Notes for 2016-10-07

1 Choice of regularization

All of the regularization methods we have discussed share a common trait: they define a parametric family of models. With more regularization, we restrict the range of models we can easily generate (adding bias), but we also reduce the sensitivity of the fit (reducing variance). The choice of the regularization parameter is a key aspect of these methods, and we now briefly discuss three different ways of systematically making that choice. In all cases, we rely on the assumption that the sample observations we use for the fit are representative of the population of observations where we might want to predict.

1.1 Morozov's discrepancy principle

Suppose that we want to fit $Ax \approx \hat{b}$ by regularized least squares, and the (noisy) observation vector $\hat{b}$ is known to be within some error bound $\|e\|$ of the true values b. The discrepancy principle says that we should choose the regularization parameter so the residual norm is approximately $\|e\|$. That is, we seek the most stable fitting problem we can get subject to the constraint that the residual error for the regularized solution (with the noisy vector $\hat{b}$) is not much bigger than we would get from unknown true solution.

One of the most obvious drawbacks of the discrepancy principle is that it requires that we have an estimate for the norm of the error in the data. Sadly, such estimates are not always available.

1.2 The L-curve

A second approach to the regularization parameter is the *L-curve*. If we draw a parametric curve of the residual error versus solution norm on a log-log plot, with $\log \|r_\lambda\|$ on the x axis and $\log \|x_\lambda\|$ on the y axis, we often see an "L" shape. In the top of the vertical bar (small λ), we find that increasing regularization decreases the solution norm significantly without significantly increasing the residual error. Along the end of the horizontal part, increasing regularization increases the residual error, but does not significantly help with the solution norm. We want the corner of the curve, where the regularization

is chosen to minimize the norm of the solution subject to the constraint that the residual is close to the smallest possible residual (which we would have without regularization).

Computing the inflection point on the L-curve is a neat calculus exercise which we will not attempt here.

1.3 Generalized cross-validation

The idea with (generalized) cross-validation is to choose the parameter by fitting the model on a subset of the data and testing on the remaining data. We may do this with multiple partitions into data used for fitting versus data reserved for checking predictions. We often choose regularization parameters to give the smallest error on the predictions in a cross-validation study.

2 Nearness problems

So far, we have considered problems of miminizing a residual error where the unknown is a vector. What if instead the unknown is a matrix? There are a variety of such *matrix nearness* problems, and this is a good place in the course for them.

2.1 Nearest symmetric matrix

As a "warm-up," we consider the problem of finding the symmetric matrix A that is nearest to some target (nonsymmetric) matrix B:

$$\text{minimize}_{A=A^T} \|B - A\|_F^2.$$

There are several ways to tackle this problem, but we want to use it as an excuse once again to relate a least squares problem to an orthogonal decomposition. In this case, we note that

$$B = H + S, \quad H \equiv \frac{1}{2}(B + B^T), \quad S \equiv \frac{1}{2}(B - B^T).$$

The matrices H and S are respectively symmetric ($H = H^T$) and skew-symmetric ($S = -S^T$). We observe that the *Frobenius inner product* of any symmetric H and skew S is

$$\langle H, S \rangle_F = \sum_{i,j} h_{ij} s_{ij} = \sum_{i>j} h_{ij}(s_{ij} + s_{ji}) = 0,$$

i.e. the set of all square matrices can be written as the direct sum of the set of symmetric matrices and an orthogonal set of skew matrices. In particular, this means that we can invoke the Pythagorean theorem:

$$\|B - A\|_F^2 = \|H - A\|_F^2 + 2\langle (H - A), S\rangle_F + \|S\|^2 = \|H - A\|_F^2 + \|S\|^2.$$

So we decompose the objective into a piece that we can set exactly equal to zero and a piece that is independent of the optimization variable. The solution to the nearest symmetric matrix problem is $H = A$, and the distance is $\|S\|_F$.

We can pose the same question in the operator two-norm (rather than the Frobenius norm), and we get mostly the same answer. The main difference is that in the case of the two-norm, the minimizing A is generally not unique.

2.2 Nearest orthogonal matrix

Now consider the problem for $B \in \mathbb{R}^{m\times n}$ and $m \geq n$

$$\text{minimize}_{Q:Q^TQ=I} \|B - Q\|_F^2.$$

This is sometimes known as the *orthogonal Procrustes problem*. If $B = U\Sigma V^T$ is a full SVD, then by invariance of the Frobenius norm under orthogonal transformations,

$$\|B - Q\|_F^2 = \|\Sigma - U^TQV\|_F^2 = \|\Sigma - \tilde{Q}\|_F^2.$$

Therefore, we reduce to the problem of finding a matrix $\tilde{Q}$ with orthonormal columns that is as close as possible to a diagonal matrix with positive diagonal entries. Considering just the first column

$$\|w - \sigma_1 e_1\|^2 = (x - \sigma_1)^2 + \|y\|^2, \quad w = \begin{bmatrix} x \\ y \end{bmatrix}$$

and expanding, we have

$$\|w - \sigma_1 e_1^2\| = x^2 - 2\sigma_1 x + \sigma_1^2 + \|y\|^2 = 1 + \sigma_1^2 - 2\sigma_1 x$$

which is minimal when x is as large as possible ($x = 1$). A similar argument shows that the closest unit length vector to column k of Σ will be e_k. Therefore, even if we only insisted that each column was unit length (rather than

insisting on orthonormality), the closest matrix to Σ would be the identity matrix. Hence $\tilde{Q} = I_k$ consists of the leading k columns of the identity, and $Q = UI_kV^T$ is the same as $Q = U_1V^T$ where U_1 is the $m \times n$ submatrix of U from the economy SVD.

Note that if $A = U\Sigma V^T$ is an economy SVD and $Q = UV^T$, we have

$$A = QH, \quad H = V\Sigma V^T.$$

The matrix H is symmetric and positive semi-definite, while Q has orthonormal columns. This decomposition is known as the *polar decomposition* of A, and it generalizes the polar decomposition of a vector into a (unit length) direction times a non-negative length.

2.3 Other matrix nearness problems

We have far from exhausted the possible matrix nearness problems with these two examples. Perhaps the obvious next one to cover, had we not run out of time, would be the Eckart-Young theorem: the nearest rank k matrix to a given matrix A (in either the Frobenius or operator 2-norm) is

$$L = A^k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

Another entertaining example is the distance to instability — that is, given a matrix A whose eigenvalues all have negative real part, what is the nearest matrix with purely imaginary eigenvalues? The Eckart-Young theorem is in any of the recommended texts; and Nick Higham has a classic paper on matrix nearness problems with these and several others.