III. Statistical Estimation and Classification

In the last chapter of this course, we considered a very specific estimation framework: the unknown variables were put through a linear operator (i.e. a matrix), perturbed, and then observed. In this chapter, we also consider estimating unknown variables from indirect observations, but here our observational model will be completely different. Our observations are samples of a random variable whose distribution is controlled (or parameterized) by the unknown $\boldsymbol{\theta}$:

$$\text{observe} \;\; Y_i \sim f_{\boldsymbol{\theta}}(y) \quad \text{for } i = 1, \ldots, M.$$

Above, $f_{\boldsymbol{\theta}}$ is the *probability density function* (pdf) for the observations — this density is different for different parameters $\boldsymbol{\theta}$. Moreover, the observations Y_i need not be scalars; they could just as easily be random vectors (or random matrices, or random function, etc). From these observations of Y, we come up with an estimate of $\boldsymbol{\theta}$.

There are two general scenarios in which this kind of problem is treated: we can treat the unknown parameters $\boldsymbol{\theta}$ as being deterministic (they are what they are, we just don't know what they are) or as being random themselves (they obey some probability law that we might know). While there are philosophical differences between these two kinds of treatment, most of the time the application we have dictates which one is more natural.

We start with one example from each of these scenarios; both of these will rely heavily on the linear algebra developed in the last chapter.

Best Linear Unbiased Estimator

Let's see what we get when we treat our canonical linear inverse problem

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

as a statistical estimation problem. Our model here is that the error vector $\boldsymbol{e}$ is random, has zero mean,

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

and covariance matrix

$$R[\ell, m] = \mathrm{E}[e[\ell]e[m]], \quad \text{or} \quad \mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}] = \boldsymbol{R}.$$

Notice that the diagonal entries of $\boldsymbol{R}$ contain the variances of the entries of $\boldsymbol{e}$, while the off-diagonal terms capture the correlations. We also note here that covariance matrices are always symmetric positive semi-definite, since

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\,\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}]\boldsymbol{x} = \mathrm{E}[\boldsymbol{x}^{\mathrm{T}}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{x}] = \mathrm{E}[|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{e}|^2] \geq 0.$$

Since $\boldsymbol{e}$ is random, our observation $\boldsymbol{y}$ is random as well. Our unknown parameters $\boldsymbol{x}_0$ effect the distribution of $\boldsymbol{y}$ in an obvious way: $\boldsymbol{y}$ has the same distribution as $\boldsymbol{e}$, only the mean is $\boldsymbol{A}\boldsymbol{x}_0$ instead of $\boldsymbol{0}$.

We can now ask what is the best way to estimate $\boldsymbol{x}_0$. There are all kinds of ways in which we might treat this problem, but for now 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,

$$\mathrm{E}[\hat{\boldsymbol{x}} - \boldsymbol{x}_0] = \mathbf{0}.$$

We will search for the best such estimator; the best linear unbiased estimator (BLUE). And by "best" we mean that the mean-squared estimation error (MSE), $\mathrm{E}[\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2]$ is minimized.

Our derivation of the BLUE will require some manipulation of covariance matrices, so let's do some warm-up exercises first.

Questions:

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

   $$\mathrm{E}[\|\boldsymbol{e}\|_2^2] \;=\; \_\_\_\_\_.$$

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

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

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

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

Calculate

$$\mathrm{E}[\|\boldsymbol{D}\boldsymbol{e}\|_2^2] \;=\; \_\_\_\_\_.$$

Answer:

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

3. Suppose that $\boldsymbol{e} \in \mathbb{R}^M$ is a random vector with zero mean and covariance matrix $\boldsymbol{R}$. Let $\boldsymbol{M}$ be an arbitrary $N \times M$ matrix, and set $\boldsymbol{g} = \boldsymbol{M}\boldsymbol{e}$. Clearly $\mathrm{E}[\boldsymbol{g}] = \boldsymbol{0}$. What is the covariance $\mathrm{E}[\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}}]$?

   Answer:

$$\begin{aligned}\mathrm{E}[\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}}] &= \mathrm{E}[\boldsymbol{M}\boldsymbol{e}(\boldsymbol{M}\boldsymbol{e})^{\mathrm{T}}] \\ &= \mathrm{E}[\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}] = \boldsymbol{M}\,\mathrm{E}[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}]\boldsymbol{M}^{\mathrm{T}} = \boldsymbol{M}\boldsymbol{R}\boldsymbol{M}^{\mathrm{T}}.\end{aligned}$$

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

$$\mathrm{E}[\|\boldsymbol{L}\boldsymbol{e}\|_2^2] \;=\; \_\_\_\_\_.$$

   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:

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

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

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

We return now to the derivation of the BLUE for $\boldsymbol{x}_0$ from observations $\boldsymbol{y} = \boldsymbol{Ax}_0 + \boldsymbol{e}$. We are restricting ourselves to linear estimators, so we can write

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

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

$$\begin{aligned}
\mathbf{0} \;=\; \mathrm{E}[\boldsymbol{x}_0 - \hat{\boldsymbol{x}}] &= \mathrm{E}[\boldsymbol{x}_0 - \boldsymbol{LAx} - \boldsymbol{Le}] \\
&= \boldsymbol{x}_0 - \boldsymbol{LAx}_0 - \mathrm{E}[\boldsymbol{Le}] \\
&= \boldsymbol{x}_0 - \boldsymbol{LAx}_0,
\end{aligned}$$

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

$$\boldsymbol{LAx}_0 = \boldsymbol{x}_0.$$

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

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

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

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 $\boldsymbol{L}$,

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

Write a candidate $\boldsymbol{L}$ as

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

Then

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

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

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

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

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

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

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

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

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

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 $(\boldsymbol{A}\boldsymbol{x}_0)[m]$

The BLUE we derived above tells us exactly how we should weight these measurements. The BLUE is

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y},$$

where

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \frac{1}{\nu_1^2} & 0 & 0 & \cdots & \\ 0 & \frac{1}{\nu_2^2} & 0 & \cdots & \\ \vdots & & \ddots & & \\ & & & & \frac{1}{\nu_M^2} \end{bmatrix}.$$

Thus we can interpret the BLUE as the solution to the least-squares problem

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

where $\boldsymbol{W}$ is a diagonal weighting matrix $W[m,m] = 1/\nu_m$. This is penalizing mismatch in the components in which we have confidence (ν_m small) more than the components in which we do not have confidence (ν_m larger).

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

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

Exercise with correlated errors: 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 $\mathrm{E}[\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_{\mathrm{blue}}\|_2^2]$.

Three closing notes on the BLUE

1. Notice that we did not need to model the entire probability distribution of $\boldsymbol{e}$, we only needed knowledge of its first two moments (mean and covariance). We got away with this because we restricted our estimator to be linear and unbiased — other classes of estimators will in general rely on higher-order moments or even knowledge of the complete distribution of $\boldsymbol{e}$.
2. The assumption that the estimator is unbiased, that $\mathrm{E}[\hat{\boldsymbol{x}}] = \boldsymbol{x}_0$, seemed natural. However, there can be significant gains (in terms of MSE) from loosening this restriction — introducing a little bias can allow us to reduce the overall error.
3. The BLUE is overall optimal (in terms of MSE) when $\boldsymbol{e}$ is a Gaussian random vector. That is, even if we allow ourselves to look at nonlinear, biased estimators, we will not find one with better performance. It is no coincidence that the probability law for Gaussian random vectors is completely determined by a mean and a covariance matrix.