ORF 526 · Homework 2 · due Sep 9

Fitting is conditioning

Lecture 2 proved one theorem: for jointly Gaussian (X, Y), the best predictor of X given Y is an orthogonal projection. This assignment turns that theorem into most of the standard apparatus of linear regression. The through-line is a dictionary:

a loss function is a noise model, a regularizer is a prior, and fitting is conditioning.

Throughout, $H \in \mathbb{R}^{n \times p}$ is a fixed design matrix and $y = Hw + \sigma\xi$ with $\xi \sim \mathcal{N}(0, I_n)$.

recalled from lecture Everything this set needs, so you can work it without the notes.
L1, Thm 1.3 (linear images). If $X \sim \mathcal{N}(\mu, \Sigma)$ then $AX + b \sim \mathcal{N}(A\mu + b,\ A\Sigma A^\top)$.
L1, Thm 1.11. If (X, Y) is *jointly* Gaussian and $\mathrm{Cov}(X, Y) = 0$, then X and Y are independent.
L2, Thm 2.3 (conditioning is projection). For jointly Gaussian centred (X, Y), the minimizer of $\mathbb{E}[(X - f(Y))^2]$ over *all* functions f is $\mathbb{E}[X \mid Y] = \Sigma_{XY}\Sigma_{YY}^{-1}Y$.
L2, Cor 2.5 (Schur complement). For jointly Gaussian (X, Y),

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

L2, Rmk 2.6. The conditional covariance above does not depend on y.

1. A loss function is a noise model

Least squares is usually introduced as an obvious thing to minimize. It is not obvious; it is a modelling assumption in disguise.

Problem 1. core Show that $-\log p(y \mid w) = \frac{1}{2\sigma^2}\|y - Hw\|^2 + c$ for a constant c not depending on w. Conclude that ordinary least squares is the maximum likelihood estimator. Then state, in one sentence, what a practitioner is assuming about the world when they choose to minimize squared error.

Problem 2. Replace Gaussian noise by Laplace noise, with density proportional to $e^{-|t|}$. Show the maximum likelihood estimator now minimizes $\|y - Hw\|_1$. One line of work, and the point of the section: the loss did not come from nowhere.

2. A regularizer is a prior

Now put a prior $w \sim \mathcal{N}(0, \tau^2 I_p)$ on the unknown parameter, independent of ξ.

Problem 3. core Show that the maximizer of the posterior density is the ridge estimator

$$\widehat{w}_\lambda = \arg\min_w\ \|y - Hw\|^2 + \lambda\|w\|^2, \qquad \lambda = \frac{\sigma^2}{\tau^2}.$$

Interpret λ: what does a large value say about your beliefs relative to your data?

3. Fitting is conditioning

Everything above went through densities and Bayes' rule. It did not have to. The pair (w, y) is a linear image of (w, ξ), hence jointly Gaussian — so Lecture 2 applies directly and no density ever appears.

Problem 4. core Compute the three blocks Σ_{ww}, Σ_{wy}, Σ_{yy} of the joint covariance of (w, y).

Problem 5. Apply the Schur complement formula to obtain $\mathbb{E}[w \mid y]$ and $\mathrm{Cov}(w \mid y)$. Verify the matrix identity

$$\tau^2 H^\top\big(\tau^2 HH^\top + \sigma^2 I\big)^{-1} = \big(H^\top H + \lambda I\big)^{-1}H^\top$$

and conclude that the posterior *mean* is exactly the ridge estimator of Problem 3. Compare the length of this derivation with the one you gave there.

Problem 6. Observe that $\mathrm{Cov}(w \mid y)$ does not depend on y. Explain what this says operationally: you have collected data and it came out surprising — by how much should you revise your uncertainty?

4. Two ways for the theorem to fail

Lecture 2 needed (X, Y) to be *jointly* Gaussian. The next two problems show what is lost without it, and they fail in different places.

Problem 7. Let $Y \sim \mathcal{N}(0,1)$ and $X = Y^2 - 1$, so both are centred. Show $\mathrm{Cov}(X, Y) = 0$, so that the best linear predictor of X is the constant 0, with mean squared error 2. Then exhibit a predictor with error 0. Linear prediction concludes that Y carries no information about X; in fact it determines it.

Problem 8. Let $g \sim \mathcal{N}(0,1)$, let ε be an independent random sign, and set $X = g$, $Y = \varepsilon g$. Show that $\mathbb{E}[Y \mid X] = 0$ — so here the conditional mean *is* linear — but that

$$\mathrm{Var}(Y \mid X = x) = x^2.$$

Which of the recalled statements does this contradict? Conclude that (X, Y) is not jointly Gaussian, and check this directly by computing $\varphi_{(X,Y)}(s, u)$ and observing that it is not exp(quadratic).

5. Shrinkage

Return to the Gaussian case and look at what ridge actually does to the least squares solution.

Problem 9. Let $H = UDV^\top$ be a singular value decomposition with singular values $d_1, \ldots, d_p$. Show that in the basis V, the i-th coordinate of $\widehat{w}_\lambda$ is the corresponding coordinate of the least squares solution multiplied by

$$\frac{d_i^2}{d_i^2 + \lambda}.$$

Which directions are shrunk most, and why is that the right behaviour?

Problem 10. Take $\tau \to \infty$ and $\tau \to 0$ in Problem 5. What estimator does each limit give? When $n < p$, show that the $\tau \to \infty$ limit of the posterior mean is the *minimum-norm* solution of $Hw = y$ rather than diverging.

Problem 11. Generate a random design H, and compute

$$\mathrm{df}(\lambda) \;=\; \mathrm{tr}\left(H\big(H^\top H + \lambda I\big)^{-1} H^\top\right)$$

numerically as λ ranges over several orders of magnitude. Plot it. Show it interpolates between p and 0, and argue why it deserves the name *effective number of parameters*.

6. Where this returns

Three of these objects come back, and it is worth knowing which.

- **Lecture 9.** Nothing in the derivation used where Σ came from. Replace Σ by a kernel matrix and Problem 5 becomes Gaussian process regression, unchanged.
- **Problem 11 is a resolvent.** The matrix $(H^\top H + \lambda I)^{-1}$ is the object that Lecture 7 studies for *random* H, and $\mathrm{df}(\lambda)$ is a trace of it. You have plotted, numerically, the quantity the Stieltjes transform computes.
- **Problem 10.** The minimum-norm interpolant in the regime $n < p$ is the estimator sitting exactly at the interpolation threshold. Why its test error behaves the way it does is a question we will be able to answer by the end of Module I.

CORE problems are short and self-contained. The first quiz, on Wednesday Sep 9, draws on the core problems of this set and of Homework 1. AI assistance is permitted and encouraged throughout.