Weighted Least Squares

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

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2.$$

Now, what if some of the measurements ore 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

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

we will minimize

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

for some $M \times M$ **weighting matrix** $\boldsymbol{W}$.

When $\boldsymbol{W}$ is a diagonal matrix,

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

then the error we are minimizing looks like

$$\|\boldsymbol{W}\boldsymbol{r}\|_2^2 = w_{11}^2 r[1]^2 + w_{22}^2 r[2]^2 + \cdots + w_{MM}^2 r[M]^2.$$

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

$$\operatorname{minimize} \|\boldsymbol{W}\boldsymbol{r}\|_2^2 = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\operatorname{minimize}} \|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{W}\boldsymbol{A}\boldsymbol{x}\|_2^2,$$

is simple. We simply use least-squares with $\boldsymbol{W}\boldsymbol{A}$ as the matrix, and $\boldsymbol{W}\boldsymbol{y}$ as the observations:

$$\hat{\boldsymbol{x}}_{\text{wls}} = (\boldsymbol{W}\boldsymbol{A})^{\dagger}\boldsymbol{W}\boldsymbol{y},$$

where $(\boldsymbol{W}\boldsymbol{A})^{\dagger}$ is the pseudo-inverse of $\boldsymbol{W}\boldsymbol{A}$.

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 $\boldsymbol{A}$ has full column rank. This allows us to write

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

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

$$y[1] = 70, \quad y[2] = 80, \quad y[3] = 120.$$

In this case, we can take

$$\boldsymbol{A} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

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

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

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

Statistical Estimation

Suppose we use the following model for our measurements:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e},$$

where $\boldsymbol{y} \in \mathbb{R}^M$, $\boldsymbol{A}$ is an $M \times N$ matrix, $\boldsymbol{x}_0 \in \mathbb{R}^N$ is what we are interested in estimating, and $\boldsymbol{e} \in \mathbb{R}^M$ is a **random error**.

We will assume that each entry of $\boldsymbol{e}$ has zero mean:

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

We will characterize $\boldsymbol{e}$ through its **covariance matrix**

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

or more compactly

$$\boldsymbol{R} = \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}].$$

The diagonal of $\boldsymbol{R}$ contains the variances of the entries of $\boldsymbol{e}$, 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

$$\mathrm{var}(e[1]) = \mathrm{E}[e[1]^2] = 3, \quad \mathrm{var}(e[2]) = \mathrm{E}[e[2]^2] = 2,$$
$$\text{and} \quad \mathrm{cov}(e[1], e[2]) = \mathrm{E}[e[1]e[2]] = -1,$$

then

$$\boldsymbol{R} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}.$$

It is always true that covariance matrices are symmetric and positive semi-definite (so their eigenvalues are ≥ 0).

A handy fact that we will use repeatedly below is that if $\boldsymbol{e}$ has covariance matrix $\boldsymbol{R}$, then for any matrix $\boldsymbol{M}$, the covariance of $\boldsymbol{M}\boldsymbol{e}$ is[1]

$$\mathrm{E}[\boldsymbol{M}\boldsymbol{e}(\boldsymbol{M}\boldsymbol{e})^{\mathrm{T}}] = \mathrm{E}[\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}] = \boldsymbol{M}\,\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}]\boldsymbol{M}^{\mathrm{T}} = \boldsymbol{M}\boldsymbol{R}\boldsymbol{M}^{\mathrm{T}}.$$

Questions:

1. Suppose that the entries of $\boldsymbol{e}$ have variances $\nu_m^2 = \mathrm{E}[e[m]^2]$. Calculate

$$\mathrm{E}[\|\boldsymbol{e}\|_2^2] \;=\; \underline{\hspace{3cm}}.$$

(the expected energy of $\boldsymbol{e}$).

Answer:

$$\begin{aligned}\mathrm{E}[\|\boldsymbol{e}\|_2^2] &= \sum_{m=1}^{M} \mathrm{E}[e[m]^2] \\ &= \sum_{m=1}^{M} \nu_m^2.\end{aligned}$$

[1] If you want to see why that second-to-last step is true more explicitly, set $\boldsymbol{Q} = \boldsymbol{M}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}$. Then if $\boldsymbol{m}_i$ is the ith row of $\boldsymbol{M}$,

$$Q[i,j] = (\boldsymbol{M}\boldsymbol{e})[i](\boldsymbol{M}\boldsymbol{e})[j] = \langle \boldsymbol{e}, \boldsymbol{m}_i\rangle\langle \boldsymbol{m}_j, \boldsymbol{e}\rangle = \sum_{\ell}\sum_{k} M[i,\ell]M[j,k]e[\ell]e[k],$$

and

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

so $\mathrm{E}[\boldsymbol{Q}] = \boldsymbol{M}\boldsymbol{R}\boldsymbol{M}^{\mathrm{T}}$.

2. Now let $\boldsymbol{D}$ be a diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_M \end{bmatrix}.$$

Calculate

$$\mathrm{E}[\|\boldsymbol{De}\|_2^2] \ = \ \underline{\hspace{3cm}}.$$

Answer:

$$\begin{aligned} \mathrm{E}[\|\boldsymbol{De}\|_2^2] &= \sum_{m=1}^{M} \mathrm{E}[d_m^2 e[m]^2] \\ &= \sum_{m=1}^{M} d_m^2 \nu_m^2. \end{aligned}$$

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

$$\mathrm{E}[\|\boldsymbol{Le}\|_2^2] \ = \ \underline{\hspace{3cm}}.$$

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

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \mathrm{trace}(\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}).$$

Second, if $\boldsymbol{Q}$ is a square matrix whose entries are random variables, then

$$\mathrm{E}[\mathrm{trace}(\boldsymbol{Q})] = \mathrm{trace}(\mathrm{E}[\boldsymbol{Q}]).$$

Then

$$
\begin{aligned}
\mathrm{E}[\|\boldsymbol{L}\boldsymbol{e}\|_2^2] &= \mathrm{E}[\langle \boldsymbol{L}\boldsymbol{e}, \boldsymbol{L}\boldsymbol{e}\rangle] \\
&= \mathrm{E}[\operatorname{trace}(\boldsymbol{L}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}})] \\
&= \operatorname{trace}\left(\mathrm{E}[\boldsymbol{L}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{T}}]\right) \\
&= \operatorname{trace}\left(\boldsymbol{L}\,\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}]\boldsymbol{L}^{\mathrm{T}}\right) \\
&= \operatorname{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}}).
\end{aligned}
$$

Uncorrelated errors

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

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

If ν_m is large, it means that we do not have much confidence in our measurement y_m. On the other hand, if ν_m is small, it means that our measurement y_m is most likely very close to the true value of $(\boldsymbol{A}\boldsymbol{x}_0)[m]$

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 $\boldsymbol{W}$ should be diagonal with

$$
W[m,m] = \frac{1}{\nu_m}, \qquad (\boldsymbol{W} = \boldsymbol{R}^{-1/2}).
$$

Then the weighted least-squares estimate is given by

$$\begin{aligned}\hat{\boldsymbol{x}}_{\text{wls}} &= (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y} \\ &= (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}.\end{aligned}$$

The reconstruction error of this estimate is

$$\begin{aligned}\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{\text{wls}} &= \boldsymbol{x}_0 - (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}) \\ &= -(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{e}\end{aligned}$$

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

$$\begin{aligned}\mathrm{E}[\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{\text{wls}}\|_2^2] &= \operatorname{trace}\left((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\right) \\ &= \operatorname{trace}\left((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\right) \\ &= \operatorname{trace}\left((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\right)\end{aligned}$$

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

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

What is the mean-square error for this estimate?

Answer: The mean-square error is

$$
\begin{aligned}
\mathrm{E}[|x_0 - \hat{x}|^2] &= \mathrm{E}\left[\left|x_0 - \frac{1}{M}\sum_{m=1}^{M}(x_0 + e[m])\right|^2\right] \\
&= \mathrm{E}\left[\left|\frac{1}{M}\sum_{m=1}^{M} e[m]\right|^2\right] \\
&= \frac{1}{M^2}\,\mathrm{E}[\langle \boldsymbol{e}, \boldsymbol{e}\rangle] \\
&= \frac{1}{M^2}\,\mathrm{E}[\mathrm{trace}(\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}})] \\
&= \frac{1}{M^2}\,\mathrm{trace}(\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}]) \\
&= \frac{\nu^2}{M},
\end{aligned}
$$

where the last step follows from the fact that the covariance matrix of the errors $\boldsymbol{e}$ 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

$$
\boldsymbol{A} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad \boldsymbol{R}^{-1} = \begin{bmatrix} 1/\nu_1^2 & & & \\ & 1/\nu_2^2 & & \\ & & \ddots & \\ & & & 1/\nu_M^2 \end{bmatrix},
$$

and

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

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

$$\operatorname{trace}((\boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}) = \left(\sum_{m=1}^{M} 1/\nu_m^2\right)^{-1}.$$

Best Linear Unbiased Estimator (BLUE)

We now return to the general estimation problem: we observe

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e},$$

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

$$\mathrm{E}[\boldsymbol{e}] = \boldsymbol{0}, \quad \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}] = \boldsymbol{R}.$$

Since $\boldsymbol{e}$ is random, the observations $\boldsymbol{y}$ are also random. We can now ask what is the best statistical estimate of $\boldsymbol{x}_0$. 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 $\boldsymbol{y}$,
$$\hat{\boldsymbol{x}} = \boldsymbol{L}\boldsymbol{y},$$
for some $N \times M$ matrix $\boldsymbol{L}$.
2. **Unbiased**. Since the estimate $\hat{\boldsymbol{x}}$ is a function of random variables, it is itself a random variable. Our estimator is unbiased if
$$\mathrm{E}[\hat{\boldsymbol{x}}] = \boldsymbol{x}_0,$$
which means the expectation of the estimation error is zero,
$$\mathrm{E}[\hat{\boldsymbol{x}} - \boldsymbol{x}_0] = \boldsymbol{0}.$$

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, $\mathrm{E}[\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2]$ is minimized.

The estimator is linear, so we can write

$$\hat{\boldsymbol{x}} = \boldsymbol{L}\boldsymbol{y} = \boldsymbol{L}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}) = \boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{L}\boldsymbol{e},$$

for some matrix $\boldsymbol{L}$ which we will optimize. We want the estimator to be unbiased, so

$$\begin{aligned}\boldsymbol{0} \;=\; \mathrm{E}[\boldsymbol{x}_0 - \hat{\boldsymbol{x}}] &= \mathrm{E}[\boldsymbol{x}_0 - \boldsymbol{L}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{L}\boldsymbol{e}] \\ &= \boldsymbol{x}_0 - \boldsymbol{L}\boldsymbol{A}\boldsymbol{x}_0 - \mathrm{E}[\boldsymbol{L}\boldsymbol{e}] \\ &= \boldsymbol{x}_0 - \boldsymbol{L}\boldsymbol{A}\boldsymbol{x}_0,\end{aligned}$$

where the last step comes from the fact that $\mathrm{E}[\boldsymbol{L}\boldsymbol{e}] = \boldsymbol{0}$, since $\mathrm{E}[\boldsymbol{e}] = \boldsymbol{0}$. Thus we need $\boldsymbol{L}$ to obey

$$\boldsymbol{L}\boldsymbol{A}\boldsymbol{x}_0 = \boldsymbol{x}_0.$$

That is, we want $\boldsymbol{L}$ to be a **left inverse** of $\boldsymbol{A}$, meaning $\boldsymbol{L}\boldsymbol{A} = \mathbf{I}$.

With these two properties in hand, the variance of our estimate for a qualifying $\boldsymbol{L}$ is

$$\begin{aligned}\mathrm{E}[\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2^2] &= \mathrm{E}[\|\boldsymbol{x}_0 - \boldsymbol{L}\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{L}\boldsymbol{e}\|_2^2] \\ &= \mathrm{E}[\|\boldsymbol{L}\boldsymbol{e}\|_2^2] \\ &= \mathrm{E}[\mathrm{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}})].\end{aligned}$$

So we would like to find the matrix which minimizes

$$\underset{\boldsymbol{L}\in\mathbb{R}^{N\times M}}{\mathrm{minimize}} \;\; \mathrm{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}}) \quad \text{subject to} \;\; \boldsymbol{L}\boldsymbol{A} = \mathbf{I}.$$

I propose that the solution to the above is

$$\boldsymbol{L}_0 = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}.$$

Let's check this. Clearly $\boldsymbol{L}_0\boldsymbol{A} = \mathbf{I}$, so $\boldsymbol{L}_0$ is a left inverse. It remains to show that for any other left inverse $\boldsymbol{L}$,

$$\operatorname{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}}) \geq \operatorname{trace}(\boldsymbol{L}_0\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}}).$$

Write a candidate $\boldsymbol{L}$ as

$$\boldsymbol{L} = \boldsymbol{L}_0 + (\boldsymbol{L} - \boldsymbol{L}_0).$$

Then

$$\begin{aligned}\operatorname{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}}) = {} & \operatorname{trace}(\boldsymbol{L}_0\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}}) + \operatorname{trace}(\boldsymbol{L}_0\boldsymbol{R}(\boldsymbol{L} - \boldsymbol{L}_0)^{\mathrm{T}}) \\ & + \operatorname{trace}((\boldsymbol{L} - \boldsymbol{L}_0)\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}}) + \operatorname{trace}((\boldsymbol{L} - \boldsymbol{L}_0)\boldsymbol{R}(\boldsymbol{L} - \boldsymbol{L}_0)^{\mathrm{T}}).\end{aligned}$$

Note that

$$\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}} = \boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1} = \boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}.$$

Thus

$$\begin{aligned}(\boldsymbol{L} - \boldsymbol{L}_0)\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}} &= (\boldsymbol{L} - \boldsymbol{L}_0)\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1} \\ &= \mathbf{0}\end{aligned}$$

since both $\boldsymbol{L}\boldsymbol{A} = \mathbf{I}$ and $\boldsymbol{L}_0\boldsymbol{A} = \mathbf{I}$. We are left with

$$\operatorname{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}}) = \operatorname{trace}(\boldsymbol{L}_0\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}}) + \operatorname{trace}((\boldsymbol{L} - \boldsymbol{L}_0)\boldsymbol{R}(\boldsymbol{L} - \boldsymbol{L}_0)^{\mathrm{T}}).$$

Since $(\boldsymbol{L} - \boldsymbol{L}_0)\boldsymbol{R}(\boldsymbol{L} - \boldsymbol{L}_0)^{\mathrm{T}}$ is symmetric and postive semi-definite, the term on the right is ≥ 0. So we conclude

$$\operatorname{trace}(\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^{\mathrm{T}}) \geq \operatorname{trace}(\boldsymbol{L}_0\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}}) \quad \text{for all left inverses } \boldsymbol{L}.$$

Best Linear Unbiased Estimator (BLUE):
From observations,

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{e}, \qquad \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}] = \boldsymbol{R},$$

the BLUE is

$$\hat{\boldsymbol{x}}_{\text{blue}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y}.$$

A quick calculation shows

$$\boldsymbol{L}_0\boldsymbol{R}\boldsymbol{L}_0^{\mathrm{T}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1},$$

and so the MSE of the BLUE is

$$\begin{aligned}\mathrm{E}[\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{\text{blue}}\|_2^2] &= \operatorname{trace}((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}) \\ &= \text{sum of eigenvalues of } (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}.\end{aligned}$$

$(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}$ is sometimes called the **information matrix**.

Exercise: We measure

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$$

with

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

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

2. Calculate $\mathrm{E}[\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{\text{blue}}\|_2^2]$.