Weighted Least Squares

Standard least-squares tries to fit a vector $\mathbf{x}$ to a set of “measurements” $\mathbf{y}$ by solving

$$\minimize_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{y} - A\mathbf{x}\|_2^2.$$

Now, what if some of the measurements are more reliable than others? Or, what if the errors are closely correlated between measurements?

There is a systematic way to treat both of these cases using weighted least-squares. Instead of minimizing the energy in the residual

$$\|\mathbf{r}\|_2^2 = \|\mathbf{y} - A\mathbf{x}\|_2^2,$$

we will minimize

$$\|W\mathbf{r}\|_2^2 = \|W\mathbf{y} - WAx\|_2^2,$$

for some $M \times M$ weighting matrix $W$.

When $W$ is a diagonal matrix,

$$W = \begin{bmatrix} w_{11} & & \\ & w_{22} & \\ & & \ddots \\ & & & w_{MM} \end{bmatrix},$$

then the error we are minimizing looks like

$$\|W\mathbf{r}\|_2^2 = w_{11}^2 r[1]^2 + w_{22}^2 r[2]^2 + \cdots + w_{MM}^2 r[M]^2.$$
By adjusting the $w_{mm}$, we can penalize some of the components of the error more than others.

By adding off-diagonal terms, we can account for correlations in the error (we will explore this further later in these notes).

Solving

$$\minimize_{x \in \mathbb{R}^N} \|W r\|_2^2 = \minimize_{x \in \mathbb{R}^N} \|W \mathbf{y} - WAx\|_2^2,$$

is simple. We simply use least-squares with $WA$ as the matrix, and $Wy$ as the observations:

$$\hat{x}_{wls} = (WA)^\dagger Wy,$$

where $(WA)^\dagger$ is the pseudo-inverse of $WA$.

For the rest of this section, we will assume that $M \geq N$ (meaning that there are at least as many observations as unknowns) and that $A$ has full column rank. This allows us to write

$$\hat{x}_{wls} = (A^T W^T W A)^{-1} A^T W^T W y.$$

**Example:** We measure a patient’s pulse 3 times, and record


In this case, we can take

$$A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$
What is the least-square estimate for the pulse rate $x_0$?

Now say that we were in a hurry when the third measurement was made, so we would like to weigh less than the others. What is the weighted least-squares estimate when

$$
W = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & w_{33}
\end{bmatrix}
$$

What about the particular case when $w_{33} = 1/2$?
Statistical Estimation

Suppose we use the following model for our measurements:

\[ y = Ax_0 + e, \]

where \( y \in \mathbb{R}^M \), \( A \) is an \( M \times N \) matrix, \( x_0 \in \mathbb{R}^N \) is what we are interested in estimating, and \( e \in \mathbb{R}^M \) is a random error.

We will assume that each entry of \( e \) has zero mean:

\[ \mathbb{E}[e[m]] = 0, \ m = 1, \ldots, M, \quad \mathbb{E}[e] = 0. \]

We will characterize \( e \) through its covariance matrix

\[ R[\ell,m] = \mathbb{E}[e[\ell]e[m]], \]

or more compactly

\[ R = \mathbb{E}[ee^T]. \]

The diagonal of \( R \) contains the variances of the entries of \( e \), while the off diagonal terms capture the correlations (which is the same as covariance, since all of the \( e[m] \) are zero mean).

For example, if two measurement errors have

\[
\begin{align*}
\text{var}(e[1]) &= \mathbb{E}[e[1]^2] = 3, \\
\text{var}(e[2]) &= \mathbb{E}[e[2]^2] = 2, \\
\text{and} \quad \text{cov}(e[1], e[2]) &= \mathbb{E}[e[1]e[2]] = -1,
\end{align*}
\]

then

\[ R = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}. \]

It is always true that covariance matrices are symmetric and positive semi-definite (so their eigenvalues are \( \geq 0 \)).
A handy fact that we will use repeatedly below is that if \( e \) has covariance matrix \( R \), then for any matrix \( M \), the covariance of \( Me \) is\(^1\)

\[
\]

**Questions:**

1. Suppose that the entries of \( e \) have variances \( \nu_m^2 = E[e[m]^2] \).
   Calculate
   \[
   E[\|e\|_2^2] = \quad \text{__________}.
   \]
   (the expected energy of \( e \)).
   **Answer:**

   \[
   E[\|e\|_2^2] = \sum_{m=1}^{M} E[e[m]^2] = \sum_{m=1}^{M} \nu_m^2.
   \]

---

\(^1\) If you want to see why that second-to-last step is true more explicitly, set \( Q = Mee^TM^T \). Then if \( m_i \) is the \( i \)th row of \( M \),

\[
Q[i, j] = (Me)[i](Me)[j] = \langle e, m_i \rangle \langle m_j, e \rangle = \sum_{\ell} \sum_{k} M[i, \ell]M[j, k]R[\ell, k] = (M^T R M)[i, j],
\]

and

\[
E[Q[i, j]] = \sum_{\ell} \sum_{k} M[i, \ell]M[j, k]R[\ell, k] = (M^T R M)[i, j],
\]

so \( E[Q] = MRM^T \).
2. Now let \( \mathbf{D} \) be a diagonal matrix

\[
\mathbf{D} = \begin{bmatrix}
d_1 & & \\
& d_2 & \\
& & \ddots \\
& & & d_M
\end{bmatrix}.
\]

Calculate

\[
\mathbb{E}[\|\mathbf{D}\mathbf{e}\|^2_2] = \rule{3cm}{1pt}.
\]

**Answer:**

\[
\mathbb{E}[\|\mathbf{D}\mathbf{e}\|^2_2] = \sum_{m=1}^{M} \mathbb{E}[d_m^2 \nu[m]^2]
= \sum_{m=1}^{M} d_m^2 \nu_m^2.
\]

3. Suppose \( \mathbf{e} \in \mathbb{R}^M \) has covariance matrix \( \mathbf{R} \). Let \( \mathbf{L} \) be an \( N \times M \) matrix. Calculate

\[
\mathbb{E}[\|\mathbf{L}\mathbf{e}\|^2_2] = \rule{3cm}{1pt}.
\]

**Answer:** We use two facts which are easily verified (do this at home). First, the inner product of two vectors \( \mathbf{u}, \mathbf{v} \in \mathbb{R}^N \) is equal to the trace of their outer product:

\[
\langle \mathbf{u}, \mathbf{v} \rangle = \text{trace}(\mathbf{u}\mathbf{v}^T).
\]

Second, if \( \mathbf{Q} \) is a square matrix whose entries are random variables, then

\[
\mathbb{E}[\text{trace}(\mathbf{Q})] = \text{trace}(\mathbb{E}[\mathbf{Q}]).
\]
Then
\[ E[\|Le\|^2_2] = E[\langle Le, Le \rangle] = E[\text{trace}(Le^TL^T)] = \text{trace}(E[Le^TL^T]) = \text{trace}(L E[ee^T]L^T) = \text{trace}(LRL^T). \]

**Uncorrelated errors**

Suppose that the random errors are uncorrelated, so that the covariance matrix is diagonal

\[ R = E[ee^T] = \begin{bmatrix} \nu_1^2 & 0 & 0 & \cdots \\ 0 & \nu_2^2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \nu_M^2 \end{bmatrix} \]

If \( \nu_m \) is large, it means that we do not have much confidence in our measurement \( y_m \). On the other hand, if \( \nu_m \) is small, it means that our measurement \( y_m \) is most likely very close to the true value of \((Ax_0)[m]\)

We will see this rigorously below, but in this case, the “correct” weighting for each component is simply the inverse of the standard deviation; the weighting matrix \( W \) should be diagonal with

\[ W[m, m] = \frac{1}{\nu_m}, \quad (W = R^{-1/2}). \]
Then the weighted least-squares estimate is given by

\[ \hat{x}_{wls} = (A^T W^T W A)^{-1} A^T W^T W y \]

\[ = (A^T R^{-1} A)^{-1} A^T R^{-1} y. \]

The reconstruction error of this estimate is

\[ x_0 - \hat{x}_{wls} = x_0 - (A^T R^{-1} A)^{-1} A^T R^{-1} (A x_0 + e) \]

\[ = - (A^T R^{-1} A)^{-1} A^T R^{-1} e \]

The mean-square error (MSE) of the error for this estimate is calculated using the result of Question 3 above:

\[ E[\| x_0 - \hat{x}_{wls} \|_2^2] = \text{trace} \left( (A^T R^{-1} A)^{-1} A^T R^{-1} R R^{-1} A (A^T R^{-1} A)^{-1} \right) \]

\[ = \text{trace} \left( (A^T R^{-1} A)^{-1} A^T R^{-1} A (A^T R^{-1} A)^{-1} \right) \]

\[ = \text{trace} \left( (A^T R^{-1} A)^{-1} \right) \]

**Example.** We take \( M \) readings of a patient’s pulse, each has an error of \( \nu^2 \). In this case, the underlying quantity (the pulse) \( x_0 \) is a scalar. The optimal estimate (no matter what \( \nu \) is) is

\[ \hat{x} = \frac{1}{M} (y[1] + y[2] + \cdots + y[M]). \]

What is the mean-square error for this estimate?
Answer: The mean-square error is

\[
E[|x_0 - \hat{x}|^2] = E \left[ \left| x_0 - \frac{1}{M} \sum_{m=1}^{M} (x_0 + e[m]) \right|^2 \right] \\
= E \left[ \left| \frac{1}{M} \sum_{m=1}^{M} e[m] \right|^2 \right] \\
= \frac{1}{M^2} E[\langle e, e \rangle] \\
= \frac{1}{M^2} E[\text{trace}(ee^T)] \\
= \frac{1}{M^2} \text{trace}(E[ee^T]) \\
= \frac{\nu^2}{M},
\]

where the last step follows from the fact that the covariance matrix of the errors \( e \) is diagonal.

Now suppose that the variance for each of the \( M \) measurements is different; \( \nu_1^2, \nu_2^2, \ldots, \nu_M^2 \).

Now what is the best estimate \( \hat{x} \)?

What is the MSE of this estimate?

Answers: We have

\[
A = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} 1/\nu_1^2 & 1/\nu_2^2 & \cdots & 1/\nu_M^2 \\ 1/\nu_2^2 & \ddots & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 1/\nu_M^2 & \cdots & \cdots & 1/\nu_M^2 \end{bmatrix},
\]
and

$$(A^T R^{-1} A)^{-1} = \left( \sum_{m=1}^{M} 1/\nu_m^2 \right)^{-1},$$

and so

$$\hat{x} = \frac{\sum_{m=1}^{M} y[m]/\nu_m^2}{\sum_{m=1}^{M} 1/\nu_m^2}.$$ 

The MSE is

$$\text{trace}((A R^{-1} A)^{-1}) = \left( \sum_{m=1}^{M} 1/\nu_m^2 \right)^{-1}.$$
Best Linear Unbiased Estimator (BLUE)

We now return to the general estimation problem: we observe

$$y = Ax_0 + e,$$

where $e \in \mathbb{R}^M$ is random with

$$E[e] = 0, \quad E[ee^T] = R.$$

Since $e$ is random, the observations $y$ are also random. We can now ask what is the best statistical estimate of $x_0$. We will restrict ourselves to estimators that have the following properties:

1. **Linearity.** That is, our estimate can be computed by applying a fixed matrix to $y$,
   $$\hat{x} = Ly,$$
   for some $N \times M$ matrix $L$.

2. **Unbiased.** Since the estimate $\hat{x}$ is a function of random variables, it is itself a random variable. Our estimator is unbiased if
   $$E[\hat{x}] = x_0,$$
   which means the expectation of the estimation error is zero,
   $$E[\hat{x} - x_0] = 0.$$

We will search for the best such estimator; the best linear unbiased estimator (BLUE).

Let’s make it clear what we mean by “best”. We mean that the MSE of the estimation error, $E[\|\hat{x} - x_0\|_2^2]$ is minimized.
The estimator is linear, so we can write

\[ \hat{x} = Ly = L(Ax + e) = LAx + Le, \]

for some matrix \( L \) which we will optimize. We want the estimator to be unbiased, so

\[ 0 = E[x_0 - \hat{x}] = E[x_0 - LAx - Le] \]
\[ = x_0 - LAx_0 - E[Le] \]
\[ = x_0 - LAx_0, \]

where the last step comes from the fact that \( E[Le] = 0 \), since \( E[e] = 0 \). Thus we need \( L \) to obey

\[ LAx_0 = x_0. \]

That is, we want \( L \) to be a left inverse of \( A \), meaning \( LA = I \).

With these two properties in hand, the variance of our estimate for a qualifying \( L \) is

\[ E[\|x_0 - \hat{x}\|_2^2] = E[\|x_0 - LAx_0 - Le\|_2^2] \]
\[ = E[\|Le\|_2^2] \]
\[ = E[\text{trace}(LRL^T)]. \]

So we would like to find the matrix which minimizes

\[ \minimize_{L \in \mathbb{R}^{N \times M}} \text{trace}(LRL^T) \quad \text{subject to} \quad LA = I. \]

I propose that the solution to the above is

\[ L_0 = (A^T R^{-1} A)^{-1} A^T R^{-1}. \]
Let’s check this. Clearly $L_0 A = I$, so $L_0$ is a left inverse. It remains to show that for any other left inverse $L$,

$$\text{trace}(LRL^T) \geq \text{trace}(L_0 RL_0^T).$$

Write a candidate $L$ as

$$L = L_0 + (L - L_0).$$

Then

$$\text{trace}(LRL^T) = \text{trace}(L_0 RL_0^T) + \text{trace}(L_0 R(L - L_0)^T) + \text{trace}((L - L_0)R(L - L_0)^T).$$

Note that

$$RL_0^T = RR^{-1} A(A^T R^{-1} A)^{-1} = A(A^T R^{-1} A)^{-1}.$$ 

Thus

$$(L - L_0)RL_0^T = (L - L_0)A(A^T R^{-1} A)^{-1} = 0$$

since both $LA = I$ and $L_0 A = I$. We are left with

$$\text{trace}(LRL^T) = \text{trace}(L_0 RL_0^T) + \text{trace}((L - L_0)R(L - L_0)^T).$$

Since $(L - L_0)R(L - L_0)^T$ is symmetric and positive semi-definite, the term on the right is $\geq 0$. So we conclude

$$\text{trace}(LRL^T) \geq \text{trace}(L_0 RL_0^T) \quad \text{for all left inverses } L.$$
**Best Linear Unbiased Estimator (BLUE):**

From observations,

\[ y = Ax_0 + e, \quad \text{E}[ee^T] = R, \]

the BLUE is

\[ \hat{x}_{\text{blue}} = (A^T R^{-1} A)^{-1} A^T R^{-1} y. \]

A quick calculation shows

\[ L_0 R L_0^T = (A^T R^{-1} A)^{-1}, \]

and so the MSE of the BLUE is

\[
\text{E}[\|x_0 - \hat{x}_{\text{blue}}\|^2] = \text{trace}((A^T R^{-1} A)^{-1}) = \text{sum of eigenvalues of } (A^T R^{-1} A)^{-1}.
\]

\((A^T R^{-1} A)^{-1}\) is sometimes called the **information matrix**.
**Exercise:** We measure

\[ y = Ax + e \]

with

\[ A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 5 \end{bmatrix}, \quad \mathbb{E}[ee^T] = R = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}. \]

1. Find the best linear unbiased estimate.  
   Hint: 
   \[ R^{-1} = \frac{1}{5} \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}. \]

2. Calculate \( \mathbb{E}[\| x_0 - \hat{x}_{\text{blue}} \|_2^2] \).