23.1 Introduction: Least Squares

In our previous exploration of linear algebra, we developed techniques such as Gaussian Elimination and matrix inversion to solve systems of linear equations exactly. Gaussian Elimination and inversion work best when we have perfect measurements of our system, i.e. when there are no errors or noise in the system. However, given a system which might be prone to noisy measurements, Gaussian Elimination might not be able to solve it; since the equations might end up being inconsistent and have no solution.

For example, in the GPS problem, we use cross-correlation to measure the distance between the satellites (beacons) and the receiver to generate the distances to feed into the trilateration algorithm. However, interference in the wireless signals or errors in correlation can easily lead to noisy measurements of the distance between receiver and transmitter.

We'd like to develop a method to solve a system of linear equations even when the system might be inconsistent due to noise. In particular, our goal is to understand how collecting more equations than unknowns can help us reduce the impact of noise and find a solution to the system of equations that might not be perfect, but is as close as possible.

This note develops the *Least Squares* technique for approximately solving systems of linear equations in the presence of noise. Our focus will be on *overdetermined* systems of equations (more equations than unknowns), so we can use the extra equations to reduce the impact of noise.

Least squares is the fundamental idea behind *data fitting* and *machine learning*: In data fitting, we find lines or curves that best match the data. In machine learning, we use a best-fit curve to make predictions about new, unseen data.

23.2 Approximate Solutions to Linear Systems

So far, we have spent a great deal of time studying problems that ultimately reduce to linear equations of the form

$$A\vec{x} = \vec{b}.$$

In other words, we are given a system of linear equations, and asked to assign values to variables that satisfy *all* of the equations *exactly*.

However, it is often the case that the equations we receive may be corrupted slightly by noise. When we have a large number of equations, the result is an *inconsistent* system, where no $\vec{x}$ exists that satisfies all the equations exactly. This sort of scenario is where techniques such as least squares come into play.

Fundamentally, least squares deals with the following problem: choosing an $\vec{x}$ that minimizes the magnitude of the *error*

$$\vec{e} = \vec{b} - A\vec{x}$$

in our solution.

Let's write out the columns of A explicitly and rearrange, to obtain

$$\begin{bmatrix} | & | & & | \\ \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} + \vec{e} = \vec{b}.$$

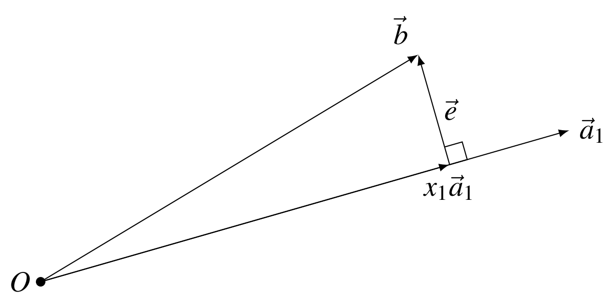
Expanding out the matrix multiplication, we obtain

$$(x_1\vec{a}_1 + x_2\vec{a}_2 + \ldots + x_m\vec{a}_m) + \vec{e} = \vec{b},$$

with A an $n \times m$ matrix. In other words, we are trying to find the "closest" vector to $\vec{b}$ in the span of the columns $\{\vec{a}_i\}$ of A.

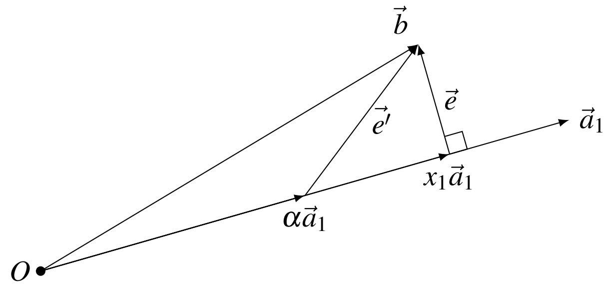
23.2.1 The 2D Special Case

How can we find such a vector? To gain some intuition, let's simplify the problem slightly, working in $n = 2$ dimensional space, where A has $m = 1$ columns. Then we want to find a vector $x_1\vec{a}_1$ that minimizes the error $\|\vec{e}\| = \left\|\vec{b} - x_1\vec{a}_1\right\|$, as shown:



Intuitively, it seems clear that we should construct $x_1\vec{a}_1$ by dropping a perpendicular from $\vec{b}$ onto $\vec{a}_1$ (known as the *projection* of $\vec{b}$ onto $\vec{a}_1$), with the error $\vec{e}$ therefore being orthogonal to $\vec{a}_1$. Let's see if we can prove this claim.

Consider another vector $\alpha\vec{a}_1$ that lies in the span of $\vec{a}_1$. Consider the error between this vector and $\vec{b}$: $\vec{e'} = \vec{b} - \alpha\vec{a}_1$. Plotting this vector on the diagram, we see that it forms the hypotenuse of a right triangle, where one leg is our original $\vec{e}$:



Thus, by the Pythagorean Theorem, $\left\|\vec{e'}\right\| > \|\vec{e}\|$, so $\alpha\vec{a}_1$ is not as close to $\vec{b}$ as our original $x_1\vec{a}_1$ was.

So we've shown that the projection of $\vec{b}$ onto $\vec{a}_1$ is the solution that minimizes our error $\vec{e}$. But what is this projection? Well, we know that $\vec{e} \perp x_1\vec{a}_1$. Thus,

$$\begin{aligned}
& \langle \vec{e}, x_1\vec{a}_1 \rangle = 0 \\
\implies \quad & x_1 \langle \vec{e}, \vec{a}_1 \rangle = 0 \\
\implies \quad & \left\langle \vec{b} - x_1\vec{a}_1, \vec{a}_1 \right\rangle = 0 \\
\implies \quad & \left\langle \vec{b}, \vec{a}_1 \right\rangle - x_1 \langle \vec{a}_1, \vec{a}_1 \rangle = 0 \\
\implies \quad & x_1 = \frac{\left\langle \vec{b}, \vec{a}_1 \right\rangle}{\langle \vec{a}_1, \vec{a}_1 \rangle}.
\end{aligned}$$

This is known as the *projection* of $\vec{b}$ onto the span of $\vec{a}_1$. It is also sometimes called an *orthogonal projection*. Essentially, what we derived above is the Least Squares algorithm in two dimensions.

23.2.2 The General Case

Now, how do we generalize beyond two dimensions? Well, we argued in 2D that, in order to minimize the magnitude of the error, we should choose x_1 such that $\vec{e} \perp x_1$ - in words, that $\vec{e}$ is orthogonal to the subspace in which $x_1\vec{a}_1$ must lie. The analogous claim in higher dimensions, given that we wish to choose $\vec{x}$ to minimize the magnitude of the error

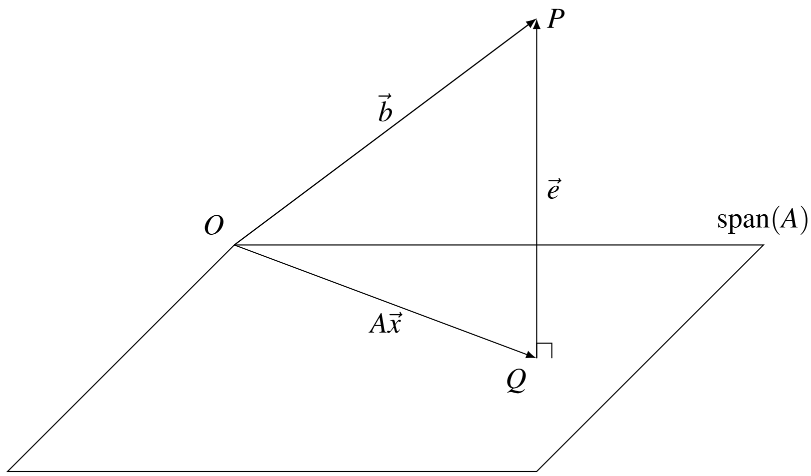
$$\vec{e} = \vec{b} - A\vec{x},$$

would be to choose an $\vec{x}$ that is orthogonal to the span of all possible $A\vec{x}$ - in other words, a $\vec{e}$ that is orthogonal to the column space of A. Can we justify this claim?

23.2.2.1 Geometric Proof

We can generalize what we did earlier in the $2D$ case to higher dimensions to derive the Least Squares algorithm more generally. In this proof, we will also see a geometric motivation for the algebraic manipulations done in the previous proof.

What are all the possible values that $A\vec{x}$ can take? $A\vec{x}$ is just the set of all linear combinations of the columns of A, i.e. it is just the columnspace of A. This is just a higher-dimensional plane in n-dimensional space that passes through the origin. To help us visualize this, we will illustrate the span of the columns of A as a 2-dimensional plane embedded in 3-dimensional space:

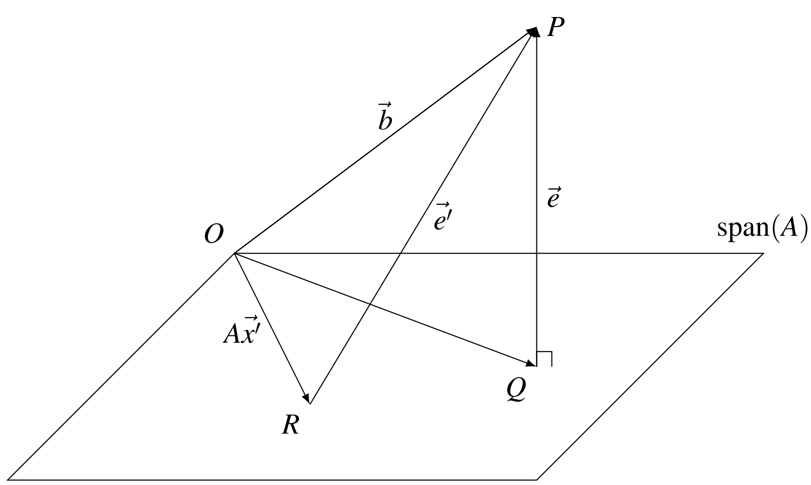


The $\vec{b}$ does not necessarily lie in the column space of A. In this case, doing Gaussian Elimination on $A\vec{x} = \vec{b}$ would lead to an inconsistent set of equations that cannot be solved.

As a result, our goal is to find that vector $A\vec{x}$ in span(A) that is closest to $\vec{b}$, i.e. we want to minimize the magnitude of the error $\|\vec{e}\| = \left\|\vec{b} - A\vec{x}\right\|$.

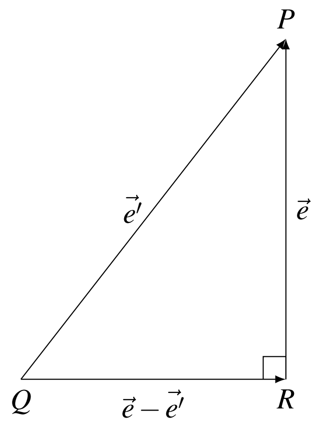
We drop a perpendicular from the tip of the vector $\vec{b}$ (denoted by P) to the plane span(A). Let this intersect the columnspace of A at point Q. Consider the vector $A\vec{x}$ that connects the origin O to point Q.

As before, our goal is to show that this choice of $\vec{x}$ minimizes the magnitude of the error $\|\vec{e}\|$. To do so, by analogy with our 2D geometric proof, we will consider any other $\vec{x'}$, where $A\vec{x'}$ is the point R on the plane span(A). Adding it to our figure, where $\vec{e'} = \vec{b} - A\vec{x'}$, we obtain



We'd like to show that $\left\|\vec{e'}\right\| \geq \|\vec{e}\|$. In the 2D case, we were able to do so by drawing a right triangle where $\vec{e}$ was a leg and $\vec{e'}$ the hypotenuse. Can we do something similar here?

From the figure, we see that the triangle formed by P, Q, and R is a triangle with the desired side lengths, as illustrated below:



But is it a *right* triangle? Observe that the bottom leg of the triangle is the vector $\vec{e}-\vec{e}' = A(\vec{x}-\vec{x'})$, which clearly lies in the column space of A. Thus, by construction, it is orthogonal to $\vec{e}$, so this is indeed a right triangle!

Therefore, we can simply apply the Pythagorean theorem to conclude that $\left\|\vec{e'}\right\| \geq \|\vec{e}\|$, as desired. So as we have shown that no $\vec{x'}$ can produce a smaller error than $\vec{x}$, we have proven that $\vec{x}$ is the optimal solution!

23.2.2.2 Algebraic Proof

We can also show this using an algebraic proof. Assume that such an $\vec{x}$ exists such that $\vec{e} = \vec{b} - A\vec{x}$ is orthogonal to all elements in the column space of A. Now, imagine for the sake of *contradiction* that there existed an alternative $\vec{x'}$ with corresponding error $\vec{e'} = \vec{b} - A\vec{x}$, such that

$$\left\|\vec{e'}\right\| < \|\vec{e}\|.$$

In other words, we are imagining that there exists a strictly *better* solution to $A\vec{x} \approx \vec{b}$ than $\vec{x}$. From the equation relating the norms of the errors, we have that

$$\begin{aligned} & \left\|\vec{e'}\right\| < \|\vec{e}\| \\ \Longrightarrow\quad & \left\|\vec{e'}\right\|^2 < \|\vec{e}\|^2 \\ \Longrightarrow\quad & \left\langle \vec{e'}, \vec{e'} \right\rangle < \langle \vec{e}, \vec{e} \rangle \end{aligned}$$

So far, all we've done is make simple manipulations to our inequality, squaring to write the norm of a vector in terms of inner products. Now, we will use a small trick. We need to use the fact that $\vec{e}$ is orthogonal to any element in the span of A *somewhere* in our proof. In other words, we want an expression of the form $\langle \vec{e}, A\vec{v} \rangle$ (for some vector $\vec{v}$) to appear somewhere, so that we can set it to 0 and apply this fact.

Where could such an expression $A\vec{v}$ come from? One way is to recognize that

$$\begin{aligned}\vec{e'} &= \vec{b} - A\vec{x'} \\ &= \vec{b} - A\vec{x} + A\vec{x} - A\vec{x'} \\ &= \vec{e} + A(\vec{x} - \vec{x'}),\end{aligned}$$

adding and subtracting the same term $A\vec{x}$ in order to bring $\vec{e}$ back into the equation. Making this substitution, we find that

$$\left\langle \vec{e} + A(\vec{x} - \vec{x'}), \vec{e} + A(\vec{x} - \vec{x'}) \right\rangle < \langle \vec{e}, \vec{e} \rangle .$$

Now, we will apply the distributive property of inner products on the left-hand-side of the inequality, to obtain

$$\langle \vec{e}, \vec{e} \rangle + \left\langle \vec{e}, A(\vec{x} - \vec{x'}) \right\rangle + \left\langle A(\vec{x} - \vec{x'}), \vec{e} \right\rangle + \left\langle A(\vec{x} - \vec{x'}), A(\vec{x} - \vec{x'}) \right\rangle < \langle \vec{e}, \vec{e} \rangle .$$

Recall that inner products of real vectors are commutative, so

$$\left\langle \vec{e}, A(\vec{x} - \vec{x'}) \right\rangle = \left\langle A(\vec{x} - \vec{x'}), \vec{e} \right\rangle .$$

Using the above identity to combine terms, and cancelling equal terms on both sides of the inequality, we find that

$$2\left\langle A(\vec{x} - \vec{x'}), \vec{e} \right\rangle + \left\langle A(\vec{x} - \vec{x'}), A(\vec{x} - \vec{x'}) \right\rangle < 0.$$

And now we can use our orthogonality assumption! Since $A(\vec{x} - \vec{x'})$ lies in the span of A, we know that it is orthogonal to $\vec{e}$, so

$$\left\langle A(\vec{x} - \vec{x'}), \vec{e} \right\rangle = 0.$$

Substituting, we find that

$$\left\langle A(\vec{x} - \vec{x'}), A(\vec{x} - \vec{x'}) \right\rangle < 0.$$

Let's examine this inequality. On the left-hand-side, we take the inner product of some vector with itself, computing its squared norm. We know that norms are always nonnegative, so the left-hand-side is greater than or equal to 0. Yet we write that it is strictly less than 0, so we've reached a contradiction!

Thus, our original assumption that an $\vec{x'}$ existed that was better than $\vec{x}$ must have been false. So we have shown *if* an $\vec{x}$ exists such that $\vec{b} - A\vec{x}$ is orthogonal to the span of A, that it must be the optimal solution to our least squares problem!

23.2.3 Least Squares

We know now that, *if* there exists an $\vec{x}$ such that $\vec{e} = \vec{b} - A\vec{x}$ is orthogonal to *every vector* in the column space of A, that such a vector would be the optimum solution to our least squares problem and minimizes $\|\vec{e}\| = \left\|\vec{b} - A\vec{x}\right\|$.

Theorem 23.1: A vector $\vec{e}$ is orthogonal to every vector in the column space of A if and only if it is orthogonal to each of the columns $\vec{a}_i$ that form the basis of its column space.

Proof. Observe that if $\vec{e}$ is orthogonal to every vector in the column space of A, then it is orthogonal to each of the $\vec{a}_i$, as each of the $\vec{a}_i$ are in the column space of A.

Now, we will try to prove the converse. Consider an arbitrary vector $\vec{v} \in \text{span}(A)$. By definition, we know that there exist coefficients α_i such that we can express

$$\vec{v} = \alpha_1\vec{a}_1 + \alpha_2\vec{a}_2 + \ldots + \alpha_m\vec{a}_m.$$

Now, consider our vector $\vec{e}$, that is orthogonal to each of the $\vec{a}_i$. In other words, we know that for any valid i,

$$\langle \vec{e}, \vec{a}_i \rangle = 0.$$

We wish to show that $\vec{e}$ is orthogonal to $\vec{v}$ as well. To show this, it is natural to try evaluating $\langle \vec{e}, \vec{v} \rangle$, in the hope that this turns out to be 0. Doing so, substituting in our various definitions, and applying the distributive property of inner products, we see that

$$\begin{aligned}
\langle \vec{e}, \vec{v} \rangle &= \langle \vec{e}, \alpha_1\vec{a}_1 + \alpha_2\vec{a}_2 + \ldots + \alpha_m\vec{a}_m \rangle \\
&= \langle \vec{e}, \alpha_1\vec{a}_1 \rangle + \langle \vec{e}, \alpha_2\vec{a}_2 \rangle + \ldots + \langle \vec{e}, \alpha_m\vec{a}_m \rangle \\
&= \alpha_1 \langle \vec{e}, \vec{a}_1 \rangle + \alpha_2 \langle \vec{e}, \vec{a}_2 \rangle + \ldots + \alpha_m \langle \vec{e}, \vec{a}_m \rangle \\
&= \alpha_1 \cdot 0 + \alpha_2 \cdot 0 + \ldots + \alpha_m \cdot 0 \\
&= 0,
\end{aligned}$$

as desired! In other words, if our $\vec{e}$ is orthogonal to all the basis vectors of a subspace, it is orthogonal to *every vector* in that subspace as well! □

We know now that, for our error vector $\vec{e}$ to be orthogonal to every vector in the column space of A, it suffices to say that it is orthogonal to each of the columns $\vec{a}_i$. Now, recall that the inner product of two vectors can be represented as a matrix multiplication, so

$$\langle \vec{e}, \vec{a}_i \rangle = 0 \iff \vec{a}_i^T \vec{e} = 0.$$

Since we have $\vec{a}_i^T \vec{e} = 0$ for all i, we can stack these scalar equations to form the single matrix equation

$$\begin{bmatrix} - & \vec{a}_1^T & - \\ - & \vec{a}_2^T & - \\ & \vdots & \\ - & \vec{a}_1^T & - \end{bmatrix} \vec{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

But we have,

$$\begin{bmatrix} - & \vec{a}_1^T & - \\ - & \vec{a}_2^T & - \\ & \vdots & \\ - & \vec{a}_1^T & - \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \vec{a}_1 & \vec{a}_2 & \cdots & \vec{a}_m \\ | & | & & | \end{bmatrix}^T = A^T,$$

so we can simplify our equation to just

$$A^T \vec{e} = \vec{0}.$$

What do we do with this equation? We're interested in solving for $\vec{x}$, so it makes sense to substitute for

$\vec{e} = \vec{b} - A\vec{x}$, to obtain

$$\begin{aligned} A^T(\vec{b} - A\vec{x}) &= \vec{0} \\ \implies \qquad A^T A\vec{x} &= A^T\vec{b}. \end{aligned}$$

Now, observe that (unlike A) A^TA is a square $m \times m$ matrix! Is it invertible? We will show below that it is, in the case where the columns of A are linearly independent. In this case A^TA is a square matrix with only a trivial nullspace, we know that it has a unique inverse $(A^TA)^{-1}$. Pre-multiplying our earlier equation by this inverse (which we know exists), we obtain

$$\begin{aligned} A^TA\vec{x} &= A^T\vec{b} \\ \implies \quad (A^TA)^{-1}A^TA\vec{x} &= (A^TA)^{-1}A^T\vec{b} \\ \implies \quad \vec{x} &= (A^TA)^{-1}A^T\vec{b}. \end{aligned}$$

This is the general least squares solution for $A\vec{x} \approx \vec{b}$ when A has independent columns, and is our final result!

23.3 Is A^TA invertible?

Let us figure out when exactly we can invert A^TA. First, we will prove a helper theorem.

Theorem 23.2: Null(A^TA) = Null(A), even when A has a nontrivial nullspace.

Proof. To see this, consider an arbitrary $\vec{v} \in \text{Null}(A^TA)$. By definition, we have that

$$A^TA\vec{v} = \vec{0}.$$

Now, we will make the "magic" step[1] of pre-multiplying by $\vec{v}^T$, to obtain

$$\vec{v}^TA^TA\vec{v} = \vec{v}^T\vec{0} = 0.$$

Rearranging the left-hand-side of the above identity, we find that

$$(A\vec{v})^T(A\vec{v}) = 0 \implies \langle A\vec{v}, A\vec{v}\rangle = \|A\vec{v}\|^2 = 0.$$

Thus, it is clear that $A\vec{v} = 0$, so $\vec{v} \in \text{Null}(A)$. Thus, we have shown that $\text{Null}(A^TA) \subseteq \text{Null}(A)$. Now, consider an arbitrary vector $\vec{v'} \in \text{Null}(A)$. Pre-multiplying by A^T, we have that

$$\begin{aligned} A\vec{v'} &= \vec{0} \\ \implies \quad A^TA\vec{v'} &= \vec{0}, \end{aligned}$$

so $\vec{v'} \in \text{Null}(A^TA)$.

[1] As an aside, make sure to understand the intuition behind our "magic" step - we wanted to write the left-hand-side as an inner product, to obtain an equation involving the norm of $A\vec{v}$. This kind of pre-multiplication in order to get an inner product is fairly common in these sorts of proofs, and is a good thing to have in your "toolbox" of algebraic tricks.

Consequently, $\text{Null}(A) \subseteq \text{Null}(A^T A)$. Combining this with the earlier result, we have that

$$\text{Null}(A) = \text{Null}(A^T A),$$

as desired! $\square$

We can now return to our least squares argument in the special case where A has linearly independent columns. Since all the $\vec{a}_i$ are independent, it is a consequence that A has only a trivial null space. From our helper theorem, we can now see that $A^T A$ also only has a trivial nullspace!

Now, observe that (unlike A) $A^T A$ is a square $m \times m$ matrix! Since it is a square matrix with only a trivial nullspace, we know that it has a unique inverse $(A^T A)^{-1}$. Pre-multiplying our earlier equation by this inverse (which we know exists), we obtain

$$\begin{aligned} A^T A\vec{x} &= A^T\vec{b} \\ \implies \quad (A^T A)^{-1} A^T A\vec{x} &= (A^T A)^{-1} A^T\vec{b} \\ \implies \quad \vec{x} &= (A^T A)^{-1} A^T\vec{b}. \end{aligned}$$

This is the general least squares solution for $A\vec{x} \approx \vec{b}$ when A has independent columns, and is our final result!

23.4 Application of Least Squares

It turns out Gauss used this technique to predict where certain planets would be in their orbit. A scientist named Piazzi made 19 observations over the period of a month in regards to the orbit of Ceres (can be viewed as equations). Gauss used some of these observations. He also knew the general shape of the orbit of planets due to Kepler's laws of planetary motion. Gauss set up equations like so:

$$\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y = \phi$$

If one divides the whole equation by ϕ, nothing significant happens so we can ignore the denominator and treat the right side of the equation as 1.

$$\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y = 1$$

We can set up a matrix like so:

$$\begin{bmatrix} x_1^2 & x_1 y_1 & y_1^2 & x_1 & y_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^2 & \cdots & \cdots & \cdots & y_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

Here, the x's and y's are known: they are the coordinates of the measured positions of Ceres. The unknowns are $\alpha, \ldots, \varepsilon$. We write the above equation with matrix/vector notation as follows

$$A\vec{v} = \vec{b}$$

where define

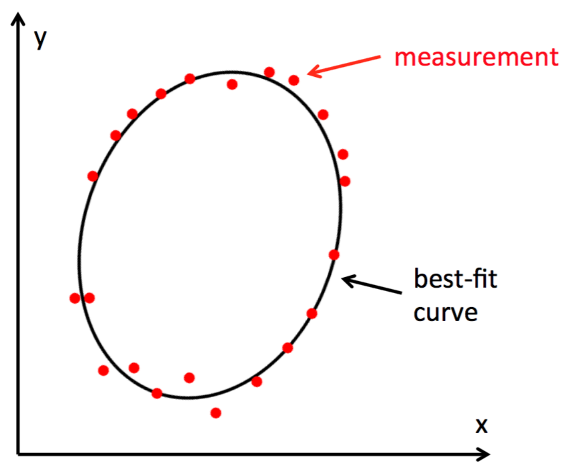
$$A = \begin{bmatrix} x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n^2 & \cdots & \cdots & \cdots & y_n \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

Now we can use the least squares formula to estimate the unknown coefficients in $\vec{v}$. From Eq. (**??**) derived earlier:

$$\vec{v} = (A^T A)^{-1} A^T \vec{b}$$

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{bmatrix} = (A^T A)^{-1} A^T \vec{b}$$

Once solved, we can now translate the coefficients, $\alpha, \ldots, \varepsilon$, back into an ellipse using the original equation $\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \varepsilon y = 1$. A possible result might look like this:



As we can see, the least squares method is useful for fitting noisy or approximate measurements to a curve, provided that we know the general shape of the curve. In addition, this example shows least squares is not limited to lines.

23.5 Practice Problems

These practice problems are also available in an interactive form on the course website.

1. True or False: Least squares is a method for solving an underdetermined system of linear equations.

2. Find the least squares solution to $\begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ 0 \\ -8 \end{bmatrix}$.

3. True or False: Least squares always has a unique solution given by $\vec{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\vec{b}$.

4. True or False: We can use the least squares method to perform regression with a sine function, solving for $y = a \cdot \sin(bx)$. Make sure to also understand the reasoning why.

5. Let's say that we have a scenario where we have a set of data which includes information about a given set of patients. We have their height, weight, age, and white blood cell count. We are trying to create a predictor function for white blood cell count by solving an equation of the form $\mathbf{A}\vec{x} = \vec{b}$. What information should be in the $\mathbf{A}$ matrix?

   (a) The white blood cell counts.

   (b) The height, weight, age, and white blood cell count of each patient.

   (c) The unknown parameters $\alpha_1, \alpha_2, \alpha_3$.

   (d) The height, weight, and age for each patient.

6. True or False: Let $\vec{x} = \text{proj}_{\text{Col}(\mathbf{A})}\vec{b}$ be the projection of $\vec{b}$ onto the column space of a matrix $\mathbf{A}$. Then, $\mathbf{A}^T(\vec{b} - \vec{x}) = \vec{0}$.

7. True or False: The projection of a vector $\vec{b}$ onto a set of vectors $\{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k\}$ is equal to $\frac{\langle \vec{b}, \vec{a}_1 \rangle}{\langle \vec{a}_1, \vec{a}_1 \rangle}\vec{a}_1 + \frac{\langle \vec{b}, \vec{a}_2 \rangle}{\langle \vec{a}_2, \vec{a}_2 \rangle}\vec{a}_2 + \cdots + \frac{\langle \vec{b}, \vec{a}_k \rangle}{\langle \vec{a}_k, \vec{a}_k \rangle}\vec{a}_k$.

8. True or False: Given an arbitrary cost function, the error vector corresponding to the best approximation of a vector $\vec{b}$ to the column space of $\mathbf{A}$ is always orthogonal to $\text{Col}(\mathbf{A})$.

9. Find the best approximation $\hat{x}$ to the system of equations $\begin{cases} a_1 x = b_1 \\ a_2 x = b_2 \end{cases}$ given the cost function $\text{cost}(x) = 2(b_1 - a_1\hat{x})^2 + (b_2 - a_2\hat{x})^2$.

   (a) $\frac{a_1 b_1 + a_2 b_2}{a_1^2 + a_2^2}$

   (b) $\frac{2a_1 b_1 + a_2 b_2}{a_1^2 + a_2^2}$

   (c) $\frac{2a_1 b_1 + a_2 b_2}{2a_1^2 + a_2^2}$

   (d) $\frac{a_1 b_1 + 2a_2 b_2}{a_1^2 + 2a_2^2}$