Week 14: Wednesday, Nov 25

A regression puzzle

John Hopcroft asked me the following question a while back. It's a little off our normal track, but given that many people are away for the break, I thought I would share.

Problem 1. Suppose $p(x)$ is a polynomial of degree at most d, and that $\{x_i\}_{i=1}^n$ are sample points in $[-1,1]$. Let $y_i = p(x_i)$, and suppose that $\hat{y}_i = y_i + e_i$ where the errors e_i are i.i.d. $N(0,\delta^2)$. Let $\hat{p}(x)$ be the polynomial of degree at most d that minimizes $\sum_{i=1}^n (\hat{p}(x_i) - \hat{y}_i)^2$. If X is drawn uniformly from $[-1,1]$, what is $E[(p(X) - \hat{p}(X))^2]$?

The beauty of this problem is that, while the statement is simple, the analysis[1] brings together many of the ideas we have discussed in class:

- Orthogonal polynomials.
- Least squares problems and problems of orthonormal bases.
- Singular values.
- The Cauchy interlace theorem.

Let us begin at the beginning, with the connection to orthogonal polynomials.

Orthogonal polynomials

The expected squared error $E[(p(X) - \hat{p}(X))^2]$ depends on the independent random variables e_i and X. If X is uniform on $[-1,1]$, we have

$$E_X[(p(X) - \hat{p}(X))^2] = \frac{1}{2}\int_- 1^1 (p(x) - \hat{p}(x))^2 \, dx = \frac{1}{2}\|p - \hat{p}\|^2_{L^2([-1,1])},$$

where $L^2([-1,1])$ is the space of square-integrable functions on $[-1,1]$, with an inner product given by

$$\langle f, g \rangle = \int_{-1}^{1} f(x)g(x) \, dx.$$

[1] It seems likely that all this is well understood by experts in regression, possibly with a simpler notation. But numerical linear algebra is the language I know, so humor me.

Multiplication by x defines a linear operaton on $L^2([-1,1])$; furthermore, this linear operation is self-adjoint, i.e.

$$\langle f, xg \rangle = \int_{-1}^{1} xf(x)g(x)\,dx = \langle xf, g \rangle.$$

Though $L^2([-1,1])$ is not a finite dimensional space, we can still apply the Lanczos algorithm to generate an orthonormal set of polynomials $\{q_0, q_1, \ldots\}$. That is, starting from $q_0 = 1/\sqrt{2}$ we each successive degree $k+1$ polynomial q_{k+1} by orthogonalizing xq_k against $q_0, q_1, \ldots, q_k$. Because of self-adjointness, we only need to orthogonalize xq_k against q_k and q_{k-1}, since for $j < k-1$ we have

$$\langle xq_k, q_j \rangle = \langle q_k, xq_j \rangle = \left\langle q_k, \sum_{i=0}^{j+1} a_i q_i \right\rangle = 0,$$

because xq_j is a degree $j+1$ polyomial and the set $\{q_0, \ldots, q_{j+1}\}$ forms a basis for the degree $j+1$ polynomials. As with the finite-dimensional procedure, we now have

$$xq_j = \beta_{j-1}q_j + \alpha_j q_j + \beta_j q_{j+1},$$

which we can rewrite as

$$\beta_{j-1}q_j + (\alpha_j - x)q_j + \beta_j q_{j+1} = 0, \tag{1}$$

where

$$\begin{aligned} \alpha_j &= \langle q_j, xq_j \rangle, \\ \beta_j &= \|q_{j+1}\|. \end{aligned}$$

Now, notice that $q_0 = 1/2$ is an even function ($q_0(x) = q_0(-x)$); Therefore, xq_0 is an odd function, and is thus automatically orthogonal to q_0, so the Lanczos algorithm does not need to subtract off a multiple of q_0. This means that q_1 ends up being odd, and so xq_1 is even; and thus xq_1 is orthogonal to q_1. Continuing in this way, we have that q_k is odd whenever k is odd, and even whenever k is even, so that $\alpha_k = \langle q_k, xq_k \rangle = 0$ for any k. Therefore, the polynomials q_k are actually generated by a recurrence of the form

$$\beta_{j-1}q_j - xq_j + \beta_j q_{j+1} = 0. \tag{2}$$

Legendre polynomials

The *Legendre polynomials* appeared in Homework 4; they are defined by the three-term recurrence

$$kp_k(x) - (2k-1)x\,p_{k-1}(x) + (k-1)p_{k-2}(x) = 0$$

with

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x. \end{aligned}$$

As we might expect from the similarity to (2) the Legendre polynomials form an orthogonal (though not orthonormal) set in $L^2([-1,1])$, i.e.

$$\langle p_j, p_k \rangle = \int_{-1}^{1} p_j(x)p_k(x)\,dx = \begin{cases} \frac{2}{2k+1}, & j = k \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the orthonormal set generated in the previous section by the Lanczos recurrence is

$$q_k(x) = \sqrt{\frac{2k+1}{2}}\,p_k(x).$$

In particular, $q_0(x), \ldots, q_d(x)$ is an orthonormal basis for the space of polynomials of degree at most d. This means we can write

$$\begin{aligned} p(x) &= \sum_j q_j(x)c_j \\ \hat{p}(x) &= \sum_j q_j(x)\hat{c}_j, \end{aligned}$$

or, in matrix notation, $p(x) = Q(x)c$ and $\hat{p}(x) = Q(x)\hat{c}$ where $Q(x) = \begin{bmatrix} q_0(x) & \ldots & q_1(x) \end{bmatrix}$. Now, note that because the columns of Q are orthonormal in the inner product associated with $L^2([-1,1])$, we can write

$$\|p - \hat{p}\|_2 = \|Q(c - \hat{c})\|_2 = \|c - \hat{c}\|_2.$$

This is useful because

$$\begin{aligned} E[(p(X) - \hat{p}(X)_2^2] &= \frac{1}{2}E\left[\int_{-1}^{1} (p(x) - \hat{p}(x))^2\,dx\right] \\ &= \frac{1}{2}E[\|p - \hat{p}\|_2^2] = \frac{1}{2}E[\|c - \hat{c}\|_2^2]. \end{aligned}$$

Therefore, we really want to find the l^2 norm of $z = c - \hat{c}$.

Enter SVD

Let us write $\hat{p}(x) = Q(x)\hat{c}$. The polynomial $\hat{p}$ is chosen to minimize the squared error

$$\sum_i \left(\hat{p}(x_i) - \hat{y}_i\right)^2 = \sum_i \left(Q(x_i)\hat{c} - \hat{y}_i\right)^2 = \|A\hat{c} - \hat{y}\|_2^2,$$

where $A \in \mathbb{R}^{n\times d+1}$ has entries $A_{ij} = q_j(x_i)$ (think of the column indices as beginning at zero). We also know that $Ac = y$, where $p(x) = Q(x)c$ is the polynomial we wish to reconstruct. Therefore,

$$\|A\hat{c} - \hat{y}\|_2 = \|A(c - \hat{c}) - (y - \hat{y})\|_2 = \|Az - e\|_2$$

is minimized over all z.

Taking the SVD $A = U\Sigma V^T$ (U rectangular) and using unitary invariance of the l^2 norm, we have

$$\|U\Sigma V^T z - e\|_2 = \|\Sigma\hat{z} - \hat{e}\|,$$

where $\hat{z} = V^T z$ and $\hat{e} = U^T e$. The minimizing $\hat{z}$ is clearly $\hat{z} = \Sigma^{-1}\hat{e}$. Again using invariance of l^2 norms under unitary transformations, we have $\|z\|_2 = \|\hat{z}\|$; and using the fact that the joint distribution of i.i.d. Gaussian random variables is also invariant under unitary transformations, we have that $\hat{e} \in \mathbb{R}^d$ again has i.i.d. Gaussian components with mean zero and variance δ^2. Thus,

$$E[(p(X) - \hat{p}(X))^2] = \frac{1}{2}E[\|\Sigma^{-1}\hat{e}\|^2] = \sum_{j=1}^{d} \frac{\delta^2}{2\sigma_j^2} = \frac{\delta^2}{2}\|A^\dagger\|_F^2.$$

where $A^\dagger = V\Sigma^{-1}U = (A^TA)^{-1}A^T$ is the Moore-Penrose pseudoinverse of A.

An instability illustrated

In the case where $n = d+1$ the matrix A becomes square, $\|A^\dagger\|_F^2 = \|A^{-1}\|_F^2$, and our problem reduces from least squares to polynomial fitting. It is well-known that fitting polynomials through equally spaced points is unstable; let us put this in concrete terms. The following MATLAB function returns the matrix A defined in the previous section:

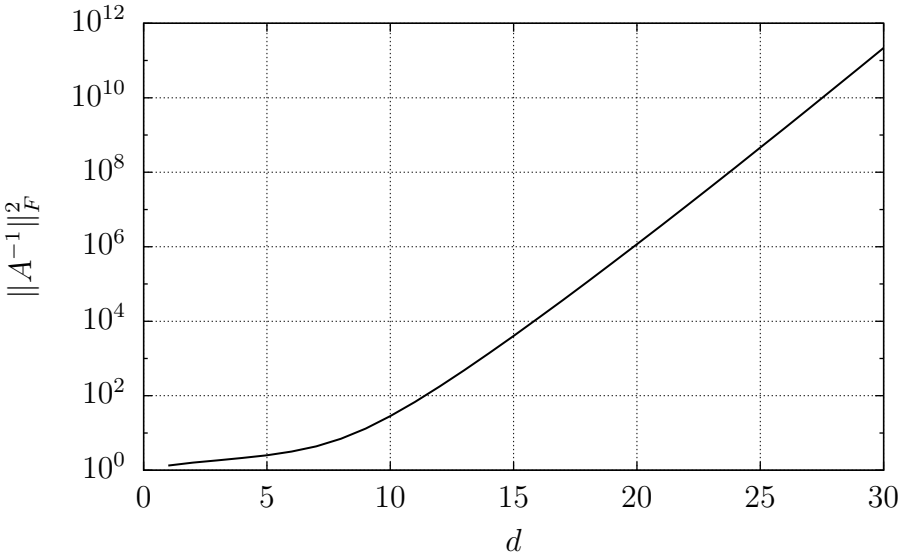

Figure 1: Growth of $\|A^{-1}\|_F^2$ for polynomial interpolation of degree d.

```
% A = lec36stab(x, d)
%
% Evaluate A(i,j) = q_{j-1}(x_i) for j = 1:d+1, where q_{j-1}
% is the normalized Legendre polynomial of degree j-1.

function A = lec36stab(x, d)

  n = length(x);
  A = zeros(n, d+1);
  pkm2 = 1;  A(:,1) = sqrt(0.5);
  pkm1 = x;  A(:,2) = sqrt(1.5)*x;
  for k = 2:d
    pk = ((2*k-1)*(x.*pkm1) - (k-1)*pkm2)/k;
    A(:,k+1) = sqrt(k+0.5)*pk;
    pkm2 = pkm1;
    pkm1 = pk;
  end
```

In Figure 1, we show $\|A^{-1}\|_F^2$ for $1 \leq d \leq 30$ on a semilogarithmic scale. The exponential growth in error amplification is clear.

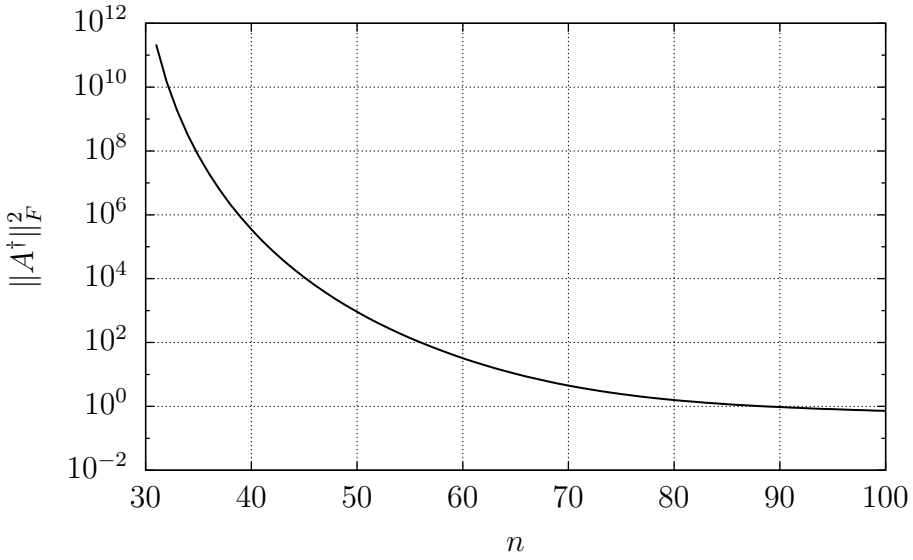

Figure 2: Decay of $\|A^\dagger\|_F^2$ for fitting degree-30 polynomials through n points.

The quadrature connection

While interpolation through $d+1$ points is risky business, fitting a degree d polynomial through $n \gg d$ equally spaced points is much safer. Let us again look at the behavior of $\|A^\dagger\|_F^2$, but this time fixing the polynomial degree at $d = 30$ and letting n vary (Figure 2). As the number of data points increases, $A^\dagger$ becomes increasingly well behaved. One way to see why this should be is to interpret the singular values in terms of numerical quadrature. Suppose $n \gg d$; then the trapezoidal rule for integration gives

$$\begin{aligned}\int_{-1}^{1} q_j(x) q_k(x)\, dx &\approx \frac{1}{n} q_j(x_1) q_k(x_1) + \frac{2}{n} \sum_{i=2}^{n-1} q_j(x_i) q_k(x_i) + \frac{1}{n} q_j(x_1) q_k(x_1) \\ &= \frac{1}{n} (A^T D A)_{jk},\end{aligned}$$

where $D = (\,1, 2, 2, \ldots, 2, 2, 1)$. Orthonormality of the q_j gives us

$$A^T D A \approx nI,$$

which means that the singular values of $\sqrt{D}A$ are all close to $\sqrt{n}$, and

$$\frac{1}{\sqrt{2}} \|\sqrt{D}Ax\|_2 \leq \|Ax\|_2 \leq \|\sqrt{D}Ax\|_2$$

so the singular values of A must lie approximately between $\sqrt{n/2}$ and $\sqrt{n}$ (and, as one might suspect, most of them will be close to $\sqrt{n/2}$). Therefore, we expect $\|A^\dagger\|_F^2$ to behave like $2d/n$ as n/d becomes sufficiently large; and this is indeed what happens.

For small values of n/d, the analysis above fails because the trapezoidal rule has a large error for the (generally highly oscillatory) polynomials $q_j(x)q_k(x)$ when $j+k$ is not much smaller than n. However, we can do an analogous analysis if we choose the nodes to be roots of the degree n Gauss-Legendre polynomial. In this case, we have an *exact* relation

$$A^T \hat{D} A = I$$

even when $n = d+1$. As before, this fact allows us to bound the singular values of A in terms of the entries of $\hat{D}$

Extensions to the analysis

The following extensions to the analysis described above are fairly easy:

1. The errors e could come from a multivariate normal distribution with known covariance. This leads to $\hat{e}$ distributed according to another multivariate normal distribution with known covariance.

2. We could choose a different distribution for the randomly chosen point X. This would lead us to choosing a different inner product

$$\langle f, g \rangle_\mu = \int f(x)g(x)\mu(x)\,dx,$$

   where $\mu(x)$ is the probability density for X. We could nonetheless construct orthogonal polynomials via Lanczos as before.

3. We could consider a weighted least square procedure to compute $\hat{c}$, corresponding to choosing a different inner product in the space of coefficients.