On the Optimal Amount of Smoothing in Penalised Spline Regression
Author(s): M. P. Wand
Source: *Biometrika*, Vol. 86, No. 4 (Dec., 1999), pp. 936-940
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2673597
Accessed: 19-02-2016 04:26 UTC

Biometrika (1999), **86**, 4, *pp.* 936–940

Printed in Great Britain

On the optimal amount of smoothing in penalised spline regression

BY M. P. WAND

Department of Biostatistics, School of Public Health, Harvard University, 665 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

mwand@hsph.harvard.edu

SUMMARY

The optimal amount of smoothing in penalised spline regression is investigated. In particular, a simple closed form approximation to the optimal smoothing parameter is derived. Comparisons with its exact counterpart show it to be a useful starting point for measuring the optimal amount of smoothing in penalised spline regression. It also lends itself to the development of quick and simple rules for automatic smoothing parameter selection.

Some key words: Asymptotic approximation; Automatic smoothing parameter selection; Nonparametric regression; Quick and simple smoothing parameter selection; Regression spline.

1. INTRODUCTION

Figure 1 shows a scatterplot smooth of data from an air quality monitoring study using the Light Detection and Ranging (LIDAR) technique. The vertical variable is a measure of cumulative particle concentration and the horizontal variable is the range of the measuring device. The full details of the data are described in Holst et al. (1996). The fitted curve is based on the penalised spline approach to smoothing which has received considerable recent attention because of its simplicity and effectiveness at tackling a wide range of semiparametric regression problems; see for example Eilers & Marx (1996) and recent unpublished work by D. Ruppert and R. J. Carroll.

The amount of smoothing in Fig. 1 was chosen not by eye, but from the data with only a minimal amount of user interaction. In particular, the choice involved nothing more than simple direct computations. No numerical minimisation or root-finding was required. This is in the spirit of quick and simple rules for selection, for example, of the histogram bin width (Scott, 1979) and the bandwidth of a kernel regression smooth (Härdle & Marron, 1995). Such rules rely on closed form approximations to the 'optimal' smoothing parameter. The purpose of this note is to show how such an approximation can be derived for penalised spline smoothing.

Section 2 presents some theory aimed at obtaining a closed-form approximation to a theoretically optimal smoothing parameter. The result is evaluated in § 3 and then applied to derive a quick and simple rule for choosing the smoothing parameter from the data.

2. THEORETICAL RESULTS

2·1. *Basic formulation*

Consider the nonparametric regression set-up

$$Y_i = m(x_i) + \varepsilon_i,$$

where (x_i, Y_i) is a set of regression data, m is the regression mean function and ε_i is the error variable for observation i. For now we will assume that the errors are uncorrelated and have

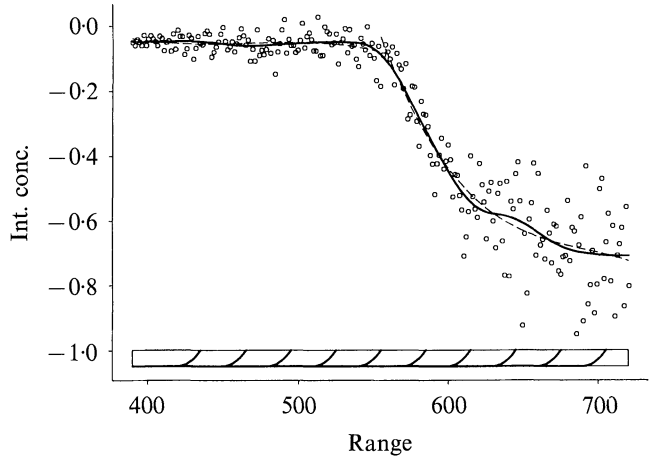

Fig. 1. Quadratic penalised spline smooth of the LIDAR data (solid curve). The dashed curves correspond to a rough initial estimate used for smoothing parameter selection described in § 3·2. The regression spline basis functions are shown at the base of the plot.

constant variance equal to σ^2. We can write this model in matrix notation as

$$Y = m + \varepsilon, \quad \mathrm{cov}(\varepsilon) = \sigma^2 I,$$

where Y denotes the vector of responses, m is the vector of means at the data and ε is the vector of errors.

A degree-p penalised spline estimate of m is

$$\hat{m}_\lambda = X(X^{\mathrm{T}}X + \lambda^{2p}D)^{-1}X^{\mathrm{T}}Y, \tag{1}$$

where

$$X = \begin{bmatrix} 1 & x_1 \cdots x_1^p & (x_1-\kappa_1)_+^p \cdots (x_1-\kappa_K)_+^p \\ \vdots & \vdots \ddots \vdots & \vdots \qquad \ddots \qquad \vdots \\ 1 & x_n \cdots x_n^p & (x_n-\kappa_1)_+^p \cdots (x_n-\kappa_K)_+^p \end{bmatrix}, \quad D = \mathrm{diag}\{0_{(p+1)\times 1}, 1_{K\times 1}\}.$$

Here $\kappa_1, \ldots, \kappa_K$ is a set of knots. In penalised spline regression, these are normally chosen to be quite 'dense' in the interval over which the x_i's range, as depicted in Fig. 1, so that the mean function can adequately be resolved. The amount of smoothing is controlled by the parameter $\lambda > 0$. The choice $\lambda = 0$ leads to an ordinary least squares fit with design matrix X, which tends to overfit the data. As λ becomes very large, $\hat{m}_\lambda$ approaches the pth degree polynomial fit. A value between these two extremes is usually desirable. The power of $2p$ on the λ ensures that λ acts as a scale parameter in that any linear transformation of the design variable should be accompanied by the same linear transformation of λ to preserve the value of $\hat{m}_\lambda$.

Note that the penalty imposed by the matrix D is one of many possible penalties. This is a very simple one, based on constraining the sum of squares of the knot coefficients and advocated by Eilers & Marx (1996) and in recent unpublished work by D. Ruppert and R. J. Carroll. The most common alternative is the smoothing spline penalty, which is related to the integrated squared derivative measure of roughness, e.g. Green & Silverman (1994, Ch. 2).

A mathematically convenient measure of the global discrepancy between $\hat{m}_\lambda$ and m is the mean average squared error,

$$\mathrm{MASE}(\hat{m}_\lambda) = E\left[\frac{1}{n}\sum_{i=1}^{n}\{\hat{m}(x_i) - m(x_i)\}^2\right].$$

It is well known that $\text{MASE}(\hat{m}_\lambda)$ has the decomposition

$$\frac{1}{n}\sum_{i=1}^{n}\text{var}\{\hat{m}_\lambda(x_i)\}+\frac{1}{n}\sum_{i=1}^{n}\{E\hat{m}_\lambda(x_i)-m(x_i)\}^2,$$

the first term representing the average variance and the second representing the average squared bias.

2·2. *Exact mean average squared error*

A useful notation is $H_\lambda = X(X^{\mathrm{T}}X+\lambda^{2p}D)^{-1}X^{\mathrm{T}}$, so that $\hat{m}_\lambda = H_\lambda Y$. This shows that $\hat{m}_\lambda$ is a linear smoother with smoother matrix equal to H_λ. If we use the notation $\|v\| = \sqrt{(v^{\mathrm{T}}v)}$, then

$$\text{MASE}(\hat{m}_\lambda) = \frac{\sigma^2}{n}\text{tr}(H_\lambda^2)+\frac{1}{n}\|(H_\lambda - I)m\|^2, \tag{2}$$

the terms representing, respectively, the average variance and average squared bias components. The theoretical optimal smoothing parameter is then

$$\lambda_{\text{MASE}} = \underset{\lambda>0}{\text{argmin}}\ \text{MASE}(\hat{m}_\lambda) \tag{3}$$

and can be found using (2) and numerical minimisation.

2·3. *Asymptotic approximation*

There are several reasons for wanting a closed-form asymptotic approximation to λ_{MASE}. The first is to provide something that is quick and simple to compute. If however λ_{MASE} is still required then an asymptotic approximation will usually provide a good starting value for the numerical minimisation problem. Finally, it can be used to develop data-driven rules for selection of the amount of smoothing, as described in § 3. Such rules can also be used to find a good starting value when computing traditional smoothing parameter selectors such as generalised crossvalidation (Craven & Wahba, 1979).

In the Appendix it is shown that the first few terms in the asymptotic expansion of $\text{MASE}(\hat{m}_\lambda)$ as $\lambda \to 0$ are

$$\begin{aligned}\text{AMASE}(\hat{m}_\lambda) = {}& \frac{\sigma^2}{n}((p+K+1)-2\lambda^{2p}\,\text{tr}\{(X^{\mathrm{T}}X)^{-1}D\}+\lambda^{4p}\,\text{tr}[\{(X^{\mathrm{T}}X)^{-1}D\}^2])\\ &+\frac{\lambda^{4p}}{n}\|X(X^{\mathrm{T}}X)^{-1}D(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}m\|^2,\end{aligned} \tag{4}$$

where AMASE stands for asymptotic MASE. The minimiser of this quantity is

$$\lambda_{\text{AMASE}} = \left(\frac{\sigma^2\,\text{tr}\{(X^{\mathrm{T}}X)^{-1}D\}}{\|X(X^{\mathrm{T}}X)^{-1}D(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}m\|^2+\sigma^2\,\text{tr}[\{(X^{\mathrm{T}}X)^{-1}D\}^2]}\right)^{1/(2p)}. \tag{5}$$

For a given m and σ^2, (5) provides an easy-to-compute approximation to the MASE-optimal smoothing parameter.

2·4. *Other bases*

The penalised spline (1) is defined with respect to the truncated polynomial basis for the space of piecewise pth degree polynomials over the knots $\kappa_1, \ldots, \kappa_K$, and represented through the matrix X. It is possible to redefine $\hat{m}_\lambda$ in terms of other bases such as the B-spline basis (Eilers & Marx, 1996) and the Demmler-Reinsch basis (Nychka & Cummins, 1996). The asymptotic approximations are easily adjusted to handle these alternative bases. All that is required is the introduction of the $(p+1+K)\times(p+1+K)$ matrix L that maps X to the corresponding X-matrix for the new basis, $X_{\text{new}} = XL$, and the substitution $X = X_{\text{new}}L^{-1}$.

2·5. *Further approximation*

The terms in (4) could be approximated further by studying the asymptotic behaviour of X as $n \to \infty$. This would presumably result in expressions that show the effect of σ^2, m and the distribution of the design variables more explicitly, in the spirit of asymptotic results for kernel smoothing, e.g. Tsybakov (1986). Preliminary investigations along these lines suggest that the first multiplier of λ^{4p} in (4) is of lower order than the second multiplier of λ^{4p} as $n \to \infty$. Therefore, from an asymptotic point of view, the first of these terms could be dropped and the one-term approximation to λ_{MASE},

$$\lambda_{\text{AMASE},1} = \left[\frac{\sigma^2 \operatorname{tr}\{(X^{\mathrm{T}} X)^{-1} D\}}{\| X(X^{\mathrm{T}} X)^{-1} D (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} m \|^2} \right]^{1/(2p)}$$

would result. While it appears that $\lambda_{\text{AMASE},1}$ is asymptotically equivalent to λ_{AMASE}, there is a distinct finite sample difference. Comparisons between the two, described in § 3·1, indicate that the two-term approximation is superior in practice.

Further asymptotic analysis along these lines may aid the interpretation of the effect of λ on the performance of m_λ. However, it is not clear that it will have any practical benefit, beyond that provided by (4).

2·6. *Heteroscedastic errors*

If the errors are heteroscedastic, or even correlated, then we have $\operatorname{cov}(\varepsilon) = V$ for some symmetric positive definite matrix V. The mathematics used to produce (4) and (5) is easily generalised to cater for this extension. In particular (4) becomes

$$\lambda_{\text{AMASE}} = \left(\frac{\operatorname{tr}\{(X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} V X (X^{\mathrm{T}} X)^{-1} D\}}{\| X(X^{\mathrm{T}} X)^{-1} D (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} m \|^2 + \operatorname{tr}[(X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} V X \{(X^{\mathrm{T}} X)^{-1} D\}^2]} \right)^{1/(2p)}$$

3. Practical implications

3·1. *Accuracy of* λ_{AMASE}

To assess the accuracy of (5) we computed it for each of the 18 homoscedastic regression settings used in Wand (1999). A reasonably meaningful measure of the quality of λ_{AMASE} is

$$\text{MASE}(m)_{\lambda_{\text{AMASE}}} \Big/ \inf_{\lambda > 0} \text{MASE}(m_\lambda).$$

The average value of this ratio for these examples is 1·31 with a standard deviation of 0·218, indicating that λ_{AMASE} provides a reasonable approximation to the optimal amount of smoothing. For the one-term approximation the average ratio is 1·41 with a standard deviation of 0·274 so it therefore seems worthwhile to use the two-term expression (5).

3·2. *A quick and simple smoothing parameter selector*

A quick and simple rule for choosing λ can be obtained by replacing m and σ^2 in (5) by $\hat{m}_{\text{init}}$ and $\hat{\sigma}^2_{\text{init}}$, representing rough but reasonable initial estimates. As suggesting by Härdle & Marron (1995), an effective means of obtaining such estimates is to divide the data into blocks and fit low-degree polynomials to each block. Obtaining a reasonable initial estimate is crucial and a small amount of user intervention, via graphical inspection, is desirable to ensure this. The dashed lines in Fig. 1 show an initial estimate of m biased on fitting cubic polynomials to the two blocks of data to the left and right of their horizontal midpoint.

Acknowlegement

This research greatly benefited from a conversation with Professor Jianqing Fan. The comments of two referees are also gratefully acknowledged. Partial support was provided by a grant from the U.S. Environmental Protection Agency.

APPENDIX

Derivation of (4)

Suppose that $\lambda \to 0$. To aid readability we let $\alpha = \lambda^{2p}$. Then using the expansion

$$(I + \alpha A)^{-1} = I - \alpha A + \alpha^2 A^2 - \ldots$$

we obtain the approximation

$$H_\alpha = H_0 - \alpha G + o(\alpha G),$$

where $G = X(X^{\mathrm{T}}X)^{-1}D(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The bias can then be written as

$$H_\alpha m - m = -\alpha G m + m_X - m + o(\alpha G), \tag{A1}$$

where $m_X = H_0 m$ is the projection of m on to the column space of X. The average squared bias therefore has the approximation

$$\frac{1}{n}\|H_\alpha m - m\|^2 = \frac{1}{n}\{\alpha^2 \|Gm\|^2 - 2\alpha(Gm)^{\mathrm{T}}(m_X - m) + \|m_X - m\|^2\}.$$

Approximation (A1) can also be used to approximate $\mathrm{tr}(H_\alpha^2)$ and obtain the first term on the right-hand side of (4). The full expression then follows by assuming that the approximation error $m_X - m$ is negligible. Such an assumption is reasonable when the knots are densely packed, relative to the curviness of m.

REFERENCES

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403.

EILERS, P. H. C. & MARX, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

GREEN, P. J. & SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

HÄRDLE, W. & MARRON, J. S. (1995). Fast and simple scatterplot smoothing. *Comp. Statist. Data Anal.* **20**, 1–17.

HOLST, U., HÖSSJER, O., BJÖRKLUND, C., RAGNARSON, P. & EDNER, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Environmetrics* **7**, 401–16.

NYCHKA, D. & CUMMINS, D. (1996). Comment on paper by P. H. C. Eilers and B. D. Marx. *Statist. Sci.* **11**, 104–5.

SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66**, 605–10.

TSYBAKOV, A. B. (1986). Robust reconstruction of functions by the local approximation method. *Prob. Info. Transm.* **22**, 133–46.

WAND, M. P. (1999). A comparison of regression spline smoothing procedures. *Comp. Statist.* To appear.

[*Received July* 1998. *Revised December* 1998]