Structured Regression Functions

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Structure assumptions and smoothing

There are several ways that have been suggested for making use of fairly low-dimensional (and thus, potentially effective) smoothing in large p problems. One of them is the "structured kernels" idea. Another concerns using various kinds of structured regression functions.

For example, one might assume **additivity** in a form

$$f(\mathbf{x}) = \alpha + \sum_{j=1}^{p} g_j(x_j) \quad (1)$$

and try to do fitting of the p functions g_j and constant α.

Additive models and "backfitting"

One more or less *ad hoc* method of fitting additive forms is the so-called "**back-fitting algorithm**." That is to (generalize form (1) slightly and) fit (under SEL)

$$f(\mathbf{x}) = \alpha + \sum_{l=1}^{L} g_l\left(\mathbf{x}^l\right) \quad (2)$$

for $\mathbf{x}^l$ some part of $\mathbf{x}$, one might set $\widehat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i$, and then cycle through $l = 1, 2, \ldots, L, 1, 2, \ldots$ by:

Backfitting algorithm

1. fitting via some appropriate (often linear) operation (e.g., spline or kernel smoothing)

$$g_l\left(\mathbf{x}^l\right) \quad \text{to} \quad \text{"data"} \quad \left\{\left(\mathbf{x}_i^l, y_i^l\right)\right\}_{i=1,2,\ldots,N}$$

for

$$y_i^l = y_i - \left(\widehat{\alpha} + \sum_{m \neq l} g_m\left(\mathbf{x}_i^m\right)\right)$$

where the g_m are the current versions of the fitted summands,

Backfitting algorithm cont.

2. setting

$$g_l = \text{the newly fitted version}$$
$$- \text{the sample mean of this newly fitted version across all } \mathbf{x}_i^l$$

(in theory this is not necessary, but it is here to prevent numerical/ round-off errors from causing the g_m to drift up and down by additive constants summing to 0 across m), and

3. iterating until convergence to, say, $\hat{f}(\mathbf{x})$.

Simultaneous fitting of additive terms

A more principled SEL fitting methodology for additive forms like that in display (2) (e.g. implemented by Wood in his `mgcf R` package) is the simultaneous fitting of α and all the functions g_l via penalized least squares. That is, using an appropriate set of basis functions for smooth functions of $\mathbf{x}^l$ (often a tensor product basis in the event that the dimension of $\mathbf{x}^l$ is more than 1) each g_l might be represented as a linear combination of those basis functions. Then form (2) is in fact a constant plus a linear combination of basis functions. So upon adopting a quadratic penalty for the coefficients, one has a kind of ridge regression problem and explicit forms for all fitted coefficients and $\widehat{\alpha}$. The practical details of making the various bases and picking ridge parameters, etc. are not trivial, but the basic idea is clear.

Scope of additive modeling

The simplest version of this line of development, based on form (1), might be termed fitting of a "main effects model." But the approach might as well be applied to fit a "main effects and two factor interactions model," using some g_ls that are functions of only one coordinate of $\mathbf{x}$ and others that depend upon only two coordinates of the input vector. One may mix types of predictors (continuous, categorical) and types of functions of them in the additive form to produce all sorts of interesting models (including semi-parametric ones and ones with low order interactions).

Other structured regression forms

Another possibility for introducing structure assumptions and making use of low-dimensional smoothing in a large p situation, is by making strong global assumptions on the forms of the influence of some input variables on the output, *but allowing parameters of those forms to vary in a flexible fashion with the values of some small number of coordinates of* $\mathbf{x}$. For sake of example, suppose that $p = 4$. One might consider predictor forms

$$f(\mathbf{x}) = \alpha(x_3, x_4) + \beta_1(x_3, x_4)\, x_1 + \beta_2(x_3, x_4)\, x_2$$

That is, one might assume that for fixed (x_3, x_4), the form of the predictor is linear in (x_1, x_2), but that the coefficients of that form may change in a flexible way with (x_3, x_4). Fitting might then be approached by locally weighted least squares, *with only* (x_3, x_4) *involved in the setting of the weights.*

Other structured regression forms cont.

That is, one might for each (x_{30}, x_{40}), let

$$\hat{y}_i = \alpha(x_{30}, x_{40}) + \beta_1(x_{30}, x_{40})\, x_{1i} + \beta_2(x_{30}, x_{40})\, x_{2i}$$

minimize over choices of $\alpha(x_{30}, x_{40})$, $\beta_1(x_{30}, x_{40})$ and $\beta_2(x_{30}, x_{40})$ the weighted sum of squares

$$\sum_{i=1}^{N} K_\lambda\left((x_{30}, x_{40}), (x_{3i}, x_{4i})\right)(y_i - \hat{y}_i)^2$$

and then employ the predictor

$$\hat{f}(\mathbf{x}_0) = \hat{\alpha}(x_{30}, x_{40}) + \hat{\beta}_1(x_{30}, x_{40})\, x_{10} + \hat{\beta}_2(x_{30}, x_{40})\, x_{20}$$

This device keeps the dimension of the space where one is doing smoothing small. But note that nothing here does any thresholding or automatic variable selection.