb{R}^n$ of dimension k and let $\mathcal{U} = \{\mathbf{u}_i\}_{i=1}^k$ be an orthonormal basis for U. Then for every $\mathbf{y} \in \mathbb{R}^n$,*

$$\operatorname{proj}_U \mathbf{y} = \sum_{i=1}^{k} \langle \mathbf{y}, \mathbf{u}_i \rangle \mathbf{u}_i$$

Example: *consider the set of orthonormal vectors*

$$\mathbf{u}_1 = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ 0 \end{bmatrix} \qquad \mathbf{u}_3 = \begin{bmatrix} -\frac{2}{\sqrt{15}} \\ \frac{1}{\sqrt{15}} \\ -\frac{1}{\sqrt{15}} \\ -\frac{3}{\sqrt{15}} \end{bmatrix}$$

Find the orthogonal projection of the vector $\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ *in the vectors subspace* $U = \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$.

By applying the theorem we get that $\operatorname{proj}_U \mathbf{y}$ *is equal to*

$$\left\langle \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{bmatrix} \right\rangle \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{bmatrix} + \left\langle \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ 0 \end{bmatrix} \right\rangle \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ 0 \end{bmatrix} + \left\langle \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -\frac{2}{\sqrt{15}} \\ \frac{1}{\sqrt{15}} \\ -\frac{1}{\sqrt{15}} \\ -\frac{3}{\sqrt{15}} \end{bmatrix} \right\rangle \begin{bmatrix} -\frac{2}{\sqrt{15}} \\ \frac{1}{\sqrt{15}} \\ -\frac{1}{\sqrt{15}} \\ -\frac{3}{\sqrt{15}} \end{bmatrix}$$

which corresponds to

$$\operatorname{proj}_U \mathbf{y} = \begin{bmatrix} \frac{4}{3} \\ 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}$$

We have seen in the homework that for any given subspace U, we can define a subspace $U^\perp$ called the orthogonal complement of U, where

$$U^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \langle \mathbf{v}, \mathbf{u} \rangle = 0 \ \text{ for every } \mathbf{u} \in U\}$$

The following is an important property of the orthogonal complement:

Theorem 2. *Let U be a subspace of $\mathbb{R}^n$, then for every $\mathbf{y}$ in $\mathbb{R}^n$ there exists a unique decomposition*

$$\mathbf{y} = \operatorname{proj}_U \mathbf{y} + \operatorname{proj}_{U^\perp} \mathbf{y}.$$

In particular, this means that $\operatorname{proj}_{U^\perp} \mathbf{y} = \mathbf{y} - \operatorname{proj}_U \mathbf{y}$. In the previous example

$$\operatorname{proj}_{U^\perp} \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{4}{3} \\ 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} \\ 1 \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}$$

Notice that in this way we can reinterpret the Gram-Schmidt algorithm on a set of independent vectors $\{\mathbf{b}_i\}_{i=1}^k$ in the following way: initialize the algorithm by normalizing the first vector. Then at each step i for $i = 1, ..., k-1$, consider $U_i = \operatorname{span}(\mathbf{u}_1, ..., \mathbf{u}_i) = \operatorname{span}(\mathbf{b}_1, ..., \mathbf{b}_i)$, find $\mathbf{u}'_{i+1}$ as

$$\mathbf{u}'_{i+1} = \mathbf{b}_{i+1} - \sum_{j \le i} \langle \mathbf{b}_{i+1}, \mathbf{u}_j \rangle \mathbf{u}_j = \mathbf{b}_{i+1} - \operatorname{proj}_{U_i} \mathbf{b}_{i+1} = \operatorname{proj}_{U_i^\perp} \mathbf{b}_{i+1}$$

and then normalize it by considering $\mathbf{u}_{i+1} = \frac{\mathbf{u}'_{i+1}}{||\mathbf{u}'_{i+1}||}$.

Finally notice that if $\{\mathbf{u}_i\}_{i=1}^n$ is an orthonormal basis of $\mathbb{R}^n$, we have that for every $\mathbf{y} \in \mathbb{R}^n$, we can write

$$\mathbf{y} = \sum_{i=1}^{n} \langle \mathbf{y}, \mathbf{u}_i \rangle \mathbf{u}_i.$$

In order to prove it, notice that since $\{\mathbf{u}_i\}_{i=1}^n$ is a basis, then there exist coefficients $\{y_i\}_{i=1}^n$ such that $\mathbf{y} = \sum_{i=1}^n y_i \mathbf{u}_i$. But then

$$\langle \mathbf{y}, \mathbf{u}_i \rangle = \langle \sum_{j=1}^{n} y_j \mathbf{u}_j, \mathbf{u}_i \rangle = \sum_{j=1}^{n} y_j \langle \mathbf{u}_j, \mathbf{u}_i \rangle = y_i.$$

so that $y_i = \langle \mathbf{y}, \mathbf{u}_i \rangle$ and so $\mathbf{y} = \sum_{i=1}^n \langle \mathbf{y}, \mathbf{u}_i \rangle \mathbf{u}_i$.

QR decomposition

The Gram-Schmidt algorithm gives rise to a way to decompose any full-rank matrix called **QR decomposition** or **QR factorization**. Consider $\{\mathbf{b}_i\}_{i=1}^k$ the columns of a full rank $n \times k$ matrix B, for $k \leq n$. If $\{\mathbf{u}_i\}_{i=1}^k$ are the orthonormal vectors obtained by applying the Gram-Schmidt algorithm to $\{\mathbf{b}_i\}_{i=1}^k$, then $\text{span}(\{\mathbf{b}_i\}_{i=1}^k) = \text{span}(\{\mathbf{u}_i\}_{i=1}^k)$. So we can write

$$\mathbf{b}_j = \sum_{i=1}^{k} \langle \mathbf{b}_j, \mathbf{u}_i \rangle \mathbf{u}_i.$$

However, $U_j = \text{span}(\mathbf{u}_1, ..., \mathbf{u}_j) = \text{span}(\mathbf{b}_1, ..., \mathbf{b}_j)$ and $\mathbf{u}_l \in U_j^{\perp}$ for $l > j$. Therefore $\langle \mathbf{b}_j, \mathbf{u}_l \rangle = 0$ for $l > j$ and we can write

$$\mathbf{b}_j = \sum_{i=1}^{j} \langle \mathbf{b}_j, \mathbf{u}_i \rangle \mathbf{u}_i$$

Now consider the following matrices:

$$Q = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & ... & \mathbf{u}_k \end{bmatrix}$$
$$R = \begin{bmatrix} \langle \mathbf{b}_1, \mathbf{u}_1 \rangle & \langle \mathbf{b}_2, \mathbf{u}_1 \rangle & ... & \langle \mathbf{b}_k, \mathbf{u}_1 \rangle \\ 0 & \langle \mathbf{b}_2, \mathbf{u}_2 \rangle & ... & \langle \mathbf{b}_k, \mathbf{u}_1 \rangle \\ ... & & & \\ 0 & 0 & ... & \langle \mathbf{b}_k, \mathbf{u}_k \rangle \end{bmatrix}$$

Then the QR decomposition of B is given by $B = QR$. Notice that if B is not a square matrix, then Q is a $n \times k$ matrix and R is a $k \times k$ matrix.

Example: consider the matrix $C = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$. We first apply the Gram-Schmidt algorithm to obtain

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad \mathbf{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \qquad \mathbf{v}_3 = \begin{bmatrix} -\sqrt{\frac{2}{5}} \\ -\frac{1}{\sqrt{10}} \\ \sqrt{\frac{2}{5}} \\ \frac{1}{\sqrt{10}} \end{bmatrix}$$

Then we find the matrix R

$$R = \begin{bmatrix} \sqrt{2} & 0 & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{\frac{5}{2}} \end{bmatrix}$$

and we can check that

$$C = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & -\sqrt{\frac{2}{5}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{10}} \\ 0 & \frac{1}{\sqrt{2}} & \sqrt{\frac{2}{5}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{\frac{5}{2}} \end{bmatrix} = QR$$

Applications of QR decomposition

Solving systems of linear equations

Consider the square system of linear equation $A\mathbf{x} = \mathbf{b}$ and suppose that A is an invertible square matrix. Then we know that the system has a unique solution. Let $A = QR$, then since Q is orthogonal, we have that $QR\mathbf{x} = \mathbf{b}$ is equivalent to solving $R\mathbf{x} = Q^T\mathbf{b}$. However, we know that R is upper triangular, so the system can be solved using back-substitution.

Example: consider the system of linear equations

$$\begin{cases} x_1 + x_2 = 1 \\ -x_1 + x_2 + x_3 = -1 \\ x_1 + x_3 = 1 \end{cases}$$

and let's try to solve it using the QR decomposition. The Gram-Schmidt algorithm gives us

$$\mathbf{u}_1 = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \qquad \mathbf{u}_3 = \begin{bmatrix} -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ \sqrt{\frac{2}{3}} \end{bmatrix}$$

Then the triangular matrix R is given by

$$R = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & 0 & \sqrt{\frac{3}{2}} \end{bmatrix}$$

and we have

$$Q^T\mathbf{b} = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} \sqrt{3} \\ 0 \\ 0 \end{bmatrix}$$

We can write this as a system

$$\begin{cases} \sqrt{3}x_1 = \sqrt{3} \\ \sqrt{2}x_2 + \frac{x_3}{\sqrt{2}} = 0 \\ \sqrt{\frac{3}{2}}x_3 = 0 \end{cases}$$

that can be solved through back-substitution:

$$\begin{cases} x_1 = 1 \\ x_2 = 0 \\ x_3 = 0 \end{cases}$$

Computing the inverse

Let A be an invertible $n \times n$ matrix. We have seen that given a QR decomposition of A, then $A^{-1} = R^{-1}Q^{-1} = R^{-1}Q^T$. Then in order to find the inverse matrix, one can solve a system $RB = Q^T$, or equivalently, for $\mathbf{b}_i$ the i-th column of B and for $\tilde{\mathbf{q}}_i$ the i-th column of Q^T, solve the system

$$R\mathbf{b}_i = \tilde{\mathbf{q}}_i \qquad \text{for } i = 1, ..., n$$

Since R is a triangular matrix, this can be solved by backsubstitution.

Example: compute the inverse of $A = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$.

From the previous exercise we already have the QR decomposition:

$$A = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & 0 & \sqrt{\frac{3}{2}} \end{bmatrix}$$

Finding the inverse B corresponds to finding the solutions to

$$\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & 0 & \sqrt{\frac{3}{2}} \end{bmatrix} \begin{bmatrix} b_{1,1} \\ b_{2,1} \\ b_{3,1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} \end{bmatrix}$$

$$\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & 0 & \sqrt{\frac{3}{2}} \end{bmatrix} \begin{bmatrix} b_{1,2} \\ b_{2,2} \\ b_{3,2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$$

$$\begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & 0 & \sqrt{\frac{3}{2}} \end{bmatrix} \begin{bmatrix} b_{1,3} \\ b_{2,3} \\ b_{3,3} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 0 \\ \sqrt{\frac{2}{3}} \end{bmatrix}$$

This can be easily done by back-substitution to obtain:

$$A^{-1} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Computational complexity

The QR decomposition for a $n \times k$ matrix (for $n \geq k$) requires a number of flops depending on the Gram-Schmidt algorithm. For every vector $\mathbf{b}_i$, for $i = 1, ..., k$ we need to compute i scalar products, i scalar multiplications and $i - 1$ sums, then the normalization takes approximately $2n$ flops, so that

$$tot = \sum_{i=1}^{k} [(2n-1)i + ni + n(i-1) + 2n] = nk + (4n-1)\sum_{i=1}^{k} i =$$
$$nk - k + (4n-1)\frac{k(k+1)}{2}$$

which is of the order of $2nk^2$, or $2n^3$ when $k = n$. The rest of the QR factorization requires calculation that have already been counted, (scalar products between the original vectors and the orthonormal basis), so the total complexity is of the order of $2nk^2$.

Least square linear model

Models of linear regression are widely used in various branches of statistics and data science. In the following section, we will introduce a general linear model of regression starting from a linear algebra application and then we will consider the statistical meaning behind it.

Consider the equation $A\mathbf{x} = \mathbf{b}$, for known $A \in \mathbb{R}^{n \times k}$ and $\mathbf{b} \in \mathbb{R}^n$ and suppose that $\mathbf{b}$ is not in the span of the columns of A, although **the columns of** A

are linearly independent. This means that the system is over-determined ($n > k$) and that it does not admit a solution. It makes sense to find a solution $\mathbf{z}$ which, in some sense, satisfies the conditions with a small margin of error. In other words, we can choose $\mathbf{z}$ which minimizes the sum of squares, that is

$$\mathbf{z} = \text{argmin}_{\mathbf{x}\in\mathbb{R}^k} ||\mathbf{b} - A\mathbf{x}||^2$$

Since we are minimizing the square of the distance between $\mathbf{b}$ and $A\mathbf{x}$, this is called a **least square method** and and $\mathbf{z}$ is called the least square solution of the system.

The problem can be solved analitically using calculus. In fact, let $lsq(\mathbf{x}) = ||\mathbf{b} - A\mathbf{x}||^2$, then

$$\frac{\partial lsq}{\partial x_m} = -2\sum_{i=1}^{n}(\mathbf{b}_i - \sum_{j=1}^{k} A_{i,j}x_j)A_{i,m}$$

which tells us that $\mathbf{z}$ is the solution of a system of linear equations

$$\sum_{j=1}^{k}\sum_{i=1}^{n} A^T_{m,i}A_{i,j}z_j = \sum_{i=1}^{n} A^T_{m,i}\mathbf{b}_i \qquad m = 1, .., n$$

which is a system of the type

$$A^T A\mathbf{z} = A^T\mathbf{b}$$

notice that this is system is obtained by multiplying the previous system on the left by A^T. Moreover, $A^T A$ is a $k \times k$ invertible matrix, since A has rank k. The solution is given by

$$\mathbf{z} = (A^T A)^{-1}A^T\mathbf{b}$$

We have already seen that the matrix $(A^T A)^{-1}A^T$ is called the pseudo-inverse of A and denoted $A^\dagger$. Its existence is guaranteed by the fact that A is a full-rank tall matrix.

Solving least square models using QR decomposition

We know that if A is a full rank $n \times k$ tall matrix, then we can find a QR decomposition $A = QR$, for $Q \in \mathbb{R}^{n\times k}$ and $R \in \mathbb{R}^{k\times k}$. Then the pseudo-inverse is given by

$$A^\dagger = (A^T A)^{-1}A^T = (R^T Q^T QR)^{-1}R^T Q^T = R^{-1}(R^T)^{-1}R^T Q^T = R^{-1}Q^T$$

To find the pseudo-inverse it therefore suffices to solve the system of linear equations $RA^\dagger = Q^T$, that reduces to $R\tilde{\mathbf{a}}_i = \tilde{\mathbf{q}}_i$ for $i = 1, ..., m$. This is particularly convenient since it's purely a matter of back-substitution, as R is a triangular matrix. This method can also be used to find the least square solution to the general linear system $\mathbf{z} = A^\dagger\mathbf{b}$, for which it's sufficient to solve $R\mathbf{z} = Q^T\mathbf{b}$.

Example: consider the system of linear equations:

$$\begin{cases} x_2 - x_3 = 2 \\ x_1 + 2x_3 = 3 \\ x_2 + x_3 = 2 \\ x_1 - 2x_3 = 0 \end{cases}$$

The matrix of the system is $A = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 0 & -2 \end{bmatrix}$. This matrix is full rank, since

$$\begin{vmatrix} 0 & 1 & -1 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \end{vmatrix} = -2 \neq 0$$

However, the augmented matrix has nonzero determinant:

$$\begin{vmatrix} 0 & 1 & -1 & 2 \\ 1 & 0 & 2 & 3 \\ 0 & 1 & 1 & 2 \\ 1 & 0 & -2 & 0 \end{vmatrix} = -\begin{vmatrix} 1 & -1 & 2 \\ 1 & 1 & 2 \\ 0 & -2 & 0 \end{vmatrix} - \begin{vmatrix} 1 & -1 & 2 \\ 0 & 2 & 3 \\ 1 & 1 & 2 \end{vmatrix} = 6 \neq 0$$

so that there is no solution to the system. However, we can compute the pseudo-inverse and the approximating solution through the QR decomposition. In fact we have

$$\mathbf{a}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{a}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{a}_3 = \begin{bmatrix} -1 \\ 2 \\ 1 \\ -2 \end{bmatrix}$$

Gram-Schmidt algorithm gives:

$$\mathbf{u}_1 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \qquad \mathbf{u}_3 = \begin{bmatrix} -\frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} \\ -\frac{2}{\sqrt{10}} \end{bmatrix}$$

so that we get

$$Q = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{2}{\sqrt{10}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{2}{\sqrt{10}} \end{bmatrix} \qquad R = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{10} \end{bmatrix}$$

The approximate solution of the system is therefore given by

$$\begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{10} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{10}} & \frac{2}{\sqrt{10}} & \frac{1}{\sqrt{10}} & -\frac{2}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 2 \\ 0 \end{bmatrix}$$

which gives

$$\begin{cases} x_1 = \frac{3}{2} \\ x_2 = 2 \\ x_3 = \frac{3}{5} \end{cases}$$

This doesn't solve the system $A\mathbf{x} = \mathbf{b}$, but minimizes the distance between $A\mathbf{x}$ and $\mathbf{b}$.

Geometrically, the problem can be treated as finding a linear combination of the columns of A that minimizes the distance from a given point $\mathbf{b}$, or equivalently, finding $\mathbf{z}$ such that $A\mathbf{z} = \text{proj}_V \mathbf{b}$ where $V = \text{span}(\mathbf{a}_1, ..., \mathbf{a}_k)$.

Least square regression

In statistics, least square regression is widely used to understand correlation between variables. In the simplest regression model, two numerical variables are compared to see if one can express correlation between them with a linear function. A priori, a decision is made on which variable is the **independent variable** x and which one is regressed onto the other. Once that decision is made, one needs to find the parameters (β, α) that best describe the relationship

$$y = \beta x + \alpha$$

The data is given by a set of observations in pair $\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^N$ for which one can find β and α that minimize the sum of the squared errors between each observation $\hat{y}_i$ and the predicted quantity $y_i = \beta\hat{x}_i + \alpha$:

$$\min_{\alpha,\beta} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

for this reason this is called **least square regression**.

The solution to this problem has an explicit form that can be obtained through calculus, as we can rewrite

$$lsq(\alpha, \beta) = \sum_{i=1}^{N} (\hat{y}_i - \beta\hat{x}_i - \alpha)^2$$

and optimize lsq in α and β. We get:

$$\frac{\partial lsq}{\partial \alpha} = -2 \sum_{i=1}^{N} (\hat{y}_i - \beta\hat{x}_i - \alpha)$$

$$\frac{\partial lsq}{\partial \beta} = -2 \sum_{i=1}^{N} (\hat{y}_i - \beta\hat{x}_i - \alpha)\hat{x}_i$$

We know we can find the optima by taking $\nabla lsq = \mathbf{0}$, so that

$$\begin{cases} \alpha = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i - \frac{\beta}{N}\sum_{i=1}^{N}\hat{x}_i \\ \sum_{i=1}^{N}(\hat{y}_i - \beta\hat{x}_i - \frac{1}{N}\sum_{j=1}^{N}\hat{y}_j + \frac{\beta}{N}\sum_{j=1}^{N}\hat{x}_j)\hat{x}_i = 0 \end{cases}$$

which gives, for the averages $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$

$$\begin{cases} \alpha = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i - \frac{\beta}{N}\sum_{i=1}^{N}\hat{x}_i \\ \sum_{i=1}^{N}\hat{y}_i\hat{x}_i - \beta\sum_{i=1}^{N}\hat{x}_i^2 - N\bar{x}\bar{y} + \beta N\bar{x}^2 = 0 \end{cases}$$

We call the quantity $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})(\hat{x}_i - \bar{x})$ the *covariance* between x and y, denoted by $COV(x, y)$ and we have

$$COV(x, y) = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i\hat{x}_i - \bar{x}\bar{y}$$

Moreover, we call $COV(x, x) = Var(x)$ the *variance* of x and we have

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_i^2 - \bar{x}^2$$

Then the solutions are given by

$$\beta = \frac{COV(x, y)}{Var(x)}$$
$$\alpha = \bar{y} - \beta\bar{x}$$

The geometric meaning of the linear regression is the following: we try to find the line $y = \beta x + \alpha$ that best approximates the data points $\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{N}$ in $\mathbb{R}^2$, like in Figure 3

In some situation, the regression line is successful in capturing the relation between two variables. However, this is not always the case, since the relation can be captured by a non-linear model.

A good way to measure whether the regression model successfully captures the relation between the variables is given by the *coefficient of determination* R^2. In fact, this is defined in the following way

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}$$

The denominator in the formula constitutes the variance of y, which explains the general variation from the average value, while the numerator is given by the sum of squares. The R^2 is meant to explain how much of the total variation is captured by the model. A low sum of squares with respect to the total variation, will give a high value of R^2, while a high such sum will give a low value for R^2, thus showing that the model fails to capture most of the variation.

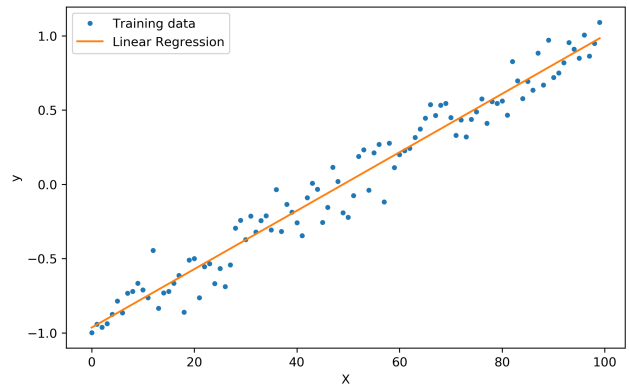

Figure 3: Linear regression on a 2-dimensional dataset

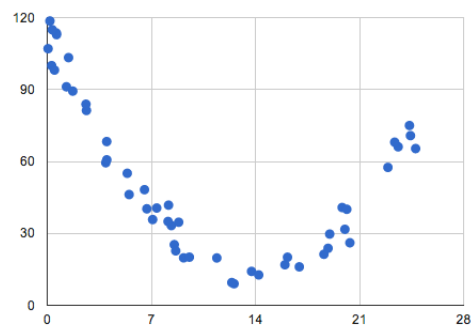

Figure 4: 2-dimensional dataset with quadratic relationship

When it comes to relating the linear regression model to our general linear least square method, notice that we can choose $\mathbf{b} = \mathbf{y}$ as a N-dimensional vector such that $\mathbf{y}_i = \hat{y}_i$ and A as a $N \times 2$ matrix having $A_{i,1} = \hat{x}_i$ and $A_{i,2} = 1$. Then finding α, β corresponds to finding the orthogonal projection of $\mathbf{y} = \mathbf{b}$ in $\text{span}(\mathbf{x}, \mathbf{1})$.

More generally, we can consider linear methods in which the dependent variable y depends on k features $(x^1, ..., x^k)$, for which we have a set of N observations $\{(\hat{x}_i^1, ..., \hat{x}_i^k, \hat{y}_i)\}_{i=1}^N$. Then the least square regression finds the $k+1$ coefficients $(\beta_i, ..., \beta_k, \alpha)$ that minimize the sum of the distances between the observation $\hat{y}_i$ and the predicted value $y_i = \sum_{i=1}^k \beta_k x_i^k + \alpha$, that is

$$(\beta_1, ..., \beta_k, \alpha) = \text{argmin}_{\beta,\alpha} \sum_{i=1}^N ||\hat{y}_i - y_i||^2$$

This can also be interpreted as a particular instance of our general linear model, in the case in which we are trying to find the linear combinations of the columns

of a $N \times (k+1)$ matrix A, with $A_{i,j} = \hat{x}_i^j$ for $j = 1, ..., k$ and $A_{i,k+1} = 1$ for $i = 1, ..., N$. Then the above problem corresponds to finding

$$\text{argmin}_{\mathbf{z} \in \mathbb{R}^{k+1}} ||A\mathbf{z} - \mathbf{y}||.$$