

Statistical models and the multivariate Gaussian

Hicham Janati

Fall 2026

Most datasets record several quantities at once: the mathematics and physics scores of each candidate in an entrance exam, blood pressure *and* cholesterol for each patient, the amount, hour and merchant category of each card transaction, the expression of thousands of genes in one cell. The prologue turned this observation into a program: treat the rows of such a table as independent draws of a random vector, fix a family of candidate distributions, let the data select a member, then interrogate it. This chapter carries out the program once, in full, with the central family for multivariate data: the multivariate normal, or Gaussian, distribution. By the end of the chapter we will know what this model can do (describe dependence, calibrate anomalies, impute missing values, quantify its own uncertainty), what it costs (a covariance matrix whose size grows quadratically with the dimension), and where it breaks (data with groups). That failure motivates the next chapter.

Throughout, the prologue’s notational convention holds: bold letters denote vectors, plain italic letters denote one-dimensional quantities. $\mathbf{X}$ is a random vector in $\mathbb{R}^d$ with components $X_1, \dots, X_d$, and Σ its covariance matrix. Two toolboxes are used constantly and collected in the appendices: the spectral decomposition of symmetric matrices (appendix A) and the basic calculus of random vectors (expectations, covariance matrices, characteristic functions; appendix B). If you have not manipulated these objects recently, read the appendices first: every argument in this chapter uses them.

1 Gaussian vectors: definition, density, geometry

The chapter works with the kind of table the course opened with: one row per observation, one column per measured quantity; in generic notation, d variables $X_1, \dots, X_d$ measured on each of n observation units,

	X_1	X_2	$\cdots$	X_d
observation 1	x_{11}	x_{12}	$\cdots$	x_{1d}
observation 2	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
observation n	x_{n1}	x_{n2}	$\cdots$	x_{nd}

so each row is a point $\mathbf{x}_i \in \mathbb{R}^d$: a customer, a patient, a transaction, an exam candidate, whatever the unit of observation is. The prologue laid out what a distribution P fitted to such rows buys (typical rows, dependence between columns, calibrated anomalies, missing entries filled in, synthetic data) and how the impossible task of guessing a whole distribution is made tractable: by fixing a family of candidate distributions in advance. One definition, met in your first-year courses in its univariate form, makes this precise.

Reminder

(*Parametric statistical model.*) A *statistical model* for observations $\mathbf{x}_1, \dots, \mathbf{x}_n$ is a family of probability distributions

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\},$$

together with the assumption that the observations are independent draws from P_{θ^*} for some unknown $\theta^* \in \Theta$. The model is *parametric* when $\Theta \subset \mathbb{R}^k$ for a finite k .

A parametric model is a drastic compression: instead of an unknown function (a full distribution on $\mathbb{R}^d$), the unknown is a vector of k numbers. Every request made of the table becomes a question about θ^* , and “learning the distribution” becomes “estimating k parameters from n rows”, a competition between k and n that will be a recurring theme of this course, starting in section 6.2.

It remains to choose the family, and this chapter develops the simplest serious candidate for multivariate data: the multivariate analogue of the distribution every first-year course reaches for first.

From one dimension to d . On the real line, the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the workhorse of your first-year statistics courses, has density $f(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}$: one center μ , one spread σ . How should one define its analogue for a random vector $\mathbf{X} = (X_1, \dots, X_d)^\top$ (a distribution for whole rows of the table rather than single numbers)? A first guess is to require each component X_i to be Gaussian. This guess is *wrong*, and seeing why it fails leads to the correct definition.

Take $X \sim \mathcal{N}(0, 1)$, flip an independent fair coin $\varepsilon \in \{-1, +1\}$, and set $Y = \varepsilon X$. Flipping the sign of a symmetric random variable does not change its distribution, so Y is also a standard Gaussian. But now consider the pair $\mathbf{Z} = (X, Y)$ and look at the sum of its components:

$$X + Y = (1 + \varepsilon)X = \begin{cases} 2X & \text{with probability } 1/2, \\ 0 & \text{with probability } 1/2. \end{cases}$$

Half of the mass of $X+Y$ sits exactly at 0: this distribution has an atom, so it cannot be Gaussian. The pair $\mathbf{Z}$ has two perfectly Gaussian coordinates, yet something un-Gaussian happens as soon as we combine them. Componentwise Gaussianity does not constrain the joint behavior, which is the object of this chapter. The correct definition demands Gaussianity of every one-dimensional projection of the vector:

Definition 1.1

A random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance Σ is a *Gaussian vector*, written $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, if every linear combination of its components is univariate Gaussian:

$$\forall \boldsymbol{\alpha} \in \mathbb{R}^d, \quad \boldsymbol{\alpha}^\top \mathbf{X} = \sum_{i=1}^d \alpha_i X_i \sim \mathcal{N}(\boldsymbol{\alpha}^\top \boldsymbol{\mu}, \boldsymbol{\alpha}^\top \Sigma \boldsymbol{\alpha}). \quad (1)$$

The parameters of $\boldsymbol{\alpha}^\top \mathbf{X}$ are forced: its mean is $\boldsymbol{\alpha}^\top \boldsymbol{\mu}$ by linearity of expectation, and its variance is $\boldsymbol{\alpha}^\top \Sigma \boldsymbol{\alpha}$ by the identity $\mathbb{V}(\boldsymbol{\alpha}^\top \mathbf{X}) = \boldsymbol{\alpha}^\top \mathbb{V}(\mathbf{X}) \boldsymbol{\alpha}$ (appendix B). The definition is legitimate (it pins down one distribution) thanks to the Cramér–Wold theorem (appendix B): the distribution of a random vector is entirely determined by the distributions of its one-dimensional projections $\mathbf{t}^\top \mathbf{X}$. In our counter-example, $\mathbf{Z} = (X, \varepsilon X)$ fails the test at $\boldsymbol{\alpha} = (1, 1)^\top$: it is *not* a Gaussian vector, although both its marginals are.

The density. When Σ is invertible, the Gaussian vector admits a density, which is the direct generalization of the 1D formula: the squared standardized deviation $(x - \mu)^2/\sigma^2$ becomes a quadratic form in the inverse covariance.

Property 1.1

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with Σ invertible, then $\mathbf{X}$ has density

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^d. \quad (2)$$

Two sanity checks help make this formula familiar. For $d = 1$, $\Sigma = \sigma^2$ and we recover the 1D density. For $\Sigma = \mathbf{I}_d$ and $\boldsymbol{\mu} = \mathbf{0}$, the density factorizes as $\prod_i (2\pi)^{-1/2} e^{-x_i^2/2}$: independent

standard Gaussian coordinates. The general case is, as we will see in section 3, nothing but a linear deformation of this product case.

Geometry: where does the mass sit? The density is constant on the level sets

$$\mathcal{E}_r = \left\{ \mathbf{x} \in \mathbb{R}^d : (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = r^2 \right\}.$$

To read this equation, diagonalize the covariance: $\Sigma = P\Lambda P^\top$ with P orthogonal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ (appendix A). In the rotated, centered coordinates $\mathbf{y} = P^\top(\mathbf{x} - \boldsymbol{\mu})$, the equation becomes

$$\sum_{i=1}^d \frac{y_i^2}{\lambda_i} = r^2,$$

the equation of an ellipsoid with axes aligned with the coordinate axes and semi-axis lengths $r\sqrt{\lambda_i}$. Undoing the rotation, $\mathcal{E}_r$ is an ellipsoid centered at $\boldsymbol{\mu}$ whose principal axes are the eigenvectors $\mathbf{e}_i$ of Σ and whose semi-axes have lengths $r\sqrt{\lambda_i}$ (figure 1). A common trap: since Σ^{-1} appears in the density, one is tempted to read the geometry off Σ^{-1} . But inverting a symmetric matrix keeps the eigenvectors and inverts the eigenvalues, so the *largest* eigenvalue of Σ (the direction of largest variance) gives the *longest* axis of the ellipsoid, as it should.

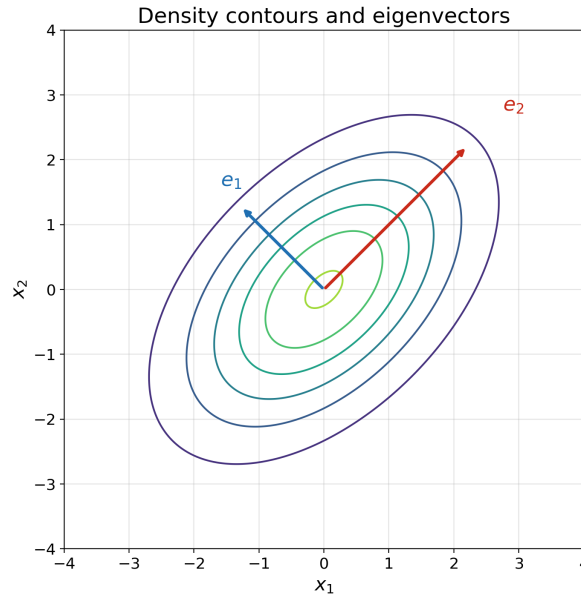


Figure 1: Density contours of $\mathcal{N}(0, \Sigma)$, with the eigenvectors of Σ (here $\Sigma_{11} = \Sigma_{22} = 2$, $\Sigma_{12} = 1$). The long axis follows the top eigenvector of Σ .

Degenerate Gaussians. Definition 1.1 does not require Σ to be invertible; only the density formula does. What happens when Σ is singular? Some direction $\boldsymbol{\alpha} \neq 0$ satisfies $\mathbb{V}(\boldsymbol{\alpha}^\top \mathbf{X}) = \boldsymbol{\alpha}^\top \Sigma \boldsymbol{\alpha} = 0$: the vector lies, with probability one, in the affine subspace $\{\mathbf{x} : \boldsymbol{\alpha}^\top (\mathbf{x} - \boldsymbol{\mu}) = 0\}$. The distribution then concentrates on $\boldsymbol{\mu} + \text{Im}(\Sigma)$, a subspace of dimension $\text{rank}(\Sigma) < d$, and no density on $\mathbb{R}^d$ exists. This is not a pathology reserved to textbooks: figure 2 shows samples from a bivariate Gaussian with unit variances and correlation ρ . As $\rho \rightarrow 1$ the ellipse flattens, and at $\rho = 1$ the cloud collapses onto the line $x_2 = x_1$: a legitimate Gaussian vector living on a one-dimensional subspace. Singular covariance matrices return in section 6.2, produced there by estimation itself when the dimension approaches the sample size.

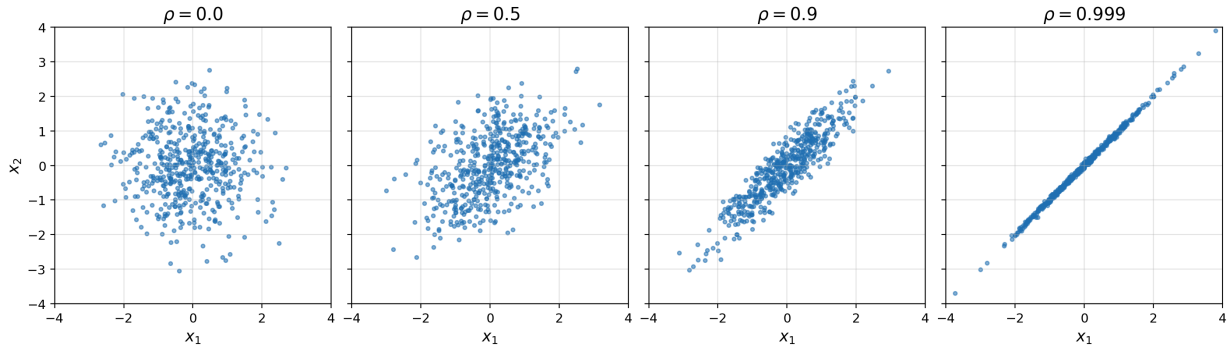


Figure 2: Samples of $\mathcal{N}(0, \Sigma)$ with unit variances and correlation ρ : the cloud flattens onto a line as $\rho \rightarrow 1$.

2 Marginals and independence

Partition the vector into two blocks: pick $1 \leq r < d$ and write

$$\mathbf{X} = \underbrace{(X_1, \dots, X_r)}_{\mathbf{X}_a \in \mathbb{R}^r}, \underbrace{(X_{r+1}, \dots, X_d)}_{\mathbf{X}_b \in \mathbb{R}^{d-r}}^\top.$$

Each block is itself a random *vector*: $\mathbf{X}_a$ of dimension r , $\mathbf{X}_b$ of dimension $d-r$, the two dimensions summing to d ; blocks of size one are just the special case $r = 1$ or $r = d - 1$. The mean and covariance split accordingly:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix},$$

where $\boldsymbol{\mu}_a \in \mathbb{R}^r$, $\Sigma_{aa} \in \mathbb{R}^{r \times r}$ and $\Sigma_{bb} \in \mathbb{R}^{(d-r) \times (d-r)}$ are the within-block means and covariances, while the rectangular block $\Sigma_{ab} = \Sigma_{ba}^\top \in \mathbb{R}^{r \times (d-r)}$ collects all the covariances *between* a coordinate of $\mathbf{X}_a$ and a coordinate of $\mathbf{X}_b$. Two natural questions arise immediately. What is the distribution of $\mathbf{X}_a$ alone (the marginal)? And when are $\mathbf{X}_a$ and $\mathbf{X}_b$ independent? For Gaussian vectors both have closed-form answers, read off directly from the blocks. A third question, what *observing* $\mathbf{X}_b$ teaches us about $\mathbf{X}_a$, has an equally explicit answer, and it is the key to filling in missing entries.

Marginals. A linear combination of the components of $\mathbf{X}_a$ is in particular a linear combination of the components of $\mathbf{X}$, so it is univariate Gaussian: by definition 1.1, $\mathbf{X}_a$ is itself a Gaussian vector, and reading off its mean and covariance gives $\mathbf{X}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \Sigma_{aa})$. Marginalizing a Gaussian is as simple as extracting sub-blocks. Figure 3 shows the phenomenon on simulated data: however correlated the cloud, its shadow on each axis is a Gaussian histogram.

Independence. The most efficient tool here is the characteristic function, which for Gaussian vectors has a remarkably simple form.

Property 2.1

For $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$,

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E} \left[e^{i \mathbf{t}^\top \mathbf{X}} \right] = \exp \left(i \mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2} \mathbf{t}^\top \Sigma \mathbf{t} \right), \quad \mathbf{t} \in \mathbb{R}^d. \quad (3)$$

Write $\mathbf{t} = (\mathbf{t}_a, \mathbf{t}_b)$ and expand the quadratic form by blocks:

$$\mathbf{t}^\top \Sigma \mathbf{t} = \mathbf{t}_a^\top \Sigma_{aa} \mathbf{t}_a + \mathbf{t}_b^\top \Sigma_{bb} \mathbf{t}_b + 2 \mathbf{t}_a^\top \Sigma_{ab} \mathbf{t}_b.$$

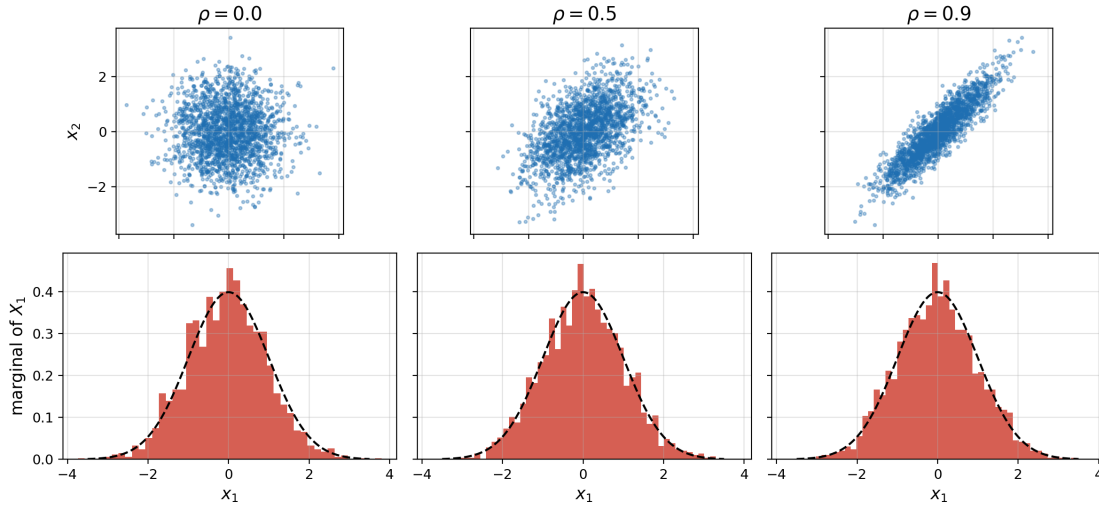


Figure 3: Correlated Gaussian samples (top row) and the marginal distribution of X_1 (bottom row): Gaussian at every correlation level.

If $\Sigma_{ab} = 0$, the cross term vanishes and $\phi_{\mathbf{X}}(\mathbf{t})$ factorizes as $\phi_{\mathbf{X}_a}(\mathbf{t}_a)\phi_{\mathbf{X}_b}(\mathbf{t}_b)$, which characterizes independence (appendix B); the converse is immediate since independence forces zero covariance. We have proved:

Theorem 2.1

For a Gaussian vector, $\mathbf{X}_a \perp\!\!\!\perp \mathbf{X}_b \iff \Sigma_{ab} = 0$: independence is equivalent to zero correlation.

This equivalence is a property of jointly Gaussian vectors, not of Gaussian marginals (the pair $(X, \varepsilon X)$ of section 1 has uncorrelated standard Gaussian coordinates, yet $|\varepsilon X| = |X|$), and a common source of mistakes outside them. In general, zero correlation is much weaker than independence: take $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 = X_1^2$. Then $\text{Cov}(X_1, X_2) = \mathbb{E}(X_1^3) = 0$ by symmetry, yet X_2 is a deterministic function of X_1 . Correlation measures only *linear* association.

Conditionals. The third question, what observing $\mathbf{X}_b$ tells us about $\mathbf{X}_a$, is answered by the following theorem.

Theorem 2.2 (Conditional distribution)

If Σ_{bb} is invertible, then

$$\mathbf{X}_a \mid \mathbf{X}_b = \mathbf{x}_b \sim \mathcal{N}\left(\boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right). \quad (4)$$

Proof. The trick is to remove from $\mathbf{X}_a$ the part of it that $\mathbf{X}_b$ “explains”. Set $\mathbf{W} = \mathbf{X}_a - \Sigma_{ab}\Sigma_{bb}^{-1}\mathbf{X}_b$. The pair $(\mathbf{W}, \mathbf{X}_b)$ is an affine transformation of $\mathbf{X}$, hence Gaussian (property 3.1 of section 3), and

$$\text{Cov}(\mathbf{W}, \mathbf{X}_b) = \Sigma_{ab} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{bb} = 0,$$

so by theorem 2.1, $\mathbf{W}$ is independent of $\mathbf{X}_b$. Now write $\mathbf{X}_a = \mathbf{W} + \Sigma_{ab}\Sigma_{bb}^{-1}\mathbf{X}_b$. Conditionally on $\mathbf{X}_b = \mathbf{x}_b$, the second term is the constant $\Sigma_{ab}\Sigma_{bb}^{-1}\mathbf{x}_b$ while $\mathbf{W}$, being independent of $\mathbf{X}_b$, keeps its unconditional distribution $\mathcal{N}(\boldsymbol{\mu}_a - \Sigma_{ab}\Sigma_{bb}^{-1}\boldsymbol{\mu}_b, \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$. Adding the constant shifts the mean and gives the result. $\square$

Two features of the formula deserve note.

- The conditional mean is *affine* in $\mathbf{x}_b$. In the bivariate case with variances σ_1^2, σ_2^2 and correlation ρ , it reads

$$\mathbb{E}(X_1 | X_2 = x_2) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2),$$

a straight line of slope $\rho \sigma_1 / \sigma_2$, and this is exactly the *regression line* you met in last year's course on linear models. What that course posited as a modeling choice, the Gaussian world derives: *if the joint distribution is Gaussian, the best prediction of X_1 from X_2 (in mean squared error) is linear*, automatically.

- The conditional covariance $\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$ does not depend on the observed value $\mathbf{x}_b$, and it is smaller (in the matrix order) than Σ_{aa} : observing $\mathbf{X}_b$ can only reduce uncertainty about $\mathbf{X}_a$.

Where is this theorem used? Its most direct application is *missing data*. Suppose one candidate's physics score was lost while the mathematics score is known: with a Gaussian model of the pair of scores, the theorem hands the entire distribution of the missing value given the observed one. The conditional mean is the natural imputation, and the conditional variance quantifies exactly how much the imputation can be trusted. The same computation is the workhorse of Bayesian statistics with Gaussian priors (the posterior of a Gaussian parameter given Gaussian data *is* such a conditional) and of time-series filtering, where each new observation updates the state by precisely this formula.

3 Affine transformations and whitening

Gaussian vectors form a family stable under affine maps: this closure property is used throughout this course.

Property 3.1

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $A \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$, then

$$A\mathbf{X} + \mathbf{b} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\top). \quad (5)$$

Proof. We compute the characteristic function of $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ and recognize a Gaussian one. For $\mathbf{t} \in \mathbb{R}^n$,

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E} \left[e^{i\mathbf{t}^\top (A\mathbf{X} + \mathbf{b})} \right] = e^{i\mathbf{t}^\top \mathbf{b}} \mathbb{E} \left[e^{i(A^\top \mathbf{t})^\top \mathbf{X}} \right] = e^{i\mathbf{t}^\top \mathbf{b}} \phi_{\mathbf{X}}(A^\top \mathbf{t}).$$

Property 2.1 evaluated at $A^\top \mathbf{t}$ gives

$$\phi_{\mathbf{X}}(A^\top \mathbf{t}) = \exp \left(i\mathbf{t}^\top A\boldsymbol{\mu} - \frac{1}{2} \mathbf{t}^\top A\Sigma A^\top \mathbf{t} \right),$$

so that

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \exp \left(i\mathbf{t}^\top (A\boldsymbol{\mu} + \mathbf{b}) - \frac{1}{2} \mathbf{t}^\top (A\Sigma A^\top) \mathbf{t} \right),$$

which is the characteristic function of $\mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\top)$. Since the characteristic function determines the distribution (appendix B), the claim follows. $\square$

Note that $A\mathbf{X} + \mathbf{b}$ is a Gaussian *vector* even when $A\Sigma A^\top$ is singular; this is another reason not to build the theory on densities.

Whitening. In one dimension there is a standard way to put a Gaussian variable on a universal scale: if $X \sim \mathcal{N}(\mu, \sigma^2)$, its *z-score*

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

(center, then divide by the spread) has one and the same distribution whatever μ and σ ; this is the standardization your first-year tables of the Gaussian are built on. How should the z-score generalize to a Gaussian *vector*? Centering is unambiguous ($\mathbf{X} - \boldsymbol{\mu}$), but there is no single number σ to divide by: the spread of a vector is the whole matrix Σ . Writing the 1D division as multiplication by $(\sigma^2)^{-1/2}$ names the right generalization: multiply by $\Sigma^{-1/2}$. Take Σ invertible and let $\Sigma^{-1/2}$ be the inverse of the matrix square root $\Sigma^{1/2} = P\Lambda^{1/2}P^\top$ (appendix A). Choosing $A = \Sigma^{-1/2}$ and $\mathbf{b} = -\Sigma^{-1/2}\boldsymbol{\mu}$ in property 3.1 gives

$$\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbf{I}_d) :$$

the multivariate z-score. Any non-degenerate Gaussian vector can be turned into the standard one by centering, rotating, rescaling and rotating back, since $\Sigma^{-1/2} = P\Lambda^{-1/2}P^\top$; the last rotation P can be skipped, as it maps $\mathcal{N}(0, \mathbf{I}_d)$ to itself. Figure 4 shows the first steps on data: the rotation $P^\top(\mathbf{X} - \boldsymbol{\mu})$ aligns the cloud with the axes (decorrelation), the rescaling by $\Lambda^{-1/2}$ makes it isotropic (whitening). Conversely, $\mathbf{X} = \boldsymbol{\mu} + \Sigma^{1/2}\mathbf{Z}$ with $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$: every Gaussian vector is a linear deformation of i.i.d. standard coordinates. This is both the standard simulation recipe (the generation purpose of the prologue: to simulate synthetic rows, draw standard normals and color them) and the promised justification that the general density is the product case seen through a linear map.

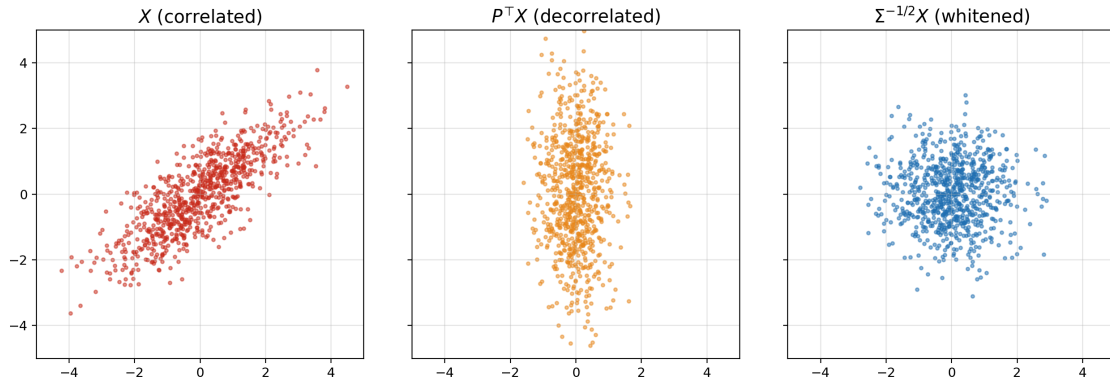


Figure 4: Whitening step by step: original, decorrelated ($P^\top(\mathbf{X} - \boldsymbol{\mu})$), and whitened ($\Lambda^{-1/2}P^\top(\mathbf{X} - \boldsymbol{\mu})$, which equals $\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ up to the rotation P) samples.

Whitening is also a common preprocessing step: several algorithms assume decorrelated inputs of comparable scales; independent component analysis requires it, and gradient-based optimizers converge faster on whitened features.

4 The Mahalanobis distance

We can now answer the anomaly question the prologue raised: is a given observation exceptional, and by how much?

Which distance should we use? Consider the entrance exam of a selective program: two papers, mathematics and physics, each marked out of 40, one row per candidate. The two scores move together: strong candidates score high on both papers, weaker ones low on both. Consider two candidates, both at Euclidean distance 10 from the mean (18, 18): candidate A near (25, 25),

along the correlation direction (strong across the board), and candidate B near (25, 11), against it: excellent in mathematics, poor in physics. The Euclidean distance rates the two candidates equally far from the average one; the data do not (figure 5): A lies inside the point cloud while B is far outside it. The Euclidean distance is the wrong measure here, and the Gaussian density suggests the right one. Model ordinary candidates as $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$: the density of section 1 is constant exactly on the ellipsoids

$$(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = r^2,$$

so two candidates are *equally plausible* under the model precisely when they share the value of this quadratic form. Candidate A lies on a small ellipsoid (high density), candidate B on a much larger one (density smaller by orders of magnitude): if “exceptional” is to mean “implausible for an ordinary candidate”, the quantity to read is not $\|\mathbf{x} - \boldsymbol{\mu}\|$ but the index r of the level set through $\mathbf{x}$. This is the definition:

Definition 4.1 (Mahalanobis distance)

For Σ positive definite, the *Mahalanobis distance* [1] from $\mathbf{x}$ to $\boldsymbol{\mu}$ is

$$d_\Sigma(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \quad (6)$$

It measures separation in “covariance units”: deviations along directions of large variance are cheap, deviations against the correlation structure are expensive, which is exactly why candidate A scores $d_\Sigma \approx 1.3$ and candidate B scores $d_\Sigma \approx 5.8$ on the same Euclidean displacement. For $\Sigma = \mathbf{I}_d$ it reduces to the Euclidean distance: the Mahalanobis distance is the Euclidean one, corrected for the shape of the cloud.

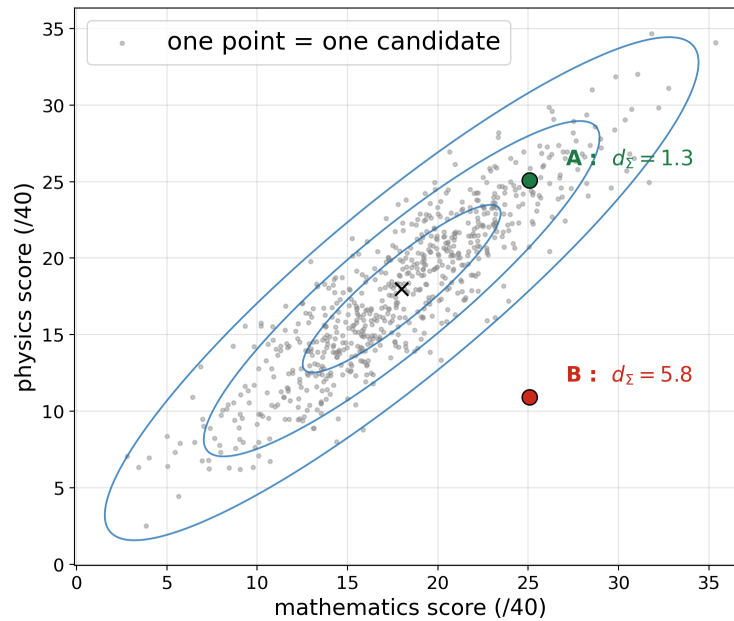


Figure 5: Two candidates at equal Euclidean distance from the mean candidate $\boldsymbol{\mu}$: the covariance structure makes A (strong in both papers) typical, $d_\Sigma \approx 1.3$, and B (excellent in mathematics, poor in physics) an extreme anomaly, $d_\Sigma \approx 5.8$.

Its distribution. What makes this distance a *statistical* tool rather than a mere score is that, under the model, its distribution is known exactly.

Reminder

If $Z_1, \dots, Z_p$ are i.i.d. $\mathcal{N}(0, 1)$, the distribution of $\sum_{i=1}^p Z_i^2$ is called chi-squared with p degrees of freedom, written $\chi^2(p)$.

Proposition 4.1

Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with Σ invertible. Then

$$d_\Sigma(\mathbf{X}, \boldsymbol{\mu})^2 = (\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(d). \quad (7)$$

Proof. Whiten as in section 3: the vector $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ is Gaussian with mean 0 and covariance $\Sigma^{-1/2} \Sigma \Sigma^{-1/2} = \mathbf{I}_d$, i.e. $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$: its d coordinates are i.i.d. $\mathcal{N}(0, 1)$. Moreover, writing $\Sigma^{-1} = \Sigma^{-1/2} \Sigma^{-1/2}$,

$$(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = (\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}))^\top (\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})) = \|\mathbf{Z}\|^2 = \sum_{i=1}^d Z_i^2,$$

a sum of d squared i.i.d. standard Gaussians: a $\chi^2(d)$ by definition. $\square$

Proposition 4.1 means Mahalanobis distances come *pre-calibrated*: quantiles of $\chi^2(d)$ convert a score into a probability. In dimension $d = 2$ the tail is even explicit, $\mathbb{P}(d_\Sigma^2 > r^2) = e^{-r^2/2}$, so candidate A's score is entirely unremarkable ($\mathbb{P}(d_\Sigma > 1.3) \approx 0.43$) while candidate B's is essentially impossible among ordinary candidates ($\mathbb{P}(d_\Sigma > 5.8) \approx 5 \cdot 10^{-8}$); such a copy calls for a second look: a marking error, an irregularity, or a genuinely exceptional profile. Thresholding a Mahalanobis distance at a chi-squared quantile is the basic recipe of multivariate anomaly detection, with the false-alarm rate chosen in advance.

A new scalar product, a new geometry. The Mahalanobis distance is not an isolated trick: it is the distance of a full Euclidean-style geometry that Σ installs on $\mathbb{R}^d$. Define

$$\langle \mathbf{u}, \mathbf{v} \rangle_\Sigma = \mathbf{u}^\top \Sigma^{-1} \mathbf{v}.$$

This is a genuine scalar product (bilinear and symmetric because Σ^{-1} is symmetric, positive definite because Σ^{-1} is), and its norm $\|\mathbf{u}\|_\Sigma = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle_\Sigma} = d_\Sigma(\mathbf{u}, 0)$ is exactly the Mahalanobis distance to the origin. All Euclidean vocabulary transfers (lengths, angles, orthogonality), but computed *anisotropically*. Figure 6 shows the two geometries side by side. In the Σ -geometry the unit ball is the covariance ellipse, and the natural coordinate grid is the image of the square grid under $\Sigma^{1/2}$, an oblique lattice whose lines are not aligned with the eigenvectors of Σ : one “unit step” is long in the direction where the data spread widely and short across it. The two red and green arrows are unit vectors of the new geometry: of unequal Euclidean length, but of unit Σ -length. The whitening map $\Sigma^{-1/2}$ of section 3 is precisely the change of coordinates that turns the right panel back into the left one: an *isometry* between the Σ -geometry and the standard Euclidean one, which is the geometric content of proposition 4.1. Several later constructions are ordinary Euclidean procedures carried out in this geometry.

5 Why the Gaussian is everywhere: the multivariate CLT

Everything so far assumed Gaussian data. The justification for this emphasis is that averages of variables from any distribution with finite variance become Gaussian once centered and rescaled by $\sqrt{n}$.

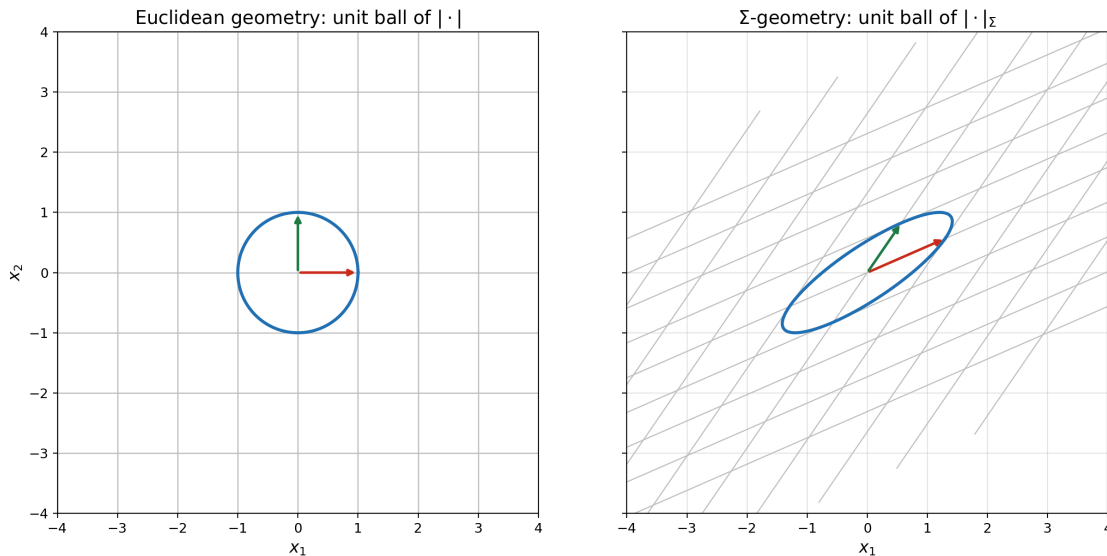


Figure 6: The scalar product $\langle \mathbf{u}, \mathbf{v} \rangle_\Sigma = \mathbf{u}^\top \Sigma^{-1} \mathbf{v}$ replaces the Euclidean geometry (left: square grid, unit circle) by an anisotropic one (right: the square grid mapped by $\Sigma^{1/2}$, an oblique lattice; the unit ball is the covariance ellipse). The arrows are the images of the standard basis: unit vectors of the new geometry.

Theorem 5.1 (Multivariate central limit theorem)

Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. random vectors from **any distribution** with mean $\boldsymbol{\mu}$ and finite covariance Σ , and let $\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$. Then

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Sigma). \quad (8)$$

The proof is a textbook application of the Cramér–Wold device (appendix B): for any fixed direction $\mathbf{a} \in \mathbb{R}^d$, the scalars $\mathbf{a}^\top \mathbf{X}_i$ are i.i.d. with mean $\mathbf{a}^\top \boldsymbol{\mu}$ and variance $\mathbf{a}^\top \Sigma \mathbf{a}$, so the one-dimensional CLT gives $\sqrt{n} \mathbf{a}^\top (\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, \mathbf{a}^\top \Sigma \mathbf{a})$; since this holds for every $\mathbf{a}$, the device gives the vector limit.

Note what the theorem does *not* assume: no density, no symmetry, no Gaussianity of the data; it needs only i.i.d. sampling and finite second moments. And note what it concludes: the limiting covariance is the covariance Σ of the data themselves, so the covariances, hence the correlations, of the coordinates survive in the limit; any nonlinear dependence between them disappears. Figure 7 illustrates both points on a deliberately skewed 2D distribution with correlated coordinates (correlation about 0.63): at $n = 1$ the cloud is asymmetric and fan-shaped; by $n = 200$ the rescaled means form a clean elliptical Gaussian cloud whose tilt reflects the correlation of the original data. This theorem justifies Gaussian modeling for aggregates: when the quantity of interest is an average, a sum, or a net effect of many contributions, the Gaussian family is asymptotically correct.

6 Estimating a Gaussian model

6.1 Maximum likelihood estimator

It remains to fit the model. Everything in this section is the univariate estimation story you rehearsed in first year, with vectors and matrices substituted for scalars, plus one new phenomenon at the end that has no univariate counterpart.

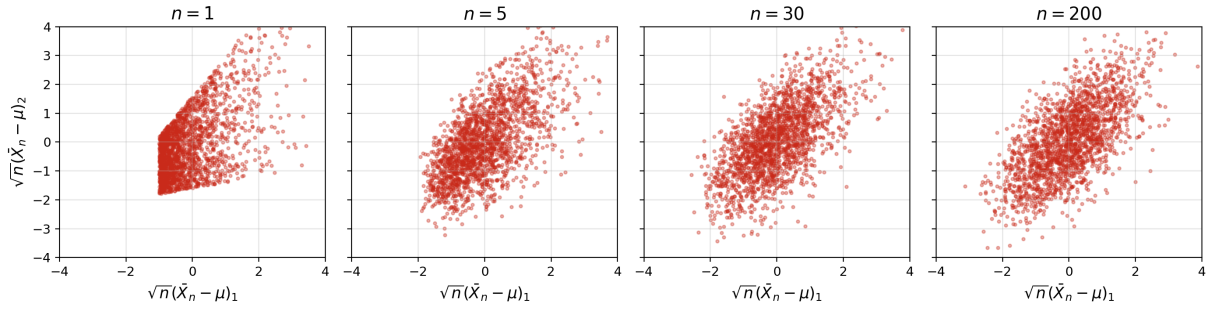


Figure 7: Rescaled means $\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu})$ of a skewed 2D distribution with correlated coordinates: an elliptical Gaussian cloud, tilted by that correlation, emerges as n grows.

Reminder

(*Likelihood and MLE in dimension 1.*) The *likelihood* of an i.i.d. sample is its joint density read as a function of the parameters, $L(\theta) = \prod_{i=1}^n f_\theta(x_i)$: it grades each candidate θ by how unsurprising it makes the observed data, and the *maximum likelihood estimator* (MLE) picks the best-graded one. For $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$: $\hat{\mu} = \bar{x}$ and $\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$; the latter is biased, $\mathbb{E}(\hat{\sigma}_{\text{MLE}}^2) = \frac{n-1}{n} \sigma^2$, whence the unbiased $\hat{\sigma}_u^2$ dividing by $n-1$. The estimators fluctuate with exactly known laws: $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, $(n-1) \hat{\sigma}_u^2 / \sigma^2 \sim \chi^2(n-1)$, and $\bar{X} \perp\!\!\!\perp \hat{\sigma}_u^2$ (Cochran's theorem).

The likelihood. Let $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with Σ invertible. By independence, the likelihood of the observed sample is

$$L(\boldsymbol{\mu}, \Sigma) = \prod_{i=1}^n \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right). \quad (9)$$

Theorem 6.1 (Gaussian MLE)

If the matrix $\hat{\Sigma}_{\text{MLE}}$ below is invertible (which requires $n \geq d+1$), the likelihood is maximized at

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i, \quad \hat{\Sigma}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \quad (10)$$

If $\hat{\Sigma}_{\text{MLE}}$ is singular, the likelihood is unbounded and has no maximizer.

The estimators have exactly the shape the reminder announces: empirical mean, empirical average of (outer-product) squared deviations. The proof is a beautiful piece of matrix calculus: the mean part works as in 1D, while the covariance part hinges on rewriting the exponent with the trace trick $\sum_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \text{tr}(\Sigma^{-1} S_\mu)$, where $S_\mu = \sum_i (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$, and optimizing over the eigenvalues of $\Sigma^{-1/2}$ -conjugated matrices.

Fluctuations. As in 1D, the estimators are random and their laws are exactly known. The law of a random *matrix* such as $\hat{\Sigma}_{\text{MLE}}$ requires one new named distribution, which plays for covariance matrices the role the chi-squared plays for variances:

Definition 6.1 (Wishart distribution)

Let $\mathbf{X}_1, \dots, \mathbf{X}_n \sim \mathcal{N}(0, \Sigma)$ be i.i.d. The distribution of the random positive semi-definite matrix $\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top$ is called the *Wishart distribution* $\mathcal{W}_d(n, \Sigma)$.

For $d = 1$ and $\Sigma = 1$ the definition reads $\sum_i X_i^2 \sim \chi^2(n)$: the Wishart is the multivariate chi-squared. The connection persists in any dimension through one-dimensional projections: for any direction $\mathbf{a}$ with $\mathbf{a}^\top \Sigma \mathbf{a} \neq 0$, and $V \sim \mathcal{W}_d(n, \Sigma)$,

$$\frac{\mathbf{a}^\top V \mathbf{a}}{\mathbf{a}^\top \Sigma \mathbf{a}} \sim \chi^2(n),$$

since $\mathbf{a}^\top V \mathbf{a} = \sum_i (\mathbf{a}^\top \mathbf{X}_i)^2$ is a sum of n squared i.i.d. Gaussians of variance $\mathbf{a}^\top \Sigma \mathbf{a}$.

Property 6.1

Under the model above, with $\hat{\Sigma}_u = \frac{1}{n-1} \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$ the unbiased covariance estimator:

$$\bar{\mathbf{X}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\Sigma}{n}\right), \quad (n-1) \hat{\Sigma}_u \sim \mathcal{W}_d(n-1, \Sigma), \quad \bar{\mathbf{X}} \perp \hat{\Sigma}_u. \quad (11)$$

Moreover $\mathbb{E}(\hat{\Sigma}_{\text{MLE}}) = \frac{n-1}{n} \Sigma$, i.e. $\hat{\Sigma}_u = \frac{n}{n-1} \hat{\Sigma}_{\text{MLE}}$ is the unbiased version: the MLE covariance is biased, for the same degree-of-freedom reason as in 1D.

The first claim is immediate from the affine stability of section 3 ($\bar{\mathbf{X}}$ is a linear map of the Gaussian vector $(\mathbf{X}_1, \dots, \mathbf{X}_n)$); the Wishart law and the independence are the multivariate Cochran theorem, which we state without proof. The table below collects the univariate–multivariate correspondence built over the last five sections:

	dimension 1 (first year)	dimension d (this chapter)
model	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$
parameters	2	$d + d(d+1)/2$
MLE of the mean	$\bar{x}$	$\bar{\mathbf{x}} \in \mathbb{R}^d$
MLE of the (co)variance	$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$	$\hat{\Sigma}_{\text{MLE}} = \frac{1}{n} \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$
unbiased estimator	$\hat{\sigma}_u^2 (\div n-1)$	$\hat{\Sigma}_u (\div n-1)$
law of the mean	$\mathcal{N}(\mu, \sigma^2/n)$	$\mathcal{N}(\boldsymbol{\mu}, \Sigma/n)$
law of the (co)variance	$(n-1)\hat{\sigma}_u^2/\sigma^2 \sim \chi^2(n-1)$	$(n-1)\hat{\Sigma}_u \sim \mathcal{W}_d(n-1, \Sigma)$
standardized deviation	$(x - \mu)^2/\sigma^2$	$d_\Sigma(\mathbf{x}, \boldsymbol{\mu})^2 \sim \chi^2(d)$

6.2 The curse of dimension

Consider the parameters row of the table. In one dimension the Gaussian model asks for two numbers. In dimension d it asks for d means plus one covariance entry per pair of coordinates, the d variances included:

$$k = d + \frac{d(d+1)}{2}, \quad \begin{array}{c|cccc} d & 2 & 10 & 100 & 1000 \\ \hline k & 5 & 65 & 5150 & 501500 \end{array} \quad (12)$$

The mean vector grows linearly; the covariance grows *quadratically*, and it dominates the parameter count from the start. This is the *curse of dimension* in its statistical guise. Your first-year course on dimension reduction showed you its geometric one (volumes and distances misbehaving as d grows); here it is estimation accuracy that suffers. Estimating Σ well is the hard part of fitting a Gaussian model, and the difficulty comes in two regimes.

If $n \leq d$, the estimate is singular, always. This is not bad luck on a particular dataset but a theorem:

Proposition 6.1

$\text{rank}(\hat{\Sigma}_{\text{MLE}}) \leq \min(d, n-1)$. In particular, whenever $n \leq d$, the sample covariance $\hat{\Sigma}_{\text{MLE}}$ is singular.

Proof. Write $\hat{\Sigma}_{\text{MLE}} = \frac{1}{n} \mathbf{C}^\top \mathbf{C}$, where $\mathbf{C} \in \mathbb{R}^{n \times d}$ stacks the centered rows $(\mathbf{x}_i - \bar{\mathbf{x}})^\top$. A matrix and its Gram matrix have the same rank, so $\text{rank}(\hat{\Sigma}_{\text{MLE}}) = \text{rank}(\mathbf{C})$, and $\text{rank}(\mathbf{C}) \leq \min(n, d)$ trivially. One more unit is lost to centering: the rows of $\mathbf{C}$ sum to $\sum_i (\mathbf{x}_i - \bar{\mathbf{x}}) = 0$, so any one row is determined by the $n - 1$ others, and $\text{rank}(\mathbf{C}) \leq n - 1$. Hence $\text{rank}(\hat{\Sigma}_{\text{MLE}}) \leq \min(d, n - 1) < d$ whenever $n \leq d$. $\square$

A singular $\hat{\Sigma}_{\text{MLE}}$ means the fitted Gaussian is *degenerate* in the sense of section 1: it concentrates on the affine subspace through $\bar{\mathbf{x}}$ spanned by the centered rows $\mathbf{x}_i - \bar{\mathbf{x}}$ (of dimension at most $n - 1$), has no density, and $\hat{\Sigma}_{\text{MLE}}^{-1}$ does not exist, so the Mahalanobis distance, the anomaly detector and the conditional formulas of theorem 2.2 (which need Σ_{bb}^{-1}) are simply unavailable. Nor is $n \leq d$ an exotic regime: a genomics dataset routinely records $d = 20\,000$ genes for a few hundred samples.

If $n > d$, the estimate is invertible, but its accuracy degrades as d grows. Figure 8 shows a simulation with the simplest possible truth: data from $\mathcal{N}(0, \mathbf{I}_d)$, whose true eigenvalues all equal 1, tracking for each dimension the *largest* and *smallest* eigenvalue of $\hat{\Sigma}_{\text{MLE}}$ as the sample grows. At $d = 2$, both are close to the truth already for modest n . At $d = 50$, a sample of $n = 100$ (twice d) still returns eigenvalues from about 0.1 to 2.8: the estimator overstates the anisotropy of the data, inflating the top of the spectrum and deflating the bottom, although all true eigenvalues are equal. At $d = 200$, thousands of observations are needed before the spectrum tightens. The reason is a counting argument: the sample provides $n \times d$ numbers and the model asks for $d(d + 1)/2$ parameters, so accuracy is governed not by n alone but by the ratio n/d . The practical consequence is that the model’s most useful formulas all use $\hat{\Sigma}_{\text{MLE}}^{-1}$, whose largest eigenvalue $1/\lambda_{\min}(\hat{\Sigma}_{\text{MLE}})$ comes from the eigenvalue the estimator underestimates most: inverting amplifies exactly the worst error.

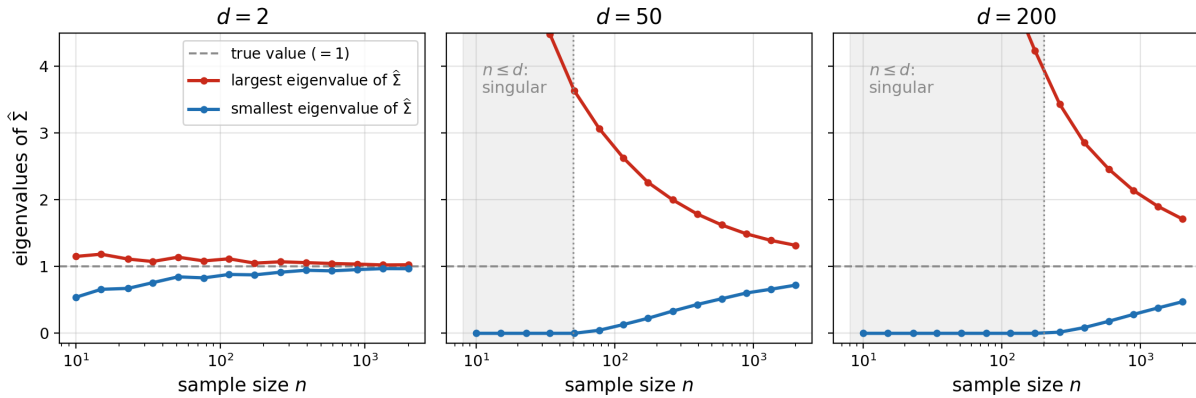


Figure 8: Largest and smallest eigenvalues of $\hat{\Sigma}_{\text{MLE}}$ (averaged over 20 replications) as the sample size grows, on data from $\mathcal{N}(0, \mathbf{I}_d)$ where every true eigenvalue is 1. Left to right: $d = 2, 50, 200$. In the shaded region $n \leq d$ the smallest eigenvalue is exactly 0 (proposition 6.1); beyond it, convergence toward the truth is slower the larger d .

6.3 Regularization

The difficulty just diagnosed is, at bottom, an *inference* problem: we want to infer $d + d(d + 1)/2$ parameters and we only have n observations to do it. The obvious solution is to get more data, and it is the right one when available: in figure 8, every curve does converge to the truth as n grows. But in practice more data is often not an option: the sample is whatever the measurement campaign, the archive, or the budget produced. Two approaches remain, and both are called regularization. Either we *reduce the parameter count*, restricting the family until n suffices for it

(regularizing the *problem*), or we keep the family and *accept a bias in the inference* in exchange for stability (regularizing the *estimator*).

Regularize the problem: fewer parameters, more assumptions. Nothing forces the family to contain every covariance matrix. Assume Σ *diagonal* and the count collapses from $d + d(d + 1)/2$ to $2d$; assume it *spherical*, $\Sigma = \sigma^2 \mathbf{I}_d$, and it collapses to $d + 1$. These are modeling assumptions in the full sense of the chapter’s opening reminder: the diagonal family asserts that the coordinates are uncorrelated. For the exam table of section 4, it asserts that knowing a candidate’s mathematics score says nothing about their physics score, which is exactly the structure the Mahalanobis section exploited. The gain is estimability at small n ; the cost is that the model can no longer represent dependence at all. Whether this is acceptable depends on the question asked.

Regularize the estimator: same model, biased but stable. The alternative keeps the full family (the parameter count stands) but replaces $\hat{\Sigma}_{\text{MLE}}$ by a blend with a fixed, well-behaved target, in its simplest form

$$\hat{\Sigma}_\alpha = (1 - \alpha) \hat{\Sigma}_{\text{MLE}} + \alpha \tau \mathbf{I}_d, \quad \alpha \in (0, 1], \quad (13)$$

with $\tau > 0$ a scale (say the average sample variance). The name *regularization* is meant literally. The defect of $\hat{\Sigma}_{\text{MLE}}$ in high dimension is its *irregularity*: it is singular or nearly so, hence not invertible, or invertible with an inverse that blows noise up without bound. The blend repairs this: since eigenvalues of a sum of symmetric matrices are bounded below by the sums of the smallest ones,

$$\lambda_{\min}(\hat{\Sigma}_\alpha) \geq (1 - \alpha) \underbrace{\lambda_{\min}(\hat{\Sigma}_{\text{MLE}})}_{\geq 0} + \alpha \tau \geq \alpha \tau > 0 :$$

$\hat{\Sigma}_\alpha$ is invertible *by construction*, for any n and d , and its inverse is stable. Pulling the estimate toward a fixed target thus turns an ill-posed estimation problem into a well-posed one. The price is a bias: $\hat{\Sigma}_\alpha$ is pulled toward $\tau \mathbf{I}_d$ even if the truth is far from it, and the weight α arbitrates between the variance of $\hat{\Sigma}_{\text{MLE}}$ and the bias of the target. Choosing α from the data can itself be automated (the Ledoit–Wolf estimator is the standard reference). How such trade-offs are tuned in general recurs throughout the course.

7 Conclusion

Figure 9 shows the same modeling act carried out in dimensions one, two, and three: a cloud of points and a fitted density. In every dimension the fit is pinned down by two objects only, a mean and a covariance. The Gaussian is the simplest multivariate distribution characterized by these two objects, and what it represents is correspondingly simple: a *single cluster* of data, symmetric around its center, whose shape is well approximated by an ellipsoid, the covariance defining the geometry of that ellipsoid, its orientation and its elongations. Everything the chapter did unfolds from this picture: the density’s level sets *are* the ellipsoids; the Mahalanobis distance measures how many ellipsoids away a point sits, calibrated by the chi-squared law; whitening is the change of variables that makes the ellipsoid a sphere; estimation fills in the two objects from data, at a price, quadratic in d , that the covariance dictates.

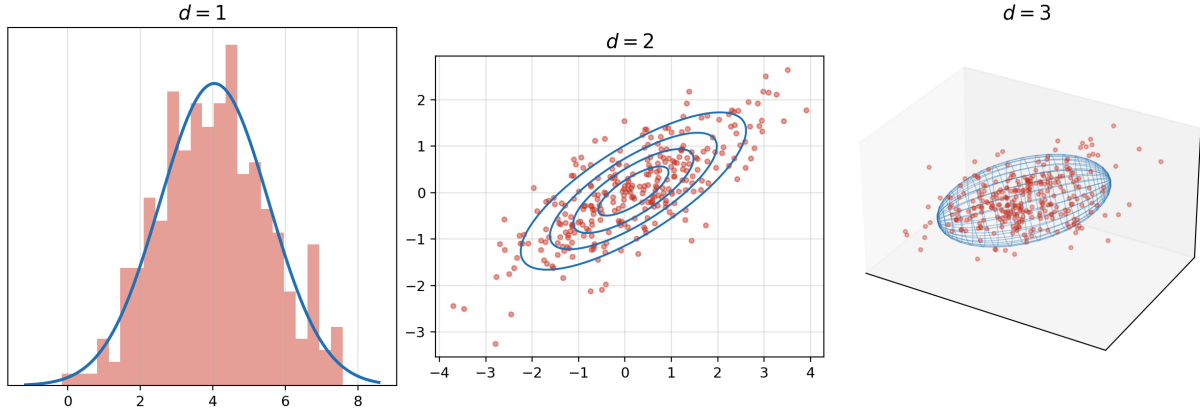


Figure 9: One Gaussian cluster in $d = 1, 2, 3$: data (red) and the fitted Gaussian (blue). In every dimension the fitted density is characterized by its mean and covariance alone: mass symmetric around the mean, level sets ellipsoids whose orientation and elongations the covariance defines.

One center, one ellipsoid: that is also the model's limit. Many datasets gather around *several* centers at once (customers of two kinds, cells of two subtypes), and no single ellipsoid can be symmetric around all of them. Describing such data with several Gaussians simultaneously, and fitting them when no column says which point belongs to which, is the subject of the next chapter.

A Linear algebra toolbox

Spectral theorem. Every symmetric matrix $\mathbf{A} \in \mathbb{S}_d$ admits d real eigenvalues $\lambda_1, \dots, \lambda_d$ with orthonormal eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_d$: $\mathbf{A}\mathbf{e}_i = \lambda_i\mathbf{e}_i$. In matrix form $\mathbf{A} = P\Lambda P^\top$ with $P = (\mathbf{e}_1 | \dots | \mathbf{e}_d)$ orthogonal ($PP^\top = \mathbf{I}_d$) and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$. Equivalently, $\mathbf{A} = \sum_i \lambda_i \mathbf{e}_i \mathbf{e}_i^\top$: a weighted sum of rank-one projectors.

Positive matrices. $\mathbf{A}$ is positive semi-definite ($\mathbf{A} \succeq 0$, all $\lambda_i \geq 0$) iff its quadratic form $Q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ is nonnegative for all $\mathbf{x}$; positive definite when strict. Covariance matrices are positive semi-definite since $\mathbf{a}^\top \Sigma \mathbf{a} = \mathbb{V}(\mathbf{a}^\top \mathbf{X}) \geq 0$. Moreover $\text{tr}(\mathbf{A}) = \sum_i \lambda_i$, $\det(\mathbf{A}) = \prod_i \lambda_i$, and $\text{rank}(\mathbf{A})$ is the number of nonzero eigenvalues.

Matrix square root. For $\mathbf{A} \succeq 0$, $\mathbf{A}^{1/2} = P\Lambda^{1/2}P^\top$ with $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_d})$; it is the unique positive semi-definite matrix whose square is $\mathbf{A}$. When $\mathbf{A} \succ 0$, $\mathbf{A}^{-1/2} = (\mathbf{A}^{1/2})^{-1} = P\Lambda^{-1/2}P^\top$.

Quadratic forms and ellipsoids. For $\mathbf{A} \succ 0$, the region $\{\mathbf{x} : (\mathbf{x} - \mathbf{x}_0)^\top \mathbf{A} (\mathbf{x} - \mathbf{x}_0) = r^2\}$ is an ellipsoid centered at $\mathbf{x}_0$ with principal axes the eigenvectors of $\mathbf{A}$ and semi-axis lengths $r/\sqrt{\lambda_i(\mathbf{A})}$. Applied to $\mathbf{A} = \Sigma^{-1}$, whose eigenvalues are $1/\lambda_i(\Sigma)$, this gives semi-axes $r\sqrt{\lambda_i(\Sigma)}$, the geometry of section 1.

B Random vectors toolbox

Expectation and covariance. For a random vector $\mathbf{X} = (X_1, \dots, X_d)^\top$, the expectation is taken componentwise, $\mathbb{E}(\mathbf{X}) = (\mathbb{E}X_1, \dots, \mathbb{E}X_d)^\top$, and for any matrix A and vector $\mathbf{b}$, $\mathbb{E}(A\mathbf{X} + \mathbf{b}) = A\mathbb{E}(\mathbf{X}) + \mathbf{b}$. The covariance matrix

$$\mathbb{V}(\mathbf{X}) = \Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] = \mathbb{E}(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

collects variances on the diagonal and covariances $\sigma_{X_i X_j}$ off it. Key identities: $\mathbb{V}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \Sigma \mathbf{a}$, $\mathbb{V}(A\mathbf{X} + \mathbf{b}) = A\Sigma A^\top$, and $\Sigma \succeq 0$ always.

Correlation. $\rho_{X_i X_j} = \sigma_{X_i X_j} / \sqrt{\mathbb{V}(X_i)\mathbb{V}(X_j)} \in [-1, 1]$. Independence implies zero correlation; the converse is false in general (it holds for Gaussian vectors, section 2).

Characteristic function. $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}^\top \mathbf{X}}]$, $\mathbf{t} \in \mathbb{R}^d$, determines the distribution of $\mathbf{X}$ entirely. Componentwise: $\phi_{X_1}(t_1) = \phi_{\mathbf{X}}(t_1, 0, \dots, 0)$. The components $X_1, \dots, X_d$ are independent iff $\phi_{\mathbf{X}}(\mathbf{t}) = \prod_j \phi_{X_j}(t_j)$ for all $\mathbf{t}$.

Cramér–Wold theorem. The distribution of $\mathbf{X} \in \mathbb{R}^d$ is entirely determined by the collection of univariate distributions of $\mathbf{t}^\top \mathbf{X}$ for all $\mathbf{t} \in \mathbb{R}^d$. It is the bridge between univariate and multivariate limit theorems, and the reason definition 1.1 characterizes a distribution at all. Its version for convergence, the *Cramér–Wold device*, reads

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \mathbf{a}^\top \mathbf{X}_n \xrightarrow{d} \mathbf{a}^\top \mathbf{X} \text{ for every } \mathbf{a} \in \mathbb{R}^d :$$

convergence in distribution of random vectors reduces to that of all their one-dimensional projections.

References

- [1] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.