12.2 Covariance Matrices and Joint Probabilities

Linear algebra enters when we run M different experiments at once. We might measure age and height and weight ($M = 3$ measurements of N people). Each experiment has its own mean value. So we have a vector $\boldsymbol{m} = (m_1, m_2, m_3)$ containing the M mean values. Those could be *sample means* of age and height and weight. Or m_1, m_2, m_3 could be *expected values* of age, height, weight based on known probabilities.

A matrix becomes involved when we look at variances. Each experiment will have a sample variance S_i^2 or an expected $\sigma_i^2 = \mathrm{E}\left[(x_i - m_i)^2\right]$ based on the squared distance from its mean. Those M numbers $\sigma_1^2, \ldots, \sigma_M^2$ will go on the main diagonal of the matrix. So far we have made no connection between the M parallel experiments. They measure M different random variables, but the experiments are not necessarily independent!

If we measure age and height and weight $(\boldsymbol{a}, \boldsymbol{h}, \boldsymbol{w})$ for children, the results will be strongly correlated. Older children are generally taller and heavier. Suppose the means m_a, m_h, m_w are known. Then $\sigma_a^2, \sigma_h^2, \sigma_w^2$ are the separate variances in age, height, weight. **The new numbers are the covariances like $\boldsymbol{\sigma_{ah}}$, where age multiplies height.**

$$\boxed{\textbf{Covariance} \qquad \boldsymbol{\sigma_{ah}} = \mathbf{E}\left[(\textbf{age} - \textbf{mean age})\ (\textbf{height} - \textbf{mean height})\right].} \qquad (1)$$

This definition needs a close look. To compute σ_{ah}, it is not enough to know the probability of each age and the probability of each height. We have to know the **joint probability of each pair** (**age and height**). This is because age is connected to height.

$\boldsymbol{p_{ah}}$ = probability that a random child has age = $\boldsymbol{a}$ **and** height = $\boldsymbol{h}$: both at once

$\boldsymbol{p_{ij}}$ = **probability that experiment 1 produces $\boldsymbol{x_i}$ and experiment 2 produces $\boldsymbol{y_j}$**

Suppose experiment 1 (age) has mean m_1. Experiment 2 (height) has mean m_2. The covariance in (1) between experiments 1 and 2 looks at **all pairs** of ages x_i, heights y_j :

$$\boxed{\textbf{Covariance} \qquad \boldsymbol{\sigma_{12}} = \sum_{\textbf{all}\, \boldsymbol{i,j}} \sum \boldsymbol{p_{ij}(x_i - m_1)(y_j - m_2)}} \qquad (2)$$

To capture this idea of "joint probability p_{ij}" we begin with two small examples.

Example 1 Flip two coins separately. With 1 for heads and 0 for tails, the results can be $(1,1)$ or $(1,0)$ or $(0,1)$ or $(0,0)$. Those four outcomes all have probability $p_{11} = p_{10} = p_{01} = p_{00} = \frac{1}{4}$. **Independent experiments have Prob of $(\boldsymbol{i, j})$ = (Prob of $\boldsymbol{i}$) (Prob of $\boldsymbol{j}$).**

Example 2 *Glue the coins together*, facing the same way. The only possibilities are $(1,1)$ and $(0,0)$. Those have probabilities $\frac{1}{2}$ and $\frac{1}{2}$. The probabilities p_{10} and p_{01} are zero. $(1,0)$ and $(0,1)$ won't happen because the coins stick together: both heads or both tails.

$$\begin{array}{l}\textbf{Probability matrices}\\ \textbf{for Examples 1 and 2}\end{array} \qquad \boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} \qquad \boldsymbol{P} = \begin{bmatrix} \frac{1}{2} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} \end{bmatrix}.$$

Let me stay longer with P, to show it in good matrix notation. The matrix shows the probability p_{ij} of each pair (x_i, y_j)—starting with (x_1, y_1) = (heads, heads) and (x_1, y_2) = (heads, tails). Notice the row sums p_i and column sums P_j and the total sum $= 1$.

$$\textbf{Probability matrix} \quad P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \qquad \begin{matrix} p_{11} + p_{12} = \boldsymbol{p_1} \\ p_{21} + p_{22} = \boldsymbol{p_2} \end{matrix} \quad \begin{pmatrix} \text{first} \\ \text{coin} \end{pmatrix}$$

$$\text{(second coin) column sums} \quad \boldsymbol{P_1} \quad \boldsymbol{P_2} \qquad\qquad 4 \ \text{entries add to } 1$$

Those numbers p_1, p_2 and P_1, P_2 are called the **marginals** of the matrix P:

$p_1 = p_{11} + p_{12}$ = chance of heads from **coin 1** (coin 2 can be heads or tails)
$P_1 = p_{11} + p_{21}$ = chance of heads from **coin 2** (coin 1 can be heads or tails)

Example 1 showed *independent* variables. Every probability p_{ij} equals p_i times p_j $\left(\frac{1}{2} \text{ times } \frac{1}{2} \text{ gave } p_{ij} = \frac{1}{4} \text{ in that example}\right)$. In this case **the covariance $\boldsymbol{\sigma}_{12}$ will be zero**. Heads or tails from the first coin gave no information about the second coin.

$$\boxed{\begin{matrix}\textbf{Zero covariance } \boldsymbol{\sigma_{12}} \\ \textbf{for independent trials}\end{matrix} \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{\sigma_1^2} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma_2^2} \end{bmatrix} = \textbf{diagonal covariance matrix}.}$$

Independent experiments have $\sigma_{12} = 0$ because every $p_{ij} = (p_i)(p_j)$ in equation (2):

$$\sigma_{12} = \sum_i \sum_j (p_i)(p_j)(x_i - m_1)(y_j - m_2) = \left[\sum_i (p_i)(x_i - m_1)\right]\left[\sum_j (p_j)(y_j - m_2)\right] = [\mathbf{0}][\mathbf{0}].$$

The glued coins show perfect correlation. Heads on one means heads on the other. The covariance σ_{12} moves from 0 to $\sigma_1\sigma_2 = \frac{1}{4}$—this is the largest possible value of σ_{12}:

$$\textbf{Means} = \frac{\mathbf{1}}{\mathbf{2}} \qquad \boldsymbol{\sigma_{12}} = \frac{\mathbf{1}}{\mathbf{2}}\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{2}\right) + \mathbf{0} + \mathbf{0} + \frac{\mathbf{1}}{\mathbf{2}}\left(0 - \frac{1}{2}\right)\left(0 - \frac{1}{2}\right) = \frac{\mathbf{1}}{\mathbf{4}}$$

Heads or tails from coin 1 gives complete information about heads or tails from coin 2:

$$\begin{matrix}\textbf{Glued coins give largest possible covariances} \\ \textbf{Singular covariance matrix: determinant} = \mathbf{0}\end{matrix} \qquad \boldsymbol{V}_{\textbf{glue}} = \begin{bmatrix} \boldsymbol{\sigma_1^2} & \boldsymbol{\sigma_1\sigma_2} \\ \boldsymbol{\sigma_1\sigma_2} & \boldsymbol{\sigma_2^2} \end{bmatrix}$$

Always $\boldsymbol{\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2}$. Thus σ_{12} is *between* $-\sigma_1\sigma_2$ *and* $\sigma_1\sigma_2$. The covariance matrix V is **positive definite** (or in this singular case of glued coins, V is **positive semidefinite**). That is an important fact about M by M covariance matrices for M experiments.

Note that the **sample covariance matrix $\boldsymbol{S}$** from N trials is certainly semidefinite. Every new sample X = (age, height, weight) contributes to the **sample mean $\overline{\boldsymbol{X}}$** and to S. Each term $(X_i - \overline{X})(X_i - \overline{X})^{\mathrm{T}}$ is positive semidefinite and we just add to reach S:

$$\overline{\boldsymbol{X}} = \frac{X_1 + \cdots + X_N}{N} \qquad \boldsymbol{S} = \frac{(X_1 - \overline{X})(X_1 - \overline{X})^{\mathrm{T}} + \cdots + (X_N - \overline{X})(X_N - \overline{X})^{\mathrm{T}}}{N - 1} \quad (3)$$

The Covariance Matrix V is Positive Semidefinite

Come back to the *expected* covariance σ_{12} between two experiments 1 and 2 (two coins):

$$\boxed{\begin{aligned}\sigma_{12} &= \text{expected value of } [(\textit{output}\,1 - \textit{mean}\,1) \text{ times } (\textit{output}\,2 - \textit{mean}\,2)] \\ &= \sum_{\text{all}}\sum_{i,j} \boldsymbol{p_{ij}\,(x_i - m_1)\,(y_j - m_2)}.\end{aligned}} \tag{4}$$

$p_{ij} \geq 0$ is the probability of seeing output x_i in experiment 1 **and** y_j in experiment 2. Some pair of outputs must appear. Therefore the N^2 probabilities p_{ij} add to 1.

$$\textbf{Total probability (all pairs) is 1} \qquad \sum_{\text{all}}\sum_{i,j} p_{ij} = 1. \tag{5}$$

Here is another fact we need. *Fix on one particular output* x_i in experiment 1. Allow *all outputs* y_j in experiment 2. Add the probabilities of $(x_i, y_1), (x_i, y_2), \ldots, (x_i, y_n)$:

$$\textbf{Row sum } \boldsymbol{p_i} \textbf{ of } \boldsymbol{P} \qquad \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{p_{ij}} = \textbf{probability } \boldsymbol{p_i} \textbf{ of } \boldsymbol{x_i} \textbf{ in experiment 1.} \tag{6}$$

Some y_j must happen in experiment 2 ! Whether the two coins are completely separate or glued together, we still get $\frac{1}{2}$ for the probability $p_H = p_{HH} + p_{HT}$ that coin 1 is heads:

$$\text{(separate) } P_{HH} + P_{HT} = \frac{1}{4} + \frac{1}{4} = \boldsymbol{\frac{1}{2}} \qquad \text{(glued) } P_{HH} + P_{HT} = \frac{1}{2} + 0 = \boldsymbol{\frac{1}{2}}.$$

That basic reasoning allows us to write one matrix formula that includes the covariance σ_{12} along with the separate variances σ_1^2 and σ_2^2 for experiment 1 and experiment 2. We get the whole covariance matrix V by adding the matrices V_{ij} for each pair (i, j):

$$\boxed{\begin{array}{c}\textbf{Covariance matrix} \\ \boldsymbol{V = \Sigma\,\Sigma\,V_{ij}}\end{array} \quad \boldsymbol{V} = \sum_{\text{all}}\sum_{i,j} p_{ij} \begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix}} \tag{7}$$

Off the diagonal, this is equation (2) for the covariance σ_{12}. On the diagonal, we are getting the ordinary variances σ_1^2 and σ_2^2. I will show in detail how we get $V_{11} = \sigma_1^2$ by using equation (6). Allowing all j just leaves the probability p_i of x_i in experiment 1:

$$\boldsymbol{V_{11}} = \sum_{\text{all}}\sum_{i,j} p_{ij}(x_i - m_1)^2 = \sum_{\text{all } i} \text{(probability of } x_i\text{)}\ (x_i - m_1)^2 = \boldsymbol{\sigma_1^2}. \tag{8}$$

Please look at that twice. It is the key to producing the whole covariance matrix by one formula (7). The beauty of that formula is that it combines 2 by 2 matrices V_{ij}. And the matrix V_{ij} in (7) for each pair of outcomes i, j is **positive semidefinite**:

V_{ij} has diagonal entries $p_{ij}(x_i - m_1)^2 \geq 0$ and $p_{ij}(y_j - m_2)^2 \geq 0$ and $\det(V_{ij}) = 0$.

That matrix V_{ij} has rank 1. Equation (7) multiplies p_{ij} *times column* $\boldsymbol{U}$ *times row* $\boldsymbol{U}^{\mathrm{T}}$:

$$\begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix} = \begin{bmatrix} x_i - m_1 \\ y_j - m_2 \end{bmatrix} \begin{bmatrix} x_i - m_1 & y_j - m_2 \end{bmatrix} \tag{9}$$

Every matrix $\boldsymbol{U}\boldsymbol{U}^{\mathbf{T}}$ *is positive semidefinite.* So the whole matrix V (combining these matrices $\boldsymbol{U}\boldsymbol{U}^{\mathbf{T}}$ with weights $p_{ij} \geq 0$) is **at least semidefinite**—and probably V is definite.

The covariance matrix V is positive definite unless the experiments are dependent.

Now we move from two variables x and y to M variables like age-height-weight. The output from each trial is a vector X with M components. (Each child has an age-height-weight vector with 3 components.) The covariance matrix V is now M by M. V is created from the output vectors $\boldsymbol{X}$ and their average $\overline{\boldsymbol{X}} = \mathbf{E}\,[\boldsymbol{X}]$:

Covariance matrix $$\boldsymbol{V} = \mathbf{E}\left[\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^{\mathrm{T}}\right] \tag{10}$$

Remember that $\boldsymbol{X}\boldsymbol{X}^{\mathbf{T}}$ and $\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathbf{T}}$ = (column) (row) are M by M matrices.

For $M = 1$ (one variable) you see that $\overline{X}$ is the mean m and V is σ^2 (Section 12.1). For $M = 2$ (two coins) you see that $\overline{X}$ is (m_1, m_2) and V matches equation (10). The expectation E always adds up outputs times their probabilities. For age-height-weight the output could be $X =$ (5 years, 31 inches, 48 pounds) and its probability is $\boldsymbol{p}_{\mathbf{5,31,48}}$.

Now comes a new idea. *Take any linear combination* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} = c_1X_1 + \cdots + c_MX_M$. With $c = (6, 2, 5)$ this would be $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} = 6\,(\text{age}) + 2\,(\text{height}) + 5\,(\text{weight})$. By linearity we know that its expected value $\mathrm{E}\,[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}]$ is $\boldsymbol{c}^{\mathrm{T}}\mathrm{E}\,[\boldsymbol{X}] = \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}}$:

$$\mathbf{E}\,[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}] = \boldsymbol{c}^{\mathrm{T}}\mathbf{E}\,[\boldsymbol{X}] = 6\,(\text{expected age}) + 2\,(\text{expected height}) + 5\,(\text{expected weight}).$$

More than that, we also know the *variance* σ^2 of that number $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}$:

$$\begin{aligned} \text{Variance of } \boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} &= \mathrm{E}\left[\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}}\right)\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}}\right)^{\mathrm{T}}\right] \\ &= \boldsymbol{c}^{\mathrm{T}}\mathrm{E}\left[\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X} - \overline{\boldsymbol{X}}\right)^{\mathrm{T}}\right]\boldsymbol{c} = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c}! \end{aligned} \tag{11}$$

Now the key point: *The variance of* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}$ *can never be negative.* So $\boldsymbol{c}^{\mathbf{T}}\boldsymbol{V}\boldsymbol{c} \geq \mathbf{0}$. *The covariance matrix V is therefore positive semidefinite by the energy test* $\boldsymbol{c}^{\mathbf{T}}\boldsymbol{V}\boldsymbol{c} \geq \mathbf{0}$.

Covariance matrices V open up the link between probability and linear algebra: V equals $Q\Lambda Q^{\mathrm{T}}$ with eigenvalues $\lambda_i \geq 0$ and orthonormal eigenvectors $\boldsymbol{q}_1$ to $\boldsymbol{q}_M$.

Diagonalizing the covariance matrix means finding *M independent* experiments as combinations of the original M experiments.

Confession I am not entirely happy with that proof based on $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \mathbf{0}$. The expectation symbol **E** is burying the key idea of **joint probability**. Allow me to show directly that V is positive semidefinite (at least for the age-height-weight example). The proof is simply that V **is the sum of the joint probability $\boldsymbol{p_{ahw}}$ of each combination (age, height, weight) times the positive semidefinite matrix $\boldsymbol{UU^{\mathrm{T}}}$**. Here U is $X - \overline{X}$:

$$\boldsymbol{V} = \sum_{\text{all } a,h,w} p_{ahw}\, \boldsymbol{U}\, \boldsymbol{U}^{\mathrm{T}} \quad \text{with} \quad \boldsymbol{U} = \begin{bmatrix} \text{age} \\ \text{height} \\ \text{weight} \end{bmatrix} - \begin{bmatrix} \text{mean age} \\ \text{mean height} \\ \text{mean weight} \end{bmatrix}. \tag{12}$$

This is exactly like the 2 by 2 coin flip matrix V in equation (7). Now $M = 3$.

The value of the expectation symbol E is that it also allows *pdf*'s (probability density functions like $p(x, y, z)$ for continuous random variables x and y and z). If we allow all numbers as ages and heights and weights, instead of age $i = 0, 1, 2, 3 \ldots$, then we need $p(x, y, z)$ instead of p_{ijk}. The sums in this section of the book would all change to integrals. But we still have $V = \mathrm{E}\,[\boldsymbol{UU}^{\mathrm{T}}]$:

$$\textbf{Covariance matrix} \quad V = \iiint p(x, y, z)\, \boldsymbol{UU}^{\mathrm{T}}\, dx\, dy\, dz \quad \text{with} \quad \boldsymbol{U} = \begin{bmatrix} x - \overline{x} \\ y - \overline{y} \\ z - \overline{z} \end{bmatrix}. \tag{13}$$

Always $\iiint p = 1$. Examples 1–2 emphasized how p can give diagonal V or singular V:

Independent variables $\boldsymbol{x, y, z}$ $\quad p(x, y, z) = \boldsymbol{p_1(x)\, p_2(y)\, p_3(z)}$.

Dependent variables $\boldsymbol{x, y, z}$ $\quad p(x, y, z) = \mathbf{0}$ except when $cx + dy + ez = 0$.

The Mean and Variance of $z = x + y$

Start with the sample mean. We have N samples of x. Their mean (= average) is m_x. We also have N samples of y and their mean is m_y. **The sample mean of $\boldsymbol{z = x + y}$ is clearly $\boldsymbol{m_z = m_x + m_y}$** :

$$\textbf{Mean of sum = Sum of means} \quad \frac{1}{N}\sum_{1}^{N}(x_i + y_i) = \frac{1}{N}\sum_{1}^{N} x_i + \frac{1}{N}\sum_{1}^{N} y_i. \tag{14}$$

Nice to see something that simple. The *expected* mean of $z = x + y$ doesn't look so simple, but it must come out as $\mathbf{E}[\boldsymbol{z}] = \mathbf{E}[\boldsymbol{x}] + \mathbf{E}[\boldsymbol{y}]$. Here is one way to see this.

The joint probability of the pair (x_i, y_j) is p_{ij}. Its value depends on whether the experiments are independent, which we don't know. But for the mean of the sum $z = x + y$, dependence or independence of x and y doesn't matter. *Expected values still add*:

$$\mathbf{E}[\boldsymbol{x}+\boldsymbol{y}]=\sum_i\sum_j p_{ij}(x_i+y_j)=\sum_i\sum_j p_{ij}x_i+\sum_i\sum_j p_{ij}y_j. \qquad (15)$$

All the sums go from 1 to N. We can add in any order. For the first term on the right side, add the p_{ij} along row i of the probability matrix P to get p_i. That double sum gives $\mathrm{E}[x]$:

$$\sum_i\sum_j p_{ij}x_i=\sum_i(p_{i1}+\cdots+p_{iN})x_i=\sum_i p_ix_i=\mathrm{E}[x].$$

For the last term, add p_{ij} down column j of the matrix to get the probability P_j of y_j. Those pairs (x_1,y_j) and (x_2,y_j) and ... and (x_N,y_j) are all the ways to produce y_j:

$$\sum_i\sum_j p_{ij}y_j=\sum_j(p_{1j}+\cdots+p_{Nj})y_j=\sum_j P_jy_j=\mathrm{E}[y].$$

Now equation (15) says that $\mathrm{E}[\boldsymbol{x}+\boldsymbol{y}]=\mathrm{E}[\boldsymbol{x}]+\mathrm{E}[\boldsymbol{y}]$.

What about the variance of $z=x+y$? The joint probabilities p_{ij} and the covariance σ_{xy} will be involved. Let me separate the variance of $x+y$ into three simple pieces:

$$\begin{aligned}\boldsymbol{\sigma_z^2}&=\sum\sum p_{ij}(x_i+y_j-m_x-m_y)^2\\&=\sum\sum p_{ij}(x_i-m_x)^2+\sum\sum p_{ij}(y_j-m_y)^2+2\sum\sum p_{ij}(x_i-m_x)(y_j-m_y)\end{aligned}$$

The first piece is $\boldsymbol{\sigma_x^2}$. The second piece is $\boldsymbol{\sigma_y^2}$. The last piece is $\boldsymbol{2\sigma_{xy}}$.

$$\textbf{The variance of } \boldsymbol{z=x+y} \quad \textbf{is} \quad \boldsymbol{\sigma_z^2=\sigma_x^2+\sigma_y^2+2\sigma_{xy}}. \qquad (16)$$

The Covariance Matrix for $Z=AX$

Here is a good way to see σ_z^2 when $z=x+y$. Think of (x,y) as a column vector $\boldsymbol{X}$. Think of the 1 by 2 matrix $A=\begin{bmatrix}1 & 1\end{bmatrix}$ multiplying that vector $\boldsymbol{X}$. Then $A\boldsymbol{X}$ is the sum $z=x+y$. The variance $\boldsymbol{\sigma_z^2}$ in equation (16) goes into matrix notation as

$$\boldsymbol{\sigma_z^2}=\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{xy} & \sigma_y^2\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix} \quad \text{which is} \quad \boldsymbol{\sigma_z^2=AVA^{\mathrm{T}}}. \qquad (17)$$

You can see that $\sigma_z^2=AVA^{\mathrm{T}}$ in (17) agrees with $\sigma_x^2+\sigma_y^2+2\sigma_{xy}$ in (16).

Now for the main point. The vector $\boldsymbol{X}$ could have M components coming from M experiments (instead of only 2). Those experiments will have an M by M covariance matrix $\boldsymbol{V_X}$. The matrix A could be K by M. Then $A\boldsymbol{X}$ is a vector with K combinations of the M outputs (instead of 1 combination $x+y$ of 2 outputs).

That vector $\boldsymbol{Z}=A\boldsymbol{X}$ of length K has a K by K covariance matrix $V_{\boldsymbol{Z}}$. Then the great rule for covariance matrices—of which equation (17) was only a 1 by 2 example—is this beautiful formula: Covariance matrix of $A\boldsymbol{X}$ is A (covariance matrix of $\boldsymbol{X}$) A^{T}:

$$\boxed{\textbf{The covariance matrix of } \boldsymbol{Z=AX} \ \textbf{ is } \ \boldsymbol{V_Z=AV_XA^{\mathrm{T}}}} \qquad (18)$$

To me, this neat formula shows the beauty of matrix multiplication. I won't prove this formula, just admire it. It is constantly used in applications—coming in Section 12.3.

The Correlation ρ

Correlation ρ_{xy} is closely related to covariance σ_{xy}. They both measure dependence or independence. Start by rescaling or "standardizing" the random variables x and y **The new $\boldsymbol{X = x/\sigma_x}$ and $\boldsymbol{Y = y/\sigma_y}$ have variance $\boldsymbol{\sigma_X^2 = \sigma_Y^2 = 1}$.** This is just like dividing a vector $\boldsymbol{v}$ by its length to produce a unit vector $\boldsymbol{v}/||\boldsymbol{v}||$ of length 1.

The correlation of $\boldsymbol{x}$ and $\boldsymbol{y}$ is the covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$. If the original covariance of x and y was σ_{xy}, then rescaling to X and Y will divide by σ_x and σ_y :

$$\boxed{\textbf{Correlation } \boldsymbol{\rho_{xy}} = \frac{\boldsymbol{\sigma_{xy}}}{\boldsymbol{\sigma_x \sigma_y}} = \textbf{covariance of } \frac{\boldsymbol{x}}{\boldsymbol{\sigma_x}} \textbf{ and } \frac{\boldsymbol{y}}{\boldsymbol{\sigma_y}}} \quad \textbf{Always } \boldsymbol{-1 \leq \rho_{xy} \leq 1}$$

Zero covariance gives zero correlation. *Independent random variables* produce $\rho_{xy} = 0$.

We know that always $\sigma_{xy}^2 \leq \sigma_x^2 \sigma_y^2$ (the covariance matrix V is at least positive semidefinite). Then $\boldsymbol{\rho_{xy}^2 \leq 1}$. Correlation near $\rho = +1$ means strong dependence in the same direction: often voting the same. Negative correlation means that y tends to be below its mean when x is above its mean: Voting in opposite directions.

Example 3 *Suppose that y is just $-x$.* A coin flip has outputs $x = 0$ or 1. The same flip has outputs $y = 0$ or -1. The mean m_x is $\frac{1}{2}$ for a fair coin, and m_y is $-\frac{1}{2}$. The covariance is $\sigma_{xy} = -\sigma_x \sigma_y$. The correlation divides by $\sigma_x \sigma_y$ to get $\boldsymbol{\rho_{xy} = -1}$. In this case the correlation matrix R has determinant zero (singular and only semidefinite):

$$\boxed{\textbf{Correlation matrix} \quad \boldsymbol{R} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\rho_{xy}} \\ \boldsymbol{\rho_{xy}} & \mathbf{1} \end{bmatrix}} \qquad \boldsymbol{R} = \begin{bmatrix} \mathbf{1} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{1} \end{bmatrix} \text{ when } \boldsymbol{y = -x}$$

R always has 1's on the diagonal because we normalized to $\sigma_X = \sigma_Y = 1$. R is the correlation matrix for x and y, and the covariance matrix for $X = x/\sigma_x$ and $Y = y/\sigma_y$.

That number ρ_{xy} is also called the Pearson coefficient.

Example 4 Suppose the random variables x, y, z are *independent. What matrix is R?*

Answer *R is the identity matrix.* All three correlations $\rho_{xx}, \rho_{yy}, \rho_{zz}$ are 1 by definition. All three cross-correlations $\rho_{xy}, \rho_{xz}, \rho_{yz}$ are zero by independence.

The correlation matrix R comes from the covariance matrix V, when we rescale every row and every column. Divide each row i and column i by the ith standard deviation σ_i.

(a) $R = DVD$ for the diagonal matrix $D = \text{diag}\,[1/\sigma_1, \ldots, 1/\sigma_M]$.

(b) If covariance V is positive definite, correlation $R = DVD$ is also positive definite.

■ WORKED EXAMPLES ■

12.2 A Suppose x and y are independent random variables with mean 0 and variance 1. Then the covariance matrix $\boldsymbol{V_X}$ for $\boldsymbol{X} = (x, y)$ is the 2 by 2 identity matrix. What are the mean $\boldsymbol{m_Z}$ and the covariance matrix $\boldsymbol{V_Z}$ for the 3-component vector $\boldsymbol{Z} = (x, y, ax+by)$?

Solution

$$\textbf{$\boldsymbol{Z}$ is connected to $\boldsymbol{X}$ by $\boldsymbol{A}$} \quad \boldsymbol{Z} = \begin{bmatrix} x \\ y \\ ax+by \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = A\boldsymbol{X}.$$

The vector $\boldsymbol{m_X}$ contains the means of the M components of $\boldsymbol{X}$. The vector $\boldsymbol{m_Z}$ contains the means of the K components of $\boldsymbol{Z} = A\boldsymbol{X}$. The matrix connection between the means of $\boldsymbol{X}$ and $\boldsymbol{Z}$ has to be linear: $\boldsymbol{m_Z} = A\,\boldsymbol{m_X}$. The mean of $ax + by$ is $am_x + bm_y$.

The covariance matrix for $\boldsymbol{Z}$ is $\boldsymbol{V_Z} = AA^{\mathrm{T}}$, when $\boldsymbol{V_X}$ is the 2 by 2 identity matrix:

$$\boldsymbol{V_Z} = \begin{matrix}\textbf{covariance matrix for} \\ \boldsymbol{Z} = (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{ax} + \boldsymbol{by})\end{matrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \boldsymbol{a} \\ \mathbf{0} & \mathbf{1} & \boldsymbol{b} \\ \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{a^2 + b^2} \end{bmatrix}.$$

Interpretation: x and y are independent so $\sigma_{xy} = 0$. Then the covariance of x with $ax + by$ is a and the covariance of y with $ax + by$ is b. Those just come from the two independent parts of $ax + by$. Finally, equation (18) gives the variance of $ax + by$:

$$\textbf{Use } \boldsymbol{V_Z} = \boldsymbol{A V_X A^{\mathrm{T}}} \qquad \sigma^2_{ax+by} = \sigma^2_{ax} + \sigma^2_{by} + 2\sigma_{ax,by} = a^2 + b^2 + 0.$$

The 3 by 3 matrix $\boldsymbol{V_Z}$ is *singular*. Its determinant is $a^2 + b^2 - a^2 - b^2 = 0$. The third component $z = ax + by$ is completely dependent on x and y. The rank of $\boldsymbol{V_Z}$ is only 2.

GPS Example The signal from a GPS satellite includes its departure time. The receiver clock gives the arrival time. The receiver multiplies the travel time by the speed of light. Then it knows the distance from that satellite. Distances from four or more satellites pinpoint the receiver position (using least squares !).

One problem: The speed of light changes in the ionosphere. But the correction will be almost the same for all nearby receivers. If one receiver stays in a known position, we can take differences from that position. **Differential GPS** reduces the error variance:

$$\begin{matrix}\textbf{Difference matrix} \\ \boldsymbol{A} = [\,\mathbf{1} \quad -\mathbf{1}\,]\end{matrix} \quad \begin{matrix}\textbf{Covariance matrix} \\ \boldsymbol{V_Z} = \boldsymbol{A V_X A^{\mathrm{T}}}\end{matrix} \quad \begin{aligned} \boldsymbol{V_Z} &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= \boldsymbol{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2} \end{aligned}$$

Errors in the speed of light are gone. Then centimeter positioning accuracy is achievable. (The key ideas are on page 320 of *Algorithms for Global Positioning* by Borre and Strang.) The GPS world is all about time and space and amazing accuracy.

Problem Set 12.2

1 (a) Compute the variance σ^2 when the coin flip probabilities are p and $1-p$ (tails $=0$, heads $=1$).

(b) The sum of N independent flips (0 or 1) is the count of heads after N tries. The rule (16-17-18) for the variance of a sum gives $\sigma^2 =$ _____ .

2 What is the covariance σ_{kl} between the results $x_1, \ldots, x_n$ of Experiment 3 and the results $y_1, \ldots, y_n$ of Experiment 5 ? Your formula will look like σ_{12} in equation (2). Then the $(3,5)$ and $(5,3)$ entries of the covariance matrix V are $\sigma_{35} = \sigma_{53}$.

3 For $M = 3$ experiments, the variance-covariance matrix V will be 3 by 3. There will be a probability p_{ijk} that the three outputs are x_i and y_j and z_k. Write down a formula like equation (7) for the matrix V.

4 What is the covariance matrix V for $M = 3$ independent experiments with means m_1, m_2, m_3 and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$?

Problems 5–9 are about the conditional probability that $\boldsymbol{Y = y_j}$ when we know $\boldsymbol{X = x_i}$. Notation: **Prob $\boldsymbol{(Y = y_j | X = x_i)}$** = probability of the outcome y_j given that $X = x_i$.

Example 1 *Coin* 1 *is glued to coin* 2. Then Prob (Y = heads when X = heads) is **1**.

Example 2 *Independent coin flips*: X gives no information about Y. Useless to know X. Then Prob (Y = heads $|X$ = heads) is the same as Prob (Y = heads).

5 Explain the **sum rule** of conditional probability:

$$\text{Prob}\,(Y = y_j) = \text{ sum over all outputs } x_i \text{ of Prob}\,(Y = y_j | X = x_i).$$

6 The n by n matrix P contains **joint probabilities $\boldsymbol{p_{ij}}$** = Prob ($X = x_i$ **and** $Y = y_j$).

Explain why the conditional Prob $(Y = y_j | X = x_i)$ equals $\dfrac{p_{ij}}{p_{i1} + \cdots + p_{in}} = \dfrac{p_{ij}}{p_i}$.

7 For this joint probability matrix with Prob $(x_1, y_2) = 0.3$, find Prob $(y_2|x_1)$ and Prob (x_1).

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix} \qquad \begin{array}{l}\text{The entries } p_{ij} \text{ add to } 1. \\ \text{Some } i, j \text{ must happen.}\end{array}$$

8 Explain the **product rule** of conditional probability:

p_{ij} = Prob ($X = x_i$ **and** $Y = y_j$) equals Prob $(Y = y_j | X = x_i)$ times Prob $(X = x_i)$.

9 Derive this **Bayes Theorem** for p_{ij} from the product rule in Problem 8:

$$\boxed{\text{Prob}\,(Y = y_j \textbf{ and } X = x_i) = \frac{\text{Prob}\,(X = x_i | Y = y_j)\ \text{Prob}\,(Y = y_j)}{\text{Prob}\,(X = x_i)}}$$

"Bayesians" use prior information. "Frequentists" only use sampling information.