Lectures 1 and 2: Gaussians

Definition, linear structure, and conditioning as projection

OUTLINE $\varphi_X = \exp(\text{quadratic}) \longrightarrow AX + b$ is Gaussian $\longrightarrow$ uncorrelated $\Rightarrow$ independent $\longrightarrow$ conditioning is orthogonal projection $\longrightarrow$ ridge regression, and denoising

HOW TO READ Sections marked $*$ are not lectured. They are complete, they are examinable, and they carry most of the exercises. Exercises are typed A (did you follow the lecture), B (did you follow the argument) and C (one idea, short once you have it). Each week's quiz has one of each.

1 Gaussians, and their linear structure

1.1 Some facts from Fourier analysis

The following two facts are standard and we do not prove them. The transform carries $e^{-ix\cdot\xi}$, the inverse $e^{+ix\cdot\xi}$.

Proposition 1.1 (Fourier analysis, imported)**.** *For $f \in L^1(\mathbb{R}^d)$ define*

$$\widehat{f}(\xi) \;=\; \int_{\mathbb{R}^d} f(x)\, e^{-ix\cdot\xi}\, dx, \qquad \xi \in \mathbb{R}^d.$$

Then:

(i) ***Inversion.*** *If f and $\widehat{f}$ both lie in $L^1(\mathbb{R}^d)$ then $f(x) = (2\pi)^{-d}\int_{\mathbb{R}^d} \widehat{f}(\xi)\, e^{ix\cdot\xi}\, d\xi$ at every point of continuity of f.*

(ii) ***Derivatives.*** *If $f, \partial_j f \in L^1$ then $\widehat{\partial_j f}(\xi) = i\xi_j \widehat{f}(\xi)$; if $x_j f \in L^1$ then $\partial_{\xi_j}\widehat{f}(\xi) = (\widehat{-ix_j f})(\xi)$. Differentiating in x is multiplication by $i\xi_j$, and multiplication by x_j is $i\partial_{\xi_j}$.*

(iii) ***Isometry.*** *For $f \in L^1 \cap L^2$ one has $\|\widehat{f}\|_{L^2}^2 = (2\pi)^d \|f\|_{L^2}^2$, and the transform extends to a bijection of $L^2(\mathbb{R}^d)$ onto itself.*

(iv) ***Uniqueness.*** *Let W, W' be random vectors in $\mathbb{R}^d$. If $\mathbb{E}[e^{-iW\cdot\xi}] = \mathbb{E}[e^{-iW'\cdot\xi}]$ for every ξ then $W \overset{d}{=} W'$. If h satisfies $\mathbb{E}|h(W)| < \infty$ and $\mathbb{E}\big[h(W)e^{-iW\cdot\xi}\big] = 0$ for every ξ, then $h(W) = 0$.*

Imported: exchanging limits with an expectation. Let Z, Z_n and Z_t be random variables.

(a) *Dominated convergence.* If $Z_n \to Z$ and $|Z_n| \le W$ with $\mathbb{E}W < \infty$, then $\mathbb{E}Z_n \to \mathbb{E}Z$. The same statement holds for integrals $\int \cdot\, dx$ in place of $\mathbb{E}$.

(b) *Fatou.* If $Z_n \ge 0$ and $Z_n \to Z$, then $\mathbb{E}Z \le \liminf_n \mathbb{E}Z_n$.

(c) *Differentiation.* If the difference quotients of $\lambda \mapsto Z_\lambda$ are dominated by a single integrable variable, then $\frac{d}{d\lambda}\mathbb{E}[Z_\lambda] = \mathbb{E}[\partial_\lambda Z_\lambda]$.

(d) *Fubini.* If $\mathbb{E}\int |Z_t|\, dt < \infty$, then $\mathbb{E}\int Z_t\, dt = \int \mathbb{E}[Z_t]\, dt$.

1.2 The definition

Definition 1.2 (Characteristic function)**.** The *characteristic function* of a random vector X in $\mathbb{R}^d$ is

$$\varphi_X(\xi) \;=\; \mathbb{E}\big[e^{-iX\cdot\xi}\big], \qquad \xi \in \mathbb{R}^d.$$

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

Definition 1.3 (Gaussian vector)**.** Let $\mu \in \mathbb{R}^d$ and let Σ be a symmetric $d \times d$ matrix. A random vector X in $\mathbb{R}^d$ is *Gaussian with mean μ and covariance Σ*, written $X \sim \mathcal{N}(\mu, \Sigma)$, if

$$\boxed{\varphi_X(\xi) \;=\; \exp\Big(- i\,\mu\cdot\xi \;-\; \tfrac{1}{2}\,\xi^{\top}\Sigma\,\xi\Big)} \qquad \text{for every } \xi \in \mathbb{R}^d.$$

By Proposition 1.1(iv) at most one law satisfies this; §1.3 settles when one does.

Proposition 1.4 (The parameters are what they are called)**.** *If $X \sim \mathcal{N}(\mu, \Sigma)$ then $\mathbb{E}|X|^2 < \infty$, and*

$$\mathbb{E}[X_j] = \mu_j, \qquad \mathbb{E}\big[(X_j - \mu_j)(X_k - \mu_k)\big] = \Sigma_{jk}.$$

In particular the law determines (μ, Σ), so two Gaussians agree in law if and only if their parameters agree.

Proof. Write $\psi(\xi) = -i\mu\cdot\xi - \frac{1}{2}\xi^{\top}\Sigma\xi$, so $\varphi_X = e^{\psi}$ is smooth and

$$\partial_j\varphi_X\big|_{\xi=0} = -i\mu_j, \qquad \partial_j\partial_k\varphi_X\big|_{\xi=0} = \big(\partial_j\partial_k\psi + \partial_j\psi\,\partial_k\psi\big)\big|_{\xi=0} = -\Sigma_{jk} - \mu_j\mu_k.$$

Finiteness of the second moment is not an assumption; it is forced by this smoothness. For $h > 0$,

$$\frac{2\varphi_X(0) - \varphi_X(he_j) - \varphi_X(-he_j)}{h^2} \;=\; \mathbb{E}\Big[\frac{2 - 2\cos(hX_j)}{h^2}\Big],$$

and the left side converges to $-\partial_j^2\varphi_X(0) = \Sigma_{jj} + \mu_j^2$ as $h \downarrow 0$. The integrand is non-negative and converges pointwise to X_j^2, so Fatou's lemma gives $\mathbb{E}[X_j^2] \le \Sigma_{jj} + \mu_j^2 < \infty$.

With $\mathbb{E}|X|^2 < \infty$ the difference quotients of $e^{-iX\cdot\xi}$ are dominated by $|X_j|$ and by $|X_jX_k|$, so we may differentiate under the expectation twice: $\partial_j\varphi_X(0) = -i\,\mathbb{E}[X_j]$ and $\partial_j\partial_k\varphi_X(0) = -\mathbb{E}[X_jX_k]$. Comparing with the display gives $\mathbb{E}[X_j] = \mu_j$ and $\mathbb{E}[X_jX_k] = \Sigma_{jk} + \mu_j\mu_k$. ∎

Theorem 1.5 (Affine images)**.** *If $X \sim \mathcal{N}(\mu, \Sigma)$ and $A \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$, then*

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

Proof. $\varphi_{AX+b}(\xi) = e^{-ib\cdot\xi}\,\mathbb{E}\big[e^{-iX\cdot(A^{\top}\xi)}\big] = e^{-ib\cdot\xi}\varphi_X(A^{\top}\xi)$. Substituting Definition 1.3 and using $\mu\cdot A^{\top}\xi = (A\mu)\cdot\xi$ gives $\exp\big(-i(A\mu+b)\cdot\xi - \frac{1}{2}\xi^{\top}A\Sigma A^{\top}\xi\big)$. ∎

Corollary 1.6 (Rotational invariance)**.** *If $X \sim \mathcal{N}(0, I_d)$ and $O \in O(d)$ then $OX \overset{d}{=} X$. More generally, for $X \sim \mathcal{N}(0, \Sigma)$ one has $OX \overset{d}{=} X$ precisely when $O\Sigma O^{\top} = \Sigma$.*

Proof. Theorem 1.5 gives $OX \sim \mathcal{N}(0, O\Sigma O^{\top})$, and by Proposition 1.4 that law equals $\mathcal{N}(0, \Sigma)$ exactly when $O\Sigma O^{\top} = \Sigma$. For $\Sigma = I_d$ this is $OO^{\top} = I_d$. ∎

Exercise 1.9 shows that this property characterizes the Gaussian among laws with independent coordinates.

Corollary 1.7 (One coordinate at a time)**.** *$X \sim \mathcal{N}(\mu, \Sigma)$ if and only if $v \cdot X \sim \mathcal{N}(v \cdot \mu,\ v^\top \Sigma v)$ for every $v \in \mathbb{R}^d$.*

Proof. Forward: take $A = v^\top$ in Theorem 1.5. Conversely, the characteristic function of the scalar $v \cdot X$ evaluated at 1 is $\mathbb{E}[e^{-i\,v\cdot X}] = \varphi_X(v)$; by hypothesis it equals $\exp(-i\,v\cdot\mu - \frac{1}{2}v^\top\Sigma v)$, which is Definition 1.3. ∎

Exercise 1.8 (TYPE A)**.** Let $X \sim \mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^2$ with $\mu = (1, -1)$ and $\Sigma = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$.

(a) Write $\varphi_X(\xi)$ out explicitly and recover $\mathbb{E}[X_1X_2]$ from it by differentiating, as in Proposition 1.4.

(b) Use Theorem 1.5 to find the laws of $X_1 + X_2$, of $X_1 - X_2$, and of the pair $(X_1 + X_2,\ X_1 - X_2)$. Are the two independent?

(c) Now let Σ be an arbitrary covariance on $\mathbb{R}^2$. For which Σ are X_1+X_2 and X_1-X_2 independent, and for which is there a $v \neq 0$ with $v \cdot X$ constant?

Exercise 1.9 (TYPE C)**.** Let $d \geq 2$ and let X have independent and identically distributed coordinates with $OX \overset{d}{=} X$ for every $O \in O(d)$. Show that $X \sim \mathcal{N}(0, cI_d)$ for some $c \geq 0$. You may use that a characteristic function is continuous.

1.3 Existence

Nothing so far exhibits a single Gaussian.

Lemma 1.10. *Let g have density $(2\pi)^{-1/2}e^{-x^2/2}$ on $\mathbb{R}$. Then $\mathbb{E}\big[e^{-i\lambda g}\big] = e^{-\lambda^2/2}$; that is, $g \sim \mathcal{N}(0, 1)$.*

Proof. Set $f(\lambda) = \mathbb{E}[e^{-i\lambda g}]$. The difference quotients of $\lambda \mapsto e^{-i\lambda g}$ are dominated by $|g|$, so we may differentiate under the expectation:

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

The observation that does the work is $x\, e^{-x^2/2} = -\frac{d}{dx}e^{-x^2/2}$, which converts the factor of x into a derivative and lets integration by parts move it onto the other factor:

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

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

Remark 1.11. The proof used $\mathbb{E}[g\, h(g)] = \mathbb{E}[h'(g)]$ with $h(x) = e^{-i\lambda x}$: against a standard Gaussian, multiplication by g acts like differentiation. It returns in Lectures 4 and 5.

If $Z = (g_1, \ldots, g_d)$ has independent standard normal coordinates then

$$\varphi_Z(\xi) = \prod_{j=1}^{d} \mathbb{E}\big[e^{-i\xi_j g_j}\big] = \prod_{j=1}^{d} e^{-\xi_j^2/2} = e^{-|\xi|^2/2},$$

so $Z \sim \mathcal{N}(0, I_d)$.

Proposition 1.12 (Well-posedness)**.** *Let $\mu \in \mathbb{R}^d$ and let Σ be symmetric. A random vector satisfying Definition 1.3 exists if and only if $\Sigma \succeq 0$, and is then unique in law. If $\Sigma = AA^\top$ one may take $X = AZ + \mu$ with $Z \sim \mathcal{N}(0, I_d)$.*

Proof. If $\Sigma \succeq 0$, factor $\Sigma = AA^\top$. Theorem 1.5 applied to $Z \sim \mathcal{N}(0, I_d)$ gives $AZ + \mu \sim \mathcal{N}(\mu, AA^\top) = \mathcal{N}(\mu, \Sigma)$. Uniqueness in law is Proposition 1.1(iv).

Conversely, suppose $v^\top \Sigma v = -c$ with $c > 0$. Along $\xi = sv$ the right side of Definition 1.3 has modulus $e^{s^2 c/2} \to \infty$, contradicting $|\varphi_X| \le 1$. ∎

The Gaussian laws on $\mathbb{R}^d$ are therefore indexed exactly by $\mathbb{R}^d \times \mathrm{Sym}_{\ge 0}(\mathbb{R}^d)$.

Example 1.13 (The degenerate case)**.** Take $d = 2$, $\mu = 0$ and $\Sigma = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, positive semidefinite of rank one. By Proposition 1.12 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, being supported on a line, and Definition 1.3 never notices; a definition through the density $\propto e^{-\frac{1}{2} x^\top \Sigma^{-1} x}$ would exclude it.

Exercise 1.14 (TYPE A)**.** Let $f(x) = (2\pi)^{-d/2} e^{-|x|^2/2}$ on $\mathbb{R}^d$. Show $\widehat{f}(\xi) = e^{-|\xi|^2/2}$, then verify Proposition 1.1(i) on this pair by evaluating $(2\pi)^{-d} \int_{\mathbb{R}^d} e^{-|\xi|^2/2} e^{ix\cdot\xi} d\xi$ and recovering f. Conclude that the standard Gaussian density is a fixed point of the Fourier transform up to the factor $(2\pi)^{d/2}$.

Exercise 1.15 (TYPE B)**.** Let $\Sigma \succ 0$, write $\Sigma = AA^\top$ with A invertible, and let $X = AZ + \mu$.

(a) Change variables in the density of Z to obtain $p_X(x) = (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \exp\big(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu)\big)$, and check that the answer does not depend on which square root A was chosen.

(b) Show that when Σ is singular no density exists: produce $v \neq 0$ with $v \cdot (X - \mu) = 0$, and conclude that X lives on an affine subspace of dimension less than d.

1.4 Uncorrelated versus independent

Theorem 1.16. *Let (X, Y) be jointly Gaussian in $\mathbb{R}^d \times \mathbb{R}^{d'}$ — the concatenated vector is Gaussian — with $\mathrm{Cov}(X, Y) = 0$. Then $X \perp\!\!\!\perp Y$.*

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

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

which is the characteristic function of a pair (X', Y') built from independent $X' \overset{d}{=} X$ and $Y' \overset{d}{=} Y$. By Proposition 1.1(iv), $(X, Y) \overset{d}{=} (X', Y')$, so $X \perp\!\!\!\perp Y$. ∎

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

Example 1.17 (Uncorrelated, standard normal, not independent)**.** Let $g \sim \mathcal{N}(0, 1)$ and let ε be an independent random sign; set $X = g$ and $Y = \varepsilon g$. 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 1.10,

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

a sum of two Gaussian transforms and not exp(quadratic). The pair fails Definition 1.3 even though each coordinate satisfies it.

Exercise 1.18 (TYPE B)**.** Continue Example 1.17.

(a) Find the law of $X+Y$ and show it is not Gaussian, by exhibiting an atom.

(b) Corollary 1.7 says a vector is Gaussian as soon as every linear functional of it is. Compute the law of $sX+uY$ for arbitrary s,u and say why there is no contradiction.

(c) Does (X,Y) have a density on $\mathbb{R}^2$?

1.5 $*$ Entropy, and why the Gaussian maximizes it

Not lectured. Exercises due at the start of Lecture 2.

Among all laws on $\mathbb{R}^d$ with a given mean and covariance $\Sigma \succ 0$, the Gaussian has the largest differential entropy. That is what this section proves; Lecture 3 uses it.

If an event of probability p occurs, the surprise $s(p)$ ought to satisfy $s(1)=0$ and $s(pq)=s(p)+s(q)$. Substituting $p=e^{-u}$ turns the second into $g(u+v)=g(u)+g(v)$ for $g(u)=s(e^{-u})$, whose only continuous solutions are linear. So $s(p)=-c\log p$, the base of the logarithm is a choice of units, and we take $c=1$.

The **entropy** of a law on a finite set is the surprise expected before looking,

$$H(p) = \mathbb{E}\big[-\log p(X)\big] = -\sum_x p(x)\log p(x),$$

and for a density p on $\mathbb{R}^d$ the same formula defines the **differential entropy** $h(p)=-\int p\log p$. The second is not the first in disguise.

Exercise 1.19 (TYPE A)**.** (a) Compute H for a coin of bias q; sketch it on $q\in[0,1]$ and find where it is largest. Compute H for the uniform law on n points.

(b) Compute h for the uniform law on $[0,a]$, the exponential of rate λ, and $\mathcal{N}(0,\sigma^2)$ — you should find $\tfrac{1}{2}\log(2\pi e\sigma^2)$ for the last. Exhibit a law with *negative* differential entropy, and say why no discrete entropy can be negative.

(c) Show $h(aX)=h(X)+\log|a|$ for $a\neq 0$. So "the entropy of a length" is not a well-defined number, while H has no units at all.

Part (c) shows h is not the right quantity to state a maximization in, so measure entropy *relative* to a reference: for densities p,q, the **relative entropy** is $D(p\,\|\,q)=\int p\log(p/q)$.

Exercise 1.20 (TYPE B)**.** Prove **Gibbs' inequality**: $D(p\,\|\,q)\geq 0$, with equality iff $p=q$. *Jensen applied to* $-\log$. Deduce the discrete instance $H(p)\leq \log n$ for every law on n points, with equality only for the uniform one, by taking q uniform. Then check that $D(p\,\|\,q)$ is unchanged when X is rescaled — unlike h.

Now fix μ and $\Sigma\succ 0$, let q be the Gaussian density, and let p be any density with that mean and covariance. Expanding $0\leq D(p\,\|\,q)$ gives $0\leq -h(p)-\int p\log q$, and everything hinges on the second term.

Exercise 1.21 (TYPE B)**.** Write out $\log q$ and observe it is a *quadratic polynomial* in x. Deduce that $\int p \log q$ depends on p only through its mean and covariance, hence equals $\int q \log q = -h(q)$. Conclude $h(p) \le h(q)$ with equality only when $p = q$, and check $h(q) = \frac{1}{2}\log\big((2\pi e)^d \det \Sigma\big)$.

The argument used only that $\log q$ is quadratic, so that $\int p \log q$ sees p through its first two moments alone.

Exercise 1.22 (TYPE C)**.** Let X, Y be i.i.d., *symmetric*, each with a density. Show

$$h\Big(\frac{X+Y}{\sqrt{2}}\Big) \;\ge\; h(X).$$

You may use subadditivity $h(U,V) \le h(U) + h(V)$, which is Gibbs' inequality applied to a joint density against the product of its marginals.

Exercise 1.23 (TYPE C)**.** Fix $S \subseteq \mathbb{R}^d$, functions $T_1, \dots, T_k$ on S and numbers $c_1, \dots, c_k$. **Assume** that for some $\lambda_1, \dots, \lambda_k$ the function $\exp(\sum_m \lambda_m T_m)$ has finite integral Z over S, so that $q = Z^{-1}\exp(\sum_m \lambda_m T_m)$ is a density, and that $\int q T_m = c_m$ for each m. Show q uniquely maximizes h among densities on S with those moments.

Recover as instances the uniform law on $[0, a]$, the exponential at fixed mean on $[0, \infty)$, and $\mathcal{N}(\mu, \Sigma)$ at fixed mean and covariance. Finally, show the assumption does real work: on $S = \mathbb{R}$ with the mean alone constrained no λ makes $e^{\lambda x}$ integrable, and there is no maximizer — exhibit densities of mean 0 with h arbitrarily large.

2 Conditioning as projection

> **Q.** We observe Y. What is our best guess for X?

2.1 Expectation is a projection

Let $Z \sim \nu$ be a random vector in $\mathbb{R}^k$ and let $f : \mathbb{R}^k \to \mathbb{R}$ satisfy $\mathbb{E}[f(Z)^2] < \infty$.

Lemma 2.1. $\mathbb{E}[f(Z)] = \arg\min_{c \in \mathbb{R}} \mathbb{E}\big[(f(Z) - c)^2\big]$, *and the minimum value is* $\mathrm{Var}(f(Z))$.

Proof. $\mathbb{E}[(f(Z) - c)^2] = \mathbb{E}\big[(f(Z) - \mathbb{E}f(Z))^2\big] + \big(\mathbb{E}f(Z) - c\big)^2$, since the cross term is $2(\mathbb{E}f(Z) - c)\,\mathbb{E}[f(Z) - \mathbb{E}f(Z)] = 0$. ∎

Two readings. First, $\mathbb{E}[f(Z)]$ is the best deterministic guess for $f(Z)$. Second, write

$$L^2(\nu) = \Big\{ f : \mathbb{R}^k \to \mathbb{R} \;\Big|\; \mathbb{E}[f(Z)^2] < \infty \Big\}, \qquad \langle f, g \rangle_\nu = \mathbb{E}\big[f(Z)g(Z)\big].$$

This is an inner product once f and g with $\mathbb{E}[(f(Z) - g(Z))^2] = 0$ are counted as the same element, which we do from here on. We use the standard properties of L^2: it is complete, and every closed subspace H carries a unique orthogonal projection Π_H, characterized by $\Pi_H f \in H$ and $f - \Pi_H f \perp H$, and equal to the unique minimizer of $\|f - h\|_\nu$ over $h \in H$.

Since $\mathbb{E}[(f(Z) - c)^2] = \|f - c\|_\nu^2$, Lemma 2.1 says $\mathbb{E}[f(Z)] = \Pi_H f$ for H the constants.

2.2 Conditional expectation

Let $Z = (X, Y)$ with $X \in \mathbb{R}$ and $Y \in \mathbb{R}^m$, and write ν_Y for the law of Y. Since $\mathbb{E}[g(Y)^2]$ is the same number computed under ν or under ν_Y, the map $g \mapsto \big((x,y) \mapsto g(y)\big)$ identifies $L^2(\nu_Y)$ with a closed subspace of $L^2(\nu)$: the functions of Y alone.

Definition 2.2 (Conditional expectation)**.** $\mathbb{E}[X \mid Y]$ is the orthogonal projection of X onto $L^2(\nu_Y)$.

It is by construction an element of $L^2(\nu_Y)$, that is, a function of Y.

Lemma 2.3. *For $m \in L^2(\nu_Y)$ the following are equivalent:*

(i) $m = \mathbb{E}[X \mid Y]$;

(ii) $\mathbb{E}\big[(X - m(Y))\, g(Y)\big] = 0$ *for every* $g \in L^2(\nu_Y)$;

(iii) m *minimizes* $\mathbb{E}\big[(X - g(Y))^2\big]$ *over* $g \in L^2(\nu_Y)$.

Moreover, for every $g \in L^2(\nu_Y)$,

$$\mathbb{E}\big[(X - g(Y))^2\big] = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2\big] + \mathbb{E}\big[(\mathbb{E}[X \mid Y] - g(Y))^2\big]. \tag{1}$$

Proof. The equivalences are the characterization of Π_H; the identity is Pythagoras. ∎

For $X \in \mathbb{R}^p$ apply the definition to each coordinate.

2.3 The tower property

Proposition 2.4. *Let $Z = (X, Y, V)$. Then*

$$\mathbb{E}\big[\,\mathbb{E}[X \mid Y, V] \,\big|\, Y\,\big] = \mathbb{E}[X \mid Y], \qquad \mathbb{E}\big[\,\mathbb{E}[X \mid Y]\,\big] = \mathbb{E}[X],$$

and $\mathbb{E}\big[g(Y)X \mid Y\big] = g(Y)\,\mathbb{E}[X \mid Y]$ for bounded g.

Proof. $L^2(\nu_Y) \subseteq L^2(\nu_{(Y,V)})$, both closed in $L^2(\nu)$. For nested closed subspaces $H_1 \subseteq H_2$ the vector $\Pi_{H_2} f - f$ is orthogonal to H_2, hence to H_1, so $\Pi_{H_1}\Pi_{H_2} f = \Pi_{H_1} f$. The first identity is this pair of subspaces; the second is H_1 the constants inside $H_2 = L^2(\nu_Y)$. For the third, $\mathbb{E}[(g(Y)X - g(Y)\mathbb{E}[X \mid Y])h(Y)] = \mathbb{E}[(X - \mathbb{E}[X \mid Y])(gh)(Y)] = 0$ for every h. ∎

Corollary 2.5 (Law of total variance)**.** *With* $\mathrm{Var}(X \mid Y) := \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2 \mid Y\big]$,

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

Proof. Take $g \equiv \mathbb{E}X$ in (1), and $\mathbb{E}\big[\mathbb{E}[\,\cdot \mid Y]\big] = \mathbb{E}[\,\cdot\,]$ for the first term. ∎

Exercise 2.6 (TYPE A)**.** Show $\mathrm{Var}\,\big(\mathbb{E}[X \mid Y]\big) \le \mathrm{Var}(X)$. Characterize equality, and characterize the case in which the first term of Corollary 2.5 carries all of the variance.

Exercise 2.7 (TYPE B)**.** Let J be a fair coin and, given J, let $X \sim \mathcal{N}(\pm 1, \sigma^2)$ accordingly. Compute both terms of Corollary 2.5 with $Y = J$ and check the total against $\mathrm{Var}(X)$. Then let W be a function of $X_1, \dots, X_n$ and suppose $\mathbb{E}[U \mid X_1, \dots, X_n] = -W/n$. Deduce $\mathbb{E}[U \mid W] = -W/n$, naming the property used.

2.4 Gaussian conditioning

Let $(X, Y) \sim \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^p \times \mathbb{R}^m$, with blocks Σ_{XX}, $\Sigma_{XY} = \mathbb{E}[XY^\top]$, $\Sigma_{YY} = \mathbb{E}[YY^\top]$.

Lemma 2.8. *There exists $A \in \mathbb{R}^{p\times m}$ with $A\Sigma_{YY} = \Sigma_{XY}$; any two solutions satisfy $A_1 Y = A_2 Y$; and if $\Sigma_{YY} \succ 0$ the solution is unique and equals $\Sigma_{XY}\Sigma_{YY}^{-1}$.*

Proof. Transposed, the equation is $\Sigma_{YY}A^\top = \Sigma_{YX}$, solvable exactly when each column $\mathbb{E}[YX_i]$ of Σ_{YX} lies in $\operatorname{range}\Sigma_{YY} = (\ker\Sigma_{YY})^\perp$. If $v \in \ker\Sigma_{YY}$ then $\operatorname{Var}(v \cdot Y) = v^\top\Sigma_{YY}v = 0$, so $v \cdot Y = 0$ and $v^\top\mathbb{E}[YX_i] = \mathbb{E}[(v \cdot Y)X_i] = 0$. If $B = A_1 - A_2$ then $\operatorname{Cov}(BY) = B\Sigma_{YY}B^\top = 0$, so $BY = 0$. ∎

Theorem 2.9. *With A as in Lemma 2.8, the vector $X - AY$ is independent of Y, and $\mathbb{E}[X \mid Y] = AY$.*

Proof. Lemma 2.3(ii) asks for a function of Y whose residual is uncorrelated with every $g(Y)$, and independence of the residual from Y would give that at once. Try AY. The pair $(X - AY,\, Y)$ is a linear image of (X, Y), hence jointly Gaussian by Theorem 1.5, and

$$\operatorname{Cov}\big(X - AY,\, Y\big) = \Sigma_{XY} - A\Sigma_{YY} = 0,$$

so $X - AY \perp\!\!\!\perp Y$ by Theorem 1.16. Hence for every $g \in L^2(\nu_Y)$,

$$\mathbb{E}\big[(X - AY)\, g(Y)\big] = \mathbb{E}[X - AY]\;\mathbb{E}[g(Y)] = 0.$$ ∎

Theorem 2.10. *Put $\Sigma_{X|Y} = \Sigma_{XX} - A\Sigma_{YX}$. Then for every $\xi \in \mathbb{R}^p$,*

$$\mathbb{E}\big[e^{-i\xi\cdot X} \;\big|\; Y\big] \;=\; \exp\Big(-i\,\xi\cdot AY - \tfrac{1}{2}\,\xi^\top\Sigma_{X|Y}\,\xi\Big),$$

that is, $X \mid Y \sim \mathcal{N}\big(AY,\ \Sigma_{X|Y}\big)$. In particular $\operatorname{Cov}(X \mid Y) = \Sigma_{X|Y}$ does not depend on Y.

Proof. Put $R = X - AY$. It is Gaussian, and using $A\Sigma_{YY} = \Sigma_{XY}$ twice,

$$\operatorname{Cov}(R) = \Sigma_{XX} - A\Sigma_{YX} - \Sigma_{XY}A^\top + A\Sigma_{YY}A^\top = \Sigma_{X|Y},$$

so $\mathbb{E}[e^{-i\xi\cdot R}] = e^{-\frac{1}{2}\xi^\top\Sigma_{X|Y}\xi}$. By Theorem 2.9 we have $R \perp\!\!\!\perp Y$, so this constant satisfies Lemma 2.3(ii) for $\mathbb{E}[e^{-i\xi\cdot R} \mid Y]$. Write $e^{-i\xi\cdot X} = e^{-i\xi\cdot AY}e^{-i\xi\cdot R}$ and take the bounded factor outside by Proposition 2.4. ∎

When $\Sigma_{YY} \succ 0$,

$$AY = \Sigma_{XY}\Sigma_{YY}^{-1}Y, \qquad \Sigma_{X|Y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX},$$

the *Schur complement* of Σ_{YY} in Σ. For uncentered (X, Y) replace AY by $\mu_X + A(Y - \mu_Y)$.

2.5 Bayesian linear models and ridge regression

Take

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

The pair (w, y) is a linear image of (w, ε), hence jointly Gaussian by Theorem 1.5, with

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

Theorem 2.10 gives the posterior; simplifying with $\tau^2 H^\top(\tau^2 HH^\top + \sigma^2 I)^{-1} = (H^\top H + \lambda I)^{-1}H^\top$:

Proposition 2.11. *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.

The exercises below are due at the start of Lecture 3.

Exercise 2.12 (TYPE A)**.** (a) Show $-\log p(y \mid w) = \frac{1}{2\sigma^2}\|y - Hw\|^2 + c$ with c free of w, so least squares is maximum likelihood. State the assumption on the noise that choosing squared error encodes.

(b) Show the maximizer of the posterior density is $\widehat{w}_\lambda = \arg\min_w \|y - Hw\|^2 + \lambda\|w\|^2$. Interpret large λ, and compute the limits $\tau \to 0$ and $\tau \to \infty$.

Exercise 2.13 (TYPE B)**.** Compute the three blocks of the joint covariance of (w, y) and apply Theorem 2.10 to obtain Proposition 2.11. Then verify that $\mathrm{Cov}(w \mid y)$ does not depend on y, and say what that means for an experimenter who observes a surprising y.

Exercise 2.14 (TYPE B)**.** Let $Y \sim \mathcal{N}(0,1)$ and $X = Y^2 - 1$. Show $\mathrm{Cov}(X, Y) = 0$, so the projection of X onto $\mathrm{span}\{1, Y\}$ is 0, with mean squared error 2. Compute $\mathbb{E}[X \mid Y]$ and its error. Which hypothesis of Theorem 2.9 fails?

2.6 ∗ The two forms of the ridge estimator

Not lectured. Proposition 2.11 inverts a $p \times p$ matrix; the derivation that produced it inverted an $n \times n$ one.

Exercise 2.15 (TYPE C)**.** Show that for $\lambda > 0$

$$\big(H^\top H + \lambda I_p\big)^{-1} H^\top \;=\; H^\top \big(H H^\top + \lambda I_n\big)^{-1},$$

and deduce that ridge regression with $p \gg n$ costs $O(n^3)$ rather than $O(p^3)$. Then take $\tau \to \infty$ with $n < p$ and H of full row rank, and show the posterior mean converges to $H^\top (H H^\top)^{-1} y$, the minimum-norm solution of $Hw = y$. Explain why the limit is visible in one of the two forms and not the other.

2.7 ∗ Denoising is conditioning

Not lectured. Let X be a random vector in $\mathbb{R}^d$ with $\mathbb{E}|X|^2 < \infty$, not assumed Gaussian, and

$$Y \;=\; X + \sigma\xi, \qquad \xi \sim \mathcal{N}(0, I_d) \text{ independent of } X.$$

Then Y has density $p_Y(y) = \mathbb{E}[\phi_\sigma(y - X)]$, where ϕ_σ is the $\mathcal{N}(0, \sigma^2 I_d)$ density. The function $\nabla \log p_Y$ is the *score* of Y.

Proposition 2.16 (Tweedie's formula)**.**

$$\mathbb{E}\big[X \mid Y\big] \;=\; Y + \sigma^2\, \nabla \log p_Y(Y).$$

Proof. By Lemma 2.3(ii) it suffices to show $\mathbb{E}\big[\langle X - Y - \sigma^2 \nabla \log p_Y(Y),\, f(Y)\rangle\big] = 0$ for every smooth compactly supported f.

Conditionally on X the vector ξ is standard Gaussian, so $\mathbb{E}[\eta F(\eta)] = \mathbb{E}[F'(\eta)]$ — Lemma 1.10, in each coordinate — gives

$$\mathbb{E}\big[\langle Y - X, f(Y)\rangle\big] = \sigma\,\mathbb{E}\big[\langle \xi, f(X+\sigma\xi)\rangle\big] = \sigma^2\,\mathbb{E}\big[\nabla\cdot f(Y)\big].$$

Integrating by parts in y,

$$\mathbb{E}\big[\nabla\cdot f(Y)\big] = \int (\nabla\cdot f)\,p_Y\,dy = -\int \langle f(y), \nabla p_Y(y)\rangle\,dy = -\,\mathbb{E}\big[\langle f(Y), \nabla \log p_Y(Y)\rangle\big]. \qquad \blacksquare$$

Exercise 2.17 (TYPE A)**.** Take $X \sim \mathcal{N}(0, \tau^2 I_d)$. Compute p_Y and $\nabla \log p_Y$, and check that Proposition 2.16 gives $\mathbb{E}[X \mid Y] = \frac{\tau^2}{\tau^2+\sigma^2} Y$, in agreement with Theorem 2.9. Which of the two derivations used that X is Gaussian?

Exercise 2.18 (TYPE B)**.** Show p_Y is smooth with $\nabla p_Y(y) = \mathbb{E}[\nabla \phi_\sigma(y - X)]$, and state what is required for the boundary term in the integration by parts to vanish. Then show that testing against smooth compactly supported f suffices in Lemma 2.3(ii).

Exercise 2.19 (TYPE C)**.** A *denoiser* is a minimizer D of $\mathbb{E}|X - D(Y)|^2$ over all functions $D : \mathbb{R}^d \to \mathbb{R}^d$. Show that D and $\nabla \log p_Y$ determine each other explicitly, and deduce that fitting a denoiser by least squares to pairs $(X, X + \sigma\xi)$ — which requires neither p_X nor a likelihood — estimates the score of the law of Y.