Note 22. Linear least squares error estimation

Alex Fu

Fall 2022

1 Problem statement

The definition of the inner product $\langle X, Y \rangle = \mathbb{E}(XY)$ in the previous note was not arbitrary — it was chosen so that the problem of *linear least squares error estimation* is precisely the problem of finding the orthogonal projection.

Definition 1 (Linear least squares error estimator)**.**

Let $Y, Z, \ldots$ be given random variables, and let X be the random variable to estimate. The **linear least squares error estimator** (LLSE) of X given $Y, Z, \ldots$ is the affine function

$$\mathbb{L}(X \mid Y, Z, \ldots) := \operatorname*{argmin}_{a+bY+cZ+\cdots} \mathbb{E}\big((X - (a + bY + cZ + \cdots))^2\big)$$

which minimizes the *mean squared error*.

We can now reformulate the problem statement in terms of the Hilbert space of random variables. As the norm-squared $\mathbb{E}(X^2) = \langle X, X \rangle = \|X\|^2$ is minimized by orthogonal projection,

$$\mathbb{L}(X \mid Y, Z, \ldots) = \mathsf{proj}_{\mathsf{span}\{1, Y, Z, \ldots\}}(X).$$

The following principle will prove useful for considering the term $X - \mathbb{L}(X \mid Y, Z, \ldots)$:

Proposition 1 (Orthogonality principle)**.**

Let $\hat{X}$ be some estimator of X that minimizes the mean squared error $\|X - \hat{X}\|^2$. Then $\hat{X}$ is the orthogonal projection of X onto some subspace of random variables U. Furthermore, the *estimation residual* $X - \hat{X}$ is orthogonal to every $Y \in U$, including $\hat{X}$ itself.

An important bit of foreshadowing:

> The LLSE estimator is really a misnomer for the **affine** least squares error estimator.

2 Solution derivation

2.1 Trivial cases

What is the LLSE of X if no other random variables are given, $\mathbb{L}(X)$ or $\mathbb{L}(X \mid c)$? Noting that the affine functions of zero or c are precisely the constant functions,

$$\mathbb{L}(X) = \operatorname*{argmin}_{a \in \mathbb{R}} \|X - a\|^2 = \mathsf{proj}_{\mathsf{span}\{1\}}(X) = \langle X, 1 \rangle \cdot 1 = \mathbb{E}(X).$$

Thus, the expectation is the best constant estimator of X without any other information. Now suppose that we are given "perfect" information, so that X is fully determined as an affine function of the given r.v.s. Then

$$\mathbb{L}(X \mid Y, Z, \ldots) = \mathsf{proj}_{\mathsf{span}\{1, Y, Z, \ldots\} \ni X}(X) = X$$

indeed behaves as expected.

2.2 One given variable

Let us now turn to the case where some Y is given, along with its covariance $\mathsf{cov}(X, Y)$ with X. Furthermore, let $\{1, Y\}$ be orthonormal, so Y is *standardized* with mean 0 and variance 1:

$$\mathbb{L}(X \mid Y) = \langle X, 1 \rangle \cdot 1 + \langle X, Y \rangle \cdot Y = \mathbb{E}(X) + \mathsf{cov}(X, Y) \cdot Y.$$

Moving up in generality, suppose that $\{1, Y\}$ is only an orthogonal set:

$$\mathbb{L}(X \mid Y) = \langle X, 1 \rangle \cdot 1 + \left\langle X, \frac{Y}{\|Y\|} \right\rangle \cdot \frac{Y}{\|Y\|} = \mathbb{E}(X) + \frac{\mathsf{cov}(X, Y)}{\mathsf{var}(Y)} \cdot Y.$$

Finally, suppose only that Y is a nonconstant random variable. By the Gram–Schmidt procedure, we can find an orthogonal basis of span $\{1, Y\}$ from the set $\{1, Y\}$, which turns out to be $\{1, \tilde{Y}\}$,

$$\tilde{Y} := \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathsf{var}(Y)}}.$$

Here, applying Gram–Schmidt is the same as standardization. Substituting $\{1, \tilde{Y}\}$ into either of the previous cases, we find the following identity.

Proposition 2 (LLSE of X given Y)**.**

Suppose that $\mathsf{var}(Y) \neq 0$. Then

$$\mathbb{L}(X \mid Y) = \mathbb{E}(X) + \frac{\mathsf{cov}(X, Y)}{\mathsf{var}(Y)} \cdot (Y - \mathbb{E}(Y)).$$

Quick concept check: what is $\mathbb{L}(X \mid Y)$ when $\mathsf{var}(Y) = 0$? In the first two steps, why were we able to substitute $\langle X, Y \rangle \leftarrow \mathsf{cov}(X, Y)$ and $\|Y\|^2 \leftarrow \mathsf{var}(Y)$?

Proposition 3 (Mean squared error of LLSE*)**.**

The mean squared error of $\mathbb{L}(X \mid Y)$ is

$$\mathbb{E}\big((X - \mathbb{L}(X \mid Y))^2\big) = \mathsf{var}(X) - \frac{\mathsf{cov}(X,Y)^2}{\mathsf{var}(Y)}.$$

Proof. We give a geometric proof, assuming without loss of generality that X and Y are zero-mean. The residual $X - \mathbb{L}(X \mid Y)$ is orthogonal to $\mathbb{L}(X \mid Y)$, which is a scalar multiple of Y. If θ is the angle between X and Y, then

$$\|X - \mathbb{L}(X \mid Y)\|^2 = \|X\|^2 \sin(\theta)^2 = \|X\|^2 \left[1 - \left(\frac{\langle X, Y\rangle}{\|X\|\,\|Y\|}\right)^2\right].$$

As a followup, find $\mathsf{var}(X - \mathbb{L}(X \mid Y))$. What is the mean of $X - \mathbb{L}(X \mid Y)$? □

2.3 General case

Finally, let multiple random variables $Y, Z, \ldots$ be given. For convenience, we will write $\mathsf{proj}_{\{*\}}$ for $\mathsf{proj}_{\mathsf{span}\{*\}}$. Let us consider orthogonalization by Gram–Schmidt once more:

$$\begin{aligned}
1 &= 1 \\
\tilde{Y} &= Y - \mathsf{proj}_{\{1\}}(Y) = Y - \mathbb{E}(Y) \\
\tilde{Z} &= Z - \mathsf{proj}_{\{1,\tilde{Y}\}}(Z) = Z - \mathbb{L}(Z \mid \tilde{Y}) \\
&\vdots
\end{aligned}$$

The term $\tilde{Z} = Z - \mathbb{L}(Z \mid \tilde{Y})$ is the **innovation** of Z, intuitively the new information that is not predictable as a linear combination of the previously given random variables. Importantly, note that the Gram–Schmidt procedure does not simply demean every given random variable.

The benefit of an orthogonal basis $\{1, \tilde{Y}, \tilde{Z}, \ldots\}$ for the subspace we are projecting X onto: we can decompose the projection as a sum of projections onto individual components.

Proposition 4 (General LLSE).

Let $\{1, \tilde{Y}, \tilde{Z}, \ldots\}$ be the innovations of $\{1, Y, Z, \ldots\}$. Then the LLSE of X given $Y, Z, \ldots$ is

$$\begin{aligned}
\mathsf{proj}_{\{1,\tilde{Y},\tilde{Z},\ldots\}}(X) &= \mathsf{proj}_{\{1\}}(X) + \mathsf{proj}_{\{\tilde{Y}\}}(X) + \mathsf{proj}_{\{\tilde{Z}\}}(X) + \cdots \\
&= \mathbb{E}(X) + \frac{\mathsf{cov}(X,\tilde{Y})}{\mathsf{var}(\tilde{Y})} \cdot \tilde{Y} + \frac{\mathsf{cov}(X,\tilde{Z})}{\mathsf{var}(\tilde{Z})} \cdot \tilde{Z} + \cdots
\end{aligned}$$

This is also an *online* formula for estimation: we can update our existing estimate to incorporate newly given random variables simply by adding. However, we find that the exact formula is not as important as the process of derivation, which will be key for further generalizations.

2.4 The affine complication

Recall the true nature of the misnamed LLSE, hinted by the extra 1 in $\mathbb{L}(X \mid *) = \text{proj}_{\{1,*\}}(X)$. Let us now consider why two seemingly equivalent methods paradoxically lead to different results.

Suppose we wish to estimate some zero-mean X given zero-mean Y and Z, where $\text{cov}(X, Y) = \text{cov}(X, Z) = 1$ and $\text{var}(Y) = \text{var}(Z) = 1$. Then

$$\mathbb{L}(X \mid Y, Z) = \text{proj}_{\{1\}}(X) + \text{proj}_{\{Y\}}(X) + \text{proj}_{\{Z\}}(X) = Y + Z.$$

A slightly different path: if LLSEs are projections, then projections should also be LLSEs.

$$\mathbb{L}(X \mid Y, Z) = \mathbb{L}(X \mid Y) + \mathbb{L}(X \mid Z) = Y + Z.$$

So far so good. However, now let $\mathbb{E}(X) = 1$, keeping Y and Z zero-mean. Then we find that

$$\begin{aligned} \text{proj}_{\{1\}}(X) + \text{proj}_{\{Y\}}(X) + \text{proj}_{\{Z\}}(X) &= 1 + Y + Z \\ \mathbb{L}(X \mid Y) + \mathbb{L}(X \mid Z) &= (1 + Y) + (1 + Z). \end{aligned}$$

The exercise for the reader: diagnose where the extra 1 comes from. It may help to expand the alternative expression as projections onto individual terms.

In general, with n given variables, the alternative path will overcount $\mathbb{E}(X)$ a total of $n-1$ times. This is fine when $n = 1$, or when $\mathbb{E}(X) = 0$, but how do we address this complication in general? A summary for whether or not to zero-mean:

- The given random variables: we recommend that you always orthogonalize $\{1, Y, Z, \ldots\}$ by Gram–Schmidt, which as a consequence demeans them as well.
- The estimated random variable: we recommend that you either never demean X, working only in terms of projections onto individual terms, or always demean X, finding $\mathbb{L}(X - \mathbb{E}(X) \mid Y, Z, \ldots)$ as a sum of projections *or* LLSEs, then add $\mathbb{E}(X)$ back at the end.

3 Estimation of multiple random variables*

A natural generalization of LLSE estimation is the problem of estimating multiple random variables $X_1, \ldots, X_m$ given multiple observations $Y_1, \ldots, Y_n$. An equivalent problem is *vector estimation*: estimating $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ by an affine transformation of $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$.

A few points requiring some care: the usual inner product of random vectors is $\langle X, Y \rangle = \mathbb{E}(X^\mathsf{T} Y)$, which one should check still defines a Hilbert space with $\mathbb{E}(\|X\|_2^2) < \infty$ for all $X \colon \Omega \to \mathbb{R}^m$. Moreover, X and Y may not even belong to the same space, but this turns out to be fine: by $\mathbb{L}(X \mid Y) = AY + b$ for $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$,

$$\mathbb{L}(X_1, \ldots, X_m \mid Y_1, \ldots, Y_n)_i = \mathbb{L}(X_i \mid Y_1, \ldots, Y_n).$$

By working componentwise, we can assume without loss of generality that $m = 1$.

The setup now: let $Y \in \mathbb{R}^n$ be zero-mean with positive definite covariance matrix $\operatorname{cov}(Y) = \Sigma_Y$, such that Σ_Y^{-1} exists. Σ_Y is positive definite iff no linear combination of the components Y_i has zero variance, so this condition expresses that "there are no redundant observations."

Let us orthonormalize the components of Y as we did previously with Gram–Schmidt, so that the innovation $\tilde{Y}$ is *orthonormal*: $\operatorname{cov}(\tilde{Y}) = \mathbb{E}(\tilde{Y}\tilde{Y}^\mathsf{T}) = I$.

1. By the spectral theorem, $\Sigma_Y = U\Lambda U^\mathsf{T}$ for an orthonormal matrix U of eigenvectors and a diagonal matrix Λ of strictly positive real (eigen)values. Λ has an invertible, diagonal square root $\Lambda^{1/2}$ given by simply entrywise square roots, so Σ^{-1} also has
$$\Sigma^{-1} = U\Lambda^{-1}U^\mathsf{T} = (U\Lambda^{-1/2})(U\Lambda^{-1/2})^\mathsf{T}.$$

2. Let us consider the random vector $\tilde{Y} = (U\Lambda^{1/2})^{-1}Y = \Lambda^{-1/2}U^\mathsf{T}Y$ with covariance matrix
$$\operatorname{cov}(\tilde{Y}) = (U\Lambda^{1/2})^{-1}(U\Lambda U^\mathsf{T})(U\Lambda^{-1/2}) = I.$$
As $\langle \tilde{Y}_i, \tilde{Y}_j \rangle = \mathbb{1}_{i=j}$, the entries of $\tilde{Y}$ are the new, orthonormalized observations.

3. Finally, we note that $Y \mapsto \tilde{Y}$ is invertible, so $\operatorname{col} Y = \operatorname{col} \tilde{Y}$. Now it's morbin' time.
$$\begin{aligned}
\mathbb{L}(X \mid Y) = \mathbb{L}(X \mid \tilde{Y}) &= \sum_{i=1}^n \langle X, \tilde{Y}_i \rangle \cdot \tilde{Y}_i \\
&= \sum_{i=1}^n \left[\left\langle X, \sum_{j=1}^n (\Lambda^{-1/2}U^\mathsf{T})_{i,j} Y_j \right\rangle \sum_{k=1}^n (\Lambda^{-1/2}U^\mathsf{T})_{i,k} Y_k \right] \\
&= \sum_{j=1}^n \sum_{k=1}^n \left[\left(\sum_{i=1}^n (U\Lambda^{-1/2})_{j,i} (\Lambda^{-1/2}U^\mathsf{T})_{i,k} \right) \langle X, Y_j \rangle \, Y_k \right] \\
&= \sum_{j=1}^n \sum_{k=1}^n (U\Lambda^{-1}U^\mathsf{T})_{j,k} \langle X, Y_j \rangle \, Y_k \\
&= \mathbb{E}\big(XY^\mathsf{T}\big) \left(\mathbb{E}\big(YY^\mathsf{T}\big)^{-1}\right) Y.
\end{aligned}$$
We made it. In the case of $n = 1$, we reduce to the familiar Proposition 2 for zero-mean X. We also find that the same equation above generalizes to all $m \geq 1$ without any change.

Proposition 5 (LLSE).

Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ be zero-mean. Define $\Sigma_Y = \mathbb{E}(YY^\mathsf{T})$ and $\Sigma_{X,Y} = \mathbb{E}(XY^\mathsf{T})$. Then
$$\mathbb{L}(X \mid Y) = \Sigma_{X,Y}\Sigma_Y^{-1}Y.$$

Proposition 6 (General mean squared error of LLSE).

We note the orthogonality principle; that the trace of a scalar is itself; the cyclic property of the trace; and the linearity of the trace, expectation, and transpose. Then

$$\begin{aligned}
\|X - \mathbb{L}(X \mid Y)\|^2 &= \mathbb{E}\big((X - \mathbb{L}(X \mid Y))^\mathsf{T}(X - \mathbb{L}(X \mid Y))\big) \\
&= \mathbb{E}\big((X - \mathbb{L}(X \mid Y))^\mathsf{T} X\big) \\
&= \mathbb{E}\big(\mathrm{tr}\big((X - \mathbb{L}(X \mid Y))^\mathsf{T} X\big)\big) \\
&= \mathbb{E}\big(\mathrm{tr}\big(X(X - \mathbb{L}(X \mid Y))^\mathsf{T}\big)\big) \\
&= \mathrm{tr}\big(\mathbb{E}\big(X(X - \mathbb{L}(X \mid Y))^\mathsf{T}\big)\big) \\
&= \mathrm{tr}\big(\mathbb{E}\big(XX^\mathsf{T} - XY^\mathsf{T}\Sigma_Y^{-1}\Sigma_{Y,X}\big)\big) \\
&= \mathrm{tr}\big(\Sigma_X - \Sigma_{X,Y}\Sigma_Y^{-1}\Sigma_{Y,X}\big).
\end{aligned}$$

4 Linear regression*

Let us now turn to describe the *non-Bayesian* perspective of estimation in *regression*, which does not assume we have knowledge about the distribution of X and Y.

- Let n be the number of *samples* or *data points* collected, and let d be the number of *features* collected for each data point.
- The *design matrix* $\mathbf{X}$ is a $(n \times d)$ matrix such that each row i corresponds to data point i, and each column j corresponds to feature j.
- The *observation vector* y is a $(n \times 1)$ vector in which each entry corresponds to an observation about the ith data point. We assume that the columns of $\mathbf{X}$ and y are zero-mean.
- The *weight vector* β is a $(d \times 1)$ vector in which each entry represents the weight we give to the jth feature. We wish to find β^* such that $\mathbf{X}\beta^*$ estimates y.

So, our problem is the following:

What is the weight vector β that minimizes the sum of squares $\|y - \mathbf{X}\beta\|_2^2$?

We can still study this problem from the Bayesian perspective. Let (X, Y) represent a *uniformly* random chosen row of the design matrix and observation vector:

$$(X, Y) \sim \text{Uniform}\left\{(x_i, y_i)\right\}_{i=1}^n.$$

Then finding the weight vector $\beta \in \mathbb{R}^d$ that minimizes the sum of squared residuals is *the same as* finding β such that $\mathbb{L}(Y \mid X) = \beta^\mathsf{T} X$!

$$\|y - \mathbf{X}\beta\|_2^2 = n \sum_{i=1}^n \frac{1}{n}(y_i - x_i^\mathsf{T}\beta)^2 = n \cdot \mathbb{E}\big((Y - X^\mathsf{T}\beta)^2\big).$$

Moreover, we know that $\beta^\mathsf{T} = \Sigma_{Y,X}\Sigma_X^{-1}$ by the LLSE, so we can find

$$\Sigma_X = \mathbb{E}\big(XX^\mathsf{T}\big) = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\mathsf{T} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$$

$$\Sigma_{Y,X} = \mathbb{E}\big(YX^\mathsf{T}\big) = \frac{1}{n}\sum_{i=1}^{n} y_i x_i^\mathsf{T} = \frac{1}{n}y^\mathsf{T}\mathbf{X}$$

$$\beta = (\Sigma_{Y,X}\Sigma_X^{-1})^\mathsf{T} = \Sigma_X^{-1}\Sigma_{X,Y} = (\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}^\mathsf{T}y$$

$$\hat{y} = \mathbf{X}(\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}^\mathsf{T}y$$

which you may be familiar with as the least squares estimate in other contexts.

This is not the last word on regression models. If we introduce noise or error terms ε_i that are modelled as i.i.d. zero-mean Gaussians $\mathcal{N}(0, \sigma^2)$, such that $y_i = x_i^\mathsf{T}\beta + \varepsilon_i$ for $i = 1, \ldots, n$, then there is a well-developed theory describing how to find the best estimator $\hat{\beta}$, the distribution and expected mean square error of $\hat{\beta}$, and confidence intervals for $\hat{\beta}$.

5 Nonlinear estimation*

So far, we have worked only in terms of linear (or affine) estimators of X. Could we generalize to different models of X, such as *quadratic mean square error estimation* (QMSE)?

$$\operatorname*{argmin}_{a+bY+cY^2} \left\|X - (a + bY + cY^2)\right\|^2.$$

Perhaps surprisingly, we already have the tools to do nonlinear estimation — in linear algebra. Even if the random variable $a + bY + cY^2$ is not linear in Y, it is linear in $\{1, Y, Y^2\}$. In general, we can handle polynomial regression of degree d by finding the orthogonal projection

$$\mathsf{proj}_{\mathsf{span}\{1,Y,\ldots,Y^d\}}(X).$$

This technique further generalizes to any set of linearly independent random variables, such as finding the optimal linear combination of 1, $\sin(Y)$, and $\cos(Y)$ to estimate X. The cost of polynomial regression, however, is more difficult equations and requiring knowledge of higher moments: we need $Y^d \in \mathcal{H}$, or $\mathbb{E}(Y^{2d}) < \infty$.

The natural limit of this generalization: what is the arbitrary (*measurable**) function of X that best estimates X? We find the answer in the following note — the minimum mean square error estimator, or *conditional expectation*.

■