4.11: Vector Spaces of Random Variables

Basic Theory

Many of the concepts in this chapter have elegant interpretations if we think of real-valued random variables as vectors in a vector space. In particular, variance and higher moments are related to the concept of norm and distance, while covariance is related to inner product. These connections can help unify and illuminate some of the ideas in the chapter from a different point of view. Of course, real-valued random variables are simply measurable, real-valued functions defined on the sample space, so much of the discussion in this section is a special case of our discussion of function spaces in the chapter on Distributions, but recast in the notation of probability.

As usual, our starting point is a random experiment modeled by a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Thus, Ω is the set of outcomes, $\mathscr{F}$ is the σ-algebra of events, and $\mathbb{P}$ is the probability measure on the sample space $(\Omega, \mathscr{F})$. Our basic *vector space* $\mathscr{V}$ consists of all real-valued random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$ (that is, defined for the experiment). Recall that random variables X_1 and X_2 are equivalent if $\mathbb{P}(X_1 = X_2) = 1$, in which case we write $X_1 \equiv X_2$. We consider two such random variables as the same vector, so that technically, our vector space consists of *equivalence classes* under this equivalence relation. The *addition operator* corresponds to the usual addition of two real-valued random variables, and the operation of *scalar multiplication* corresponds to the usual multiplication of a real-valued random variable by a real (non-random) number. These operations are compatible with the equivalence relation in the sense that if $X_1 \equiv X_2$ and $Y_1 \equiv Y_2$ then $X_1 + Y_1 \equiv X_2 + Y_2$ and $c X_1 \equiv c X_2$ for $c \in \mathbb{R}$. In short, the vector space $\mathscr{V}$ is well-defined.

Norm

Suppose that $k \in [1, \infty)$. The k *norm* of $X \in \mathscr{V}$ is defined by

$$\|X\|_k = \left[\mathbb{E}\left(|X|^k\right)\right]^{1/k} \tag{4.11.1}$$

Thus, $\|X\|_k$ is a measure of the size of X in a certain sense, and of course it's possible that $\|X\|_k = \infty$. The following theorems establish the fundamental properties. The first is the *positive property*.

Suppose again that $k \in [1, \infty)$. For $X \in \mathscr{V}$,

1. $\|X\|_k \geq 0$
2. $\|X\|_k = 0$ if and only if $\mathbb{P}(X = 0) = 1$ (so that $X \equiv 0$).

Proof

These results follow from the basic inequality properties of expected value. First $|X|^k \geq 0$ with probability 1, so $\mathbb{E}\left(|X|^k\right) \geq 0$. In addition, $\mathbb{E}\left(|X|^k\right) = 0$ if and only if $\mathbb{P}(X = 0) = 1$.

The next result is the *scaling property*.

Suppose again that $k \in [1, \infty)$. Then $\|cX\|_k = |c| \, \|X\|_k$ for $X \in \mathscr{V}$ and $c \in \mathbb{R}$.

Proof

$$\|cX\|_k = [\mathbb{E}\left(|cX|^k\right]^{1/k} = \left[\mathbb{E}\left(|c|^k |X|^k\right)\right]^{1/k} = \left[|c|^k \mathbb{E}\left(|X|^k\right)\right]^{1/k} = |c| \left[\mathbb{E}\left(|X|^k\right)\right]^{1/k} = |c| \, \|X\|_k \tag{4.11.2}$$

The next result is *Minkowski's inequality*, named for Hermann Minkowski, and also known as the *triangle inequality*.

Suppose again that $k \in [1, \infty)$. Then $\|X + Y\|_k \leq \|X\|_k + \|Y\|_k$ for $X, Y \in \mathscr{V}$.

Proof

The first quadrant $S = \left\{(x, y) \in \mathbb{R}^2 : x \geq 0, \; y \geq 0\right\}$ is a convex set and $g(x, y) = \left(x^{1/k} + y^{1/k}\right)^k$ is concave on S. From Jensen's inequality, if U and V are nonnegative random variables, then

$$\mathbb{E}\left[(U^{1/k} + V^{1/k})^k\right] \leq \left([\mathbb{E}(U)]^{1/k} + [\mathbb{E}(V)]^{1/k}\right)^k \tag{4.11.3}$$

Letting $U = |X|^k$ and $V = |Y|^k$ and simplifying gives the result. To show that g really is concave on S, we can compute the second partial derivatives. Let $h(x, y) = x^{1/k} + y^{1/k}$ so that $g = h^k$. Then

$$g_{xx} = \frac{k-1}{k} h^{k-2} x^{1/k-2} \left(x^{1/k} - h\right) \tag{4.11.4}$$

$$g_{yy} = \frac{k-1}{k} h^{k-2} y^{1/k-2} \left(y^{1/k} - h\right) \tag{4.11.5}$$

$$g_{xy} = \frac{k-1}{k} h^{k-2} x^{1/k-1} y^{1/k-1} \tag{4.11.6}$$

Clearly $h(x,y) \geq x^{1/k}$ and $h(x,y) \geq y^{1/k}$ for $x \geq 0$ and $y \geq 0$, so $g_{xx}$and g_{yy}, the diagonal entries of the second derivative matrix, are nonpositive on S. A little algebra shows that the determinant of the second derivative matrix $g_{xx} g_{yy} - g_{xy}^2 = 0$ on S. Thus, the second derivative matrix of g is negative semi-definite.

It follows from the last three results that the set of random variables (again, modulo equivalence) with finite k norm forms a *subspace* of our parent vector space $\mathscr{V}$, and that the k norm really is a norm on this vector space.

For $k \in [1, \infty)$, $\mathscr{L}_k$ denotes the vector space of $X \in \mathscr{V}$ with $\|X\|_k < \infty$, and with norm $\|\cdot\|_k$.

In analysis, p is often used as the index rather than k as we have used here, but p seems too much like a probability, so we have broken with tradition on this point. The $\mathscr{L}$ is in honor of Henri Lebesgue, who developed much of this theory. Sometimes, when we need to indicate the dependence on the underlying σ-algebra $\mathscr{F}$, we write $\mathscr{L}_k(\mathscr{F})$. Our next result is *Lyapunov's inequality*, named for Aleksandr Lyapunov. This inequality shows that the k-norm of a random variable is increasing in k.

Suppose that $j,\ k \in [1, \infty)$ with $j \leq k$. Then $\|X\|_j \leq \|X\|_k$ for $X \in \mathscr{V}$.

Proof

Note that $S = \{x \in \mathbb{R} : x \geq 0\}$ is convex and $g(x) = x^{k/j}$ is convex on S. From Jensen's inequality, if U is a nonnegative random variable then $[\mathbb{E}(U)]^{k/j} \leq \mathbb{E}\left(U^{k/j}\right)$. Letting $U = |X|^j$ and simplifying gives the result.

Lyapunov's inequality shows that if $1 \leq j \leq k$ and $\|X\|_k < \infty$ then $\|X\|_j < \infty$. Thus, $\mathscr{L}_k$ is a subspace of $\mathscr{L}_j$.

Metric

The k norm, like any norm on a vector space, can be used to define a metric, or distance function; we simply compute the norm of the difference between two vectors.

For $k \in [1, \infty)$, the k *distance* (or k *metric*) between $X, Y \in \mathscr{V}$ is defined by

$$d_k(X,Y) = \|X - Y\|_k = \left[\mathbb{E}\left(|X - Y|^k\right)\right]^{1/k} \tag{4.11.7}$$

The following properties are analogous to the properties in norm properties (and thus very little additional work is required for the proofs). These properties show that the k metric really is a metric on $\mathscr{L}_k$ (as always, modulo equivalence). The first is the *positive property*.

Suppose again that $k \in [1, \infty)$ $X,\ Y \in \mathscr{V}$. Then

1. $d_k(X,Y) \geq 0$
2. $d_k(X,Y) = 0$ if and only if $\mathbb{P}(X = Y) = 1$ (so that $X \equiv Y$ and Y).

Proof

These results follow directly from the positive property.

Next is the obvious *symmetry property*:

$d_k(X,Y) = d_k(Y,X)$ for $X,\ Y \in \mathscr{V}$.

Next is the distance version of the *triangle inequality*.

$d_k(X,Z) \leq d_k(X,Y) + d_k(Y,Z)$ for $X,\ Y,\ Z \in \mathscr{V}$

Proof

From Minkowski's inequality,

$$d_k(X,Z) = \|X - Z\|_k = \|(X - Y) + (Y - Z)\|_k \leq \|X - Y\|_k + \|Y - Z\|_k = d_k(X,Y) + d_k(Y,Z) \tag{4.11.8}$$

The last three properties mean that d_k is indeed a metric on $\mathscr{L}_k$ for $k \geq 1$. In particular, note that the standard deviation is simply the 2-distance from X to its mean $\mu = \mathbb{E}(X)$:

$$\text{sd}(X) = d_2(X, \mu) = \|X - \mu\|_2 = \sqrt{\mathbb{E}\left[(X - \mu)^2\right]} \tag{4.11.9}$$

and the variance is the square of this. More generally, the kth moment of X about a is simply the kth power of the k-distance from X to a. The 2-distance is especially important for reasons that will become clear below, in the discussion of inner product. This distance is also called the *root mean square distance*.

Center and Spread Revisited

Measures of center and measures of spread are best thought of together, in the context of a *measure of distance*. For a real-valued random variable X, we first try to find the constants $t \in \mathbb{R}$ that are closest to X, as measured by the given distance; any such t is a *measure of center* relative to the distance. The minimum distance itself is the corresponding *measure of spread*.

Let us apply this procedure to the 2-distance.

For $X \in \mathscr{L}_2$, define the *root mean square* error function by

$$d_2(X, t) = \|X - t\|_2 = \sqrt{\mathbb{E}\left[(X - t)^2\right]}, \quad t \in \mathbb{R} \tag{4.11.10}$$

For $X \in \mathscr{L}_2$, $d_2(X, t)$ is minimized when $t = \mathbb{E}(X)$ and the minimum value is $\text{sd}(X)$.

Proof

Note that the minimum value of $d_2(X, t)$ occurs at the same points as the minimum value of $d_2^2(X, t) = \mathbb{E}\left[(X - t)^2\right]$ (this is the *mean square* error function). Expanding and taking expected values term by term gives

$$\mathbb{E}\left[(X - t)^2\right] = \mathbb{E}\left(X^2\right) - 2t\mathbb{E}(X) + t^2 \tag{4.11.11}$$

This is a quadratic function of t and hence the graph is a parabola opening upward. The minimum occurs at $t = \mathbb{E}(X)$, and the minimum value is $\text{var}(X)$. Hence the minimum value of $t \mapsto d_2(X, t)$ also occurs at $t = \mathbb{E}(X)$ and the minimum value is $\text{sd}(X)$.

We have seen this computation several times before. The best constant predictor of X is $\mathbb{E}(X)$, with mean square error $\text{var}(X)$. The physical interpretation of this result is that the moment of inertia of the mass distribution of X about t is minimized when $t = \mu$, the center of mass. Next, let us apply our procedure to the 1-distance.

For $X \in \mathscr{L}_1$, define the *mean absolute error* function by

$$d_1(X, t) = \|X - t\|_1 = \mathbb{E}\left[|X - t|\right], \quad t \in \mathbb{R} \tag{4.11.12}$$

We will show that $d_1(X, t)$ is minimized when t is any median of X. (Recall that the set of medians of X forms a closed, bounded interval.) We start with a discrete case, because it's easier and has special interest.

Suppose that $X \in \mathscr{L}_1$ has a discrete distribution with values in a finite set $S \subseteq \mathbb{R}$. Then $d_1(X, t)$ is minimized when t is any median of X.

Proof

Note first that $\mathbb{E}(|X - t|) = \mathbb{E}(t - X,\ X \leq t) + \mathbb{E}(X - t,\ X > t)$. Hence $\mathbb{E}(|X - t|) = a_t\, t + b_t$, where $a_t = 2\,\mathbb{P}(X \leq t) - 1$ and where $b_t = \mathbb{E}(X) - 2\,\mathbb{E}(X,\ X \leq t)$. Note that $\mathbb{E}(|X - t|)$ is a continuous, piecewise linear function of t, with corners at the values in S. That is, the function is a *linear spline*. Let m be the smallest median of X. If $t < m$ and $t \notin S$, then the slope of the linear piece at t is negative. Let M be the largest median of X. If $t > M$ and $t \notin S$, then the slope of the linear piece at t is positive. If $t \in (m, M)$ then the slope of the linear piece at t is 0. Thus $\mathbb{E}(|X - t|)$ is minimized for every t in the median interval $[m, M]$.

The last result shows that mean absolute error has a couple of basic deficiencies as a measure of error:

- The function may not be smooth (differentiable).
- The function may not have a unique minimizing value of t.

Indeed, when X does not have a unique median, there is no compelling reason to choose one value in the median interval, as the measure of center, over any other value in the interval.

Suppose now that $X \in \mathscr{L}_1$ has a general distribution on $\mathbb{R}$. Then $d_1(X, t)$ is minimized when t is any median of X.

Proof

Let $s,\ t \in \mathbb{R}$. Suppose first that $s < t$. Computing the expected value over the events $X \leq s$, $s < X \leq t$, and $X \geq t$, and simplifying gives

$$\mathbb{E}(|X-t|) = \mathbb{E}(|X-s|) + (t-s)\,[2\,\mathbb{P}(X \le s) - 1] + 2\,\mathbb{E}(t - X,\ s < X \le t) \tag{4.11.13}$$

Suppose next that $t < s$. Using similar methods gives

$$\mathbb{E}(|X-t|) = \mathbb{E}(|X-s|) + (t-s)\,[2\,\mathbb{P}(X < s) - 1] + 2\,\mathbb{E}(X - t,\ t \le X < s) \tag{4.11.14}$$

Note that the last terms on the right in these equations are nonnegative. If we take s to be a median of X, then the middle terms on the right in the equations are also nonnegative. Hence if s is a median of X and t is any other number then $\mathbb{E}(|X-t|) \ge \mathbb{E}(|X-s|)$.

Convergence

Whenever we have a measure of distance, we automatically have a criterion for convergence.

Suppose that $X_n \in \mathscr{L}_k$ for $n \in \mathbb{N}_+$ and that $X \in \mathscr{L}_k$, where $k \in [1, \infty)$. Then $X_n \to X$ as $n \to \infty$ in *kth mean* if $X_n \to X$ as $n \to \infty$ in the vector space $\mathscr{L}_k$. That is,

$$d_k(X_n, X) = \|X_n - X\|_k \to 0 \text{ as } n \to \infty \tag{4.11.15}$$

or equivalently $\mathbb{E}\left(|X_n - X|^k\right) \to 0$ as $n \to \infty$.

When $k = 1$, we simply say that $X_n \to X$ as $n \to \infty$ *in mean*; when $k = 2$, we say that $X_n \to X$ as $n \to \infty$ *in mean square*. These are the most important special cases.

Suppose that $1 \le j \le k$. If $X_n \to X$ as $n \to \infty$ in kth mean then $X_n \to X$ as $n \to \infty$ in jth mean.

Proof

This follows from Lyanpuov's inequality. Note that $0 \le d_j(X_n, X) \le d_k(X_n, X) \to 0$ as $n \to \infty$.

Convergence in kth mean implies that the k norms converge.

Suppose that $X_n \in \mathscr{L}_k$ for $n \in \mathbb{N}_+$ and that $X \in \mathscr{L}_k$, where $k \in [1, \infty)$. If $X_n \to X$ as $n \to \infty$ in kth mean then $\|X_n\|_k \to \|X\|_k$ as $n \to \infty$. Equivalently, if $\mathbb{E}(|X_n - X|^k) \to 0$ as $n \to \infty$ then $\mathbb{E}(|X_n|^k) \to \mathbb{E}(|X|^k)$ as $n \to \infty$.

Proof

This is a simple consequence of the reverse triangle inequality, which holds in any normed vector space. The general result is that if a sequence of vectors in a normed vector space converge then the norms converge. In our notation here,

$$|\|X_n\|_k - \|X\|_k| \le \|X_n - X\|_k \tag{4.11.16}$$

so if the right side converges to 0 as $n \to \infty$, then so does the left side.

The converse is not true; a counterexample is given below. Our next result shows that convergence in mean is stronger than convergence in probability.

Suppose that $X_n \in \mathscr{L}_1$ for $n \in \mathbb{N}_+$ and that $X \in \mathscr{L}_1$. If $X_n \to X$ as $n \to \infty$ in mean, then $X_n \to X$ as $n \to \infty$ in probability.

Proof

This follows from Markov's inequality. For $\epsilon > 0$, $0 \le \mathbb{P}(|X_n - X| > \epsilon) \le \mathbb{E}(|X_n - X|) / \epsilon \to 0$ as $n \to \infty$.

The converse is not true. That is, convergence with probability 1 does not imply convergence in kth mean; a counterexample is given below. Also convergence in kth mean does not imply convergence with probability 1; a counterexample to this is given below. In summary, the implications in the various modes of convergence are shown below; no other implications hold in general.

- Convergence with probability 1 implies convergence in probability.
- Convergence in kth mean implies convergence in jth mean if $j \le k$.
- Convergence in kth mean implies convergence in probability.
- Convergence in probability implies convergence in distribution.

However, the next section on uniformly integrable variables gives a condition under which convergence in probability implies convergence in mean.

Inner Product

The vector space $\mathscr{L}_2$ of real-valued random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ (modulo equivalence of course) with finite second moment is special, because it's the only one in which the norm corresponds to an inner product.

The *inner product* of $X, Y \in \mathscr{L}_2$ is defined by

$$\langle X, Y \rangle = \mathbb{E}(XY) \tag{4.11.17}$$

The following results are analogous to the basic properties of covariance, and show that this definition really does give an inner product on the vector space

For X, Y, $Z \in \mathscr{L}_2$ and $a \in \mathbb{R}$,

1. $\langle X, Y \rangle = \langle Y, X \rangle$, the *symmetric property*.
2. $\langle X, X \rangle \geq 0$ and $\langle X, X \rangle = 0$ if and only if $\mathbb{P}(X = 0) = 1$ (so that $X \equiv 0$), the *positive property*.
3. $\langle aX, Y \rangle = a \langle X, Y \rangle$, the *scaling property*.
4. $\langle X + Y, Z \rangle = \langle X, Z \rangle + \langle Y, Z \rangle$, the *additive property*.

Proof

1. This property is trivial from the definition.
2. Note that $\mathbb{E}(X^2) \geq 0$ and $\mathbb{E}(X^2) = 0$ if and only if $\mathbb{P}(X = 0) = 1$.
3. This follows from the scaling property of expected value: $\mathbb{E}(aXY) = a\mathbb{E}(XY)$
4. This follows from the additive property of expected value: $\mathbb{E}[(X + Y)Z] = \mathbb{E}(XZ) + \mathbb{E}(YZ)$.

From parts (a), (c), and (d) it follows that inner product is *bi-linear*, that is, linear in each variable with the other fixed. Of course bi-linearity holds for any inner product on a vector space. Covariance and correlation can easily be expressed in terms of this inner product. The covariance of two random variables is the inner product of the corresponding *centered* variables. The correlation is the inner product of the corresponding standard scores.

For X, $Y \in \mathscr{L}_2$,

1. $\cov(X, Y) = \langle X - \mathbb{E}(X), Y - \mathbb{E}(Y) \rangle$
2. $\cor(X, Y) = \left\langle [X - \mathbb{E}(X)] / \sd(X), [Y - \mathbb{E}(Y)] / \sd(Y) \right\rangle$

Proof

1. This is simply a restatement of the definition of covariance.
2. This is a restatement of the fact that the correlation of two variables is the covariance of their corresponding standard scores.

Thus, real-valued random variables X and Y are uncorrelated if and only if the centered variables $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$ are *perpendicular* or *orthogonal* as elements of $\mathscr{L}_2$.

For $X \in \mathscr{L}_2$, $\langle X, X \rangle = \|X\|_2^2 = \mathbb{E}\left(X^2\right)$.

Thus, the norm associated with the inner product is the 2-norm studied above, and corresponds to the *root mean square* operation on a random variable. This fact is a fundamental reason why the 2-norm plays such a special, honored role; of all the k-norms, only the 2-norm corresponds to an inner product. In turn, this is one of the reasons that *root mean square difference* is of fundamental importance in probability and statistics. Technically, the vector space $\mathscr{L}_2$ is a *Hilbert space*, named for David Hilbert.

The next result is *Hölder's inequality*, named for Otto Hölder.

Suppose that j, $k \in [1, \infty)$ and $\frac{1}{j} + \frac{1}{k} = 1$. For $X \in \mathscr{L}_j$ and $Y \in \mathscr{L}_k$,

$$\langle |X|, |Y| \rangle \leq \|X\|_j \|Y\|_k \tag{4.11.18}$$

Proof

Note that $S = \left\{(x, y) \in \mathbb{R}^2 : x \geq 0,\ y \geq 0\right\}$ is a convex set and $g(x, y) = x^{1/j} y^{1/k}$ is concave on S. From Jensen's inequality, if U and V are nonnegative random variables then $\mathbb{E}\left(U^{1/j} V^{1/k}\right) \leq [\mathbb{E}(U)]^{1/j} [\mathbb{E}(V)]^{1/k}$. Substituting $U = |X|^j$ and $V = |Y|^k$ gives the result.

To show that g really is concave on S, we compute the second derivative matrix:

$$\begin{bmatrix} (1/j)(1/j - 1) x^{1/j - 2} y^{1/k} & (1/j)(1/k) x^{1/j - 1} y^{1/k - 1} \\ (1/j)(1/k) x^{1/j - 1} y^{1/k - 1} & (1/k)(1/k - 1) x^{1/j} y^{1/k - 2} \end{bmatrix} \tag{4.11.19}$$

Since $1/j < 1$ and $1/k < 1$, the diagonal entries are negative on S. The determinant simplifies to

$$(1/j)(1/k) x^{2/j - 2} y^{2/k - 2} [1 - (1/j + 1/k)] = 0 \tag{4.11.20}$$

In the context of the last theorem, j and k are called *conjugate exponents*. If we let $j = k = 2$ in Hölder's inequality, then we get the *Cauchy-Schwarz inequality*, named for Augustin Cauchy and Karl Schwarz: For X, $Y \in \mathscr{L}_2$,

$$\mathbb{E}(|X||Y|) \le \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)} \tag{4.11.21}$$

In turn, the Cauchy-Schwarz inequality is equivalent to the basic inequalities for covariance and correlations: For $X, Y \in \mathscr{L}_2$,

$$|\text{cov}(X,Y)| \le \text{sd}(X)\text{sd}(Y), \quad |\text{cor}(X,Y)| \le 1 \tag{4.11.22}$$

If $j, k \in [1, \infty)$ are conjugate exponents then

1. $k = \frac{j}{j-1}$.
2. $k \downarrow 1$ as $j \uparrow \infty$.

The following result is an equivalent to the identity $\text{var}(X+Y) + \text{var}(X-Y) = 2[\text{var}(X) + \text{var}(Y)]$ that we studied in the section on covariance and correlation. In the context of vector spaces, the result is known as the *parallelogram rule*:

If $X, Y \in \mathscr{L}_2$ then

$$\|X+Y\|_2^2 + \|X-Y\|_2^2 = 2\|X\|_2^2 + 2\|Y\|_2^2 \tag{4.11.23}$$

Proof

This result follows from the bi-linearity of inner product:

$$\|X+Y\|_2^2 + \|X-Y\|_2^2 = \langle X+Y, X+Y\rangle + \langle X-Y, X-Y\rangle \tag{4.11.24}$$

$$= (\langle X, X\rangle + 2\langle X, Y\rangle + \langle Y, Y\rangle) + (\langle X, X\rangle - 2\langle X, Y\rangle + \langle Y, Y\rangle) = 2\|X\|^2 + 2\|Y\|^2 \tag{4.11.25}$$

The following result is equivalent to the statement that the variance of the sum of uncorrelated variables is the sum of the variances, which again we proved in the section on covariance and correlation. In the context of vector spaces, the result is the famous *Pythagorean theorem*, named for Pythagoras of course.

If $(X_1, X_2, \ldots, X_n)$ is a sequence of random variables in $\mathscr{L}_2$ with $\langle X_i, X_j\rangle = 0$ for $i \ne j$ then

$$\left\|\sum_{i=1}^n X_i\right\|_2^2 = \sum_{i=1}^n \|X_i\|_2^2 \tag{4.11.26}$$

Proof

Again, this follows from the bi-linearity of inner product:

$$\left\|\sum_{i=1}^n X_i\right\|_2^2 = \left\langle \sum_{i=1}^n X_i, \sum_{j=1}^n X_j \right\rangle = \sum_{i=1}^n \sum_{j=1}^n \langle X_i, X_j\rangle \tag{4.11.27}$$

The terms with $i \ne j$ are 0 by the orthogonality assumption, so

$$\left\|\sum_{i=1}^n X_i\right\|_2^2 = \sum_{i=1}^n \langle X_i, X_i\rangle = \sum_{i=1}^n \|X_i\|_2^2 \tag{4.11.28}$$

Projections

The best linear predictor studied in the section on covariance and correlation and conditional expected values have nice interpretation in terms of projections onto subspaces of $\mathscr{L}_2$. First let's review the concepts. Recall that $\mathscr{U}$ is a *subspace* of $\mathscr{L}_2$ if $\mathscr{U} \subseteq \mathscr{L}_2$ and $\mathscr{U}$ is also a vector space (under the same operations of addition and scalar multiplication). To show that $\mathscr{U} \subseteq \mathscr{L}_2$ is a subspace, we just need to show the *closure properties* (the other axioms of a vector space are inherited).

- If $U, V \in \mathscr{U}$ then $U + V \in \mathscr{U}$.
- If $U \in \mathscr{U}$ and $c \in \mathbb{R}$ then $cU \in \mathscr{U}$.

Suppose now that $\mathscr{U}$ is a subspace of $\mathscr{L}_2$ and that $X \in \mathscr{L}_2$. Then the *projection* of X onto $\mathscr{U}$ (if it exists) is the vector $V \in \mathscr{U}$ with the property that $X - V$ is perpendicular to $\mathscr{U}$:

$$\langle X - V, U\rangle = 0, \quad U \in \mathscr{U} \tag{4.11.29}$$

The projection has two critical properties: It is unique (if it exists) and it is the vector in $\mathscr{U}$ closest to X. If you look at the proofs of these results, you will see that they are essentially the same as the ones used for the best predictors of X mentioned at the beginning of this subsection. Moreover, the proofs use only vector space concepts—the fact that our vectors are random variables on a probability space plays no special role.

The projection of X onto $\mathscr{U}$ (if it exists) is unique.

Proof

Suppose that V_1 and V_2 satisfy the definition. then

$$\|V_1 - V_2\|_2^2 = \langle V_1 - V_2, V_1 - V_2 \rangle = \langle V_1 - X + X - V_2, V_1 - V_2 \rangle = \langle V_1 - X, V_1 - V_2 \rangle + \langle X - V_2, V_1 - V_2 \rangle = 0 \tag{4.11.30}$$

Hence $V_1 \equiv V_2$. The last equality in the displayed equation holds by assumption and the fact that $V_1 - V_2 \in \mathscr{U}$

Suppose that V is the projection of X onto $\mathscr{U}$. Then

1. $\|X - V\|_2^2 \leq \|X - U\|_2^2$ for all $U \in \mathscr{U}$.
2. Equality holds in (a) if and only if $U \equiv V$

Proof

1. If $U \in \mathscr{U}$ then

$$\|X - U\|_2^2 = \|X - V + V - U\|_2^2 = \|X - V\|_2^2 + 2\langle X - V, V - U \rangle + \|V - U\|_2^2 \tag{4.11.31}$$

   But the middle terms is 0 so

$$\|X - U\|_2^2 = \|X - V\|_2^2 + \|V - U\|_2^2 \geq \|X - V\|_2^2 \tag{4.11.32}$$

2. Equality holds if and only if $\|V - U\|_2^2 = 0$, if and only if $V \equiv U$.

Now let's return to our study of best predictors of a random variable.

If $X \in \mathscr{L}_2$ then the set $\mathscr{W}_X = \{a + bX : a \in \mathbb{R},\ b \in \mathbb{R}\}$ is a subspace of $\mathscr{L}_2$. In fact, it is the subspace *generated* by X and 1.

Proof

Note that $\mathscr{W}_X$ is the set of all *linear combinations* of the vectors 1 and X. If U, $V \in \mathscr{W}_X$ then $U + V \in \mathscr{W}_X$. If $U \in \mathscr{W}_X$ and $c \in \mathbb{R}$ then $cU \in \mathscr{W}_X$.

Recall that for X, $Y \in \mathscr{L}_2$, the best linear predictor of Y based on X is

$$L(Y \mid X) = \mathbb{E}(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)}[X - \mathbb{E}(X)] \tag{4.11.33}$$

Here is the meaning of the predictor in the context of our vector spaces.

If X, $Y \in \mathscr{L}_2$ then $L(Y \mid X)$ is the projection of Y onto $\mathscr{W}_X$.

Proof

Note first that $L(Y \mid X) \in \mathscr{W}_X$. Thus, we just need to show that $Y - L(Y \mid X)$ is perpendicular to $\mathscr{W}_X$. For this, it suffices to show

1. $\langle Y - L(Y \mid X), X \rangle = 0$
2. $\langle Y - L(Y \mid X), 1 \rangle = 0$

We have already done this in the earlier sections, but for completeness, we do it again. Note that $\mathbb{E}(X[X - \mathbb{E}(X)]) = \text{var}(X)$. Hence $\mathbb{E}[XL(Y \mid X)] = \mathbb{E}(X)\mathbb{E}(Y) + \text{cov}(X, Y) = \mathbb{E}(XY)$. This gives (a). By linearity, $\mathbb{E}[L(Y \mid X)] = \mathbb{E}(Y)$ so (b) holds as well.

The previous result is actually just the random variable version of the standard formula for the projection of a vector onto a space spanned by two other vectors. Note that 1 is a unit vector and that $X_0 = X - \mathbb{E}(X) = X - \langle X, 1 \rangle 1$ is perpendicular to 1. Thus, $L(Y \mid X)$ is just the sum of the projections of Y onto 1 and X_0:

$$L(Y \mid X) = \langle Y, 1 \rangle 1 + \frac{\langle Y, X_0 \rangle}{\langle X_0, X_0 \rangle} X_0 \tag{4.11.34}$$

Suppose now that $\mathscr{G}$ is a sub σ-algebra of $\mathscr{F}$. Of course if $X : \Omega \to \mathbb{R}$ is $\mathscr{G}$-measurable then X is $\mathscr{F}$-measurables, so $\mathscr{L}_2(\mathscr{G})$ is a subspace of $\mathscr{L}_2(\mathscr{F})$.

If $X \in \mathscr{L}_2(\mathscr{F})$ then $\mathbb{E}(X \mid \mathscr{G})$ is the projection of X onto $\mathscr{L}_2(\mathscr{G})$.

Proof

This is essentially the definition of $\mathbb{E}(X \mid \mathscr{G})$ as the only (up to equivalence) random variable in $\mathscr{L}_2(\mathscr{G})$ with $\mathbb{E}[\mathbb{E}(X \mid \mathscr{G})U] = \mathbb{E}(XU)$ for every $U \in \mathscr{L}_2(\mathscr{G})$.

But remember that $\mathbb{E}(X \mid \mathscr{G})$ is defined more generally for $X \in \mathscr{L}_1(\mathscr{F})$. Our final result in this discussion concerns convergence.

Suppose that $k \in [1, \infty)$ and that $\mathscr{G}$ is a sub σ-algebra of $\mathscr{F}$.

1. If $X \in \mathscr{L}_k(\mathscr{F})$ then $\mathbb{E}(X \mid \mathscr{G}) \in \mathscr{L}_k(\mathscr{G})$
2. If $X_n \in \mathscr{L}_k(\mathscr{F})$ for $n \in \mathbb{N}_+$, $X \in \mathscr{L}_k(\mathscr{F})$, and $X_n \to X$ as $n \to \infty$ in $\mathscr{L}_k(\mathscr{F})$ then $\mathbb{E}(X_n \mid \mathscr{G}) \to \mathbb{E}(X \mid \mathscr{G})$ as $n \to \infty$ in $\mathscr{L}_k(\mathscr{G})$

Proof

1. Note that $|\mathbb{E}(X \mid \mathscr{G})| \le \mathbb{E}(|X| \mid \mathscr{G})$. Since $t \mapsto t^k$ is increasing and convex on $[0, \infty)$ we have

$$|\mathbb{E}(X \mid \mathscr{G})|^k \le [\mathbb{E}(|X| \mid \mathscr{G})]^k \le \mathbb{E}\left(|X|^k \mid \mathscr{G}\right) \tag{4.11.35}$$

   The last step uses Jensen's inequality. Taking expected values gives

$$\mathbb{E}[|\mathbb{E}(X \mid \mathscr{G})|^k] \le \mathbb{E}(|X|^k) < \infty \tag{4.11.36}$$

2. Using the same ideas,

$$\mathbb{E}\left[\left|\mathbb{E}(X_n \mid \mathscr{G}) - \mathbb{E}(X \mid \mathscr{G})\right|^k\right] = \mathbb{E}\left[\left|\mathbb{E}(X_n - X \mid \mathscr{G})\right|^k\right] \le E[|X_n - X|^k] \tag{4.11.37}$$

   By assumption, the right side converges to 0 as $n \to \infty$ and hence so does the left side.

Examples and Applications

App Exercises

In the error function app, select the root mean square error function. Click on the x-axis to generate an empirical distribution, and note the shape and location of the graph of the error function.

In the error function app, select the mean absolute error function. Click on the x-axis to generate an empirical distribution, and note the shape and location of the graph of the error function.

Computational Exercises

Suppose that X is uniformly distributed on the interval $[0, 1]$.

1. Find $\|X\|_k$ for $k \in [1, \infty)$.
2. Graph $\|X\|_k$ as a function of $k \in [1, \infty)$.
3. Find $\lim_{k \to \infty} \|X\|_k$.

Answer

1. $\frac{1}{(k+1)^{1/k}}$
3. 1

Suppose that X has probability density function $f(x) = \frac{a}{x^{a+1}}$ for $1 \le x < \infty$, where $a > 0$ is a parameter. Thus, X has the Pareto distribution with shape parameter a.

1. Find $\|X\|_k$ for $k \in [1, \infty)$.
2. Graph $\|X\|_k$ as a function of $k \in (1, a)$.
3. Find $\lim_{k \uparrow a} \|X\|_k$.

Answer

1. $\left(\frac{a}{a-k}\right)^{1/k}$ if $k < a$, ∞ if $k \ge a$
3. ∞

Suppose that (X, Y) has probability density function $f(x, y) = x + y$ for $0 \le x \le 1$, $0 \le y \le 1$. Verify Minkowski's inequality.

Answer

1. $\|X + Y\|_k = \left(\frac{2^{k+2} - 2}{(k+2)(k+3)}\right)^{1/k}$
2. $\|X\|_k + \|Y\|_k = 2\left(\frac{1}{k+2} + \frac{1}{2(k+1)}\right)^{1/k}$

Let X be an indicator random variable with $\mathbb{P}(X = 1) = p$, where $0 \le p \le 1$. Graph $\mathbb{E}(|X - t|)$ as a function of $t \in \mathbb{R}$ in each of the cases below. In each case, find the minimum value of the function and the values of t where the minimum occurs.

1. $p < \frac{1}{2}$
2. $p = \frac{1}{2}$
3. $p > \frac{1}{2}$

Answer

1. The minimum is p and occurs at $t = 0$.
2. The minimum is $\frac{1}{2}$ and occurs for $t \in [0, 1]$
3. The minimum is $1 - p$ and occurs at $t = 1$

Suppose that X is uniformly distributed on the interval $[0, 1]$. Find $d_1(X, t) = \mathbb{E}(|X - t|)$ as a function of t and sketch the graph. Find the minimum value of the function and the value of t where the minimum occurs.

Suppose that X is uniformly distributed on the set $[0, 1] \cup [2, 3]$. Find $d_1(X, t) = \mathbb{E}(|X - t|)$ as a function of t and sketch the graph. Find the minimum value of the function and the values of t where the minimum occurs.

Suppose that (X, Y) has probability density function $f(x, y) = x + y$ for $0 \le x \le 1$, $0 \le y \le 1$. Verify Hölder's inequality in the following cases:

1. $j = k = 2$
2. $j = 3$, $k = \frac{3}{2}$

Answer

1. $\|X\|_2 \|Y\|_2 = \frac{5}{12}$
2. $\|X\|_3 + \|Y\|_{3/2} \approx 0.4248$

Counterexamples

The following exercise shows that convergence with probability 1 does not imply convergence in mean.

Suppose that $(X_1, X_2, \ldots)$ is a sequence of independent random variables with

$$\mathbb{P}\left(X = n^3\right) = \frac{1}{n^2}, \; \mathbb{P}(X_n = 0) = 1 - \frac{1}{n^2}; \quad n \in \mathbb{N}_+ \tag{4.11.38}$$

1. $X_n \to 0$ as $n \to \infty$ with probability 1.
2. $X_n \to 0$ as $n \to \infty$ in probability.
3. $\mathbb{E}(X_n) \to \infty$ as $n \to \infty$.

Proof

1. This follows from the basic characterization of convergence with probability 1: $\sum_{n=1}^\infty \mathbb{P}(X_n > \epsilon) = \sum_{n=1}^\infty 1/n^2 < \infty$ for $0 < \epsilon < 1$.
2. This follows since convergence with probability 1 implies convergence in probability.
3. Note that $\mathbb{E}(X_n) = n^3 / n^2 = n$ for $n \in \mathbb{N}_+$.

The following exercise shows that convergence in mean does not imply convergence with probability 1.

Suppose that $(X_1, X_2, \ldots)$ is a sequence of independent indicator random variables with

$$\mathbb{P}(X_n = 1) = \frac{1}{n}, \; \mathbb{P}(X_n = 0) = 1 - \frac{1}{n}; \quad n \in \mathbb{N}_+ \tag{4.11.39}$$

1. $\mathbb{P}(X_n = 0 \text{ for infinitely many } n) = 1$.
2. $\mathbb{P}(X_n = 1 \text{ for infinitely many } n) = 1$.
3. $\mathbb{P}(X_n \text{ does not converge as } n \to \infty) = 1$.
4. $X_n \to 0$ as $n \to \infty$ in kth mean for every $k \ge 1$.

Proof

1. This follows from the second Borel-Cantelli lemma since $\sum_{n=1}^\infty \mathbb{P}(X_n = 1) = \sum_{n=1}^\infty 1/n = \infty$
2. This also follows from the second Borel-Cantelli lemma since $\sum_{n=1}^\infty \mathbb{P}(X_n = 0) = \sum_{n=1}^\infty (1 - 1/n) = \infty$.
3. This follows from parts (a) and (b).
4. Note that $\mathbb{E}(X_n) = 1/n \to 0$ as $n \to \infty$.

The following exercise show that convergence of the kth means does not imply convergence *in* kth mean.

Suppose that U has the Bernoulli distribution with parmaeter $\frac{1}{2}$, so that $\mathbb{P}(U = 1) = \mathbb{P}(U = 0) = \frac{1}{2}$. Let $X_n = U$ for $n \in \mathbb{N}_+$ and let $X = 1 - U$. Let $k \in [1, \infty)$. Then

1. $\mathbb{E}(X_n^k) = \mathbb{E}(X^k) = \frac{1}{2}$ for $n \in \mathbb{N}_+$, so $\mathbb{E}(X_n^k) \to \mathbb{E}(X^k)$ as $n \to \infty$
2. $\mathbb{E}(|X_n - X|^k) = 1$ for $n \in \mathbb{N}$ so X_n does not converge to X as $n \to \infty$ in $\mathscr{L}_k$.

Proof

1. Note that $X_n^k = U^k = U$ for $n \in \mathbb{N}_+$, since U just takes values 0 and 1. Also, U and $1 - U$ have the same distribution so $\mathbb{E}(U) = \mathbb{E}(1 - U) = \frac{1}{2}$.
2. Note that $X_n - X = U - (1 - U) = 2U - 1$ for $n \in \mathbb{N}_+$. Again, U just takes values 0 and 1, so $|2U - 1| = 1$.