Variance, covariance and correlation

Ten problems on variance, covariance and correlation: the shortcut formulas, the variance of a sum, bilinearity and the quadratic form $a^T\Sigma a$, the bounds on a correlation, a dependent pair with zero covariance, the covariance matrix as a positive semidefinite matrix, the covariance of a linear map and whitening, the variance of an average of correlated variables, the sample covariance as X^THX, and the correlation matrix with the bound on a common correlation, with worked solutions and the mistakes that add variances of correlated terms or read zero covariance as independence.

BEFORE YOU START

Variance and covariance are the second moments that nearly every calculation in machine learning runs on: the variance of a weighted sum is what initialization schemes control, the covariance matrix is what PCA diagonalizes and what the multivariate Gaussian is built from, and the variance of an average of correlated predictions is why ensembles stop improving. All of it follows from one algebraic fact, that covariance is bilinear, and from one inequality, that a variance is never negative. These ten problems derive the shortcut formulas, the variance of a sum and of an arbitrary linear combination, the bounds on a correlation, a pair of variables that are dependent but uncorrelated, the properties of a covariance matrix, how it transforms under a linear map, the variance of a mean of correlated terms, the sample covariance in matrix form and the correlation matrix, ending with the smallest correlation that d variables can all share. The five mistakes at the end are the ones that produce a plausible number: variances added across correlated terms, a linear map's covariance with its transposes swapped, zero covariance read as independence, a variance scaled by a factor instead of its square, and a sample covariance built from uncentered data.

- $\mathbb{E}$ is expectation, which is linear: $\mathbb{E}[au + bv] = a\,\mathbb{E}[u] + b\,\mathbb{E}[v]$ for constants a, b. A constant c has $\mathbb{E}[c] = c$.
- The variance of a random variable u with mean $m_u = \mathbb{E}[u]$ is $\mathrm{Var}(u) = \mathbb{E}[(u - m_u)^2]$, and its standard deviation is $\sigma_u = \sqrt{\mathrm{Var}(u)}$. The covariance of u and v is $\mathrm{Cov}(u, v) = \mathbb{E}[(u - m_u)(v - m_v)]$, so $\mathrm{Var}(u) = \mathrm{Cov}(u, u)$. When $\sigma_u, \sigma_v > 0$ the correlation is $\rho = \mathrm{Cov}(u, v)/(\sigma_u \sigma_v)$.
- u and v are uncorrelated if $\mathrm{Cov}(u, v) = 0$. They are independent if every event about u is independent of every event about v; independence gives $\mathbb{E}[g(u)h(v)] = \mathbb{E}[g(u)]\,\mathbb{E}[h(v)]$ for any functions g, h.
- A random vector $x \in \mathbb{R}^d$ has mean $\mu = \mathbb{E}[x]$ (entry by entry) and covariance matrix $\Sigma = \mathbb{E}[(x - \mu)(x - \mu)^\top]$, with $\Sigma_{ij} = \mathrm{Cov}(x_i, x_j)$ and $\Sigma_{ii} = \mathrm{Var}(x_i)$.

- Data are $X \in \mathbb{R}^{N \times d}$ with rows $x_n^\top$ (examples), sample mean $\bar{x} = \frac{1}{N}\sum_n x_n = \frac{1}{N} X^\top \mathbf{1}$ with $\mathbf{1}$ the all-ones vector, and sample covariance $S = \frac{1}{N-1}\sum_n (x_n - \bar{x})(x_n - \bar{x})^\top$, which is what `np.cov` computes; the maximum-likelihood page's estimate divides by N instead.
- A symmetric matrix M is positive semidefinite, written $M \succeq 0$, if $a^\top M a \geq 0$ for every a, equivalently if its eigenvalues are all ≥ 0 (the eigenvalues page).

Problems

1. Show that $\operatorname{Var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2$, that $\operatorname{Cov}(u, v) = \mathbb{E}[uv] - \mathbb{E}[u]\,\mathbb{E}[v]$, and that $\operatorname{Var}(au + b) = a^2 \operatorname{Var}(u)$ for constants a, b. difficulty ·

2. Show that $\operatorname{Var}(u + v) = \operatorname{Var}(u) + \operatorname{Var}(v) + 2\operatorname{Cov}(u, v)$ and $\operatorname{Var}(u - v) = \operatorname{Var}(u) + \operatorname{Var}(v) - 2\operatorname{Cov}(u, v)$. When do the variances simply add? difficulty ·

3. Show that $\operatorname{Cov}\left(\sum_i a_i u_i, \sum_j b_j v_j\right) = \sum_{i,j} a_i b_j \operatorname{Cov}(u_i, v_j)$. Deduce that for a random vector x with covariance Σ and constant vectors a, b, $\operatorname{Cov}(a^\top x, b^\top x) = a^\top \Sigma b$ and $\operatorname{Var}(a^\top x) = a^\top \Sigma a$. difficulty ··

4. Show that $-1 \leq \rho \leq 1$, that $\rho = \pm 1$ exactly when $v = au + b$ for constants with $a \neq 0$ having the sign of ρ, and that $\operatorname{corr}(au + b, cv + d) = \operatorname{sgn}(ac)\,\rho$ for $ac \neq 0$. difficulty ··

5. Let u be uniform on $\{-1, 0, 1\}$ and $v = u^2$. Compute $\operatorname{Cov}(u, v)$, show that u and v are not independent, and compute $\operatorname{Var}(u + v)$ directly and by Problem 2. difficulty ··

6. For a random vector $x \in \mathbb{R}^d$ with mean μ and covariance Σ, show that $\Sigma = \mathbb{E}[xx^\top] - \mu\mu^\top$, that Σ is symmetric and positive semidefinite, and that Σ is singular exactly when some $a \neq 0$ makes $a^\top x$ almost surely constant. difficulty ··

7. Let $y = Ax + b$ with $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ constant. Show that $\mathbb{E}[y] = A\mu + b$ and $\operatorname{Cov}(y) = A\Sigma A^\top$. Then, with $\Sigma = Q\Lambda Q^\top$ and every eigenvalue positive, show that $W = \Lambda^{-1/2} Q^\top$ gives $\operatorname{Cov}(Wx) = I$. difficulty ··

8. Let $x_1, \dots, x_N$ each have variance σ^2, and $\bar{x} = \frac{1}{N}\sum_n x_n$. Show that $\mathrm{Var}(\bar{x}) = \frac{1}{N^2}\mathbf{1}^\top \Sigma \mathbf{1}$ in general, that it is σ^2/N when the x_n are uncorrelated, and that it is $\sigma^2\big(\rho + \frac{1-\rho}{N}\big)$ when every pair has correlation ρ. What happens as $N \to \infty$ in the last case? `difficulty ··`

9. Let $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ ($N \times N$). Show that $HX = X - \mathbf{1}\bar{x}^\top$, the centered data, that H is a symmetric projection ($H^\top = H$ and $H^2 = H$), and that $S = \frac{1}{N-1}X^\top H X$. `difficulty ··`

10. The correlation matrix is $R = D^{-1/2}\Sigma D^{-1/2}$ with $D = \mathrm{diag}(\Sigma_{11}, \dots, \Sigma_{dd})$. Show that $R_{ij} = \mathrm{corr}(x_i, x_j)$, that R has unit diagonal and $R \succeq 0$. Then, for $R = (1-r)I + r\mathbf{1}\mathbf{1}^\top$, where every pair has the same correlation r, find the eigenvalues of R and deduce that it is a valid correlation matrix exactly when $-\frac{1}{d-1} \le r \le 1$. `difficulty ···`

Answers

1. $\operatorname{Var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2$; $\operatorname{Cov}(u, v) = \mathbb{E}[uv] - \mathbb{E}[u]\,\mathbb{E}[v]$; $\operatorname{Var}(au + b) = a^2 \operatorname{Var}(u)$
2. $\operatorname{Var}(u \pm v) = \operatorname{Var}(u) + \operatorname{Var}(v) \pm 2 \operatorname{Cov}(u, v)$; the variances add exactly when u and v are uncorrelated, which independent variables are
3. $\operatorname{Cov}\left(\sum_i a_i u_i, \sum_j b_j v_j\right) = \sum_{i,j} a_i b_j \operatorname{Cov}(u_i, v_j)$; $\operatorname{Cov}(a^\top x, b^\top x) = a^\top \Sigma b$; $\operatorname{Var}(a^\top x) = a^\top \Sigma a$
4. $-1 \le \rho \le 1$; $\rho = \pm 1$ exactly when $v = au + b$ with $\operatorname{sgn}(a) = \rho$; $\operatorname{corr}(au + b, cv + d) = \operatorname{sgn}(ac)\,\rho$
5. $\operatorname{Cov}(u, u^2) = 0$ although $v = u^2$ is a function of u; $\operatorname{Var}(u) = \frac{2}{3}$, $\operatorname{Var}(v) = \frac{2}{9}$ and $\operatorname{Var}(u + v) = \frac{8}{9} = \operatorname{Var}(u) + \operatorname{Var}(v)$, the cross term being 0
6. $\Sigma = \mathbb{E}[xx^\top] - \mu\mu^\top$; $\Sigma = \Sigma^\top$; $a^\top \Sigma a = \operatorname{Var}(a^\top x) \ge 0$ for all a; Σ is singular exactly when some $a \neq 0$ has $a^\top x$ almost surely constant
7. $\mathbb{E}[Ax + b] = A\mu + b$; $\operatorname{Cov}(Ax + b) = A\Sigma A^\top$; $W = \Lambda^{-1/2} Q^\top$ whitens: $\operatorname{Cov}(Wx) = I$
8. $\operatorname{Var}(\bar{x}) = \frac{1}{N^2} \mathbf{1}^\top \Sigma \mathbf{1}$; σ^2 / N when uncorrelated; $\sigma^2\left(\rho + \frac{1-\rho}{N}\right)$ under a common correlation ρ, which tends to $\rho\sigma^2$ rather than 0 as $N \to \infty$
9. $HX = X - \mathbf{1}\bar{x}^\top$; $H^\top = H = H^2$; $S = \frac{1}{N-1} X^\top H X$, and the maximum-likelihood page's $\frac{1}{N}$ version is $\frac{1}{N} X^\top H X$
10. $R_{ij} = \operatorname{corr}(x_i, x_j)$, $R_{ii} = 1$, $R \succeq 0$; $(1 - r)I + r\mathbf{1}\mathbf{1}^\top$ has eigenvalues $1 + (d-1)r$ (once, eigenvector $\mathbf{1}$) and $1 - r$ ($d - 1$ times), so it is a correlation matrix exactly when $-\frac{1}{d-1} \le r \le 1$

Worked solutions

Problem 1

> Show that $\operatorname{Var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2$, that $\operatorname{Cov}(u, v) = \mathbb{E}[uv] - \mathbb{E}[u]\,\mathbb{E}[v]$, and that $\operatorname{Var}(au + b) = a^2 \operatorname{Var}(u)$ for constants a, b.

With $m = \mathbb{E}[u]$: $(u - m)^2 = u^2 - 2mu + m^2$.	Expand the square; m is a number.
$\operatorname{Var}(u) = \mathbb{E}[u^2] - 2m\,\mathbb{E}[u] + m^2 = \mathbb{E}[u^2] - 2m^2 + m^2 = \mathbb{E}[u^2] - m^2$.	Linearity of $\mathbb{E}$, then $\mathbb{E}[u] = m$.
$(u - m_u)(v - m_v) = uv - m_u v - m_v u + m_u m_v$, with expectation $\mathbb{E}[uv] - m_u m_v - m_v m_u + m_u m_v = \mathbb{E}[uv] - m_u m_v$.	Expand, then linearity with $\mathbb{E}[v] = m_v$ and $\mathbb{E}[u] = m_u$.
$\mathbb{E}[au + b] = am + b$, so $au + b - (am + b) = a(u - m)$ and $\operatorname{Var}(au + b) = \mathbb{E}[a^2(u - m)^2] = a^2 \operatorname{Var}(u)$.	The shift b cancels in the deviation, and the constant a^2 comes out of the expectation.
$\operatorname{Var}(u) = \mathbb{E}[u^2] - (\mathbb{E}[u])^2$; $\operatorname{Cov}(u, v) = \mathbb{E}[uv] - \mathbb{E}[u]\,\mathbb{E}[v]$; $\operatorname{Var}(au + b) = a^2 \operatorname{Var}(u)$	The variance formula is the covariance formula with $v = u$. A shift leaves spread unchanged, and scaling by a scales the standard deviation by $\lvert a \rvert$, so the variance by a^2 (Mistake 4). Since a variance is an expectation of a square, $\mathbb{E}[u^2] \geq (\mathbb{E}[u])^2$ always, which is the initialization page's starting point.

Problem 2

> Show that $\operatorname{Var}(u+v) = \operatorname{Var}(u) + \operatorname{Var}(v) + 2\operatorname{Cov}(u,v)$ and $\operatorname{Var}(u-v) = \operatorname{Var}(u) + \operatorname{Var}(v) - 2\operatorname{Cov}(u,v)$. When do the variances simply add?

$\mathbb{E}[u+v] = m_u + m_v$, so $(u+v) - (m_u + m_v) = (u - m_u) + (v - m_v)$.	Linearity of $\mathbb{E}$.
$\big((u-m_u)+(v-m_v)\big)^2 = (u-m_u)^2 + 2(u-m_u)(v-m_v) + (v-m_v)^2$.	Expand the square.
$\operatorname{Var}(u+v) = \operatorname{Var}(u) + 2\operatorname{Cov}(u,v) + \operatorname{Var}(v)$.	Take expectations term by term; each is a definition.
$\operatorname{Cov}(u,-v) = -\operatorname{Cov}(u,v)$ and $\operatorname{Var}(-v) = \operatorname{Var}(v)$, so step 3 with $-v$ in place of v gives the difference.	$-v$ has mean $-m_v$ and deviation $-(v - m_v)$, which flips the sign of the cross term and leaves the square alone; Problem 1 with $a = -1$.
$\operatorname{Var}(u \pm v) = \operatorname{Var}(u) + \operatorname{Var}(v) \pm 2\operatorname{Cov}(u,v)$; **the variances add exactly when u and v are uncorrelated, which independent variables are**	Independence gives $\mathbb{E}[uv] = \mathbb{E}[u]\,\mathbb{E}[v]$, so $\operatorname{Cov} = 0$ by Problem 1; the converse fails (Problem 5). A sum of correlated variables can have variance anywhere between $(\sigma_u - \sigma_v)^2$ and $(\sigma_u + \sigma_v)^2$ (Problem 4), and adding variances regardless is Mistake 1. Problem 8 is this formula applied to N terms.

Problem 3

> Show that $\operatorname{Cov}\left(\sum_i a_i u_i, \sum_j b_j v_j\right) = \sum_{i,j} a_i b_j \operatorname{Cov}(u_i, v_j)$. Deduce that for a random vector x with covariance Σ and constant vectors a, b, $\operatorname{Cov}(a^\top x, b^\top x) = a^\top \Sigma b$ and $\operatorname{Var}(a^\top x) = a^\top \Sigma a$.

Let $U = \sum_i a_i u_i$ and $V = \sum_j b_j v_j$. Then $U - \mathbb{E}[U] = \sum_i a_i (u_i - \mathbb{E}[u_i])$ and likewise $V - \mathbb{E}[V] = \sum_j b_j (v_j - \mathbb{E}[v_j])$.	Linearity of $\mathbb{E}$ gives $\mathbb{E}[U] = \sum_i a_i \mathbb{E}[u_i]$; subtract.

$(U - \mathbb{E}[U])(V - \mathbb{E}[V]) = \sum_i \sum_j a_i b_j (u_i - \mathbb{E}[u_i])(v_j - \mathbb{E}[v_j])$.	Multiply the two sums term by term.
$\operatorname{Cov}(U, V) = \sum_{i,j} a_i b_j \, \mathbb{E}[(u_i - \mathbb{E}[u_i])(v_j - \mathbb{E}[v_j])] = \sum_{i,j} a_i b_j \operatorname{Cov}(u_i, v_j)$.	Linearity of $\mathbb{E}$ over the double sum.
With $u = v = x$: $\operatorname{Cov}(a^\top x, b^\top x) = \sum_{i,j} a_i \Sigma_{ij} b_j = a^\top \Sigma b$, and $b = a$ gives $\operatorname{Var}(a^\top x) = a^\top \Sigma a$.	$\Sigma_{ij} = \operatorname{Cov}(x_i, x_j)$, and $a^\top \Sigma b = \sum_{i,j} a_i \Sigma_{ij} b_j$ is the definition of the matrix product.
$\operatorname{Cov}\left(\sum_i a_i u_i, \sum_j b_j v_j\right) = \sum_{i,j} a_i b_j \operatorname{Cov}(u_i, v_j)$; $\operatorname{Cov}(a^\top x, b^\top x) = a^\top \Sigma b$; $\operatorname{Var}(a^\top x) = a^\top \Sigma a$	Covariance is bilinear, so it behaves like an inner product on random variables, and Σ is the matrix of that inner product on the coordinates of x. The quadratic form $a^\top \Sigma a$ is never negative because it is a variance, which is Problem 6, and it is the quantity PCA maximizes over unit vectors a on the PCA page.

Problem 4

> Show that $-1 \le \rho \le 1$, that $\rho = \pm 1$ exactly when $v = au + b$ for constants with $a \neq 0$ having the sign of ρ, and that $\operatorname{corr}(au + b, cv + d) = \operatorname{sgn}(ac)\, \rho$ for $ac \neq 0$.

Let $\tilde{u} = (u - m_u)/\sigma_u$ and $\tilde{v} = (v - m_v)/\sigma_v$. Then $\mathbb{E}[\tilde{u}] = \mathbb{E}[\tilde{v}] = 0$, $\operatorname{Var}(\tilde{u}) = \operatorname{Var}(\tilde{v}) = 1$ and $\operatorname{Cov}(\tilde{u}, \tilde{v}) = \operatorname{Cov}(u, v)/(\sigma_u \sigma_v) = \rho$.	Problem 1 with $a = 1/\sigma_u$ and $b = -m_u/\sigma_u$ for the variance; Problem 3 for the covariance.
$0 \le \operatorname{Var}(\tilde{u} - \tilde{v}) = 1 + 1 - 2\rho$, so $\rho \le 1$.	A variance is the expectation of a square; Problem 2 for the difference.
$0 \le \operatorname{Var}(\tilde{u} + \tilde{v}) = 2 + 2\rho$, so $\rho \ge -1$.	Problem 2 for the sum.

If $\rho = 1$ then $\operatorname{Var}(\tilde{u} - \tilde{v}) = 0$, so $\tilde{u} - \tilde{v}$ equals its mean 0 and $v = m_v + (\sigma_v/\sigma_u)(u - m_u)$, a line of positive slope; $\rho = -1$ gives $\tilde{v} = -\tilde{u}$, a line of negative slope. Conversely, if $v = au + b$ then $\operatorname{Cov}(u, v) = a \operatorname{Var}(u)$ and $\sigma_v = \lvert a \rvert \sigma_u$, so $\rho = a/\lvert a \rvert = \operatorname{sgn}(a)$.	A random variable with zero variance is constant; Problems 3 and 1 for the converse.
$\operatorname{Cov}(au + b, cv + d) = ac \operatorname{Cov}(u, v)$ and $\sigma_{au+b}\, \sigma_{cv+d} = \lvert a \rvert \lvert c \rvert \sigma_u \sigma_v$, so the correlation is $\dfrac{ac}{\lvert ac \rvert} \rho$.	Problem 3 for the covariance, with the shifts dropping out as in Problem 1; Problem 1 for the standard deviations.
$-1 \le \rho \le 1$; $\rho = \pm 1$ **exactly when** $v = au + b$ **with** $\operatorname{sgn}(a) = \rho$; $\operatorname{corr}(au + b, cv + d) = \operatorname{sgn}(ac)\, \rho$	This is the Cauchy–Schwarz inequality for the inner product of Problem 3. Correlation measures only linear association and ignores units: meters to feet or Celsius to Fahrenheit leaves it unchanged, and negating one variable flips its sign. $\rho = 0$ says nothing about nonlinear dependence (Problem 5).

Problem 5

> Let u be uniform on $\{-1, 0, 1\}$ and $v = u^2$. Compute $\operatorname{Cov}(u, v)$, show that u and v are not independent, and compute $\operatorname{Var}(u + v)$ directly and by Problem 2.

$\mathbb{E}[u] = 0$ and $\mathbb{E}[u^2] = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3}$, so $\operatorname{Var}(u) = \frac{2}{3}$.	Each value has probability $\frac{1}{3}$; Problem 1.
v is 1 with probability $\frac{2}{3}$ and 0 with probability $\frac{1}{3}$, so $\mathbb{E}[v] = \frac{2}{3}$, $\mathbb{E}[v^2] = \mathbb{E}[v] = \frac{2}{3}$ and $\operatorname{Var}(v) = \frac{2}{3} - \frac{4}{9} = \frac{2}{9}$.	$v^2 = v$ for a 0–1 variable; Problem 1.

$\mathbb{E}[uv] = \mathbb{E}[u^3] = \frac{1}{3}(-1+0+1) = 0$, so $\operatorname{Cov}(u, v) = 0 - 0 \cdot \frac{2}{3} = 0$.	Problem 1's shortcut.
$P(v = 1 \mid u = 0) = 0$ while $P(v = 1) = \frac{2}{3}$, so u and v are dependent.	Knowing $u = 0$ fixes $v = 0$; independence would need every conditional probability to equal the marginal (the Bayes page).
$u + v$ takes the values 0 (at $u = -1$), 0 (at $u = 0$) and 2 (at $u = 1$), so $\mathbb{E}[u+v] = \frac{2}{3}$, $\mathbb{E}[(u+v)^2] = \frac{4}{3}$ and $\operatorname{Var}(u+v) = \frac{4}{3} - \frac{4}{9} = \frac{8}{9}$.	Enumerate the three cases; Problem 1.
$\operatorname{Cov}(u, u^2) = 0$ **although** $v = u^2$ **is a function of** u; $\operatorname{Var}(u) = \frac{2}{3}$, $\operatorname{Var}(v) = \frac{2}{9}$ **and** $\operatorname{Var}(u+v) = \frac{8}{9} = \operatorname{Var}(u) + \operatorname{Var}(v)$, **the cross term being** 0	Covariance sees only linear association, and v is an even function of a symmetric u, so the positive and negative sides cancel. Zero correlation is enough for variances to add (Problem 2) but not for independence (Mistake 3): $\mathbb{E}[u^2 v] = \mathbb{E}[u^4] = \frac{2}{3}$ while $\mathbb{E}[u^2]\,\mathbb{E}[v] = \frac{4}{9}$. For jointly Gaussian variables the two notions coincide, which is the case people generalize from.

Problem 6

> For a random vector $x \in \mathbb{R}^d$ with mean μ and covariance Σ, show that $\Sigma = \mathbb{E}[xx^\top] - \mu\mu^\top$, that Σ is symmetric and positive semidefinite, and that Σ is singular exactly when some $a \neq 0$ makes $a^\top x$ almost surely constant.

$(x-\mu)(x-\mu)^\top = xx^\top - x\mu^\top - \mu x^\top + \mu\mu^\top$, with expectation $\mathbb{E}[xx^\top] - \mu\mu^\top - \mu\mu^\top + \mu\mu^\top = \mathbb{E}[xx^\top] - \mu\mu^\top$.	Expand the outer product; $\mathbb{E}$ applies entry by entry and $\mathbb{E}[x] = \mu$.
$\Sigma^\top = \mathbb{E}\big[\big((x-\mu)(x-\mu)^\top\big)^\top\big] = \Sigma$.	$(ab^\top)^\top = ba^\top$, and here $a = b$.
For every $a \in \mathbb{R}^d$, $a^\top \Sigma a = \operatorname{Var}(a^\top x) \geq 0$.	Problem 3; a variance is nonnegative.

If $\Sigma a = 0$ for some $a \neq 0$ then $a^\top \Sigma a = 0$, so $\mathrm{Var}(a^\top x) = 0$ and $a^\top x$ equals its mean almost surely. Conversely, if $\mathrm{Var}(a^\top x) = 0$ then $a^\top \Sigma a = 0$; writing $\Sigma = Q \Lambda Q^\top$, that is $\sum_j \lambda_j (q_j^\top a)^2 = 0$ with every $\lambda_j \geq 0$, so $\lambda_j q_j^\top a = 0$ for each j and $\Sigma a = \sum_j \lambda_j q_j (q_j^\top a) = 0$.	A singular matrix has a nonzero null vector; the eigenvalues page's decomposition of a symmetric matrix, with step 3 giving $\lambda_j \geq 0$.
$\Sigma = \mathbb{E}[xx^\top] - \mu\mu^\top$; $\Sigma = \Sigma^\top$; $a^\top \Sigma a = \mathrm{Var}(a^\top x) \geq 0$ **for all** a; Σ **is singular exactly when some** $a \neq 0$ **has** $a^\top x$ **almost surely constant**	A singular covariance says the data lie on a hyperplane. One-hot encoding a category with all K columns does this, since the columns sum to 1 on every example, which is the dummy-variable trap of linear regression. The multivariate-Gaussian page needs Σ invertible to write its density, and the PCA page reads the eigenvalues of Σ as variances along the q_j.

Problem 7

> Let $y = Ax + b$ with $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ constant. Show that $\mathbb{E}[y] = A\mu + b$ and $\mathrm{Cov}(y) = A \Sigma A^\top$. Then, with $\Sigma = Q \Lambda Q^\top$ and every eigenvalue positive, show that $W = \Lambda^{-1/2} Q^\top$ gives $\mathrm{Cov}(Wx) = I$.

$\mathbb{E}[y] = A\,\mathbb{E}[x] + b = A\mu + b.$	Linearity, entry by entry: $y_i = \sum_j A_{ij} x_j + b_i$.
$y - \mathbb{E}[y] = A(x - \mu).$	Subtract step 1 from y; b cancels.
$\mathrm{Cov}(y) = \mathbb{E}\big[A(x-\mu)(x-\mu)^\top A^\top\big] = A\,\mathbb{E}[(x-\mu)(x-\mu)^\top]\,A^\top = A \Sigma A^\top.$	$(Av)(Av)^\top = Avv^\top A^\top$, and the constant matrices come out of the expectation.
$\mathrm{Cov}(Wx) = W \Sigma W^\top = \Lambda^{-1/2} Q^\top \, Q \Lambda Q^\top \, Q \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I.$	Step 3 with $A = W$; $Q^\top Q = I$ because Q is orthogonal, and diagonal matrices multiply entry by entry.

$\mathbb{E}[Ax+b] = A\mu + b$; $\mathrm{Cov}(Ax+b) = A\Sigma A^\top$; $W = \Lambda^{-1/2}Q^\top$ **whitens:** $\mathrm{Cov}(Wx) = I$	Shapes: $(m \times d)(d \times d)(d \times m) = m \times m$, symmetric and positive semidefinite as a covariance must be; the version with the transposes swapped is Mistake 2. The multivariate-Gaussian page proves the same for Gaussian x and adds that y is Gaussian; nothing here needed a distribution. Any W with $W\Sigma W^\top = I$ whitens, for instance L^{-1} from the Cholesky factor $\Sigma = LL^\top$; the PCA page's scores are $Q^\top(x-\mu)$, whitened by $\Lambda^{-1/2}$.

Problem 8

> Let $x_1, \dots, x_N$ each have variance σ^2, and $\bar{x} = \frac{1}{N}\sum_n x_n$. Show that $\mathrm{Var}(\bar{x}) = \frac{1}{N^2}\mathbf{1}^\top\Sigma\mathbf{1}$ in general, that it is σ^2/N when the x_n are uncorrelated, and that it is $\sigma^2\big(\rho + \frac{1-\rho}{N}\big)$ when every pair has correlation ρ. What happens as $N \to \infty$ in the last case?

$\bar{x} = a^\top x$ with $a = \frac{1}{N}\mathbf{1}$, so $\mathrm{Var}(\bar{x}) = a^\top\Sigma a = \frac{1}{N^2}\mathbf{1}^\top\Sigma\mathbf{1} = \frac{1}{N^2}\sum_{n,m}\Sigma_{nm}$.	Problem 3; $\mathbf{1}^\top\Sigma\mathbf{1}$ adds up every entry of Σ.
Uncorrelated: $\Sigma = \sigma^2 I$, so $\mathbf{1}^\top\Sigma\mathbf{1} = N\sigma^2$ and $\mathrm{Var}(\bar{x}) = \sigma^2/N$.	Only the N diagonal entries are nonzero.
Common correlation: $\Sigma_{nn} = \sigma^2$ and $\Sigma_{nm} = \rho\sigma^2$ for $n \neq m$, so $\mathbf{1}^\top\Sigma\mathbf{1} = N\sigma^2 + N(N-1)\rho\sigma^2$.	N diagonal entries and $N(N-1)$ off-diagonal ones.
$\mathrm{Var}(\bar{x}) = \sigma^2\,\dfrac{N + N(N-1)\rho}{N^2} = \sigma^2\,\dfrac{1 + (N-1)\rho}{N} = \sigma^2\Big(\rho + \dfrac{1-\rho}{N}\Big)$.	Divide by N^2, then write $1 + (N-1)\rho = N\rho + (1-\rho)$.

$\mathrm{Var}(\bar{x}) = \frac{1}{N^2}\mathbf{1}^\top\Sigma\mathbf{1}$; σ^2/N **when uncorrelated;** $\sigma^2\left(\rho + \frac{1-\rho}{N}\right)$ **under a common correlation** ρ**, which tends to** $\rho\sigma^2$ **rather than** 0 **as** $N \to \infty$	Averaging removes only the uncorrelated part of the variance. This is the ensemble calculation: averaging N models whose errors have correlation ρ can reduce the variance by at most a factor ρ however many models are used, which is why random forests decorrelate their trees by subsampling features, and why Mistake 1 at scale, σ^2/N for correlated models, overstates the gain. Problem 10 says which values of ρ are possible.

Problem 9

> Let $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ ($N \times N$). Show that $HX = X - \mathbf{1}\bar{x}^\top$, the centered data, that H is a symmetric projection ($H^\top = H$ and $H^2 = H$), and that $S = \frac{1}{N-1}X^\top HX$.

$\mathbf{1}^\top X = N\bar{x}^\top$, so $\frac{1}{N}\mathbf{1}\mathbf{1}^\top X = \mathbf{1}\bar{x}^\top$ and $HX = X - \mathbf{1}\bar{x}^\top$.	$\mathbf{1}^\top X$ is the row of column sums, which is N times the row of column means; $\mathbf{1}\bar{x}^\top$ repeats $\bar{x}^\top$ in every row, so row n of HX is $(x_n - \bar{x})^\top$.
$H^\top = I - \frac{1}{N}(\mathbf{1}\mathbf{1}^\top)^\top = H$.	$(\mathbf{1}\mathbf{1}^\top)^\top = \mathbf{1}\mathbf{1}^\top$.
$H^2 = I - \frac{2}{N}\mathbf{1}\mathbf{1}^\top + \frac{1}{N^2}\mathbf{1}(\mathbf{1}^\top\mathbf{1})\mathbf{1}^\top = I - \frac{2}{N}\mathbf{1}\mathbf{1}^\top + \frac{1}{N}\mathbf{1}\mathbf{1}^\top = H$.	Expand; $\mathbf{1}^\top\mathbf{1} = N$.
$\sum_n (x_n - \bar{x})(x_n - \bar{x})^\top = (HX)^\top(HX) = X^\top H^\top HX = X^\top HX$.	$M^\top M$ is the sum of the outer products of the rows of M, and row n of HX is $(x_n - \bar{x})^\top$; then steps 2 and 3.
$HX = X - \mathbf{1}\bar{x}^\top$; $H^\top = H = H^2$; $S = \frac{1}{N-1}X^\top HX$, **and the maximum-likelihood page's** $\frac{1}{N}$ **version is** $\frac{1}{N}X^\top HX$	H projects onto the vectors whose entries sum to zero (the least-squares page's projections), and centering is that projection applied to each column. The two normalizations differ by a scalar, so they share eigenvectors; `np.cov(X.T)` divides by $N-1$, `bias=True` by N. Skipping the H is Mistake 5.

Problem 10

> The correlation matrix is $R = D^{-1/2}\Sigma D^{-1/2}$ with $D = \operatorname{diag}(\Sigma_{11}, \dots, \Sigma_{dd})$. Show that $R_{ij} = \operatorname{corr}(x_i, x_j)$, that R has unit diagonal and $R \succeq 0$. Then, for $R = (1-r)I + r\mathbf{1}\mathbf{1}^\top$, where every pair has the same correlation r, find the eigenvalues of R and deduce that it is a valid correlation matrix exactly when $-\frac{1}{d-1} \le r \le 1$.

$(D^{-1/2}\Sigma D^{-1/2})_{ij} = \dfrac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}} = \dfrac{\operatorname{Cov}(x_i, x_j)}{\sigma_i \sigma_j}$, which is 1 when $i = j$.	A diagonal matrix on the left scales row i by its entry and on the right scales column j.
$R = \operatorname{Cov}(D^{-1/2}x) \succeq 0$.	Problem 7 with $A = D^{-1/2}$, which is symmetric: R is the covariance of the standardized variables x_i/σ_i, and every covariance is positive semidefinite by Problem 6.
$R\mathbf{1} = (1-r)\mathbf{1} + r\mathbf{1}(\mathbf{1}^\top\mathbf{1}) = \big(1 + (d-1)r\big)\mathbf{1}$.	$\mathbf{1}^\top\mathbf{1} = d$; so $\mathbf{1}$ is an eigenvector with eigenvalue $1 + (d-1)r$.
For any a with $\mathbf{1}^\top a = 0$: $Ra = (1-r)a + r\mathbf{1}(\mathbf{1}^\top a) = (1-r)a$.	The zero-sum vectors form a $(d-1)$-dimensional space, so $1-r$ is an eigenvalue with multiplicity $d-1$, and with step 3 all d eigenvalues are found.
$R \succeq 0$ exactly when $1 - r \ge 0$ and $1 + (d-1)r \ge 0$, that is when $-\frac{1}{d-1} \le r \le 1$.	A symmetric matrix is positive semidefinite exactly when its eigenvalues are nonnegative (the eigenvalues page).
$R_{ij} = \operatorname{corr}(x_i, x_j)$, $R_{ii} = 1$, $R \succeq 0$; $(1-r)I + r\mathbf{1}\mathbf{1}^\top$ **has eigenvalues** $1 + (d-1)r$ **(once, eigenvector** $\mathbf{1}$**) and** $1 - r$ **(**$d-1$ **times), so it is a correlation matrix exactly when** $-\frac{1}{d-1} \le r \le 1$	Three variables cannot all be pairwise correlated at -0.9: the lower bound is $-\frac{1}{2}$ for $d = 3$ and tends to 0 as d grows, so a large set of variables can be only slightly negatively correlated on average. At $r = -\frac{1}{d-1}$ the sum $\mathbf{1}^\top x$ has variance $\mathbf{1}^\top R\mathbf{1} = d\big(1 + (d-1)r\big) = 0$, Problem 6's singular case. Problem 8's variance of the mean is nonnegative on exactly this range.

Where this goes wrong

1. Adding the variances of correlated terms

Variances add for independent variables, which is the case every first example uses.

$\mathrm{Var}(u+v) = \mathrm{Var}(u) + \mathrm{Var}(v) + 2\,\mathrm{Cov}(u,v)$	Right so far: Problem 2.
“Variances add.”	The habit that causes the mistake: additivity learned on independent variables and carried over to a pair that is correlated.
$\mathrm{Var}(u+v) = \mathrm{Var}(u) + \mathrm{Var}(v)$ ✗	With $\rho = \frac{1}{2}$ and equal variances σ^2 the true variance is $3\sigma^2$, not $2\sigma^2$, and with $v = c - u$ the sum is constant while the formula says $2\sigma^2$. At scale it is Problem 8 with the correlations dropped: an average of N correlated models is credited with variance σ^2/N, which it cannot reach, because $\rho\sigma^2$ remains however large N is.

2. Covariance of a linear map written with the transposes swapped

Quadratic forms are written $a^\top \Sigma a$, and the matrix version looks like the same thing with A in place of a.

$y = Ax + b$ with $\mathrm{Cov}(y) = \mathbb{E}[(y - \mathbb{E}[y])(y - \mathbb{E}[y])^\top]$	Right so far: the definition, as in Problem 7.
“$\mathrm{Var}(a^\top x) = a^\top \Sigma a$, so $\mathrm{Cov}(Ax) = A^\top \Sigma A$.”	The analogy that causes the mistake: the vector formula has its transpose on the left because the map is $a^\top x$, a row times x; for Ax the map's transpose lands on the right.

$\mathrm{Cov}(Ax + b) = A^\top \Sigma A$ ✗	For $A \in \mathbb{R}^{m \times d}$ with $m \neq d$ the product $A^\top \Sigma A$ is not even defined, since $A^\top$ is $d \times m$ and Σ is $d \times d$. When $m = d$ it is a different matrix: for $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\Sigma = I$, $A\Sigma A^\top = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ while $A^\top \Sigma A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. The vector case is the $m = 1$ instance of Problem 7 with $A = a^\top$, and $(a^\top)\Sigma(a^\top)^\top = a^\top \Sigma a$ agrees.

3. Zero covariance read as independence

Independence implies zero covariance, and the implication is usually learned in that direction only.

$\mathrm{Cov}(u, v) = 0$	Right so far: a computed fact, as for u and u^2 in Problem 5.
"Zero covariance means no relation, so u and v are independent."	The converse that causes the mistake: true for jointly Gaussian variables, and generalized from there.
$P(v = 1 \mid u = 0) = P(v = 1)$ for $v = u^2$ ✗	Problem 5: the left side is 0 and the right is $\frac{2}{3}$. Covariance is one number measuring linear association, and a symmetric nonlinear dependence leaves it at 0. The consequences are not only probabilistic: $\mathbb{E}[uv]$ factorizes but $\mathbb{E}[u^2 v] = \frac{2}{3} \neq \mathbb{E}[u^2]\,\mathbb{E}[v] = \frac{4}{9}$, so any calculation that uses independence beyond the second moment is wrong.

4. Scaling a variance by the factor instead of its square

Expectations scale linearly, and a variance is an expectation.

$\mathbb{E}[au] = a\,\mathbb{E}[u]$	Right so far: linearity.
"Variance is an expectation, so it scales the same way."	The analogy that causes the mistake: variance is the expectation of a square, and the factor sits inside the square.

$\mathrm{Var}(au) = a\,\mathrm{Var}(u)$ ✗	Problem 1: $\mathrm{Var}(au) = a^2\,\mathrm{Var}(u)$. Converting meters to centimeters multiplies a variance by 10^4, not 100, and halving a layer's outputs quarters their variance, which is the square in the initialization page's $\mathrm{Var}(wx) = \sigma_w^2\,\mathbb{E}[x^2]$. For $a < 0$ the wrong formula makes a variance negative. It is correct for the standard deviation, which scales by $\lvert a \rvert$, and that is where it comes from.

5. Sample covariance from uncentered data

$X^\top X$ is the Gram matrix of least squares and of every normal equation, and the covariance looks like it with a $\frac{1}{N-1}$ in front.

$S = \frac{1}{N-1}\sum_n (x_n - \bar{x})(x_n - \bar{x})^\top$	Right so far: the definition, as in Problem 9.
"The sum of the outer products of the rows is $X^\top X$."	The shortcut that causes the mistake: that is right for the rows of X, but the rows summed here are those of the centered HX.
$S = \frac{1}{N-1} X^\top X$ ✗	Expanding $x_n = (x_n - \bar{x}) + \bar{x}$ gives $\sum_n x_n x_n^\top = \sum_n (x_n - \bar{x})(x_n - \bar{x})^\top + N\bar{x}\bar{x}^\top$, so the line computes $S + \frac{N}{N-1}\bar{x}\bar{x}^\top$: the covariance plus a rank-one term that dominates whenever the means are large next to the spread. Its leading eigenvector then points toward $\bar{x}$ rather than along the direction of greatest spread, which is the PCA page's first mistake. The fix is $X^\top H X$, or subtracting the column means before forming the product.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 9 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/variance-covariance-and-correlation