University of Illinois | Department of Economics
Spring 2017 | Econ 574 | Roger Koenker

Lecture 17
"Smoothing Things Over: Some Notes on Cross-Validated Smoothing Splines"

1. In which we narrow the topic

   We will consider the simplest Gaussian non-parametric regression problem, i.e.,

   $$y_i = g(x_i) + u_i \qquad i = 1, \ldots, n \tag{1.1}$$

   where $\{x_i\}$ is a know sequence of scalars in [0, 1] g is unknown, but pleasantly smooth function, and $\{u_i\}$ are independent and identically distributed Gaussian random variables. Our task is to produce a credible estimate of g on [0, 1].

   A nice approach to this problem is offered by Reinsch (1967) who suggests solving

   $$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_0^1 (g''(x))^2 dx \tag{1.2}$$

   where $G = \{g | g, g' \text{ absolutely continuous, and } g'' \in L_2[0,1]\}$. (This is a Sobelev space.) Evidently, this approach balances a desire for the estimate to exhibit some fidelity-to-the-data represented by the first term, while achieving a reasonable degree of smoothness expressed by the second term. The parameter λ controls the significance attached to the roughness penalty. As $\lambda \to \infty, g$ is required to approach linearity and thus the solution approaches the least squares solution to a bivariate linear model.

2. In which we review what you may already know

   The solution to the problem posed in (1.2) is a cubic smoothing spline. See Reinsch (1967) or DeBoor (1978) for discussions in the "How to" spirit of the present exposition. See Wahba (1978) for a beautiful treatment from a somewhat Bayesian (Reproducing Kernel Hilbert Space) point of view.

   A cubic spline is a piecewise cubic polynomial: a function with continuous first and second derivatives, whose third derivative may take discrete jumps at designated points, called knots, or breakpoints. Since the function is cubic in each subinterval all higher order derivatives (than the third) vanish. Why? Because the Euler condition for the solution to (1.2) says,

   $$(y_i - g(x_i))\delta_{x_i}(x) + \frac{d^2}{dx^2} g''(x) = 0 \tag{2.1}$$

   which implies that $\psi''''(x) = 0$ for almost all $x \in [0, 1]$. Thus a solution takes the form

   $$\hat{g}(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

   for $x \in [x_i, x_{i+1})$ and $i = 1, \ldots n - 1$. The following counting exercise may serve to clarify the essentially finite dimensional nature of the estimate $\hat{g}$. We have $4(n - 1)$ unknown

coefficients in the previous expression, and if we wish to extend $\hat{g}$ beyond $x_{(1)}$ and $x_{(n)}$ we have another 4 parameters, say, a_0, b_0, a_n, b_n. Note that the c's and d's in these outer regions are zero; were they not, the roughness penalty could be reduced without disturbing the value of the first term in the objective function. Now at each of the design points, x_i, we have 3 linear continuity restrictions on the coefficients, so we are left with a problem in n parameters.

Following Reinsch and DeBoor we may express the objective function as

$$(y-a)'(y-a) + \lambda c' R c \tag{2.3}$$

where $a = (a_i, \ldots, a_n), c = (c_2, \ldots, c_{n-1})$, and R is a $(n-2)^2$ tridiagonal matrix with entries $r_{ii} = 2(h_i + h_{i+1})/3, r_{i,i+1} = r_{i+1,i} = h_{i+1}/3$, and $h_i = x_{i+1} - x_i$. We will assume that the observations are ordered so that $0 \leq x_1 < x_2 < \ldots < x_n \leq 1$. The continuity restrictions imply that

$$Rc = Q'a \tag{2.4}$$

where Q is a $n \times (n-2)$ tridiagonal matrix with entries $q_{i,i+1} = 1/h_{i+1}, q_{i+1,i} = 1/h_{i+1}$, and $q_{ii} = -(1/h_i + 1/h_{i+1})$. Thus we may write (2.3) as,

$$(y-a)'(y-a) + \lambda a' Q R^{-1} Q' a \tag{2.5}$$

This is obviously a garden-variety, finite-dimensional, quadratic optimization problem with solution,

$$a = (I + \lambda Q R^{-1} Q')^{-1} y \tag{2.6}$$

Premultiplying $(I + \lambda Q R^{-1} Q')a = y$ by Q' and using (2.4) gives,

$$c = (R + \lambda Q'Q)^{-1} Q' y$$

and we may write,

$$\begin{aligned} a &= y + \lambda Q c \\ &= (I + \lambda Q (R + \lambda Q'Q)^{-1} Q') y \\ &= A(\lambda) y. \end{aligned} \tag{2.7}$$

Note that (2.7) is preferable to (2.6) since linear system to be solved has a simpler banded structure.

Given the a's it is a simple matter to compute the b's, c's, and d's. In R everything is handled nicely with the function `smooth.spline`.

The question remains how should we choose λ? Ideally we would minimize,

$$R(\lambda) = n^{-1} E \parallel g - \hat{g} \parallel^2 = n^{-1} \parallel (I - A) g \parallel^2 + \frac{\sigma^2}{n} Tr(A)^2 \tag{2.8}$$

but this is unrealistic since g is unknown. A possible approach if σ^2 is known would be to minimize,

$$\begin{aligned} \hat{R}(\lambda) &= n^{-1} \parallel (I - A(\lambda)) y \parallel^2 - \frac{\sigma^2}{n} [Tr(I-A)^2 - TrA^2] \\ &= n^{-1} \parallel (I - A) y \parallel^2 + \frac{2\sigma^2}{n} TrA - \sigma^2 \end{aligned} \tag{2.9}$$

This yields a version of Mallow's C^p criterion. A natural alternative for unknown σ^2 is to minimize the cross-validation criterion,

$$V(\lambda) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{g}^{(i)}(x_i))^2 \tag{2.10}$$

where $\hat{g}^{(i)}$ is the estimate of g omitting the i^{th} observation. It is well known, see for example Efron (1982) that,

$$V(\lambda) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{g}(x_i))^2 / (1 - A_{ii}(\lambda))^2. \tag{2.11}$$

Craven and Wahba (1978) suggest that instead one should minimize

$$V(\lambda) = \frac{n^{-1} \parallel (I - A)y \parallel^2}{(n^{-1} Tr(I - A))^2} \tag{2.12}$$

This is a variant of the cross-validation approach in which the quantities in the denominator of (2.11) are replaced by their sample average. They call it generalized cross-validation, or GCV.

3. In which we interpret GCV and suggest a modification

For convenience we will work with the reciprocal of λ, say μ, so (2.6) becomes,

$$a = A(\mu)y = (I - Q(\mu R + Q'Q)^{-1}Q')y. \tag{3.1}$$

Modifying $V(\lambda)$ accordingly, and taking the logarithm of its square-root we have,

$$\nu(\mu) = \log(\hat{\sigma}(\mu)) - \log(1 - n^{-1} TrA(\mu)) \tag{3.2}$$

where $\hat{\sigma}^2 = n^{-1} \parallel (I - A(\mu))y \parallel^2$. Now $A(\mu)$ plays the role of the "hat" matrix in the theory of linear regression. In the present case $TrA(\mu)$ varies from 2, when $\mu = 0$ to n as $\mu \to \infty$. Thus $k(\mu) = TrA(\mu)$ may be interpreted, loosely, as the "effective dimensionality" of the estimate $\hat{g}$. As long as this is small relative to the sample size, n, we have the approximation,

$$\nu(\mu) \simeq \log(\hat{\sigma}(\mu)) + n^{-1} k(\mu) \tag{3.3}$$

which is simply an "interpretation" of Akaike's (1974) information criterion for the present context. A similar point was made by Stone (1977), and Terasvirta (1985) has also emphasized the close relationship between smoothing and model selection. This interpretation immediately suggests several modifications of the dimensionality penalty in accordance with the proposals of Schwartz (1978) and others.

To motivate such modifications, I would like to reconsider an example from Craven and Wahba (1979). We have exactly the model of (1.1) with g chosen to be a mixture of beta density functions illustrated in Figure 1 by the solid line, and given by,

$$g(x) = \sum_{i=1}^{n} w_i \beta_{p_j, q_j}(x)$$

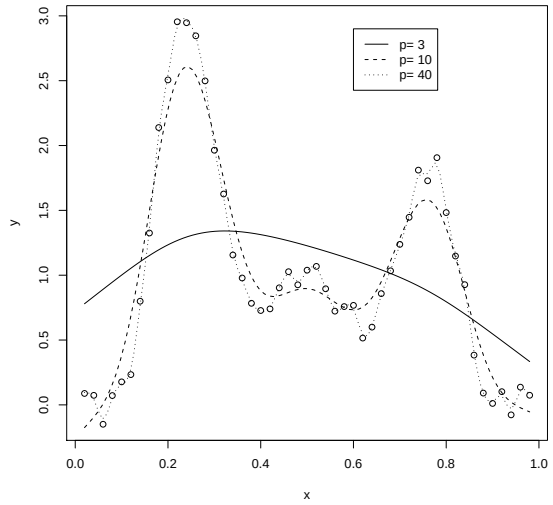

Figure 1: The Cubic Smoothing Spline: The figure illustrates an example from Craven and Wahba (1979) fitting $n = 49$ observations. Three fitted curves are superimposed on the observed points corresponding to effective dimension of the model: 3, 10 and 30.

with $w = (.5, .2, .3), p = (10, 20, 30), q = (30, 20, 10), u \sim \mathcal{N}(0, \sigma^2), \sigma^2 = .01$, and $n = 50$. The design points are equally spaced on $[0, 1]$. The y observations plotted in Figure 1 were generated by adding normal noise with $\sigma = .1$ to $g(x_i)$, as in Figure 2, Example III of Craven and Wahba.

In Figure 2 we offer three distinct replications of this example of Craven and Wahba. We plot three functions for each replication, (i.) the true mean-squared-error function, (ii.) $\nu(\mu)$ as defined in (3.2) above, and (iii.) $\nu^*(\mu)$ which is identical to $\nu(\mu)$ except that we have multiplied the dimensionality penalty by 1.5. These curves are designated: mse, gcv, bic respectively. The horizontal axis represents the effective dimension of the fitted function, Trace$(A(\mu)$, rather than μ itself. This number is somewhat easier to interpret since it must vary from 2 to n. There are several possible rationalizations for the factor 1.5. My initial view was simply to replace the first term mle of σ^2 by the corresponding unbiased estimator, however, a more cogent argument may be that 1.5 happens to fall between the numbers $1/2 \log(50) \simeq 1.95$ and $\log \log 50 \simeq 1.36$ suggested by Schwarz (1987) and Hannan and Quinn (1979) respectively for the conventional model selection problem. In Figure 2 we can see that the true mse function has a minimum at about dimension 20 in all three realizations. The bic curves are also minimized at roughly the same value. In the third panel of the figure the gcv criteria also has a distinct minimum at about dimension 20. However, in the middle panel we discern only a slight upward tendency in the curve beyond dimension 20. In the first panel the gcv curve declines over the entire range and this produces an estimate of the fitted function that is "as rough as possible" interpolating every point. This seems to be characteristic of the gcv criteria – in a small proportion of cases it fails quite dramatically, yielding a λ that is zero. This suggests that alternative

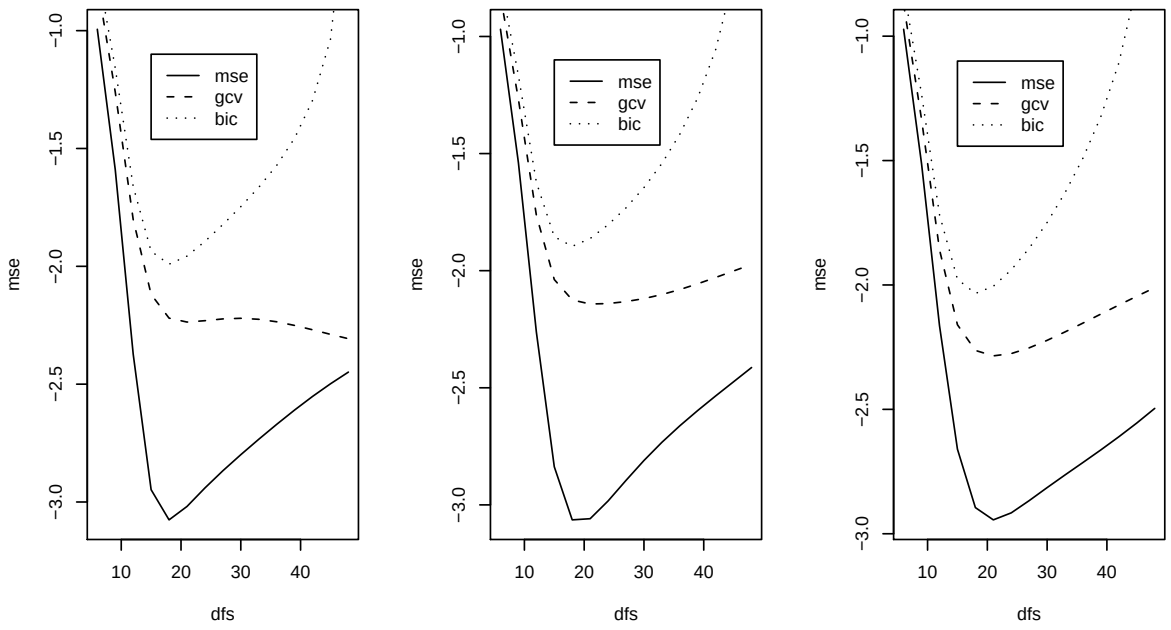

Figure 2: λ-Selection for Smoothing Splines

criteria that assign more weight to the penalty may be preferable. Note that in these examples at least the modified gcv criteria performs quite reasonably.

Finally, I would like to emphasize that the three figures appearing above constitute a very small sample and considerable further work would be needed to draw any very convincing conclusions. Repeating this exercise one sees that GCV fails in around five percent of cases, this is not terrible, but it isn't exactly encouraging either. An excellent recent reference on related smoothing problems is Wood (2006).

Since these notes were originally written the appearance of Hastie and Tibshirani (1990) and Green and Silverman (1994) have rendered them nearly superfluous.

GCV and SURE

Li (1985) established a nice connection betweenWahba's GCV and Stein's unbiased risk estimator (SURE) which we will now briefly describe. Consider a general model

$$y_i = \mu_i + \epsilon_i$$

with associated linear estimator

$$\hat{\mu}(\lambda) = A(\lambda)y$$

encompassing many situations including classical least squares regression and spline smoothing.

The GCV criterion is

$$\mathrm{GCV}(\lambda) = \frac{n^{-1} \parallel y - \hat{\mu}(\lambda) \parallel^2}{(1 - n^{-1} \text{ Tr } A(\lambda))^2}$$

In regression, we can interpret λ as controlling the dimension of the parameter β, say p, and then $\text{Tr } A(\lambda) = p$, so we have the criteria

$$\mathrm{GCV}(\lambda) = \frac{\hat{\sigma}^2(p)}{(1 - p/n)^2}$$

As we have already noted, if we take logs and approximate the denominator we obtain

$$\log \sqrt{\mathrm{GCV}(\lambda)} \approx \log \hat{\sigma} - p/n$$

which is the AIC criterion.

Stein (1981) showed that for Gaussian ϵ_i and squared error loss on μ, the estimator

$$\tilde{\mu}(\lambda) = y - \left[\frac{\sigma^2}{y'B(\lambda)y}\right] M(\lambda)y$$

dominates y and $\hat{\mu}$. Here $M(\lambda) = I - A(\lambda)$ and

$$B(\lambda) = (\text{Tr } M(\lambda)I - 2M(\lambda))^{-1} M(\lambda)^2$$

The factor in square brackets is called the "shrinkage factor", since if it were 1, $\tilde{\mu}$ would reduce to $\hat{\mu}$, whereas when it is less than 1 we can interpret $\tilde{\mu}$ as a compromise between $\hat{\mu}$ and y,

$$\tilde{u} = (1 - v)y + v\hat{\mu}.$$

where v denotes the scalar shrinkage factor.

Stein showed that

$$\mathrm{SURE}(\lambda) = \sigma^2 - \frac{\sigma^4 \parallel M(\lambda)y \parallel^2}{n(y'B(\lambda)y)^2}$$

is an unbiased estimate for the risk of $\tilde{\mu}(\lambda)$, that is

$$E \text{ SURE}(\lambda) = En^{-1} \parallel \mu - \tilde{\mu}(\lambda) \parallel^2$$

for any $\mu \in \mathcal{R}^n$, so to select a good λ it is natural to minimize SURE. This is obviously equivalent to minimizing

$$S(\lambda) = \frac{n(y'B(\lambda)y)^2}{\parallel M(\lambda)y \parallel^2}$$

Now, in the same spirit as the AIC approximation, for n sufficiently large and $M(\lambda)$ balanced so its largest eigenvalue is small compared to its trace, we can approximate $B(\lambda)$ by $(\text{Tr } M(\lambda))^{-1} M^2(\lambda)$ and then

$$
\begin{aligned}
S(\lambda) &\approx \frac{n(\text{Tr } M(\lambda))^{-2} \parallel M(\lambda)y \parallel^4}{\parallel M(\lambda)y \parallel^2} \\
&= \frac{\parallel M(\lambda)y \parallel^2}{n^{-1}(\text{Tr } M(\lambda))^2} \\
&= \text{GCV } (\lambda).
\end{aligned}
$$

References

Akaike, H. (1974). A new look at the statistical identification model, *IEEE Transactions in Automatic Control*, 19, 716-25.

Becker, R.A., and J.M. Chamers (1984). *S: An Interactive Environment for Data Analysis and Graphics*, Belmont, CA: Wadsworth.

Cox, D.D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function, *Annals of the Institute of Statistical Mathematics*, 37, 271-88.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions, *Numerische Mathematik*, 31, 377-403.

DeBoor, C. (1978). *A Practical Guide to Splines*, New York: Springer-Verlag.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: SIAM.

Green, P.J. and B.W. Silverman, (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman-Hall.

Hannan, E.J. and B.G. Quinn (1979). The determination of the order of autoregression, *Journal of the Royal Statistical Society*, 41, 190-196.

Hastie, T. and R. Tibshirani, (1990). *Generalized Additive Models*, Chapman-Hall.

Li, Ker-Chau (1985) From Stein's Unbiased risk estimates to the method of GCV, *Annals of Statistics*, 13, 1352-1377.

Reinsch, C.M. (1967). Smoothing by Spline Functions, *Numerische Mathematik*, 10, 177-83.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461-64.

Stein, C. (1981) Estimation of the meanof a multivariate normal distribution, *Annals of Statistics*, 9, 1135-55.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society*, 39, 44-47.

Terasvirta, T. (1985). Smoothness in regression: Asymptotic considerations, UC-San Diego Technical Report.

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression, *Journal of the Royal Statistical Society (B)*, 40, 364-72.

Wood, S. (2006) *Generalized Additive Models: An Introduction with R*, Chapman and Hall.