RANDOM VARIABLES AS VECTORS: GEOMETRÍA DE L^2, PROYECCIONES, VARIANZA Y SESGO

NOTAS A PARTIR DE UN TRANSCRIPT DE YOUTUBE

Resumen. Estas notas convierten la intuición del video en una explicación formal. La idea central es que una variable aleatoria cuadrado-integrable puede verse como un vector en un espacio de Hilbert. El valor esperado es la proyección ortogonal sobre el subespacio de variables aleatorias constantes, la varianza es el cuadrado de la longitud residual y la descomposición sesgo-varianza es el teorema de Pitágoras en L^2.

Índice

1. Glosario mínimo 2
2. Problema formal 2
3. Variables aleatorias: no son números al azar, son funciones 2
4. Por qué las variables aleatorias forman un espacio vectorial 3
5. Por qué necesitamos L^2 y no solamente L^1 3
5.1. Qué es L^1 3
5.2. Qué es L^2 3
5.3. Respuesta corta: por qué no L^1 3
6. Por qué el producto interno es $\mathbb{E}[XY]$ 4
7. El espacio de Hilbert $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P})$ 4
8. Qué significa span$\{1\}$ 5
9. La esperanza como proyección ortogonal 5
10. La varianza como longitud residual 6
11. Descomposición sesgo-varianza 6
12. Por qué se resta y se suma la esperanza 7
13. Comparación con regresión lineal 7
14. Mapa entre disciplinas 8
15. Límites y sutilezas 8
16. Resumen simbólico 8
17. Bias–variance en regresión: MSE $= \text{Bias}^2 + \text{Var} + \sigma^2$ 8
17.1. Datos de entrenamiento y el subespacio $\mathcal{H}_D$ 8
17.2. Definiciones 9
17.3. Descomposición ortogonal 9
18. Esperanza condicional como proyección ortogonal 10
18.1. El subespacio generado por Y 10
18.2. Condición de ortogonalidad 11
18.3. Ley de la esperanza total: propiedad torre paso a paso 11
19. Chapman–Kolmogorov como propiedad de torre 13
19.1. Probabilidades como esperanzas de indicadores 13
19.2. Introducir un tiempo intermedio 13
19.3. Uso de la propiedad de Markov 13

Date: 17 de junio de 2026.

19.4. Abrir la esperanza condicional en el caso discreto 14
19.5. Lectura geométrica 14
19.6. Notación con operadores de transición 15
19.7. Resumen conceptual 15
19.8. Sacar lo dado 15
20. Ley de la varianza total como Pitágoras 16
20.1. Demostración geométrica 16
21. Ejemplo: suma aleatoria 17

1. Glosario mínimo

Ω: Espacio muestral: conjunto de resultados posibles de un experimento.

$\mathcal{F}$: Sigma-álgebra de eventos: colección de subconjuntos de Ω a los que podemos asignar probabilidad.

$\mathbb{P}$: Medida de probabilidad: función que asigna a cada evento $A \in \mathcal{F}$ un número $\mathbb{P}(A) \in [0,1]$.

X: Variable aleatoria real: función medible $X : \Omega \to \mathbb{R}$.

$L^p(\Omega, \mathcal{F}, \mathbb{P})$: Espacio de variables aleatorias con $\mathbb{E}|X|^p < \infty$, identificadas salvo igualdad casi segura.

L^2: Espacio de variables aleatorias con segundo momento finito: $\mathbb{E}[X^2] < \infty$.

$\langle X, Y \rangle$: Producto interno en L^2: $\langle X, Y \rangle = \mathbb{E}[XY]$.

$\|X\|$: Norma inducida por el producto interno: $\|X\| = \sqrt{\mathbb{E}[X^2]}$.

span$\{1\}$: Recta vectorial formada por las variables aleatorias constantes: $\{c \cdot 1 : c \in \mathbb{R}\}$.

Proyección ortogonal:
Punto de un subespacio que minimiza la distancia cuadrática al vector original.

Casi seguro:
Una propiedad vale casi seguro si falla sólo en un conjunto de probabilidad cero.

2. Problema formal

Queremos entender geométricamente tres fórmulas:

$$\mathbb{E}[X] = \arg\min_{c \in \mathbb{R}} \mathbb{E}[(X-c)^2], \quad \mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2], \quad \mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2. \tag{1}$$

La interpretación es:

valor esperado = proyección; varianza = longitud residual al cuadrado; sesgo-varianza = Pitágoras.

3. Variables aleatorias: no son números al azar, son funciones

Definición 3.1 (Espacio de probabilidad)**.** Un espacio de probabilidad es una terna

$$(\Omega, \mathcal{F}, \mathbb{P}),$$

donde Ω es el conjunto de resultados, $\mathcal{F}$ es la colección de eventos medibles y $\mathbb{P}$ asigna probabilidades a esos eventos.

Definición 3.2 (Variable aleatoria)**.** Una variable aleatoria real es una función medible

$$X : \Omega \to \mathbb{R}.$$

No es "una variable que cambia al azar"; es una función determinista evaluada sobre un resultado incierto $\omega \in \Omega$.

Ejemplo 3.3 (Dos monedas)**.** Si

$$\Omega = \{HH, HT, TH, TT\},$$

y $X(\omega)$ cuenta el número de caras, entonces

$$X(HH) = 2, \qquad X(HT) = 1, \qquad X(TH) = 1, \qquad X(TT) = 0.$$

La aleatoriedad no está en la función X sino en qué resultado ω ocurre.

4. Por qué las variables aleatorias forman un espacio vectorial

Si $X, Y : \Omega \to \mathbb{R}$ son variables aleatorias y $a, b \in \mathbb{R}$, definimos

$$(aX + bY)(\omega) = aX(\omega) + bY(\omega).$$

Entonces $aX + bY$ también es una variable aleatoria real. Por eso las variables aleatorias se pueden sumar y escalar como vectores.

Proposición 4.1. *El conjunto de variables aleatorias reales medibles sobre* $(\Omega, \mathcal{F}, \mathbb{P})$ *forma un espacio vectorial real bajo suma puntual y multiplicación escalar puntual.*

Demostración. La suma y la multiplicación escalar se definen punto por punto. Las propiedades de espacio vectorial se reducen a las propiedades de los números reales para cada $\omega \in \Omega$. La única sutileza técnica es que las combinaciones lineales de funciones medibles son medibles. □

5. Por qué necesitamos L^2 y no solamente L^1

5.1. Qué es L^1.

Definición 5.1 (L^1)**.** Decimos que $X \in L^1$ si

$$\mathbb{E}|X| < \infty.$$

Esto basta para que el valor esperado $\mathbb{E}[X]$ esté bien definido.

Pero L^1 no basta para hacer geometría euclidiana. En L^1 tenemos una norma

$$\|X\|_1 = \mathbb{E}|X|,$$

pero no hay, en general, un producto interno natural cuya norma sea $\|X\|_1$. Sin producto interno no hay una noción lineal de ángulo, perpendicularidad ni proyección ortogonal.

5.2. Qué es L^2.

Definición 5.2 (L^2)**.** Decimos que $X \in L^2$ si

$$\mathbb{E}[X^2] < \infty.$$

En ese caso definimos

$$\langle X, Y \rangle = \mathbb{E}[XY].$$

El punto crucial es que si $X, Y \in L^2$, entonces $XY \in L^1$ por Cauchy-Schwarz:

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]} < \infty.$$

Así, $\mathbb{E}[XY]$ existe y se puede usar como producto interno.

5.3. Respuesta corta: por qué no L^1. Usamos L^2 porque queremos minimizar errores cuadráticos y hablar de proyecciones ortogonales. Eso requiere un producto interno. En este contexto el producto interno natural es

$$\langle X, Y \rangle = \mathbb{E}[XY].$$

L^1 sirve para esperanza absoluta, pero no da la geometría de Pitágoras que aparece en varianza, MSE y regresión.

6. Por qué el producto interno es $\mathbb{E}[XY]$

En $\mathbb{R}^n$, el producto punto usual es

$$x \cdot y = \sum_{i=1}^{n} x_i y_i.$$

Si los puntos i tienen pesos probabilísticos p_i, la versión ponderada es

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i p_i.$$

Pero si X e Y son funciones sobre $\Omega = \{\omega_1, \ldots, \omega_n\}$, esto se escribe como

$$\langle X, Y \rangle = \sum_{i=1}^{n} X(\omega_i) Y(\omega_i) \mathbb{P}(\{\omega_i\}) = \mathbb{E}[XY].$$

Por eso aparece el valor esperado del producto: es el análogo probabilístico del producto punto ponderado.

Proposición 6.1. *En L^2, la función*

$$\langle X, Y \rangle = \mathbb{E}[XY]$$

es un producto interno, salvo la identificación natural de variables iguales casi seguramente.

Demostración. Para $a, b \in \mathbb{R}$,

$$\langle aX + bY, Z \rangle = \mathbb{E}[(aX + bY)Z] = a\mathbb{E}[XZ] + b\mathbb{E}[YZ].$$

Además,

$$\langle X, Y \rangle = \mathbb{E}[XY] = \mathbb{E}[YX] = \langle Y, X \rangle .$$

Finalmente,

$$\langle X, X \rangle = \mathbb{E}[X^2] \geq 0.$$

Si $\mathbb{E}[X^2] = 0$, entonces $X = 0$ casi seguramente. Por eso, rigurosamente, L^2 identifica variables aleatorias que son iguales excepto en conjuntos de probabilidad cero. □

7. El espacio de Hilbert $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P})$

Definición 7.1 (Espacio de Hilbert)**.** Un espacio de Hilbert es un espacio vectorial con producto interno que es completo respecto a la norma inducida.

En nuestro caso,

$$\mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P}), \qquad \langle X, Y \rangle = \mathbb{E}[XY], \qquad \|X\| = \sqrt{\mathbb{E}[X^2]}.$$

La distancia entre dos variables aleatorias es

$$d(X, Y) = \|X - Y\| = \sqrt{\mathbb{E}[(X - Y)^2]}.$$

Esta distancia es exactamente la raíz del error cuadrático medio entre X e Y.

8. Qué significa span{1}

La notación 1 significa la variable aleatoria constante

$$1(\omega) = 1 \qquad \text{para todo } \omega \in \Omega.$$

Entonces

$$\text{span}\{1\} = \{c \cdot 1 : c \in \mathbb{R}\}.$$

Cada elemento de span{1} es una variable aleatoria constante:

$$(c \cdot 1)(\omega) = c.$$

Geométricamente, span{1} es una recta dentro de L^2. Estadísticamente, es el conjunto de predictores que no dependen del resultado ω.

9. La esperanza como proyección ortogonal

Queremos aproximar X por una constante c. El error cuadrático es

$$g(c) = \mathbb{E}[(X - c)^2].$$

Expandimos:

$$\begin{aligned} g(c) &= \mathbb{E}[X^2 - 2cX + c^2] \\ &= \mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2. \end{aligned}$$

Derivando,

$$g'(c) = -2\mathbb{E}[X] + 2c.$$

La condición $g'(c) = 0$ da

$$c = \mathbb{E}[X].$$

Por tanto,

$$\mathbb{E}[X] = \arg\min_{c \in \mathbb{R}} \mathbb{E}[(X - c)^2].$$

Proposición 9.1 (Esperanza como proyección)**.** *Si $X \in L^2$, entonces la proyección ortogonal de X sobre* span{1} *es*

$$\text{Proj}_{\text{span}\{1\}} X = (\mathbb{E}X)1.$$

Demostración. Un punto genérico de span{1} es $c1$. La proyección debe satisfacer que el residual sea ortogonal al subespacio:

$$X - c1 \perp \text{span}\{1\}.$$

Basta exigir ortogonalidad contra el generador 1:

$$\langle X - c1, 1 \rangle = 0.$$

Como

$$\langle X - c1, 1 \rangle = \mathbb{E}[(X - c)1] = \mathbb{E}[X] - c,$$

obtenemos $c = \mathbb{E}[X]$. □

10. La varianza como longitud residual

El residual después de proyectar X sobre las constantes es

$$X - \mathbb{E}[X].$$

Su longitud en L^2 es

$$\|X - \mathbb{E}[X]\| = \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]} = \sqrt{\mathrm{Var}(X)}.$$

Por tanto,

$$\mathrm{Var}(X) = \|X - \mathbb{E}[X]\|^2 .$$

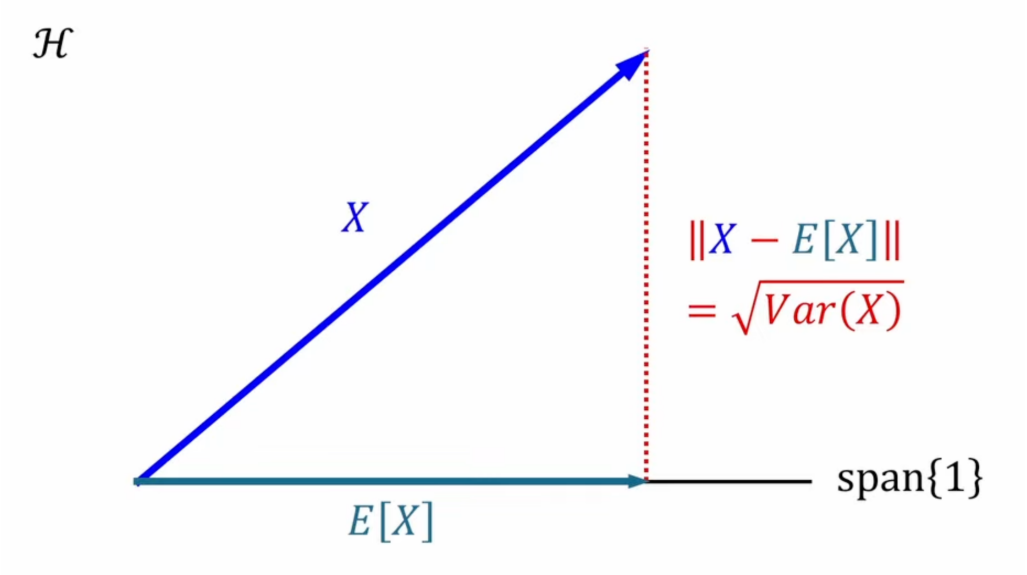

Figura 1. La esperanza es la proyección de X sobre span$\{1\}$; la desviación estándar es la longitud del residual.

11. Descomposición sesgo-varianza

Supongamos que $\theta \in \mathbb{R}$ es un parámetro fijo desconocido y que $\hat{\theta}$ es un estimador, es decir, una variable aleatoria. Como θ es constante, lo identificamos con $\theta 1 \in \mathrm{span}\{1\}$.

Definición 11.1 (Sesgo)**.** El sesgo del estimador $\hat{\theta}$ para θ es

$$\mathrm{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta.$$

Definición 11.2 (Error cuadrático medio)**.** El error cuadrático medio es

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

Descomponemos el error total como

$$\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta).$$

El primer término es aleatorio y tiene media cero. El segundo es constante. Además son ortogonales:

$$\begin{aligned} \left\langle \hat{\theta} - \mathbb{E}[\hat{\theta}], \mathbb{E}[\hat{\theta}] - \theta \right\rangle &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta) \right] \\ &= (\mathbb{E}[\hat{\theta}] - \theta)\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] \\ &= 0. \end{aligned}$$

Por Pitágoras,

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \left\|\hat{\theta} - \theta\right\|^2 \\ &= \left\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\right\|^2 + orm\mathbb{E}[\hat{\theta}] - \theta^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2.\end{aligned}$$

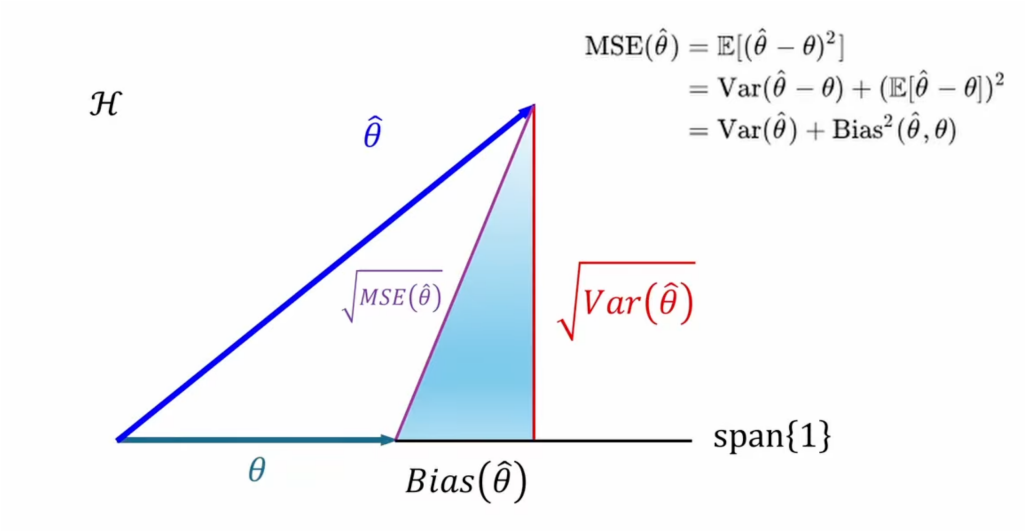

FIGURA 2. La descomposición sesgo-varianza es Pitágoras en L^2.

12. Por qué se resta y se suma la esperanza

En demostraciones usuales aparece el paso

$$\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta).$$

Algebraicamente parece un truco: restar y sumar lo mismo. Geométricamente no es un truco; es una descomposición ortogonal:

$$\text{error total} = exterroraleatoriocentrado + \text{error sistemático constante}.$$

La parte aleatoria centrada vive perpendicularmente a span{1}. La parte constante vive dentro de span{1}. Por eso el término cruzado se anula.

13. Comparación con regresión lineal

En regresión lineal finita, con vector de datos $y \in \mathbb{R}^n$, se proyecta y sobre un subespacio generado por columnas de una matriz de diseño X_{mat}. En el caso más simple, si sólo hay intercepto, el subespacio es

$$\mathrm{span}\{(1, 1, \ldots, 1)\}.$$

La proyección de y sobre esa recta es

$$(\bar{y}, \bar{y}, \ldots, \bar{y}).$$

En probabilidad, el vector y se reemplaza por una variable aleatoria X, y la recta de vectores constantes se reemplaza por span{1}. La muestra finita usa sumas; L^2 usa esperanzas.

14. Mapa entre disciplinas

Álgebra lineal	Probabilidad	Estadística
Vector $x \in \mathbb{R}^n$	Variable aleatoria $X \in L^2$	Estimador $\hat{\theta}$
Producto punto $x \cdot y$	Producto interno $\mathbb{E}[XY]$	Covarianza si están centradas
Norma $\|x\|$	$\sqrt{\mathbb{E}[X^2]}$	Raíz de segundo momento
Recta de constantes	$\text{span}\{1\}$	Estimadores constantes
Proyección	Esperanza condicional/proyección	Mejor predictor cuadrático
Residual	$X - \mathbb{E}[X]$	Error aleatorio centrado
Pitágoras	Ortogonalidad en L^2	Sesgo-varianza

15. Límites y sutilezas

1. En L^2 no distinguimos variables que difieren sólo en un evento de probabilidad cero.
2. Si $X \in L^1$ pero $X \notin L^2$, su esperanza puede existir, pero su varianza puede ser infinita.
3. La geometría de proyecciones ortogonales depende del error cuadrático. Si usamos error absoluto $\mathbb{E}|X - c|$, el minimizador es una mediana, no la media.
4. En espacios más generales, la proyección sobre subespacios cerrados de Hilbert existe y es única. Esa propiedad no se conserva igual en todos los espacios normados.
5. La esperanza condicional $\mathbb{E}[X \mid \mathcal{G}]$ es una proyección de X sobre el subespacio de variables $\mathcal{G}$-medibles. Esta es la extensión natural de la idea del video.

16. Resumen simbólico

$$\boxed{\begin{array}{c} \mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P}), \qquad \langle X, Y \rangle = \mathbb{E}[XY], \qquad \|X\| = \sqrt{\mathbb{E}[X^2]}, \\ \text{span}\{1\} = \{c1 : c \in \mathbb{R}\}, \qquad \text{Proj}_{\text{span}\{1\}} X = (\mathbb{E}X)1, \\ \text{Var}(X) = \|X - \mathbb{E}X\|^2, \qquad \text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta}, \theta)^2. \end{array}}$$

17. Bias–variance en regresión: $\text{MSE} = \text{Bias}^2 + \text{Var} + \sigma^2$

La primera descomposición sesgo–varianza estimaba una constante θ. En regresión, el objeto verdadero ya no es un número sino una función evaluada en un punto fijo x. El modelo es

$$Y = f(x) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \text{Var}(\varepsilon) = \sigma^2.$$

Aquí Y es una nueva respuesta, $f(x)$ es la relación verdadera en el punto de predicción x, y ε es el error no observable del nuevo dato. El número σ^2 se llama *error irreducible*: aun si conociéramos exactamente f, no podríamos predecir Y sin error porque permanece el ruido ε.

17.1. Datos de entrenamiento y el subespacio $\mathcal{H}_D$. Sea

$$D = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$$

el conjunto de entrenamiento. En el análisis usual de regresión, los x_i se tratan como fijos y los Y_i como aleatorios. Un estimador entrenado, por ejemplo una recta de mínimos cuadrados, se escribe

$$\hat{f}(x; D).$$

Para un x fijo, $\hat{f}(x; D)$ es una variable aleatoria porque depende del conjunto aleatorio D. Definimos entonces

$$\mathcal{H}_D = \{Z \in \mathcal{H} : Z \text{ depende solamente de } D\}.$$

Así,

$$\hat{f}(x; D) \in \mathcal{H}_D.$$

El nuevo ruido ε se supone independiente, o al menos no correlacionado, con el entrenamiento. En forma geométrica, esto se expresa como

$$\varepsilon \perp \mathcal{H}_D.$$

En efecto, si $Z \in \mathcal{H}_D$, entonces

$$\langle \varepsilon, Z \rangle = \mathbb{E}[\varepsilon Z] = \mathbb{E}\big[Z\mathbb{E}(\varepsilon \mid D)\big] = 0.$$

La independencia estadística se convierte en ortogonalidad geométrica.

17.2. Definiciones. El error cuadrático medio de predicción es

$$\mathrm{MSE}(\hat{f}; x) = \mathbb{E}\big[(Y - \hat{f}(x; D))^2\big] = \Big\|Y - \hat{f}(x; D)\Big\|^2.$$

La esperanza se toma sobre dos fuentes de aleatoriedad: el nuevo dato Y y el conjunto de entrenamiento D.

El sesgo en el punto x es

$$\mathrm{Bias}(\hat{f}; x) = f(x) - \mathbb{E}_D[\hat{f}(x; D)].$$

La varianza del estimador en x es

$$\mathrm{Var}_D(\hat{f}(x; D)) = \mathbb{E}_D\left[\big(\hat{f}(x; D) - \mathbb{E}_D[\hat{f}(x; D)]\big)^2\right].$$

17.3. Descomposición ortogonal. Descomponemos el vector de error como

$$Y - \hat{f} = \underbrace{Y - f(x)}_{\varepsilon} + \underbrace{f(x) - \mathbb{E}_D[\hat{f}]}_{\text{sesgo}} + \underbrace{\mathbb{E}_D[\hat{f}] - \hat{f}}_{\text{fluctuación del estimador}}.$$

Los tres términos son ortogonales: el primero es perpendicular a $\mathcal{H}_D$; el segundo pertenece a $\mathrm{span}\{1\} \subset \mathcal{H}_D$; y el tercero es el residuo de proyectar $\hat{f}$ sobre $\mathrm{span}\{1\}$, por lo que es perpendicular a las constantes. Por Pitágoras en L^2,

$$\Big\|Y - \hat{f}\Big\|^2 = \|\varepsilon\|^2 + \Big\|f(x) - \mathbb{E}_D[\hat{f}]\Big\|^2 + \Big\|\hat{f} - \mathbb{E}_D[\hat{f}]\Big\|^2.$$

Por tanto,

$$\boxed{\mathrm{MSE}(\hat{f}; x) = \mathrm{Bias}(\hat{f}; x)^2 + \mathrm{Var}_D(\hat{f}(x; D)) + \sigma^2.}$$

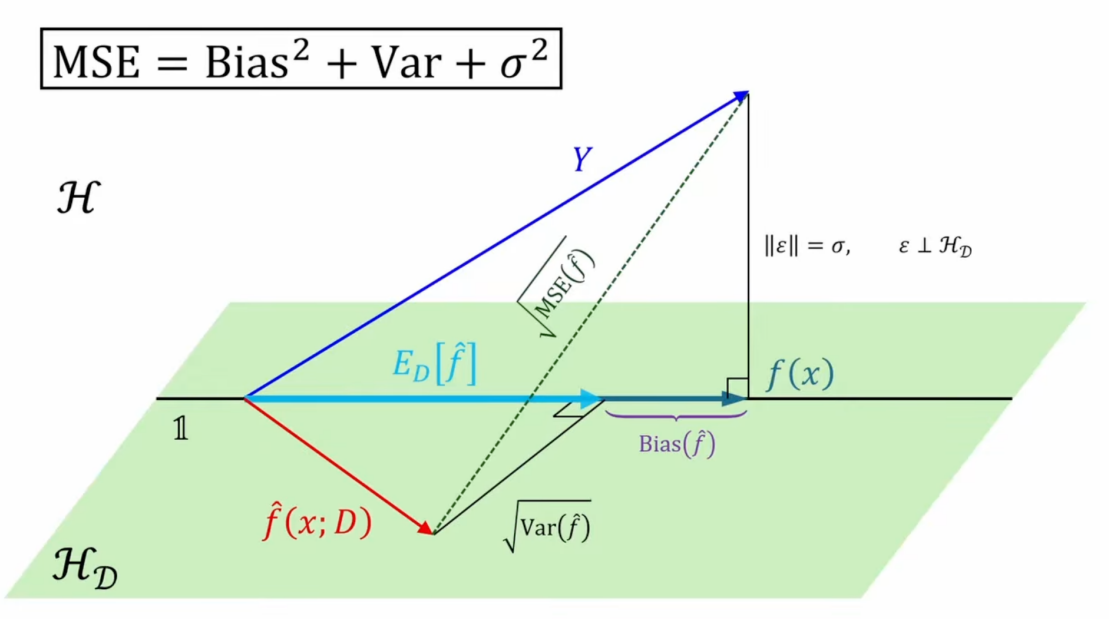

FIGURA 3. La descomposición bias–variance en regresión se interpreta como una descomposición ortogonal en $\mathcal{H} = L^2$. El espacio $\mathcal{H}_D$ contiene las variables que dependen del entrenamiento; el ruido nuevo queda fuera de ese espacio y es ortogonal a él.

18. Esperanza condicional como proyección ortogonal

La esperanza condicional tiene una interpretación geométrica más profunda que la esperanza ordinaria. Si $X, Y \in L^2$, entonces $\mathbb{E}[X \mid Y]$ es la mejor aproximación cuadrática de X usando solamente información contenida en Y.

18.1. El subespacio generado por Y.

Definimos

$$\mathcal{H}_Y = \{g(Y) : g(Y) \in L^2\}.$$

Este es el subespacio de variables aleatorias que dependen únicamente de Y. Como las constantes también son funciones triviales de Y, se tiene

$$\operatorname{span}\{1\} \subseteq \mathcal{H}_Y \subseteq \mathcal{H}.$$

Geométricamente, condicionar en Y significa proyectar sobre $\mathcal{H}_Y$:

$$\boxed{\mathbb{E}[X \mid Y] = \operatorname{Proj}_{\mathcal{H}_Y}(X).}$$

Equivalente: $\mathbb{E}[X \mid Y]$ es la variable de la forma $g(Y)$ que minimiza

$$\mathbb{E}\big[(X - g(Y))^2\big].$$

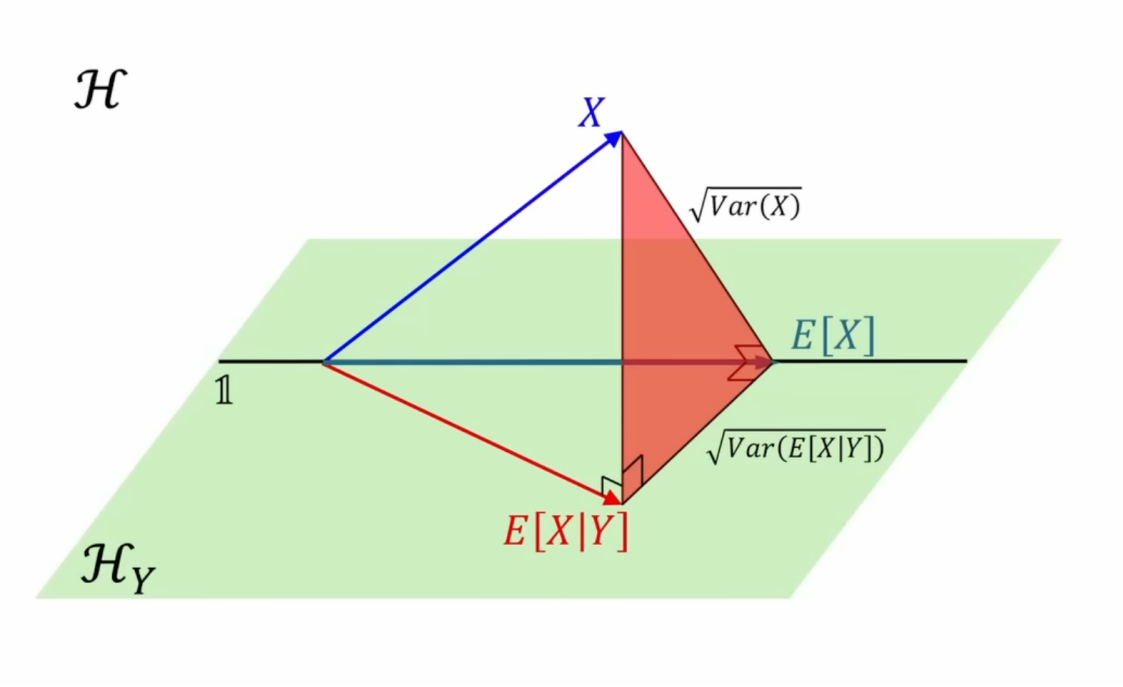

FIGURA 4. La esperanza condicional $\mathbb{E}[X \mid Y]$ es la proyección de X sobre el subespacio $\mathcal{H}_Y$ de variables que sólo dependen de Y.

18.2. Condición de ortogonalidad. La caracterización fundamental es

$$\mathbb{E}[X \mid Y] = g(Y) \quad \Longleftrightarrow \quad X - g(Y) \perp \mathcal{H}_Y.$$

Es decir, para todo $h(Y) \in \mathcal{H}_Y$,

$$\mathbb{E}\big[(X - g(Y))h(Y)\big] = 0.$$

La demostración es la misma que en mínimos cuadrados. Para $h(Y) \in \mathcal{H}_Y$, definimos

$$\phi(t) = \mathbb{E}\big[(X - g(Y) - th(Y))^2\big].$$

Si $g(Y)$ minimiza, entonces $\phi'(0) = 0$, y por tanto

$$-2\mathbb{E}\big[(X - g(Y))h(Y)\big] = 0.$$

Esto prueba la ortogonalidad.

18.3. Ley de la esperanza total: propiedad torre paso a paso. La propiedad torre afirma que si primero promediamos X usando la información de Y, y después promediamos de nuevo sin condicionar, obtenemos el mismo promedio total:

$$\boxed{\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big].}$$

También se conoce como ley de la esperanza total, ley de Adam o *tower property*.

La forma geométrica de leerla es la siguiente. La variable X vive en el espacio grande $\mathcal{H} = L^2(\Omega, \mathcal{F}, P)$. La variable $\mathbb{E}[X \mid Y]$ vive en el subespacio $\mathcal{H}_Y$, porque depende solamente de Y. Como

$$\mathbb{E}[X \mid Y] = \mathrm{Proj}_{\mathcal{H}_Y}(X),$$

el vector residual

$$R := X - \mathbb{E}[X \mid Y]$$

es perpendicular a todo el subespacio $\mathcal{H}_Y$:

$$R \perp \mathcal{H}_Y.$$

Ahora usamos un hecho muy simple pero importante: la variable constante 1 pertenece a $\mathcal{H}_Y$. En efecto,

$$1(\omega) = 1$$

es una función constante de Y. Por tanto,

$$1 \in \operatorname{span}\{1\} \subseteq \mathcal{H}_Y.$$

Como R es perpendicular a todo $\mathcal{H}_Y$, en particular es perpendicular a 1:

$$\langle R, 1\rangle = 0.$$

Sustituyendo $R = X - \mathbb{E}[X \mid Y]$, tenemos

$$\langle X - \mathbb{E}[X \mid Y], 1\rangle = 0.$$

Por definición del producto interno en L^2,

$$\langle A, B\rangle = \mathbb{E}[AB].$$

Entonces

$$\langle X - \mathbb{E}[X \mid Y], 1\rangle = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y]) \cdot 1\right].$$

Como multiplicar por 1 no cambia nada,

$$\mathbb{E}\left[(X - \mathbb{E}[X \mid Y]) \cdot 1\right] = \mathbb{E}\left[X - \mathbb{E}[X \mid Y]\right].$$

Luego la ortogonalidad dice

$$\mathbb{E}\left[X - \mathbb{E}[X \mid Y]\right] = 0.$$

Usando linealidad de la esperanza,

$$\mathbb{E}[X] - \mathbb{E}\left[\mathbb{E}[X \mid Y]\right] = 0.$$

Por tanto,

$$\boxed{\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X \mid Y]\right].}$$

Interpretación intuitiva. La variable $\mathbb{E}[X \mid Y]$ es un promedio local: para cada valor de Y, calcula el promedio de X dentro del grupo compatible con ese valor de Y. Después, $\mathbb{E}[\mathbb{E}[X \mid Y]]$ promedia esos promedios locales ponderándolos por la probabilidad de cada valor de Y. El resultado debe coincidir con el promedio global de X.

En el caso discreto, si Y toma valores $y_1, \ldots, y_k$, entonces

$$\mathbb{E}[X \mid Y] = \sum_{j=1}^{k} \mathbb{E}[X \mid Y = y_j]\mathbf{1}_{\{Y=y_j\}}.$$

Al tomar esperanza total,

$$\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \sum_{j=1}^{k} \mathbb{E}[X \mid Y = y_j]\mathbb{P}(Y = y_j).$$

Pero por la fórmula elemental de partición,

$$\mathbb{E}[X] = \sum_{j=1}^{k} \mathbb{E}[X \mid Y = y_j]\mathbb{P}(Y = y_j).$$

Así obtenemos de nuevo

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big].$$

Lectura geométrica final. La propiedad torre no es un truco algebraico. Es la afirmación de que proyectar X sobre $\mathcal{H}_Y$ no cambia su componente constante. Dicho de otro modo, X y su proyección $\mathbb{E}[X \mid Y]$ tienen la misma sombra sobre la recta de constantes $\operatorname{span}\{1\}$:

$$\operatorname{Proj}_{\operatorname{span}\{1\}}(X) = \operatorname{Proj}_{\operatorname{span}\{1\}}(\mathbb{E}[X \mid Y]).$$

Como proyectar sobre $\operatorname{span}\{1\}$ equivale a tomar esperanza,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

19. Chapman–Kolmogorov como propiedad de torre

La ecuación de Chapman–Kolmogorov puede leerse como una versión dinámica de la propiedad torre. La idea es la misma que antes: condicionar equivale a proyectar sobre un subespacio de información. Si proyectamos primero sobre la información de un tiempo intermedio y después sobre la información inicial, obtenemos lo mismo que proyectar directamente sobre la información inicial.

19.1. Probabilidades como esperanzas de indicadores. Sea $(X_u)_{u\geq 0}$ un proceso estocástico con espacio de estados discreto S. Para un estado $j \in S$, definimos la función indicadora

$$f_j(x) = \mathbf{1}_{\{x=j\}}.$$

Entonces el evento "el proceso está en j al tiempo t" se escribe como

$$\{X_t = j\},$$

y su indicador es

$$f_j(X_t) = \mathbf{1}_{\{X_t=j\}}.$$

Por la regla básica

$$\mathbb{P}(A) = \mathbb{E}[\mathbf{1}_A],$$

la probabilidad condicional de transición puede escribirse como una esperanza condicional:

$$\boxed{\mathbb{P}(X_t = j \mid X_s = i) = \mathbb{E}[\mathbf{1}_{\{X_t=j\}} \mid X_s = i].}$$

Esta identidad es el puente entre probabilidades de transición y geometría de L^2.

19.2. Introducir un tiempo intermedio. Tomemos tres tiempos ordenados

$$s < r < t.$$

Queremos descomponer el movimiento desde s hasta t pasando por el tiempo intermedio r.

La propiedad torre, aplicada a la variable aleatoria $\mathbf{1}_{\{X_t=j\}}$, dice

$$\mathbb{E}[\mathbf{1}_{\{X_t=j\}} \mid X_s] = \mathbb{E}\left[\mathbb{E}[\mathbf{1}_{\{X_t=j\}} \mid X_r, X_s] \mid X_s\right].$$

La esperanza interna mira primero desde el tiempo intermedio r hacia el futuro t. La esperanza externa promedia después todos los posibles valores intermedios compatibles con el estado en s.

19.3. Uso de la propiedad de Markov. La propiedad de Markov afirma que, dado el presente, el futuro no depende del pasado. En símbolos, si $s < r < t$, entonces

$$\mathbb{P}(X_t \in A \mid X_r, X_s) = \mathbb{P}(X_t \in A \mid X_r)$$

para todo conjunto de estados A.

Aplicada al evento $\{X_t = j\}$, esto implica

$$\mathbb{E}[\mathbf{1}_{\{X_t=j\}} \mid X_r, X_s] = \mathbb{E}[\mathbf{1}_{\{X_t=j\}} \mid X_r] = \mathbb{P}(X_t = j \mid X_r).$$

Sustituyendo en la propiedad torre,

$$\boxed{\mathbb{P}(X_t = j \mid X_s) = \mathbb{E}\left[\mathbb{P}(X_t = j \mid X_r) \mid X_s\right].}$$

Esta es la forma de Chapman–Kolmogorov escrita como esperanza condicional.

19.4. Abrir la esperanza condicional en el caso discreto. Si además condicionamos en $X_s = i$, entonces

$$\mathbb{P}(X_t = j \mid X_s = i) = \mathbb{E}\left[\mathbb{P}(X_t = j \mid X_r) \mid X_s = i\right].$$

Como X_r toma valores $k \in S$, la variable

$$\mathbb{P}(X_t = j \mid X_r)$$

es una función de X_r. Específicamente,

$$\mathbb{P}(X_t = j \mid X_r) = \sum_{k \in S} \mathbb{P}(X_t = j \mid X_r = k)\mathbf{1}_{\{X_r = k\}}.$$

Tomando esperanza condicional dado $X_s = i$,

$$\begin{aligned}
\mathbb{P}(X_t = j \mid X_s = i) &= \mathbb{E}\left[\sum_{k \in S} \mathbb{P}(X_t = j \mid X_r = k)\mathbf{1}_{\{X_r = k\}} \mid X_s = i\right] \\
&= \sum_{k \in S} \mathbb{P}(X_t = j \mid X_r = k)\mathbb{E}[\mathbf{1}_{\{X_r = k\}} \mid X_s = i] \\
&= \sum_{k \in S} \mathbb{P}(X_t = j \mid X_r = k)\mathbb{P}(X_r = k \mid X_s = i).
\end{aligned}$$

Por tanto,

$$\boxed{\mathbb{P}(X_t = j \mid X_s = i) = \sum_{k \in S} \mathbb{P}(X_t = j \mid X_r = k)\mathbb{P}(X_r = k \mid X_s = i).}$$

Esta es la ecuación clásica de Chapman–Kolmogorov.

19.5. Lectura geométrica. Definamos los subespacios de información

$$\mathcal{H}_s \subseteq \mathcal{H}_r \subseteq \mathcal{H}_t \subseteq \mathcal{H},$$

donde $\mathcal{H}_u$ representa las variables aleatorias cuadrado-integrables determinadas por la información disponible hasta el tiempo u.

La esperanza condicional es una proyección ortogonal:

$$\mathbb{E}[Z \mid \mathcal{H}_u] = \mathrm{Proj}_{\mathcal{H}_u}(Z).$$

Entonces,

$$\mathbb{E}\left[\mathbb{E}[Z \mid \mathcal{H}_r] \mid \mathcal{H}_s\right] = \mathrm{Proj}_{\mathcal{H}_s}\big(\mathrm{Proj}_{\mathcal{H}_r}(Z)\big).$$

Como $\mathcal{H}_s \subseteq \mathcal{H}_r$, proyectar primero sobre $\mathcal{H}_r$ y luego sobre $\mathcal{H}_s$ equivale a proyectar directamente sobre $\mathcal{H}_s$:

$$\mathrm{Proj}_{\mathcal{H}_s}\,\mathrm{Proj}_{\mathcal{H}_r} = \mathrm{Proj}_{\mathcal{H}_s}.$$

Por consiguiente,

$$\boxed{\mathbb{E}\left[\mathbb{E}[Z \mid \mathcal{H}_r] \mid \mathcal{H}_s\right] = \mathbb{E}[Z \mid \mathcal{H}_s].}$$

Chapman–Kolmogorov es esta misma identidad aplicada a

$$Z = \mathbf{1}_{\{X_t = j\}}.$$

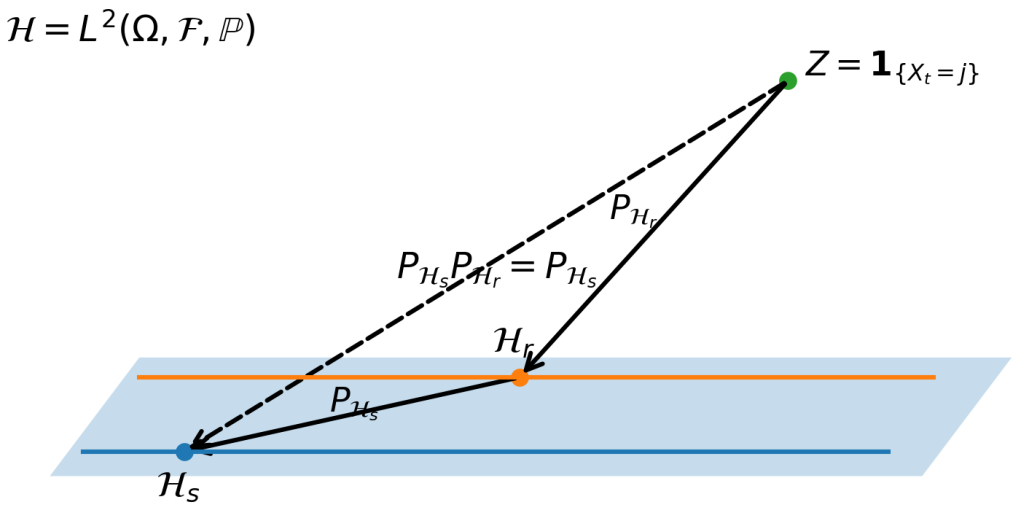

FIGURA 5. Lectura geométrica de Chapman–Kolmogorov: proyectar la información futura primero al tiempo intermedio r y después al tiempo inicial s produce la misma sombra que proyectar directamente al tiempo s.

19.6. Notación con operadores de transición. Para una cadena de Markov homogénea, definimos el operador de transición P_u por

$$(P_u f)(i) = \mathbb{E}[f(X_u) \mid X_0 = i].$$

La ecuación de Chapman–Kolmogorov se escribe entonces como

$$\boxed{P_{t-s} = P_{r-s} P_{t-r}.}$$

Dependiendo de la convención de composición de operadores, también puede escribirse como

$$P_{s,t} = P_{s,r} P_{r,t}.$$

La interpretación es la misma: una transición larga se obtiene componiendo dos transiciones cortas.

19.7. Resumen conceptual. La cadena de ideas es

$$\boxed{\mathbb{P}(X_t = j \mid X_s = i) = \mathbb{E}[\mathbf{1}_{\{X_t=j\}} \mid X_s = i] \xrightarrow{\text{torre}} \mathbb{E}\left[\mathbb{P}(X_t = j \mid X_r) \mid X_s = i\right] \xrightarrow{\text{abrir en } k} \sum_k p_{s,r}(i,k) p_{r,t}(k,j).}$$

Así, Chapman–Kolmogorov no aparece como una fórmula aislada: es la propiedad torre aplicada a indicadores, más la propiedad de Markov.

19.8. Sacar lo dado. Si $h(Y)$ depende sólo de Y, entonces

$$\boxed{\mathbb{E}[h(Y)X \mid Y] = h(Y)\mathbb{E}[X \mid Y].}$$

La razón geométrica es que multiplicar por una función de Y no introduce información externa a $\mathcal{H}_Y$. Para verificarlo, basta revisar que el residuo es ortogonal a todo $g(Y) \in \mathcal{H}_Y$:

$$\mathbb{E}\left[\big(h(Y)X - h(Y)\mathbb{E}[X \mid Y]\big)g(Y)\right] = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])h(Y)g(Y)\right] = 0,$$

porque $h(Y)g(Y) \in \mathcal{H}_Y$.

20. Ley de la varianza total como Pitágoras

La varianza condicional se define por

$$\mathrm{Var}(X \mid Y) = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2 \mid Y\right].$$

La identidad central es

$$\boxed{\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).}$$

También se conoce como ley de la varianza total, ley de Eve o identidad ANOVA.

20.1. Demostración geométrica. Primero descomponemos

$$X = \mathbb{E}[X \mid Y] + \big(X - \mathbb{E}[X \mid Y]\big),$$

donde

$$X - \mathbb{E}[X \mid Y] \perp \mathcal{H}_Y.$$

Restando $\mathbb{E}[X]$,

$$X - \mathbb{E}[X] = \underbrace{X - \mathbb{E}[X \mid Y]}_{\text{residuo}} + \underbrace{\mathbb{E}[X \mid Y] - \mathbb{E}[X]}_{\text{parte explicada}}.$$

El segundo término pertenece a $\mathcal{H}_Y$, mientras que el primero es ortogonal a $\mathcal{H}_Y$. Luego son ortogonales entre sí. Por Pitágoras,

$$\|X - \mathbb{E}[X]\|^2 = \|X - \mathbb{E}[X \mid Y]\|^2 + \|\mathbb{E}[X \mid Y] - \mathbb{E}[X]\|^2.$$

Interpretando cada término:

$$\|X - \mathbb{E}[X]\|^2 = \mathrm{Var}(X),$$

$$\|X - \mathbb{E}[X \mid Y]\|^2 = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2\right] = \mathbb{E}[\mathrm{Var}(X \mid Y)],$$

por la propiedad torre, y

$$\|\mathbb{E}[X \mid Y] - \mathbb{E}[X]\|^2 = \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

Así obtenemos

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

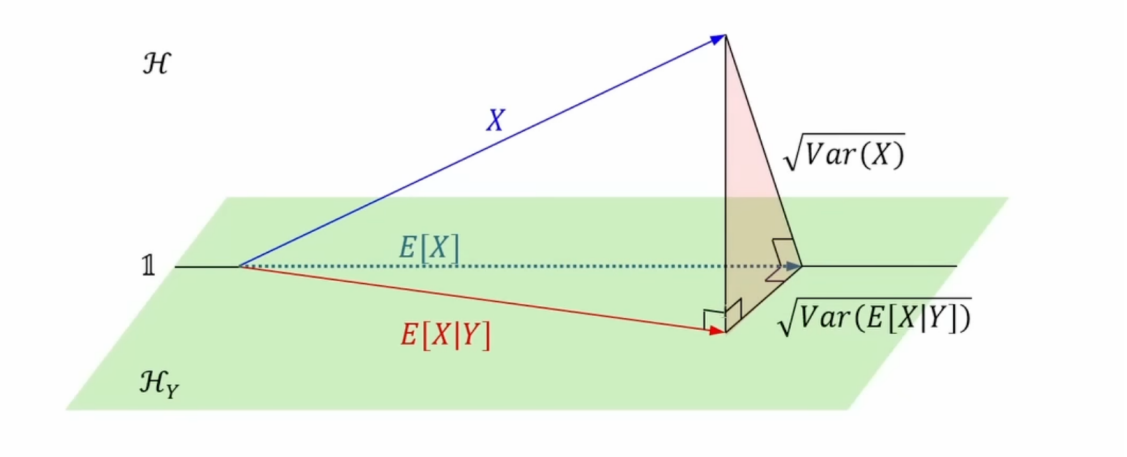

Figura 6. La ley de la varianza total es Pitágoras: la variación total de X se divide en variación residual $\mathbb{E}[\mathrm{Var}(X \mid Y)]$ y variación explicada $\mathrm{Var}(\mathbb{E}[X \mid Y])$.

21. Ejemplo: suma aleatoria

Sea N una variable aleatoria con valores en $\{0, 1, 2, \ldots\}$, y sean $X_1, X_2, \ldots$ variables i.i.d. independientes de N, con

$$\mathbb{E}[X_i] = \mu_X, \qquad \operatorname{Var}(X_i) = \sigma_X^2.$$

Definimos la suma aleatoria

$$S = \sum_{i=1}^{N} X_i,$$

con la convención $S = 0$ si $N = 0$.

Condicionando en N, el número de sumandos queda fijo. Entonces

$$\mathbb{E}[S \mid N] = N\mu_X.$$

Por la ley de la esperanza total,

$$\mathbb{E}[S] = \mathbb{E}[N\mu_X] = \mu_X \mathbb{E}[N].$$

Para la varianza,

$$\operatorname{Var}(S \mid N) = N\sigma_X^2.$$

Usando la ley de la varianza total,

$$\operatorname{Var}(S) = \mathbb{E}[N\sigma_X^2] + \operatorname{Var}(N\mu_X) = \sigma_X^2 \mathbb{E}[N] + \mu_X^2 \operatorname{Var}(N).$$

Por tanto,

$$\boxed{\mathbb{E}[S] = \mu_X \mathbb{E}[N], \qquad \operatorname{Var}(S) = \sigma_X^2 \mathbb{E}[N] + \mu_X^2 \operatorname{Var}(N).}$$