Least-Squares Approximation and Bilinear Forms

The Normal Equations:
Suppose a column vector \( \mathbf{g} \) is given in an Euclidean space into which a given matrix \( \mathbf{F} \) maps another real space of column vectors \( \mathbf{u} \). In the Euclidean space, length \( ||\mathbf{g}|| := \sqrt{\mathbf{g}^T \mathbf{g}} \). Usually matrix \( \mathbf{F} \) is rectangular with more rows than columns. Our task is to choose \( \mathbf{u} \) to minimize \( ||\mathbf{Fu} - \mathbf{g}|| \), which will then be the distance from \( \mathbf{g} \) to \( \text{Range}(\mathbf{F}) \). This task is called “Least-Squares” because the \( \mathbf{u} \) we seek would minimize the sum of squares of the elements of \( \mathbf{Fu} - \mathbf{g} \).

Differentiating the sum of squares produces \( d ((\mathbf{Fu} - \mathbf{g})^T(\mathbf{Fu} - \mathbf{g})) = 2(\mathbf{Fu} - \mathbf{g})^T\mathbf{F} d\mathbf{u} \), which vanishes for all (infinitesimal) perturbations \( d\mathbf{u} \) if and only if \( (\mathbf{Fu} - \mathbf{g})^T\mathbf{F} = 0^T \). (Do you see why?) Transposed, this becomes the “Normal Equations” of the Least-Squares Problem,

\[
(\mathbf{F}^T\mathbf{F})\mathbf{u} = \mathbf{F}^T\mathbf{g}.
\]

\( \mathbf{u} \) must satisfy them to minimize \( ||\mathbf{Fu} - \mathbf{g}|| \) and so best approximate \( \mathbf{g} \) by a vector in \( \text{Range}(\mathbf{F}) \).

Do the Normal Equations have at least one solution \( \mathbf{u} = \hat{\mathbf{u}} \)?

If so, is \( ||\mathbf{F}\hat{\mathbf{u}} - \mathbf{g}|| \leq ||\mathbf{Fu} - \mathbf{g}|| \) for all \( \mathbf{u} \)? i.e., does \( \hat{\mathbf{u}} \) minimize rather than maximize?
These questions among others are addressed in this note.

At first sight one might think the Normal Equations’ solution should be \( \hat{\mathbf{u}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{g} \). But this formula fails if the columns of \( \mathbf{F} \) are linearly dependent. To see why, observe that

\[
\mathbf{Fz} = 0 \quad \iff \quad (\mathbf{Fz})^T(\mathbf{Fz}) = 0 \quad \iff \quad (\mathbf{F}^T\mathbf{F})\mathbf{z} = 0,
\]

so \( \mathbf{F}^T\mathbf{F} \) is invertible (nonsingular) if and only if the columns of \( \mathbf{F} \) are linearly independent. This hasn’t been assumed; in fact matrix \( \mathbf{F} \) could be rectangular with more columns than rows.

**Exercise:** Show that, if the rows of \( \mathbf{F} \) are linearly independent, a solution \( \hat{\mathbf{u}} = \mathbf{F}^T(\mathbf{F}\mathbf{F}^T)^{-1}\mathbf{g} \) and, if not the only solution of the Normal Equations, it is the only solution that minimizes \( \hat{\mathbf{u}}^T\hat{\mathbf{u}} \) too.

In short, if neither the rows nor the columns of \( \mathbf{F} \) are linearly independent, neither \( \mathbf{F}\mathbf{F}^T \) nor \( \mathbf{F}^T\mathbf{F} \) need be invertible, and then the existence of a minimizing solution \( \mathbf{u} = \hat{\mathbf{u}} \) is in question.

**Warning:** Even when an indicated inverse exists, neither formula \( \hat{\mathbf{u}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{g} \) nor \( \hat{\mathbf{u}} = \mathbf{F}^T(\mathbf{F}\mathbf{F}^T)^{-1}\mathbf{g} \) should be used with numerical data unless the computer’s arithmetic carries at least twice as many sig. digits as are trusted in the data \([\mathbf{F}, \mathbf{g}]\) or desired in the result \( \hat{\mathbf{u}} \). Otherwise roundoff will degrade the result \( \hat{\mathbf{u}} \) too badly whenever \( \mathbf{F} \) is too near a matrix of lower rank. The reason behind this warning will become clear after *Singular Values* have been discussed. If the arithmetic carries barely more sig. digits than are trusted in the data or desired in the result, it should be computed by means of a *QR factorization*, which will also be discussed later. Matlab uses such a factorization to compute \( \hat{\mathbf{u}} \), which Matlab calls “\( \mathbf{F}\backslash\mathbf{g} \)”, whenever \( \mathbf{F} \) is not square. Least-Squares is built into Matlab.

Existence and Uniqueness of a Minimizing Solution \( \hat{\mathbf{u}} \):
We shall use Fredholm’s Alternatives (*q.v.*) to deduce that the Normal Equations always have at least one solution \( \hat{\mathbf{u}} \), and to determine when it is unique. At least one solution exists if and only if \( \mathbf{w}^T(\mathbf{F}^T\mathbf{g}) = 0 \) whenever \( \mathbf{w}^T(\mathbf{F}^T\mathbf{F}) = 0^T \), so consider any row \( \mathbf{w} \) that satisfies the last equation. It must satisfy also \( 0 = \mathbf{w}^T(\mathbf{F}^T\mathbf{F})\mathbf{w} = (\mathbf{Fw})^T(\mathbf{Fw}) \), which implies \( \mathbf{Fw} = \mathbf{0} \), which implies \( \mathbf{w}^T(\mathbf{F}^T\mathbf{g}) = (\mathbf{Fw})^T\mathbf{g} = 0 \), whereupon Fredholm’s Alternative (1) implies that the Normal Equations have at least one solution \( \hat{\mathbf{u}} \). It is unique if and only if the columns of \( \mathbf{F} \) are linearly independent; otherwise add any nonzero solution \( \mathbf{z} \) of \( \mathbf{Fz} = \mathbf{0} \) to one \( \hat{\mathbf{u}} \) to get another.
How do we know that setting \( u = \hat{u} \) minimizes \( \|Fu - g\| \)? For every \( u \) we find
\[
\|Fu - g\|^2 - \|F\hat{u} - g\|^2 = \|F(u - \hat{u}) + (F\hat{u} - g)\|^2 - \|F\hat{u} - g\|^2 \\
= \|F(u - \hat{u})\|^2 + 2(u - \hat{u})^T(F\hat{u} - g) \\
= \|F(u - \hat{u})\|^2 + 2(u - \hat{u})^TF\hat{u} - 2g \\
\]
with equality instead of inequality just when \( u \) is another solution of the Normal Equations.

When the Normal Equations have many solutions \( \hat{u} \), which does Matlab choose for \( F\backslash g \)? It has a near minimal number of nonzero elements. A different solution minimizes \( \|\hat{u}\|^2 := \hat{u}^T\hat{u} \), as if also the space of vectors \( u \) were Euclidean. This doubly minimizing solution \( \hat{u} \) satisfies both the Normal Equations \((F^TF)\hat{u} = F^Tg\) and an auxiliary equation \( \hat{u} = F^TFv \) for some vector \( v \) of “Lagrange Multipliers.” In consequence \( v \) satisfies \((F^TF)^2v = F^Tg\), an equation with least one solution \( v \) whose existence is assured by an application of Fredholm’s first alternative very much like before except that the hypothesis \( w^T(F^TF) = o^T \) is replaced by \( w^T(F^TF)^2 = o^T \). (Can you carry out this inference?) Every other solution \( u \) of the Normal Equations satisfies \( \|u\|^2 - \|\hat{u}\|^2 = \|(u - \hat{u}) + \hat{u}\|^2 \geq 0 \), so this \( \hat{u} = F^TFv \) really is doubly minimizing; moreover it is determined uniquely by the data \([F, g]\). (Can you see why?) As we shall see later after Singular Values have been discussed, there is a matrix \( F^\dagger \) called the “Moore-Penrose Pseudo-Inverse” of \( F \) such that the doubly minimizing \( \hat{u} = F^\dagger g \) is a linear function of \( g \). (Matlab’s name for \( F^\dagger \) is pinv(F).) However, whenever neither \( F^TF \) nor \( FF^T \) is invertible, so \( F^\dagger \) is interesting, it turns out to be a violently discontinuous function of \( F \). This renders the doubly minimizing \( \hat{u} \) doubly dubious because the space of vectors \( u \) need not be Euclidean. Matlab’s \( F\backslash g \) can be discontinuous too, even when \( FF^T \) is invertible and the doubly minimizing \( \hat{u} \) is continuous.

Linear Regression:
Least-Squares approximation has been applied to statistical estimation for over two centuries. An \( m \times n \) matrix \( F \) is assumed given with linearly independent columns (so \( m \geq n \)); and a given \( m \)-vector \( g = y + q \) of “data” is thought to include a systematic contribution \( y \) and a “random error” \( q \). The question is how near is \( y \) to \( \text{Range}(F) \)? The answer is obscured by the random error. The elements of this error \( q \) are assumed independently distributed with mean 0 and known variance \( \beta^2 \). These terms are given meaning by an Averaging or Expectation operator \( \mathbb{E} \) which acts upon every random variable \( r \) linearly to produce \( \mathbb{E}r \), the average or mean of the population of values of \( r \). Thus \( \mathbb{E}q = o \) because every element of \( q \) has mean 0; and \( q \) has covariance matrix \( \mathbb{E}(q - \mathbb{E}q)(q - \mathbb{E}q)^T = \beta^2I \) since the square of every element of \( q \) has mean \( \beta^2 \) but every product of different elements of \( q \) has mean 0 because they are independent. The smaller is \( \beta \), the less uncertainty does random error \( q \) introduce into the data \( g \).

Define \( x := (F^TF)^{-1}F^Ty \) to minimize \( \|Fx - y\| \) although neither \( y \) nor \( x \) can be known. As the known \( g \) approximates \( y \), so is \( x \) approximated by whatever \( \hat{u} \) minimizes \( \|F\hat{u} - g\| \). Get \( \hat{u} = (F^TF)^{-1}F^Tg \); how well can it approximate \( x \)? Since \( \mathbb{E}g = y \), we find that \( \mathbb{E}\hat{u} = x \), so \( \hat{u} \) is an unbiased estimate of \( x \). The covariance matrix of \( \hat{u} \) is computable too; it is
\[
\mathbb{E}(\hat{u} - x)(\hat{u} - x)^T = \mathbb{E}(F^TF)^{-1}F^Tqq^TF(F^TF)^{-1} = (F^TF)^{-1}F^TF\mathbb{E}(qq^T)F(F^TF)^{-1} = \beta^2(F^TF)^{-1}. \\
\]
The smaller this is, the better does \( \hat{u} \) approximate \( x \) on average. The smaller is \( \|Fx - y\| \), the smaller do we expect \( \|F\hat{u} - g\| \) to be. How small should we expect it to be? A calculation below shows that \( \mathbb{E}(\|F\hat{u} - g\|^2) = \|Fx - y\|^2 + (m-n)\beta^2 \). It means that \( \|F\hat{u} - g\| \) is unlikely to exceed \( \beta\sqrt{m-n} \) much if \( y \) lies in or very near \( \text{Range}(F) \); conversely, \( \|Fx - y\| \) is unlikely to be much smaller than \( \|F\hat{u} - g\| \) if this is many times bigger than \( \beta\sqrt{m-n} \). Explanation follows.
Proof that \( \mathbb{E}(\|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2) = \|\mathbf{F} \mathbf{x} - \mathbf{y}\|^2 + (m-n)\beta^2 \): The Trace of a square matrix is defined to be the sum of its diagonal elements; evaluate this sum to confirm that Trace\(\mathbf{A}^T \mathbf{B}\) = Trace\(\mathbf{B}^T \mathbf{A}\) for any matrices \(\mathbf{A}^T\) and \(\mathbf{B}\) whose products \(\mathbf{A}^T \mathbf{B}\) and \(\mathbf{B}^T \mathbf{A}\) are both square, though perhaps of different dimensions. Next define \( \mathbf{H} := \mathbf{F}^T \mathbf{F} \mathbf{F}^{-1} \mathbf{F} \) and confirm that \( \mathbf{H}^T = \mathbf{H} \) = \( \mathbf{H}^2 \). ( \( \mathbf{H} \) is the orthogonal projector onto Range(\(\mathbf{F}\)) because \( \mathbf{p} = \mathbf{F} \mathbf{z} \) for some \( \mathbf{z} \) \( \iff \mathbf{p} = \mathbf{H} \mathbf{p} \), so Range(\(\mathbf{F}\)) = Range(\(\mathbf{H}\)), and \( \mathbf{H} \mathbf{z} = \mathbf{0} \) \( \iff \mathbf{z}^T \mathbf{H} = \mathbf{0}^T \), so Nullspace(\(\mathbf{H}\)) = Range(\(\mathbf{H}^T\)).)

Proof that \( \mathbb{E}(\|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2) \) is unlikely to be many times bigger than its mean \( \mathbb{E}(\|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2) \): More precisely, we shall deduce that \( \|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2 \) exceeds \( \lambda \mathbb{E}(\|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2) \) with probability less than \( 1/\lambda \), for every \( \lambda > 1 \). This deduction is an instance of Tchebyshev’s Inequality: If a positive random variable \( \mathbf{p} \) has mean \( \mathbb{E}(\mathbf{p}) = \mu \), then the probability that \( \mathbf{p} \geq \lambda \mu \) cannot exceed 1/\( \lambda \) for any \( \lambda > 1 \). Here is a proof of Tchebyshev’s Inequality. Let \( \rho(\xi) \) be the probability that \( \rho \) \leq \( \xi \). This \( \rho(\xi) \) is a nondecreasing function increasing from \( \rho(0) = 0 \) to \( \rho(\infty) = 1 \), and \( \mu = \int_{\xi=0}^{\infty} \xi \rho(\xi) d\xi \) by virtue of the definition of \( \mathbb{E} \). We seek an overestimate for \( \int_{\xi=0}^{\infty} \xi \rho(\xi) d\xi \), which is the probability that \( \rho \geq \lambda \mu \). We find that \( \int_{\xi=0}^{\infty} \xi \rho(\xi) d\xi \leq \int_{\xi=0}^{\infty} \xi \rho(\xi) d\xi / (\lambda \mu) \leq \int_{\xi=0}^{\infty} \xi \rho(\xi) d\xi / (\lambda \mu) = \mu / (\lambda \mu) \), which yields the result claimed. (This can be a gross overestimate because it uses almost no information about \( \rho \). For almost all values of \( \lambda > 1 \), and for all values of \( \lambda > 1 \) for almost all probability functions \( \rho \), the probability that \( \rho \geq \lambda \mu \) is actually far tinier than \( 1/\lambda \).)

Thus the computed \( \|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2 \) is unlikely to be many times bigger than \( \|\mathbf{F} \mathbf{x} - \mathbf{y}\|^2 + \beta^2(m-n) \) in which \( \beta^2(m-n) \) is given and \( \|\mathbf{F} \mathbf{x} - \mathbf{y}\|^2 \) is unknown, whence something probabilistic can be inferred about the unknown. Another similar application of Least-Squares is to the assumption that \( \mathbf{y} = \mathbf{F} \mathbf{x} + \mathbf{q} \) for a random error \( \mathbf{q} \) about which \( \beta^2 \) is unknown but estimated from \( \|\mathbf{F} \mathbf{u} - \mathbf{g}\|^2/(m-n) \). These applications are treated in Statistics courses.

Abstract Least-Squares:
Suppose a column vector \( \mathbf{g} \) is given in an Euclidean space into which a given linear operator \( \mathbf{F} \) maps a real space of abstract vectors \( \mathbf{u} \). In the Euclidean space, length \( ||\mathbf{g}|| := \sqrt{\mathbf{g}^T \mathbf{g}} \), but no such length is defined (yet) for Domain(\(\mathbf{F}\)). Again our task is to choose \( \mathbf{u} \) to minimize \( ||\mathbf{F} \mathbf{u} - \mathbf{g}|| \), which will then be the distance from \( \mathbf{g} \) to Range(\(\mathbf{F}\)). Differentiating the sum of squares \( ||\mathbf{F} \mathbf{u} - \mathbf{g}||^2 = (\mathbf{F} \mathbf{u} - \mathbf{g})^T (\mathbf{F} \mathbf{u} - \mathbf{g}) \) produces \( d (\mathbf{F} \mathbf{u} - \mathbf{g})^T (\mathbf{F} \mathbf{u} - \mathbf{g}) = 2(\mathbf{F} \mathbf{u} - \mathbf{g})^T \mathbf{F} d\mathbf{u} \), which vanishes for all (infinitesimal) perturbations \( d\mathbf{u} \) if and only if \( (\mathbf{F} \mathbf{u} - \mathbf{g})^T \mathbf{F} = \mathbf{o}^T \). This \( \mathbf{o}^T \) is the linear functional that annihilates Domain(\(\mathbf{F}\)). The last equation says that when \( ||\mathbf{F} \mathbf{u} - \mathbf{g}|| \) is minimized the residual \( \mathbf{F} \mathbf{u} - \mathbf{g} \) must be normal (perpendicular, orthogonal) to Range(\(\mathbf{F}\)). (This explains the word “Normal” in “Normal Equations” and removes any suggestion that other equations are abnormal.) Drawing a picture helps; imagine Range(\(\mathbf{F}\)) to be a plane in Euclidean 3-space containing a vector \( \mathbf{F} \mathbf{u} \) which, when it comes closest to a given vector \( \mathbf{g} \) not in the plane, comes to that point in the plane reached by dropping a perpendicular from \( \mathbf{g} \).

We could transpose \( (\mathbf{F} \mathbf{u} - \mathbf{g})^T \mathbf{F} = \mathbf{o}^T \) to \( (\mathbf{F}^T \mathbf{F}) \mathbf{u} = \mathbf{F}^T \mathbf{g} \) if we knew what “\( \mathbf{F}^T \mathbf{F} \)” meant.
The trouble with the expression “$F^T F$” is that it is not what it first seems; if $F$ were a matrix then $F^T F$ would map $\text{Domain}(F)$ to itself, but a change of basis in $\text{Domain}(F)$ does not change $F^T F$ to the expected similar matrix. Here is what happens instead:

Let $B$ be a basis for $\text{Domain}(F)$. Then abstract vector $u = Bu$ for some column vector $u$, and $Fu = FBu = Fu$ for a matrix $F = FB$. The Normal Equations “$(Fu - g)^T F = o^T$” turn into “$(Fu - g)^T F = o^T$” which becomes “$(F^T Fu) = F^T g$” after matrix transposition. $BC$ is a new basis for $\text{Domain}(F)$, and $u = Bu$ for matrix $F = FC$, where $C$ is any invertible matrix of the same dimension as $\text{Domain}(F)$. What was “$(F^T F)u = F^T g$” in the old basis becomes “$(F^T F)u = F^T g$” in the new, replacing matrix $F^T F$ by $BC = CTFC$. This differs from $C^{-1} F^T FC$, which is how the change in basis would have changed $F^T F$ if it were the matrix of a map from $\text{Domain}(F)$ to itself. Instead, $F^T F$ is the matrix of a map from $\text{Domain}(F)$ to its own dual space.

If you doubt that these choices of basis matter, try the following example: Let $g := 10101$, a scalar, and suppose $F = [1, 10, 100]$ in some coordinate system. Then get Matlab to compute $u = F \backslash g$ to solve the least-squares problem. Next change to a new basis using a diagonal matrix $C = \text{diag}(10, 1, 1/16)$. It changes $F$ to $F = FC$ and thus changes the solution of the least-squares problem to $u = F \backslash g$. This maps back to $C^{-1} u = C^{-1} (F \backslash g)$ in the old basis. Compare with the old solution $u$. Try again with 6-vectors $g$ and 6-by-3 matrices $F$ at random.

Bilinear Forms:

There is no uniquely defined operator $F^T F$ just as there is no functional $u^T$ determined uniquely by vector $u$ in a non-Euclidean space. The matrices that appear in the Normal Equations are not all matrices that represent linear maps from one space of column vectors to another or itself; matrix $F^T F$ belongs to a Symmetric Bilinear Form that maps column vectors to row vectors.

Consider $(Fu)^T Fv$. It maps pairs $\{u, v\}$ of vectors from $\text{Domain}(F)$ to real scalars, and does so as a linear function of each vector separately; this is the definition of a Bilinear Form. And since $(Fu)^T Fv$ is unaltered when $u$ and $v$ are swapped, it is a Symmetric Bilinear Form.

There are many notations for bilinear forms: $H_{uv}$, $H(u, v)$, $(v, Hu)$, … . They all mean this: $H_{uv}$ is a linear functional in the space dual to vectors $v$, and $H_{uv}$ is its scalar value; $H_{v,u}$ is a linear functional in the space dual to vectors $u$, and $H_{uv}$ is its scalar value; Given a basis $B$ for vectors $u = Bu$, and a basis $E$ for vectors $v = Ev$, there is a matrix $H$ for which $H_{uv} = (HBu)^T Ev = (Hu)^T v = v^T Hu$;

Changing bases from $B$ to $BC$ and $E$ to $ED$ changes $u$ to $u' = C^{-1} u$, $v$ to $v' = D^{-1} v$, and $H$ to $\Xi = D^T HC$ so that $H_{uv} = v^T Hu = \nu^T Huv$. 

Exercise: Express the elements of matrix $H$ in terms of the effect $H$ has upon the elements of bases $B$ and $E$.

A Symmetric bilinear form maps vectors $u$ and $v$ from the same space to scalars, and does so in a way independent of the order of $u$ and $v$ thus: $H_{uv} = H_{vu}$. A symmetric bilinear form has a symmetric matrix $H = H^T$ in any basis. (Why?) Changing the basis changes $H$ to matrix $\Xi = C^T HC$ for some invertible $C$; the two matrices $\Xi$ and $H$ are called “Congruent.” This congruence is an Equivalence, so it preserves rank; i.e., $\text{rank}(\Xi) = \text{rank}(H)$. Congruence also preserves a thing called “Signature” as we’ll see when we come to Sylvester’s Inertia Theorem.