The Hilbert Space of Random Variables

EECS 126 (UC Berkeley)

Spring 2019

1 Outline

Fix a probability space and consider the set

$$\mathcal{H} := \{X : X \text{ is a real-valued random variable with } \mathbb{E}[X^2] < \infty\}.$$

There are natural notions of *length* and *orthogonality* for objects in $\mathcal{H}$, which allow us to work with random variables *geometrically*, as if they were vectors in Euclidean space. Such a space is known as a *Hilbert space*.

Geometric reasoning leads to an insightful view of mean square error estimation as a *projection* onto an appropriate space.

First, we will review linear algebra and explain what it means for $\mathcal{H}$ to be a Hilbert space. Then, we will study projections in detail and solve the constrained optimization problem of finding the closest point on a linear space to a given point. Using the ideas of projection and orthogonality, we will derive the linear least squares estimator (LLSE). We will then extend the ideas to the non-linear case, arriving at the minimum mean square error (MMSE) estimator.

2 Vector Spaces

A **(real) vector space** V is a collection of objects, including a zero vector $0 \in V$, equipped with two operations, **vector addition** (which allows us to add two vectors $u, v \in V$ to obtain another vector $u + v \in V$) and **scalar multiplication** (which allows us to "scale" a vector $v \in V$ by a real number c to obtain a vector cv), satisfying the following axioms: for all $u, v, w \in V$ and all $a, b \in \mathbb{R}$,

- vector addition is associative, commutative, and 0 is the identity element, that is, $u + (v + w) = (u + v) + w$, $u + v = v + u$, and $u + 0 = u$;

- scalar multiplication is compatible with vector operations: $1v = v$, $a(bv) = (ab)v$, $a(u + v) = au + av$, and $(a + b)u = au + bu$.

It is not important to memorize all of the axioms; however, it is important to recognize that the axioms capture a lot of useful mathematical structures, including the space of random variables, which is why linear algebra plays a key role in many disciplines.

To gain intuition for these axioms, the most natural example of a vector space is $\mathbb{R}^n$, for any positive integer n. The space $\mathbb{R}^n$ consists of n-tuples of real numbers. Vector addition and scalar multiplication are defined componentwise:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),$$
$$c(x_1, \ldots, x_n) = (cx_1, \ldots, cx_n).$$

Given a set $S \subseteq V$, the **span** of S is the set of vectors we can reach from vectors in S using a finite number of vector addition and scalar multiplication operations:

$$\operatorname{span} S := \{c_1 v_1 + \cdots + c_m v_m : m \in \mathbb{N},\ v_1, \ldots, v_m \in S,\ c_1, \ldots, c_m \in \mathbb{R}\}.$$

We say that an element of $\operatorname{span} S$ is a **linear combination** of the vectors in S. Also, $\operatorname{span} S$ is itself a vector space. Whenever we have a subset $U \subseteq V$ such that U is also a vector space in its own right, then we call U a **subspace** of V.

Notice that if any vector $v \in S$ can be written as a linear combination of the other vectors in S, then we can safely remove v from S without decreasing the span of the vectors, i.e., $\operatorname{span} S = \operatorname{span}(S \setminus \{v\})$. This is because any linear combination using v can be rewritten as a linear combination using only vectors in $S \setminus \{v\}$. From the perspective of figuring out which vectors lie in $\operatorname{span} S$, v is redundant. Thus, we say that S is **linearly independent** if it contains no redundant vectors, i.e., if no vector in S can be written as a linear combination of the other vectors in S.

A set S of vectors which is both linearly independent and has $\operatorname{span} S = V$ is called a **basis** of V. The significance of a basis S is that any element of V can be written as a *unique* linear combination of elements of S. One of the most fundamental results in linear algebra says that for any vector space V, a basis always exists, and moreover, the cardinality of any basis is the same. The size of any basis of V is called the **dimension** of V, denoted $\dim V$.

Here is a fact: *any finite-dimensional vector space is essentially identical to* $\mathbb{R}^n$, which means that $\mathbb{R}^n$ is truly a model vector space. However, in this note, we will have need for infinite-dimensional vector spaces too.

2.1 Inner Product Spaces & Hilbert Spaces

We have promised to talk geometrically, but so far, the definition of an vector space does not have any geometry built into it. For this, we need another definition.

For a real vector space V, a map $\langle \cdot, \cdot \rangle : V \times V \to [0, \infty)$ satisfying, for all $u, v, w \in V$ and $c \in \mathbb{R}$,

- (symmetry) $\langle u, v \rangle = \langle v, u \rangle$,
- (linearity) $\langle u + cv, w \rangle = \langle u, w \rangle + c\langle v, w \rangle$, and
- (positive definiteness) $\langle u, u \rangle > 0$ if $u \neq 0$

is called a **(real) inner product** on V. Then, V along with the map $\langle \cdot, \cdot \rangle$ is called a **(real) inner product space**. Note that combining symmetry and linearity gives us linearity in the second argument too, $\langle u, v + cw \rangle = \langle u, v \rangle + c\langle u, w \rangle$.

The familiar inner product on Euclidean space $\mathbb{R}^n$ is $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$, also sometimes called the dot product.

The first bit of geometry that the inner product gives us is a **norm** map

$$\|\cdot\| : V \to [0, \infty), \qquad \text{given by} \qquad \|v\| := \sqrt{\langle v, v \rangle}.$$

By analogy to Euclidean space, we can consider the norm to be the *length* of a vector.

The second bit of geometry is the notion of an *angle* θ between vectors u and v, which we can define via the formula $\langle u, v \rangle = \|u\|\|v\| \cos\theta$. We are only interested in the case when $\cos\theta = 0$, which tells us when u and v are orthogonal. Precisely, we say that u and v are **orthogonal** if $\langle u, v \rangle = 0$.

Now, it is your turn! Do the following exercise.

Exercise 1. Prove that $\langle X, Y \rangle := \mathbb{E}[XY]$ makes $\mathcal{H}$ into a real inner product space. (*Hint*: You must first show that $\mathcal{H}$ is a real vector space, which requires $\mathcal{H}$ to be closed under vector addition, i.e., if $X, Y \in \mathcal{H}$, then $X + Y \in \mathcal{H}$. For this, use the Cauchy-Schwarz inequality, which says that for random variables X and Y, $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$.)

To motivate the definition of the inner product given above, first consider the case when the probability space Ω is finite. Then, $\mathbb{E}[XY] = \sum_{\omega \in \Omega} X(\omega)Y(\omega)\mathbb{P}(\omega)$, which bears resemblance to the Euclidean inner product $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$. However, $\mathbb{E}[XY]$ is a sum in which we *weight* each sample point $\omega \in \Omega$ by its probability, which makes sense in a probabilistic context. In the case when X and Y have joint density f, then $\mathbb{E}[XY] = \int_{-\infty}^{\infty} xyf(x, y)\,\mathrm{d}x\,\mathrm{d}y$, which is again similar but with an integral

replacing the summation and $f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ standing in as the "probability" of the point (x,y).

Finally, we are not quite at the definition of a Hilbert space yet. A *(real) Hilbert space* is a real inner product space which satisfies an additional analytic property called *completeness*, which we will not describe (for this, you will have to take a course in functional analysis).

If Ω is finite, then $\mathcal{H}$ is finite-dimensional. Indeed, a basis is given by the indicators $\{\mathbb{1}_\omega\}_{\omega\in\Omega}$. However, in general $\mathcal{H}$ is infinite-dimensional, and we will soon run into analytical issues which obscure the core ideas. *Therefore, from this point forward, we will behave as if $\mathcal{H}$ were finite-dimensional, when in reality it is not.*

3 Projection

Now that we know that $\mathcal{H}$ is a Hilbert space, we would like to apply our knowledge to the problem of *estimating* a random variable $Y \in \mathcal{H}$. Clearly, if we could directly observe Y, then estimation would not be a difficult problem. However, we are often not given direct access to Y, and we are only allowed to observe some other random variable X which is correlated with Y. The problem is then to find the best estimator of Y, if we restrict ourselves to using only functions of our observation X. Even still, finding the best function of X might be computationally prohibitive, and it may be desired to only use linear functions, i.e., functions of the form $a + bX$ for $a, b \in \mathbb{R}$. Notice that "linear functions of the form $a + bX$" can also be written as $\operatorname{span}\{1, X\}$, a subspace of V, so we may formulate our problem more generally as the following:

Given $y \in V$ and a subspace $U \subseteq V$, find the closest point $x \in U$ to y.

The answer will turn out to be the *projection* of y onto the subspace U. We will explain this concept now.

Given a set $S \subseteq V$, we define the **orthogonal complement** of S, denoted $S^\perp$:

$$S^\perp := \{v \in V : \langle u, v\rangle = 0 \text{ for all } u \in S\}.$$

That is, $S^\perp$ is the set of vectors which are orthogonal to everything in S. Check for yourself that $S^\perp$ is a subspace.

Given a subspace $U \subseteq V$, what does it mean to "project" y onto U? To get a feel for the idea, imagine a slanted pole in broad daylight. One might say that the shadow it casts on the ground is a "projection" of the streetlight onto the ground. From this visualization, you might also realize that there are different types of projections, depending on the location of the sun in the sky. The projection we are interested in is

the shadow cast when the sun is directly overhead, because this projection minimizes the distance from the tip of the pole to the tip of the shadow; this is known as an orthogonal projection.

Formally, the **orthogonal projection** onto a subspace U is the map $P : V \to U$ such that $Py := \arg\min_{x \in U} \|y - x\|$. In words, given an input y, Py is the closest point in U to y. We claim that Py satisfies the following two conditions (see Figure 1):

$$Py \in U \qquad \text{and} \qquad y - Py \in U^{\perp}. \tag{1}$$

Why? Suppose that (1) holds. Then, for any $x \in U$, since $Py - x \in U$,

$$\begin{aligned} \|y - x\|^2 &= \|y - Py + Py - x\|^2 = \|y - Py\|^2 + 2\langle y - Py, Py - x\rangle + \|Py - x\|^2 \\ &= \|y - Py\|^2 + \|Py - x\|^2 \geq \|y - Py\|^2, \end{aligned}$$

with equality if and only if $x = Py$, i.e., Py is the minimizer of $\|y - x\|^2$ over $x \in U$.

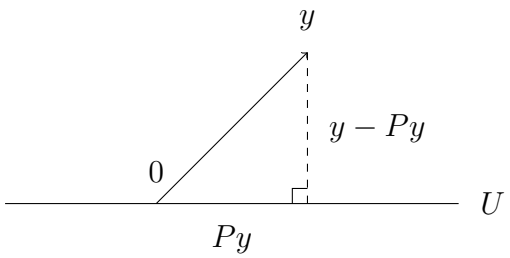

Figure 1: $y - Py$ is orthogonal to U.

We now invite you to further explore the properties of P.

Exercise 2. (a) A map $T : V \to V$ is called a **linear transformation** if for all $u, v \in V$ and all $c \in \mathbb{R}$, $T(u + cv) = Tu + cTv$. Prove that P is a linear transformation. (*Hint*: Apply the same method of proof used above.)

(b) Suppose that U is finite-dimensional, $n := \dim U$, with basis $\{v_i\}_{i=1}^n$. Suppose that the basis is **orthonormal**, that is, the vectors are pairwise orthogonal and $\|v_i\| = 1$ for each $i = 1, \ldots, n$. Show that $Py = \sum_{i=1}^n \langle y, v_i\rangle v_i$. (Note: If we take $U = \mathbb{R}^n$ with the standard inner product, then P can be represented as a matrix in the form $P = \sum_{i=1}^n v_i v_i^{\mathsf{T}}$.)

3.1 Gram-Schmidt Process

Let us see what (1) tells us in the case when we have a finite basis $\{v_i\}_{i=1}^n$ for U. The condition $Py \in U$ says that Py is a linear combination $\sum_{i=1}^n c_i v_i$ of the basis $\{v_i\}_{i=1}^n$. The condition $y - Py \in U^\perp$ is equivalent to saying that $y - Py$ is orthogonal to v_i for $i = 1, \ldots, n$. These two conditions gives us a system of equations which we can, in principle, solve:

$$\left\langle y - \sum_{i=1}^n c_i v_i, v_j \right\rangle = 0, \qquad j = 1, \ldots, n. \tag{2}$$

However, what if the basis $\{v_i\}_{i=1}^n$ is orthonormal, as in Exercise 2(b)? Then, the computation of Py is simple: for each $i = 1, \ldots, n$, $\langle y, v_i \rangle$ gives the component of the projection in the direction v_i.

Fortunately, there is a simple procedure for taking a basis and converting it into an orthonormal basis. It is known as the **Gram-Schmidt process**. The algorithm is iterative: at step $j \in \{1, \ldots, n\}$, we will have an orthonormal set of vectors $\{u_i\}_{i=1}^j$ so that $\text{span}\{u_i\}_{i=1}^j = \text{span}\{v_i\}_{i=1}^j$. To start, take $u_1 := v_1/\|v_1\|$. Now, at step $j \in \{1, \ldots, n-1\}$, consider $P_{u_1,\ldots,u_j} v_{j+1}$, where $P_{u_1,\ldots,u_j}$ is the orthogonal projection onto $\text{span}\{u_1, \ldots, u_j\}$. Because of (1), we know that $w_{j+1} := v_{j+1} - P_{u_1,\ldots,u_j} v_{j+1}$ lies in $(\text{span}\{u_i\}_{i=1}^j)^\perp$, so that w_{j+1} is orthogonal to $u_1, \ldots, u_j$. Also, we know that $w_{j+1} \neq 0$ because if $v_{j+1} = P_{u_1,\ldots,u_j} v_{j+1}$, then $v_{j+1} \in \text{span}\{u_1, \ldots, u_j\}$, but this is impossible since we started with a basis. Thus, we may take $u_{j+1} := w_{j+1}/\|w_{j+1}\|$ and add it to our orthonormal set.

Computation of the Gram-Schmidt process is not too intensive because subtracting the projection onto $\text{span}\{u_i\}_{i=1}^j$ only requires computing the projection onto an orthonormal basis. From Exercise 2(b), we can explicitly describe the procedure:

1. Let $u_1 := v_1/\|v_1\|$.
2. For $j = 1, \ldots, n-1$:
   (a) Set $w_{j+1} := v_{j+1} - \sum_{i=1}^j \langle v_{j+1}, u_i \rangle u_i$.
   (b) Set $u_{j+1} := w_{j+1}/\|w_{j+1}\|$.

4 Linear Least Squares Estimation (LLSE)

Finally, we are ready to solve the problem of linear least squares estimation. Formally, the problem is:

Given $X, Y \in \mathcal{H}$, minimize $\mathbb{E}[(Y-a-bX)^2]$ over $a, b \in \mathbb{R}$. The solution to this problem is called the **linear least squares estimator (LLSE)**.

Using our previous notation, $\mathbb{E}[(Y-a-bX)^2] = \|Y-a-bX\|^2$ and thus the solution is $\arg\min_{\hat{Y}\in U}\|Y-\hat{Y}\|^2$, where U is the subspace $U = \operatorname{span}\{1, X\}$. We have already solved this problem! If we apply (2) directly, then we obtain the equations:

$$\begin{aligned}\mathbb{E}[Y-a-bX] &= 0,\\ \mathbb{E}[(Y-a-bX)X] &= 0.\end{aligned}$$

We can solve for a and b. Alternatively, we can apply the Gram-Schmidt process to $\{1, X\}$ (assuming X is not constant) to convert it to the basis $\{1, (X-\mathbb{E}[X])/\sqrt{\operatorname{var} X}\}$. Now, applying Exercise 2(b),

$$L[Y \mid X] := \mathbb{E}[Y] + \mathbb{E}\Big[Y\Big(\frac{X-\mathbb{E}[X]}{\sqrt{\operatorname{var} X}}\Big)\Big]\frac{X-\mathbb{E}[X]}{\sqrt{\operatorname{var} X}} = \mathbb{E}[Y] + \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}(X-\mathbb{E}[X]).$$

This is a nice result, so let us package it so:

Theorem 1 (LLSE)**.** *For $X, Y \in \mathcal{H}$, where X is not a constant, the LLSE of Y given X is*

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

Furthermore, the squared error of the LLSE is

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

Proof. Since we have already proven the formula for the LLSE, let us prove the second assertion, geometrically. Notice that both sides of the equation for the squared error are unaffected if we replace X and Y with $X-\mathbb{E}[X]$ and $Y-\mathbb{E}[Y]$, so we may assume that X and Y are zero mean. Now consider Figure 2. The angle at 0 is

$$\cos\theta = \frac{\langle X, Y\rangle}{\|X\|\|Y\|}.$$

Thus, from geometry,

$$\begin{aligned}\mathbb{E}[(Y-L[Y \mid X])^2] = \|Y-L[Y \mid X]\|^2 &= \|Y\|^2(\sin\theta)^2 = \|Y\|^2\Big(1-\frac{\langle X,Y\rangle^2}{\|X\|^2\|Y\|^2}\Big)\\ &= \operatorname{var} Y - \frac{\operatorname{cov}(X,Y)^2}{\operatorname{var} X}.\end{aligned}$$

□

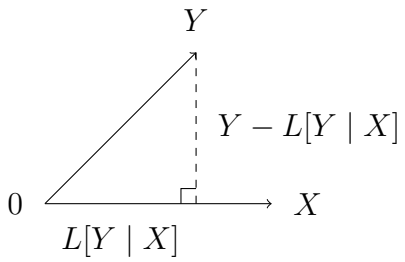

Figure 2: The variance of the error of the LLSE can be found geometrically.

4.1 Orthogonal Updates

Given $X, Y, Z \in \mathcal{H}$, how do we calculate $L[Y \mid X, Z]$? Here is where the Gram-Schmidt process truly shines. The key idea in the Gram-Schmidt algorithm is to take a new basis vector and subtract the orthogonal projection of the vector onto the previous basis vectors (the part where we normalize the vector is just icing on the cake). The point is that we can compute the projection onto an orthogonal basis, *one component* at a time (by Exercise 2(b)).

Now, the "basis vectors" are interpreted as new *observations* in the context of estimation. The projection of Z onto X is $L[Z \mid X]$, so the new orthogonal observation is $Z - L[Z \mid X]$. This is called the **innovation**, because it represents the new information that was not already predictable from the previous observations. Now, since X and $Z - L[Z \mid X]$ are orthogonal, we have the following:

Theorem 2 (Orthogonal LLSE Update)**.** *Let $X, Y, Z \in \mathcal{H}$, where X and Z are not constant. Then, $L[Y \mid X, Z] = L[Y \mid X] + L[Y \mid \tilde{Z}]$, where $\tilde{Z} := Z - L[Z \mid X]$.*

This observation is crucial for the design of online algorithms, because it tells us how to *recursively* update our LLSE after collecting new observations. We will revisit this topic when we discuss tracking and the Kalman filter.

4.2 Non-Linear Estimation

Since we have placed such an emphasis on the techniques of linear algebra, you may perhaps think that quadratic estimation might be far harder, but this is not true. If we are looking for the best $a, b, c \in \mathbb{R}$ to minimize $\mathbb{E}[(Y - a - bX - cX^2)]$, then this is again the projection of Y onto the subspace $\text{span}\{1, X, X^2\}$, so the methods we developed apply to this situation as well. In general, we can easily handle polynomial regression of degree d, for any positive integer d, by projecting Y

onto span$\{1, X, \ldots, X^d\}$. The equations become more difficult to solve and require knowledge of higher moments of X and Y (in fact, we must now have $X^d \in \mathcal{H}$, i.e., $\mathbb{E}[X^{2d}] < \infty$), but nonetheless it is possible.

The same is true as long as we are projecting onto a linearly independent set of random variables. The same techniques can be used to compute the best linear combination of 1, $\sin X$, and $\cos X$ to estimate Y.

4.3 Vector Case

[1]

More generally, what if we have n observations, where n is a positive integer, and we would like to calculate $L[Y \mid X_1, \ldots, X_n]$? We could try to apply the orthogonal update method, but if we are not interested in an online algorithm, then we may as well try to solve the problem once and for all if we can.

First, assume that all random variables are centered. Let Σ_X denote the covariance matrix of $(X_1, \ldots, X_n)$: $\Sigma_X := \mathbb{E}[XX^\mathsf{T}]$. Since Σ_X is symmetric, it has a decomposition $\Sigma_X = U\Lambda U^\mathsf{T}$ by the spectral theorem, where U is an orthogonal matrix of eigenvectors ($U^\mathsf{T}U = I$) and Λ is a diagonal matrix of real eigenvalues. Furthermore, Σ_X is positive semi-definite: for any $v \in \mathbb{R}^n$,

$$v^\mathsf{T}\Sigma_X v = \mathbb{E}[v^\mathsf{T}(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\mathsf{T}v] = \mathbb{E}\left[\left((X - \mathbb{E}[X])^\mathsf{T}v\right)^2\right] \geq 0.$$

Thus, Σ_X has only non-negative eigenvalues; we will assume that Σ_X is invertible so that it has strictly positive eigenvalues (it is positive definite). Therefore, Λ has a real square root $\Lambda^{1/2}$ defined by $(\Lambda^{1/2})_{i,i} := \sqrt{\Lambda_{i,i}}$ for each $i = 1, \ldots, n$. Now, notice that the covariance of the random vector $Z := \Lambda^{-1/2}U^\mathsf{T}X$ is $\Lambda^{-1/2}U^\mathsf{T}(U\Lambda U^\mathsf{T})U\Lambda^{-1/2} = I$, so the components of Z are orthonormal. Moreover, $L[Y \mid X] = L[Y \mid Z]$ because left multiplication by $\Lambda^{-1/2}U^\mathsf{T}$ is an invertible linear map (with inverse $y \mapsto U\Lambda^{1/2}y$).

After performing this transformation,

$$\begin{aligned} L[Y \mid Z] &= \sum_{i=1}^{n} \langle Y, Z_i \rangle Z_i = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} (\Lambda^{-1/2}U^\mathsf{T})_{i,j}(\Lambda^{-1/2}U^\mathsf{T})_{i,k}\langle Y, X_j\rangle X_k \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} (\Sigma_X^{-1})_{j,k}\langle Y, X_j\rangle X_k = \Sigma_{Y,X}\Sigma_X^{-1}X, \end{aligned}$$

[1] This section requires more advanced linear algebra concepts. Do not worry if you do not yet have the required linear algebra background to understand all of the arguments. They are presented here to offer a different perspective on the material; see also the treatment in Walrand's textbook, which is more direct.

where $\Sigma_{Y,X} := \mathbb{E}[YX^\mathsf{T}]$.

What if we wish to predict more than one observation, i.e., $Y = (Y_1, \ldots, Y_m)$ where m is a positive integer? Then, the LLSE of the vector Y is the vector whose components are the estimates of the corresponding components of Y, so

$$L[(Y_1, \ldots, Y_m) \mid X] = (L[Y_1 \mid X], \ldots, L[Y_m \mid X]);$$

the resulting formula is still $L[Y \mid X] = \Sigma_{Y,X}\Sigma_X^{-1}X$.

We can also calculate the squared error of the LLSE. To prepare for the derivation, recall that for two matrices A and B of compatible dimensions (for positive integers m and n, A is $m \times n$ and B is $n \times m$), then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Also, since the trace is a *linear* function of the entries of the matrix, by linearity of expectation, the trace commutes with expectation.

$$\begin{aligned}\mathbb{E}[\|Y - L[Y \mid X]\|_2^2] &= \mathbb{E}[(Y - \Sigma_{Y,X}\Sigma_X^{-1}X)^\mathsf{T}(Y - \Sigma_{Y,X}\Sigma_X^{-1}X)] \\ &= \mathbb{E}[Y^\mathsf{T}Y - (\Sigma_{Y,X}\Sigma_X^{-1}X)^\mathsf{T}Y] = \mathrm{tr}(\Sigma_Y - \Sigma_{Y,X}\Sigma_X^{-1}\Sigma_{X,Y}),\end{aligned}$$

where $\Sigma_Y := \mathbb{E}[YY^\mathsf{T}]$ and $\Sigma_{X,Y} = \Sigma_{Y,X}^\mathsf{T} = \mathbb{E}[XY^\mathsf{T}]$. In the first line, we use the fact that $Y - L[Y \mid X]$ is orthogonal to $L[Y \mid X]$. In the second line, we use

$$\begin{aligned}\mathbb{E}[Y^\mathsf{T}Y - (\Sigma_{Y,X}\Sigma_X^{-1}X)^\mathsf{T}Y] &= \mathbb{E}\big[\mathrm{tr}\big(Y^\mathsf{T}Y - (\Sigma_{Y,X}\Sigma_X^{-1}X)^\mathsf{T}Y\big)\big] \\ &= \mathbb{E}[\mathrm{tr}(YY^\mathsf{T} - YX^\mathsf{T}\Sigma_X^{-1}\Sigma_{X,Y})] \\ &= \mathrm{tr}\,\mathbb{E}[YY^\mathsf{T} - YX^\mathsf{T}\Sigma_X^{-1}\Sigma_{X,Y}] \\ &= \mathrm{tr}(\Sigma_Y - \Sigma_{Y,X}\Sigma_X^{-1}\Sigma_{X,Y}).\end{aligned}$$

4.4 Non-Bayesian Perspective: Linear Regression

So far, the method of estimation that we have developed is a *Bayesian* method because it assumes that we have knowledge of the distributions of X and Y. Here, we will explain how one can study the non-Bayesian perspective of *regression*, which can be formulated without any mention of probabilities at all, using the results we have already obtained.

Let n and d be positive integers, where n represents the number of samples collected and d represents the number of *features*. Typically the data is organized into a $n \times d$ matrix $\mathbf{X}$ called the **design matrix**, where the entry in the ith row and jth column (for $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$) is the value of the jth feature for the ith data point. Thus, the *rows* of the design matrix contain the feature values for a single data point, and we typically denote the rows by x_i^T, for $i = 1, \ldots, n$.

We are also given an $n \times 1$ *observation vector* y. We may assume for simplicity that the design matrix and observation vector have been *centered*, i.e., subtract appropriate matrices from $\mathbf{X}$ and y such that for each $j \in \{1, \ldots, d\}$, $\sum_{i=1}^{n} \mathbf{X}_{i,j} = 0$, and $\sum_{i=1}^{n} y_i = 0$. The problem is the following:

Find the $d \times 1$ *weight vector* β to minimize the sum of squares $\|y - \mathbf{X}\beta\|_2^2$.

In other words, we would like to estimate y using a linear combination of the features, where β represents the weight we place on each feature.

To study the problem of regression from the Bayesian perspective, let X be a $d \times 1$ random vector and Y be a scalar random variable with joint distribution

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

In other words, (X, Y) represents a uniformly randomly chosen row of the design matrix and observation vector. Now, for $\beta \in \mathbb{R}^d$, observe:

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

Hence, *finding the weight vector β that minimizes the sum of squared residuals in the non-Bayesian formulation is the same as finding the weight vector β such that* $L[Y \mid X] = \beta^\mathsf{T} X$. However, we already know that $L[Y \mid X] = \Sigma_{Y,X}\Sigma_X^{-1} X$, so the solution is given by $\beta = \Sigma_X^{-1}\Sigma_{X,Y}$. Moreover,

$$\Sigma_X = \mathbb{E}[XX^\mathsf{T}] = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\mathsf{T} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T},$$
$$\Sigma_{Y,X} = \mathbb{E}[YX^\mathsf{T}] = \frac{1}{n}\sum_{i=1}^{n} y_i x_i^\mathsf{T} = \frac{1}{n} y^\mathsf{T}\mathbf{X}.$$

Therefore, $\beta = (\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}^\mathsf{T} y$ and the optimal estimate is $\hat{y} := \mathbf{X}\beta = \mathbf{X}(\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}^\mathsf{T} y$.

This is not the last word on regression models. If one assumes that for each $i = 1, \ldots, n$, $y_i = x_i^\mathsf{T}\beta + \varepsilon_i$ for a true parameter vector β and $(\varepsilon_i)_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, then there is a well-developed theory that describes how to find the best estimator $\hat{\beta}$, the distribution and expected mean squared error of $\hat{\beta}$, confidence intervals for $\hat{\beta}$, etc.

5 Minimum Mean Square Estimation (MMSE)

We will now drop the restriction to linear functions of X or linear functions of observations $X_1, \ldots, X_n$, and we will instead find the best *arbitrary* function of X to estimate Y.

> Given $X, Y \in \mathcal{H}$, find the best function ϕ to minimize $\mathbb{E}[(Y - \phi(X))^2]$. The solution to this problem is called the **minimum mean square error (MMSE) estimator**.

As before, we can write down the orthogonality condition:

$$Y - \phi(X) \text{ should be orthogonal to all other functions of } X.$$

Unlike the case of linear estimation, where we looked at the span of a finite number of random variables, we are now looking at the projection onto the subspace of all functions of X, which is quite difficult to visualize or even imagine. In fact, you might wonder whether such a function ϕ is always guaranteed to exist. The answer is *yes*, such a function exists and is essentially unique, although the details are technical.

The **conditional expectation** of Y given X is formally defined as the function of X, denoted $\mathbb{E}(Y \mid X)$, such that for all bounded continuous functions ϕ,

$$\mathbb{E}\big[\big(Y - \mathbb{E}(Y \mid X)\big)\phi(X)\big] = 0. \tag{3}$$

It is the solution to our estimation problem.

Interestingly, although we started out by looking only at random variables in $\mathcal{H}$, the definition (3) does not require X and Y to be in $\mathcal{H}$. Even if they have infinite second moments, as long as they have a well-defined first moment we can still define the conditional expectation. A full exploration of the definition, properties, and applications of conditional expectation would take up another note by itself, so we will stop here.