Note 23. Minimum mean square error estimation

Alex Fu

Fall 2022

1 Problem statement

Definition 1 (Minimum mean square error estimator).

The **minimum mean square error** (MMSE) estimator of X given Y is the function of Y that minimizes its mean squared error to X:

$$\phi(Y) := \operatorname*{argmin}_{f(Y) \in \mathcal{H}} \mathbb{E}\big((X - f(Y))^2\big).$$

The subspace of all functions of Y is quite difficult to describe explicitly: it includes all continuous functions of Y at the very minimum. Instead, we will find an equivalent definition given by the **orthogonality principle**. $\phi(Y)$ is the MMSE of X given Y iff for every $f(Y) \in \mathcal{H}$,

$$\mathbb{E}((X - \phi(Y)) \cdot f(Y)) = 0.$$

Note that it is not enough for the residual $X - \phi(Y)$ to be orthogonal to Y. The LLSE is defined by $\mathbb{L}(X \mid Y) \in \operatorname{span}\{1, Y\}$ and $X - \mathbb{L}(X \mid Y) \perp \{1, Y\}$, but the MMSE may not be linear.

2 Solution derivation

Proposition 1 (MMSE is conditional expectation).

The MMSE of X given Y is precisely the conditional expectation $\mathbb{E}(X \mid Y)$.

Definition 2 (Conditional expectation I).

The **conditional expectation** of discrete X given event A with nonzero probability is

$$\mathbb{E}(X \mid A) := \sum_{x \in S} x \cdot \mathbb{P}(X = x \mid A) = \frac{1}{\mathbb{P}(A)} \mathbb{E}(X \cdot \mathbb{1}_A).$$

$\mathbb{E}(X \cdot \mathbb{1}_A)$ is also written $\mathbb{E}(X|_A)$, where $X|_A$ is X *restricted* to the event A. In particular, taking $A = \{Y = y\}$ defines $\mathbb{E}(X \mid Y = y)$. Then the conditional expectation of discrete X given discrete Y is the function of Y

$$\mathbb{E}(X \mid Y) \colon y \longmapsto \mathbb{E}(X \mid Y = y) = \frac{1}{p_Y(y)} \sum_{x \in S} x \cdot p_{X,Y}(x, y).$$

The definition of $\mathbb{E}(X \mid Y)$ for continuous X and Y follows analogously.

Graphically, we can imagine the conditional expectation given Y to restrict the two-dimensional joint density $f_{X,Y}$ to a one-dimensional slice along the line $Y = y$. Then $\mathbb{E}(X \mid Y = y)$ is the center of this slice distribution, normalized by its total mass of $f_Y(y)$.

Let us consider the two extreme examples. If Y is independent of X, then $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$: "$Y$ gives no information about X." On the other hand, if Y is some nonzero multiple of X, then $\mathbb{E}(X \mid cX) = X$: "$Y$ gives all of the information needed to determine X."

In general, the "information" given by a random variable Y is captured by the events $\{Y = y\}$. If $\mathbb{E}(X \mid Y = y)$ is the best determination of the value of X at $Y = y$, then $\mathbb{E}(X \mid Y)$ is the best determination of the random variable X given information about Y, which matches the MMSE.

Proposition 2 (Law of total expectation).

Also known as the law of iterated expectation (LIE), the tower rule, or the smoothing theorem.

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X).$$

We will leave you to verify this law. For now, if we rearrange the equation slightly,

$$\mathbb{E}((X - \mathbb{E}(X \mid Y)) \cdot 1) = 0$$

says that the residual $X - \mathbb{E}(X \mid Y)$ is orthogonal to 1. In general,

$$\begin{aligned}
\mathbb{E}(f(Y)\,\mathbb{E}(X \mid Y)) &= \sum_y \mathbb{P}(Y = y) \cdot f(y)\,\mathbb{E}(X \mid Y = y) \\
&= \sum_y \mathbb{P}(Y = y) \cdot f(y) \left[\sum_x x\,\mathbb{P}(X = x \mid Y = y) \right] \\
&= \sum_{x,y} \mathbb{P}(X = x, Y = y) \cdot f(y) x \\
&= \mathbb{E}(f(Y) X)
\end{aligned}$$

shows that $X - \mathbb{E}(X \mid Y)$ is orthogonal to $f(Y)$. Thus the minimum mean square error estimator of X given Y is $\phi(Y) = \mathbb{E}(X \mid Y)$, which exists and is unique by construction!

There is an intuitive interpretation of the orthogonality principle: the residual $X - \mathbb{E}(X \mid Y)$ has no "remnant model" of X dependent on Y. For instance, let $X = 3Y^2 + Y + Z$ for some Z orthogonal to all functions of Y, and suppose that $\mathbb{E}(X \mid Y) = Y$. Then

$$\left\langle X - \mathbb{E}(X \mid Y),\ 3Y^2 \right\rangle = \left\langle 3Y^2, 3Y^2 \right\rangle + \left\langle Z, 3Y^2 \right\rangle = \left\|3Y^2\right\|^2 > 0$$

indicating some remaining "quadratic model," namely $3Y^2$, which we can add to improve the estimate $\mathbb{E}(X \mid Y)$. The MMSE is the best estimate in the sense that if it is linear in Y, then it equals the LLSE; if it is quadratic, then it equals the QMSE.

3 Properties of conditional expectation

Conditional expectation inherits the properties of both (unconditional) expectation and orthogonal projections, so the following is only an incomplete list of identities for your reference.

a. **Law of total expectation**. See Proposition 2.

b. **Independence**. If X is independent of Y, then $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$.

c. **Known functions**. For any $f(X)$, $\mathbb{E}(f(X) \mid X) = f(X)$.

d. **Indicator**. $\mathbb{E}(\mathbb{1}_A \mid B) = \mathbb{P}(A \mid B)$ for any event B.

e. **Linearity**. $\mathbb{E}(aX + bY \mid Z) = a \cdot \mathbb{E}(X \mid Z) + b \cdot \mathbb{E}(Y \mid Z)$.

f. **Monotonicity**. If $X \leq Y$, then $\mathbb{E}(X \mid Z) \leq \mathbb{E}(Y \mid Z)$.

g. **Monotone convergence***. If $X_n \uparrow X$ are nonnegative, then $\mathbb{E}(X_n \mid Y) \uparrow \mathbb{E}(X \mid Y)$.

h. **Dominated convergence***. If $X_n \overset{\text{a.s.}}{\to} X$, and there exists Y with finite mean so that $|X_n| \leq Y$ and $|X| \leq Y$, then $\mathbb{E}(X_n \mid Z) \overset{\text{a.s.}}{\to} \mathbb{E}(X \mid Z)$.

i. **Cauchy-Schwarz inequality***. $\mathbb{E}(XY \mid Z)^2 \leq \mathbb{E}(X \mid Z)^2 \cdot \mathbb{E}(Y \mid Z)^2$.

j. **Jensen's inequality***. If $\varphi\colon \mathbb{R} \to \mathbb{R}$ is convex, then $\varphi(\mathbb{E}(X \mid Y)) \leq \mathbb{E}(\varphi(X) \mid Y)$, with equality iff $\varphi(X)$ is not strictly convex.

k. **Tower property** I. $\mathbb{E}(\mathbb{E}(X \mid Y) \mid f(Y)) = \mathbb{E}(X \mid f(Y))$.

l. **Tower property** II. $\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Y) = \mathbb{E}(X \mid Y)$.

We also recommend you try the following exercises.

m. **Wald's identity**. Suppose that i.i.d. X_i and independent N have finite expectations. Then

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}(N) \cdot \mathbb{E}(X_1).$$

n. Let $X \sim$ Uniform$([0, 1])$ and $Y \sim$ Uniform$([0, X])$. Find $\mathbb{E}(Y)$, $\mathbb{L}(X \mid Y)$, and $\mathbb{E}(X \mid Y)$.

A common technique: the law of total expectation is often used to find the value of $\mathbb{E}(Y)$ when Y is dependent on some other X.

o. The MMSE of jointly Gaussian random variables equals the LLSE. Let X, Y, Z be i.i.d. as standard normals. Find $\mathbb{E}(X \mid X+Y, X+Z, Y-Z)$, considering redundancy and symmetry.

Finally, we will briefly consider the two following conditional definitions.

Definition 3 (Conditional variance; conditional covariance).

The natural functions $y \mapsto \text{var}(X \mid Y = y)$ and $z \mapsto \text{cov}(X, Y \mid Z = z)$ define

$$\begin{aligned} \text{var}(X \mid Y) &= \mathbb{E}\big((X - \mathbb{E}(X \mid Y))^2 \mid Y\big) \\ &= \mathbb{E}(X^2 \mid Y) - \mathbb{E}(X \mid Y)^2. \\ \text{cov}(X, Y \mid Z) &= \mathbb{E}((X - \mathbb{E}(X \mid Z)) \cdot (Y - \mathbb{E}(Y \mid Z)) \mid Z) \\ &= \mathbb{E}(XY \mid Z) - \mathbb{E}(X \mid Z) \cdot \mathbb{E}(Y \mid Z). \end{aligned}$$

Proposition 3 (Law of total variance).

"The variance of X is the expected variance of X given all sorts of information about Y, plus on average, how much uncertainty still remains in X given we know Y."

$$\text{var}(X) = \mathbb{E}(\text{var}(X \mid Y)) + \text{var}(\mathbb{E}(X \mid Y)).$$

Proposition 4 (Law of total covariance*).

Interpreted similarly as the law of total variance.

$$\text{cov}(X, Y) = \mathbb{E}(\text{cov}(X, Y \mid Z)) + \text{cov}(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)).$$

4 Formal definition of conditional expectation*

Here, we will explain a more involved definition of conditional expectation you may find in other contexts: what issues does it address, and how is the extra generality useful?

Definition 4 (Conditional expectation II).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X : \Omega \to \mathbb{R}^n$ be a real-valued random variable with finite expectation, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then the **conditional expectation** of X given $\mathcal{G}$ is a $\mathcal{G}$-measurable random variable $\mathbb{E}(X \mid \mathcal{G}) : \Omega \to \mathbb{R}^n$ so that for every $G \in \mathcal{G}$,

$$\int_G \mathbb{E}(X \mid \mathcal{G})\, d\,\mathbb{P} = \int_G X\, d\,\mathbb{P}\,.$$

We recall, or learn, the following result:

Proposition 5 (Function of a random variable).

A random variable Y is a measurable function $f(X)$ of X iff Y is $\sigma(X)$-measurable.

In probability theory, we are not working with completely arbitrary functions of random variables, but rather *measurable* functions, which have the same nature as random variables. The result tells us that the σ-algebra generated by X,

$$\sigma(X) := \left\{X^{-1}(B) : B \in \Sigma\right\}$$

truly captures the idea of "all information given by X," so that Y is fully dependent on X iff its distribution is fully determined by $\sigma(X)$.

Also, Y being $\sigma(X)$-measurable is equivalent to $\sigma(Y) \subseteq \sigma(X)$: the minimal collection of events $\sigma(Y)$ needed to "measure all probabilities of Y" are all known if X is given. For example, every random variable is $\mathcal{F}$-measurable, and every indicator $\mathbb{1}_A$ is $\{\varnothing, A, A^c, \mathcal{F}\}$-measurable.

In short, $\mathbb{E}(X \mid \mathcal{G})$ being a $\mathcal{G}$-measurable random variable means that its distribution is determined by the events in $\mathcal{G}$. As a special case of Definition 4,

Definition 5 (Conditional expectation III).

$\mathbb{E}(X \mid Y) := \mathbb{E}(X \mid \sigma(Y)).$

So, defining measurability addressed the impreciseness of "arbitrary" functions of random variables, and using σ-algebras to encode "information" means that conditioning on events and random variables alike fall out as special cases of this more powerful theory.

Let us now consider the defining property of $\mathbb{E}(X \mid \mathcal{G})$. Recall that in Definition 2,

$$\mathbb{E}(X \mid G) \cdot \mathbb{P}(G) = \mathbb{E}(X|_G) = \int_\Omega X \cdot \mathbb{1}_G \, d\,\mathbb{P} = \int_G X \, d\,\mathbb{P}$$

for any event G, such as $G = \{Y \in C\}$. Now, we can note that since $\mathbb{E}(X \mid G)$ is constant,

$$\int_G \mathbb{E}(X \mid G) \, d\,\mathbb{P} = \mathbb{P}(G) \cdot \mathbb{E}(X \mid G) = \mathbb{E}(X|_G) = \int_G X \, d\,\mathbb{P}$$

which is precisely the defining property of $\mathbb{E}(X \mid \mathcal{G})$. On every event $G \in \mathcal{G}$, the random variable $\mathbb{E}(X \mid \mathcal{G})$ should act exactly like the restriction $\mathbb{E}(X|_G)$ normalized by $\mathbb{P}(G)$.

In the Hilbert space $\mathcal{H}$, we also find that Definition 4 is the same as the orthogonality principle:

$$\langle X - \mathbb{E}(X \mid \mathcal{G}), \mathbb{1}_G \rangle = 0 \quad \forall G \in \mathcal{G}.$$

Why might we care about the indicator functions $\{\mathbb{1}_G : G \in \mathcal{G}\}$? Consider the special case where $\mathcal{G} = \sigma(Y)$. The expectations of $\mathcal{G}$-measurable random variables are really (differences of suprema of) finite linear combinations of indicators of disjoint events, which is really not too farfetched:

$$\begin{aligned}\mathbb{E}(Y^2) &= 1 \cdot \mathbb{P}(Y = 1) + 4 \cdot \mathbb{P}(Y = 2) + \cdots + n^2 \cdot \mathbb{P}(Y = n) \\ &= \mathbb{E}\big(1 \cdot \mathbb{1}\{Y = 1\} + 4 \cdot \mathbb{1}\{Y = 2\} + \cdots + n^2 \cdot \mathbb{1}\{Y = n\}\big)\end{aligned}$$

for a random variable Y taking finitely many values. So, the indicators of events $\mathbb{1}_G$ informally span the $\mathcal{G}$-measurable random variables, and they carry the orthogonality of $X - \mathbb{E}(X \mid \mathcal{G})$ to the whole subspace of functions.

A major problem with Definition 4 is that it fails to actually construct any version of $\mathbb{E}(X \mid \mathcal{G})$. Fortunately, we have the following result.

Theorem 1 (Hilbert projection theorem).

For every element X and for every nonempty, closed, convex subset $\mathcal{C}$ of a Hilbert space $\mathcal{H}$, there exists a unique element $X^* \in \mathcal{C}$ such that

$$X^* = \operatorname*{argmin}_{C \in \mathcal{C}} \|X - C\|^2 .$$

Moreover, if $\mathcal{C}$ is also a subspace of $\mathcal{H}$, then X^* is the unique vector in $\mathcal{H}$ such that $X - X^*$ is orthogonal to $\mathcal{C}$.

We can simply take $\mathcal{C}$ to be the space of all functions of Y, so that $X^* = \mathbb{E}(X \mid Y)$ exists and is unique in the Hilbert space of random variables.

This section was only a quick primer on the measure-theoretic definition of conditional expectation, so don't worry if the explanations seem obtuse at the moment: understanding will naturally come with developing proper foundations and repeated exposure.

■