Lectures 1 and 2: Gaussians

Definition, linear structure, and conditioning as projection

---

1 Gaussians, and their linear structure

1.1 What is a Gaussian?

Here are three answers, each of which is taught somewhere.

(a) X has density proportional to $\exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))$.

(b) $X = AZ + \mu$, where Z has independent standard normal coordinates.

(c) Every linear functional $\langle v, X\rangle$ is a one-dimensional Gaussian.

They are not equivalent. Answer (a) requires Σ to be invertible, so it refuses to call $X = (g, g)$ Gaussian — a vector obtained from a standard normal by a linear map, all of whose linear functionals are Gaussian. A definition that has to apologize for its most obvious example is the wrong definition.

> **Q.** What should we ask of a definition of the Gaussian?
> That the class be closed under the operations we care about — and the operation we care about is applying a linear map.

We take a fourth answer, which implies the others and costs nothing at the degenerate boundary.

1.2 What we import

Definition 1.1 (Characteristic function)**.** For a random vector X in $\mathbb{R}^n$,

$$\varphi_X(t) \;=\; \mathbb{E}\big[e^{i\langle t, X\rangle}\big], \qquad t \in \mathbb{R}^n.$$

It is defined for every X with no integrability assumption, because $|e^{i\theta}| = 1$. Immediately: $\varphi_X(0) = 1$, $|\varphi_X(t)| \le 1$, and $\varphi_X(-t) = \overline{\varphi_X(t)}$.

Imported: Fourier uniqueness and inversion. If $\varphi_X = \varphi_Y$ then $X \overset{d}{=} Y$. If moreover $\varphi_X \in L^1(\mathbb{R}^n)$, then X has a bounded continuous density

$$p(x) \;=\; \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\langle t, x\rangle} \varphi_X(t)\, dt.$$

This is standard Fourier analysis and the only thing we do not prove.

1.3 The definition

Definition 1.2 (Gaussian vector)**.** Let $\mu \in \mathbb{R}^n$ and let Σ be symmetric $n \times n$. We say $X \sim \mathcal{N}(\mu, \Sigma)$ if

$$\boxed{\varphi_X(t) \;=\; \exp\Big(i\langle t, \mu\rangle - \tfrac{1}{2}\, t^\top \Sigma t\Big)}$$

for all $t \in \mathbb{R}^n$.

No inverse of Σ appears, so nothing prevents Σ from being singular. The definition describes a law rather than constructing one, so we owe an existence proof; we give it in §1.5, where it will take a single line.

1.4 Linear structure

Theorem 1.3 (Linear images)**.** *If $X \sim \mathcal{N}(\mu, \Sigma)$ and $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, then*

$$AX + b \;\sim\; \mathcal{N}\big(A\mu + b,\; A\Sigma A^\top\big).$$

Proof. $\varphi_{AX+b}(t) \;=\; e^{i\langle t,b\rangle}\mathbb{E}[e^{i\langle A^\top t, X\rangle}] \;=\; e^{i\langle t,b\rangle}\varphi_X(A^\top t)$. Substituting Definition 1.2 and using $\langle A^\top t, \mu\rangle = \langle t, A\mu\rangle$ gives $\exp\big(i\langle t, A\mu + b\rangle - \frac{1}{2} t^\top A\Sigma A^\top t\big)$. ∎

Remark 1.4. It is $A\Sigma A^\top$, not $A^\top \Sigma A$.

Corollary 1.5 (Answer (c), recovered)**.** *$X \sim \mathcal{N}(\mu, \Sigma)$ if and only if $\langle v, X\rangle \sim \mathcal{N}(\langle v, \mu\rangle,\, v^\top \Sigma v)$ for every $v \in \mathbb{R}^n$.*

Proof. Forward: take $A = v^\top$ in Theorem 1.3. Conversely, evaluating the one-dimensional transform of $\langle v, X\rangle$ at 1 gives $\varphi_X(v) = \exp(i\langle v, \mu\rangle - \frac{1}{2} v^\top \Sigma v)$, which is Definition 1.2. ∎

1.5 The standard normal, and existence

One computation is still missing: we have never exhibited a single Gaussian.

Lemma 1.6. *If g is standard normal on $\mathbb{R}$ then $\mathbb{E}[e^{i\lambda g}] = e^{-\lambda^2/2}$; equivalently $g \sim \mathcal{N}(0, 1)$.*

Proof. Set $f(\lambda) = \mathbb{E}[e^{i\lambda g}]$. Differentiating under the integral,

$$f'(\lambda) \;=\; \frac{1}{\sqrt{2\pi}} \int i x\, e^{i\lambda x} e^{-x^2/2}\, dx.$$

The key observation is that $x\, e^{-x^2/2} = -\frac{d}{dx} e^{-x^2/2}$, so integrating by parts moves the derivative onto the exponential:

$$f'(\lambda) \;=\; \frac{i}{\sqrt{2\pi}} \int \Big(\frac{d}{dx} e^{i\lambda x}\Big) e^{-x^2/2}\, dx \;=\; i \cdot i\lambda\, f(\lambda) \;=\; -\lambda f(\lambda).$$

With $f(0) = 1$, $f(\lambda) = e^{-\lambda^2/2}$. ∎

Remark 1.7. The proof used $\mathbb{E}[g\, h(g)] = \mathbb{E}[h'(g)]$ with $h(x) = e^{i\lambda x}$: multiplication by g acts like differentiation. This identity returns in Lecture 4 under its own name.

If $Z = (g_1, \dots, g_n)$ has independent standard normal coordinates then $\varphi_Z(t) = \prod_j e^{-t_j^2/2} = e^{-|t|^2/2}$, so $Z \sim \mathcal{N}(0, I_n)$. Existence is now immediate.

Proposition 1.8 (Well-posedness)**.** *A random vector satisfying Definition 1.2 exists if and only if $\Sigma \succeq 0$, and is then unique in law. If $\Sigma = AA^\top$, one may take $X = AZ + \mu$.*

Proof. If $\Sigma \succeq 0$, factor $\Sigma = AA^\top$. Since $Z \sim \mathcal{N}(0, I)$, Theorem 1.3 gives $AZ + \mu \sim \mathcal{N}(\mu, AA^\top) = \mathcal{N}(\mu, \Sigma)$. Uniqueness in law is Fourier uniqueness.

Conversely, suppose $v^\top \Sigma v = -c$ with $c > 0$. Along $t = sv$ the claimed transform has modulus $e^{s^2 c/2} \to \infty$, contradicting $|\varphi_X| \le 1$. ∎

The family of Gaussian laws on $\mathbb{R}^n$ is therefore indexed exactly by $\mathbb{R}^n \times \{\Sigma \succeq 0\}$. Answer (b) has become a construction inside a proof rather than a rival definition.

Corollary 1.9 (Rotational invariance)**.** *If $Z \sim \mathcal{N}(0, I_n)$ and O is orthogonal then $OZ \overset{d}{=} Z$.*

Proof. $\varphi_{OZ}(t) = \varphi_Z(O^\top t) = e^{-|O^\top t|^2/2} = e^{-|t|^2/2}$. ∎

Invariance is what singles the Gaussian out among product measures. In November, invariance under orthogonal conjugation will single out a particular matrix ensemble in the same way.

Example 1.10 (The degenerate case)**.** Take $n = 2$, $\mu = 0$, $\Sigma = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, which is positive semidefinite of rank one. By Proposition 1.8 the law $\mathcal{N}(0, \Sigma)$ exists, and $A = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$ realizes it as $X = (g, g)$. It has no density — it is supported on a line — and Definition 1.2 never notices.

1.6 Uncorrelated versus independent

For general random variables zero covariance says very little. For jointly Gaussian ones it says everything.

Theorem 1.11. *Let (X, Y) be jointly Gaussian in $\mathbb{R}^n \times \mathbb{R}^m$ — the concatenated vector is Gaussian — with $\operatorname{Cov}(X, Y) = 0$. Then $X \perp\!\!\!\perp Y$.*

Proof. Write $t = (s, u)$. The covariance of the concatenated vector is block diagonal, so $t^\top \Sigma t = s^\top \Sigma_{XX} s + u^\top \Sigma_{YY} u$ and $\langle t, \mu \rangle = \langle s, \mu_X \rangle + \langle u, \mu_Y \rangle$. Hence

$$\varphi_{(X,Y)}(s, u) \;=\; \varphi_X(s)\, \varphi_Y(u),$$

which is the characteristic function of the product law $\mathcal{L}(X) \otimes \mathcal{L}(Y)$. By Fourier uniqueness the joint law is that product. ∎

The hypothesis that fails in practice is joint Gaussianity, not Gaussianity of the marginals.

Example 1.12 (Uncorrelated, standard normal, not independent)**.** Let $g \sim \mathcal{N}(0, 1)$ and let ε be an independent random sign; set $X = g$, $Y = \varepsilon g$. Conditionally on $\varepsilon = \pm 1$ we have $Y = \pm g \overset{d}{=} g$, so both marginals are standard normal. They are uncorrelated, since $\mathbb{E}[XY] = \mathbb{E}[\varepsilon]\mathbb{E}[g^2] = 0$. And $|X| = |Y|$ always, so X determines Y up to a sign.

Conditioning on ε and applying Lemma 1.6,

$$\varphi_{(X,Y)}(s, u) \;=\; \tfrac{1}{2} e^{-(s+u)^2/2} + \tfrac{1}{2} e^{-(s-u)^2/2},$$

a sum of two Gaussian transforms and not exp(quadratic). The pair fails Definition 1.2 even though each coordinate satisfies it.

1.7 What we have not proved

Maximum entropy. Among all laws on $\mathbb{R}^n$ with mean μ and covariance $\Sigma \succ 0$, the Gaussian uniquely maximizes differential entropy: it is the least committal law consistent with a given second moment. Homework 1 proves it, and Lecture 3 uses it.

The moment generating function. The computation of Lemma 1.6 with a real exponent gives $\mathbb{E}[e^{\lambda g}] = e^{\lambda^2/2}$, from which Gaussian tail bounds follow by Chernoff's method. In Module II this becomes the engine of large deviations.

2 Conditioning as projection

> **Q.** We observe Y. What is our best guess for X?
>
> Before answering we must say what *best* means.

Fix jointly Gaussian (X, Y) in $\mathbb{R} \times \mathbb{R}^m$, both centered for readability.

2.1 What would count as an answer

Measure a guess by mean squared error. A predictor is any function of the observation, and we want the one minimizing $\mathbb{E}[(X - f(Y))^2]$.

Definition 2.1 (Conditional expectation)**.** $\mathbb{E}[X \mid Y]$ is the minimizer of $\mathbb{E}[(X - f(Y))^2]$ over all measurable f with $\mathbb{E}f(Y)^2 < \infty$.

One could instead define $\mathbb{E}[X \mid Y]$ as the orthogonal projection of X onto the span of the coordinates of Y. That would make Theorem 2.3 a tautology. Under Definition 2.1 the competitors are all functions of Y, including nonlinear ones.

2.2 The right inner product

On the linear span of the coordinates of X and Y — a finite-dimensional space of random variables — put

$$\langle U, V \rangle \;:=\; \mathbb{E}[UV].$$

Inner product is covariance; orthogonal is uncorrelated. Let $H_Y = \operatorname{span}\{Y_1, \dots, Y_m\}$ and let Π be orthogonal projection onto H_Y.

Lemma 2.2 (Best linear predictor)**.** $\Pi X = \Sigma_{XY}\Sigma_{YY}^{-1}Y$*, with* $\Sigma_{XY} = \mathbb{E}[XY^\top]$ *and* $\Sigma_{YY} = \mathbb{E}[YY^\top]$*. If* Σ_{YY} *is singular the same formula holds with a pseudo-inverse, and* ΠX *remains unique.*

Proof. Write $\Pi X = \sum_j c_j Y_j$. Orthogonality of the residual to each Y_k reads $\mathbb{E}[(X - \sum_j c_j Y_j)Y_k] = 0$, that is $\Sigma_{YY}c = \Sigma_{YX}$. ∎

Lemma 2.2 holds for any square-integrable pair.

2.3 The theorem

Theorem 2.3 (Conditioning is projection)**.** *Let* (X, Y) *be jointly Gaussian and centered. Then*

$$\mathbb{E}[X \mid Y] \;=\; \Pi X \;=\; \Sigma_{XY}\Sigma_{YY}^{-1}Y.$$

Proof. Let $R = X - \Pi X$. By construction $\mathrm{Cov}(R, Y) = \Sigma_{XY} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YY} = 0$, and (R, Y) is a linear image of (X, Y), hence jointly Gaussian by Theorem 1.3. So Theorem 1.11 gives $R \perp\!\!\!\perp Y$. For any competitor f,

$$\mathbb{E}\big[(X - f(Y))^2\big] = \mathbb{E}[R^2] + \mathbb{E}\big[(\Pi X - f(Y))^2\big] + 2\,\mathbb{E}\big[R\,(\Pi X - f(Y))\big].$$

The cross term vanishes: $\Pi X - f(Y)$ is a function of Y, and R is independent of Y with mean zero, so the expectation factors. Hence $\mathbb{E}[(X - f(Y))^2] \geq \mathbb{E}[R^2]$, with equality exactly when $f(Y) = \Pi X$ almost surely. ∎

Remark 2.4. Uncorrelatedness of R and Y does not suffice to kill the cross term; independence does, and independence comes from Theorem 1.11.

2.4 The conditional law

Since $X = \Pi X + R$ with $R \perp\!\!\!\perp Y$, conditioning on $Y = y$ leaves R untouched.

Corollary 2.5 (Schur complement)**.** *For jointly Gaussian (X, Y),*

$$X \mid Y = y \;\sim\; \mathcal{N}\Big(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y),\;\; \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Big).$$

Remark 2.6. The conditional covariance does not depend on y.

Remark 2.7. Conditioning on $\{Y = y\}$ conditions on an event of probability zero. The decomposition $X = \Pi X + R$ with $R \perp\!\!\!\perp Y$ constructs a conditional law directly, so Corollary 2.5 is a definition justified by Theorem 2.3.

2.5 Bayesian linear models and ridge regression

Take

$$y \;=\; Hw + \sigma\xi, \qquad w \sim \mathcal{N}(0, \tau^2 I_p), \quad \xi \sim \mathcal{N}(0, I_n) \text{ independent of } w.$$

The pair (w, y) is a linear image of (w, ξ), hence jointly Gaussian by Theorem 1.3. Reading off blocks,

$$\Sigma_{ww} = \tau^2 I, \qquad \Sigma_{wy} = \tau^2 H^\top, \qquad \Sigma_{yy} = \tau^2 H H^\top + \sigma^2 I,$$

and Corollary 2.5 gives the posterior. Simplifying with $\tau^2 H^\top(\tau^2 H H^\top + \sigma^2 I)^{-1} = (H^\top H + \lambda I)^{-1} H^\top$:

Proposition 2.8. *With $\lambda = \sigma^2/\tau^2$,*

$$\mathbb{E}[w \mid y] = \big(H^\top H + \lambda I\big)^{-1} H^\top y, \qquad \mathrm{Cov}(w \mid y) = \sigma^2\big(H^\top H + \lambda I\big)^{-1}.$$

The posterior mean is ridge regression, with regularization parameter the ratio of noise variance to prior variance.

Exercise 2.9. Verify the matrix identity above. Examine the limits $\tau \to \infty$ and $\tau \to 0$ in Proposition 2.8: what estimator does each give, and does the first exist when $n < p$?

Homework 1 (entropy, and why the Gaussian maximizes it) is due at the start of Lecture 2. Homework 2 develops §2.5 in full.