Penalized *N*-space Fitting

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Penalized optimization problem

In abstraction of the smoothing spline development, suppose that $\{\mathbf{u}_j\}$ is a set of $M \leq N$ orthonormal N-vectors, $\lambda \geq 0, \eta_j \geq 0$ for $j = 1, 2, \ldots, M$, and consider the optimization problem

$$\underset{\mathbf{v} \in \operatorname{span}\{\mathbf{u}_j\}}{\text{minimize}} \left((\mathbf{Y} - \mathbf{v})' (\mathbf{Y} - \mathbf{v}) + \lambda \sum_{j=1}^{M} \eta_j \langle \mathbf{u}_j, \mathbf{v} \rangle^2 \right)$$

For $\mathbf{v} = \sum_{j=1}^{M} c_j \mathbf{u}_j \in \operatorname{span}\{\mathbf{u}_j\}$, the penalty is

$$\lambda \sum_{j=1}^{M} \eta_j \langle \mathbf{u}_j, \mathbf{v} \rangle^2 = \lambda \sum_{j=1}^{M} \eta_j c_j^2$$

and in this penalty, $\lambda \eta_j$ is a multiplier of the squared length of the component of $\mathbf{v}$ in the direction of $\mathbf{u}_j$.

Solution to the optimization problem

The optimization criterion is thus

$$(\mathbf{Y}-\mathbf{v})'(\mathbf{Y}-\mathbf{v}) + \lambda \sum_{j=1}^{M} \eta_j \langle \mathbf{u}_j, \mathbf{v} \rangle^2 = \sum_{j=1}^{M} (\langle \mathbf{u}_j, \mathbf{Y} \rangle - c_j)^2 + \lambda \sum_{j=1}^{M} \eta_j c_j^2$$

and it is then easy to see (via simple calculus) that

$$c_j^{\text{opt}} = \frac{\langle \mathbf{u}_j, \mathbf{Y} \rangle}{1 + \lambda \eta_j}$$

i.e.

$$\hat{\mathbf{Y}} = \mathbf{v}^{\text{opt}} = \sum_{j=1}^{M} \frac{\langle \mathbf{u}_j, \mathbf{Y} \rangle}{1 + \lambda \eta_j} \mathbf{u}_j$$

From this it's clear how the penalty structure dictates optimally shrinking the components of the projection of $\mathbf{Y}$ onto $\text{span}\{\mathbf{u}_j\}$.

“Smoother” matrix and application

$\hat{\mathbf{Y}}$ can be represented in the form $\mathbf{SY}$ for

$$\mathbf{S} = \sum_{j=1}^{M} d_j \mathbf{u}_j \mathbf{u}_j' = \mathbf{U}\mathbf{diag}\left(\frac{1}{1+\lambda\eta_1}, \ldots, \frac{1}{1+\lambda\eta_M}\right)\mathbf{U}'$$

with $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M)$. (It's easy to see that $\mathbf{S}$ is a rank M matrix for which $\hat{\mathbf{Y}} = \mathbf{SY}$.)

One context in which this material might find immediate application is where some set of basis functions $\{h_j\}$ are increasingly "wiggly" with increasing j and the vectors $\mathbf{u}_j$ come from applying the Gram-Schmidt process to the vectors

$$\mathbf{h}_j = (h_j(\mathbf{x}_1), \ldots, h_j(\mathbf{x}_N))'$$

In this context, it would be very natural to penalize the later $\mathbf{u}_j$ more severely than the early ones.