# Nonparametric regression using kernel and spline methods

Jean D. Opsomer* F. Jay Breidt †

March 23, 2016

## 1 The statistical model

When applying nonparametric regression methods, the researcher is interested in estimating the relationship between one dependent variable, $Y$, and one or several covariates, $X_1, \ldots, X_q$. We discuss here the situation with one covariate, $X$ (the case with multiple covariates is addressed in the references provided below). The relationship between $X$ and $Y$ can be expressed as the conditional expectation

$$\mathrm{E}(Y|X = x) = f(x).$$

Unlike in parametric regression, the shape of the function $f(\cdot)$ is not restricted to belong to a specific parametric family such as polynomials.

This representation for the mean function is the key difference between parametric and nonparametric regression, and the remaining aspects of the statistical model for $(X, Y)$ are similar between both regression approaches. In particular, the random variable $Y$ is often assumed to have a constant (conditional) variance, $\mathrm{Var}(Y|X) = \sigma^2$, with $\sigma^2$ unknown. The constant variance and other common regression model assumptions, such as independence, can be relaxed just as in parametric regression.

## 2 Kernel methods

Suppose that we have a dataset available with observations $(x_1, y_1), \ldots, (x_n, y_n)$. A simple kernel-based estimator of $f(x)$ is the *Nadaraya-Watson kernel regression* estimator, defined as

$$\hat{f}_h(x) = \frac{\sum_{i=1}^n K_h(x_i - x) y_i}{\sum_{i=1}^n K_h(x_i - x)}, \tag{1}$$

*Department of Statistics, Colorado State University, Fort Collins, CO, USA. Email: jopsomer@stat.colostate.edu.

†Department of Statistics, Colorado State University, Fort Collins, CO, USA. Email: jbreidt@stat.colostate.edu

with $K_h(\cdot) = K(\cdot/h)/h$ for some kernel function $K(\cdot)$ and bandwidth parameter $h > 0$. The function $K(\cdot)$ is usually a symmetric probability density and examples of commonly used kernel functions are the Gaussian kernel $K(t) = (\sqrt{2\pi})^{-1}\exp(-t^2/2)$ and the *Epanechnikov* kernel $K(t) = \max\{\frac{3}{4}(1-t^2), 0\}$.

Generally, the researcher is not interested in estimating the value of $f(\cdot)$ at a single location $x$, but in estimating the curve over a range of values, say for all $x \in [a_x, b_x]$. In principle, kernel regression requires computing (1) for any value of interest. In practice, $\hat{f}_h(x)$ is calculated on a sufficiently fine grid of $x$-values and the curve is obtained by interpolation.

We used the subscript $h$ in $\hat{f}_h(x)$ in (1) to emphasize the fact that the bandwidth $h$ is the main determinant of the shape of the estimated regression, as demonstrated in Figure 1. When $h$ is small relative to the range of the data, the resulting fit can be highly variable and look "wiggly." When $h$ is chosen to be larger, this results in a less variable, more smooth fit, but it makes the estimator less responsive to local features in the data and introduces the possibility of bias in the estimator. Selecting a value for the bandwidth in such a way that it balances the variance with the potential bias is therefore a crucial decision for researchers who want to apply nonparametric regression on their data. Data-driven bandwidth selection methods are available in the literature, including in the references provided below.

A class of kernel-based estimators that generalizes the Nadaraya-Watson estimator in (1) is referred to as *local polynomial regression* estimators. At each location $x$, the estimator $\hat{f}_h(x)$ is obtained as the estimated intercept, $\hat{\beta}_0$, in the weighted least squares fit of a polynomial of degree $p$,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p\right) K_h(x_i - x).$$

This estimator can be written explicitly in matrix notation as

$$\hat{f}_h(x) = (1, 0, \ldots, 0)\left(\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y}, \tag{2}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{W}_x = \text{diag}\{K_h(x_1 - x), \ldots, K_h(x_n - x)\}$ and

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{bmatrix}.$$

It should be noted that the Nadaraya-Watson estimator (1) is a special case of the local polynomial regression estimator with $p = 0$. In practice, the local linear ($p = 1$) and local quadratic estimators ($p = 2$) are frequently used.

An extensive literature on kernel regression and local polynomial regression exists, and their theoretical properties are well understood. Both kernel regression and local polynomial regression estimators are biased but consistent

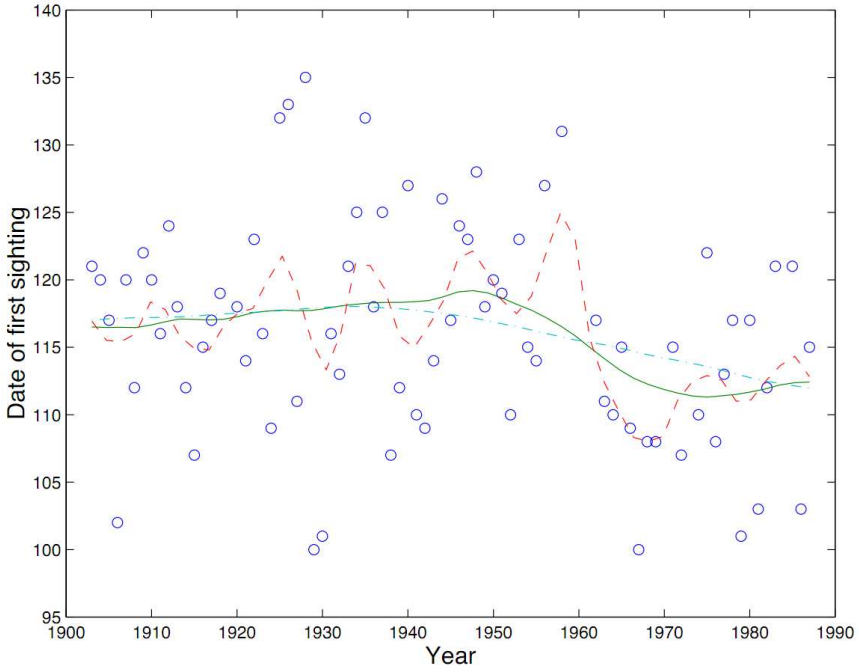


Figure 1: Dates (Julian days) of first sightings of bank swallows in Cayuga Lake basin, with three kernel regressions using bandwidth values $h$ calculated as the range of years multiplied by 0.05 (−−), 0.2 (–) and 0.4 (−·).

estimators of the unknown mean function, when that function is continuous and sufficiently smooth. For further information on these methods, we refer to reader to the monographs by Wand and Jones (1995) and Fan and Gijbels (1996).

# 3 Spline methods

In the previous section, the unknown mean function was assumed to be *locally* well approximated by a polynomial, which led to local polynomial regression. An alternative approach is to represent the fit as a *piecewise* polynomial, with the pieces connecting at points called *knots*. Once the knots are selected, such an estimator can be computed globally in a manner similar to that for a parametrically specified mean function, as will be explained below. A fitted mean function represented by a piecewise continuous curve only rarely provides a satisfactory fit, however, so that usually the function and at least its first derivative are constrained to be continuous everywhere, with only the second or higher derivatives allowed to be discontinuous at the knots. For historical reasons, these constrained piecewise polynomials are referred to as *splines*, leading to

the name *spline regression* or *spline smoothing* for this type of nonparametric regression.

Consider the following simple type of polynomial spline of degree $p$:

$$\beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - \kappa_k)_+^p, \tag{3}$$

where $p \geq 1$, $\kappa_1, \ldots, \kappa_K$ are the knots and $(\cdot)_+^p = \max\{(\cdot)^p, 0\}$. Clearly, (3) has continuous derivatives up to degree $(p-1)$, but the $p$th derivative can be discontinuous at the knots. Model (3) is constructed as a linear combination of *basis functions* $1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p$. This basis is referred to as the *truncated power basis*. A popular set of basis functions are the so-called *B-splines*. Unlike the truncated power splines, the B-splines have compact support and are numerically more stable, but they span the same function space. In what follows, we will write $\psi_j(x), j = 1, \ldots, J$ for a set of (generic) basis functions used in fitting regression splines, and replace (3) by $\beta_1\psi_1(x) + \cdots + \beta_J\psi_J(x)$.

For fixed knots, a regression spline is linear in the unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$ and can be fitted parametrically using least squares techniques. Under the homoskedastic model described in Section 1, the *regression spline* estimator for $f(x)$ is obtained by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{J} \beta_j \psi_j(x_i) \right)^2 \tag{4}$$

and setting $\hat{f}(x) = \sum_{j=1}^{J} \hat{\beta}_j \psi_j(x)$. Since deviations from the parametric shape can only occur at the knots, the amount of smoothing is determined by the degree of the basis and the location and number of knots. In practice, the degree is fixed (with $p = 1, 2$ or 3 as common choices) and the knot locations are usually chosen to be equally-spaced over the range of the data or placed at regularly spaced data quantiles. Hence, the number of knots $K$ is the only remaining smoothing parameter for the spline regression estimator. As $K$ (and therefore $J$) is chosen to be larger, increasingly flexible estimators for $f(\cdot)$ are produced. This reduces the potential bias due to approximating the unknown mean function by a spline function, but increases the variability of the estimators.

The *smoothing spline* estimator is an important extension of the regression spline estimator. The smoothing spline estimator for $f(\cdot)$ for a set of data generated by the statistical model described in Section 1 is defined as the minimizer of

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{a_x}^{b_x} (f^{(p)}(t))^2 dt, \tag{5}$$

over the set of all functions $f(\cdot)$ with continuous $(p-1)$th derivative and square integrable $p$th derivative, and $\lambda > 0$ is a constant determining the degree of smoothness of the estimator. Larger values of $\lambda$ correspond to smoother fits. The choice $p = 2$ leads to the popular *cubic smoothing splines*. While not

immediately obvious from the definition, the function minimizing (5) is exactly equal to a special type of regression spline with knots at each of the observation points $x_1, \ldots, x_n$ (assuming each of the locations $x_i$ is unique).

Traditional regression spline fitting as in (4) is usually done using a relatively small number of knots. By construction, smoothing splines use a large number of knots (typically, $n$ knots), but the smoothness of the function is controlled by a penalty term and the smoothing parameter $\lambda$. The *penalized spline* estimator represents a compromise between these two approaches. It uses a moderate number of knots and puts a penalty on the coefficients of the basis functions. Specifically, a simple type of penalized spline estimator for $m(\cdot)$ is obtained by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{J} \beta_j \psi_j(x_i) \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \tag{6}$$

and setting $\hat{f}_\lambda(x) = \sum_{j=1}^{J} \hat{\beta}_j \psi_j(x)$ as for regression splines. Penalized splines combine the advantage of a parametric fitting method, as for regression splines, with the flexible adjustment of the degree of smoothness as in smoothing splines. Both the basis function and the exact form of the penalization of the coefficients can be varied to accommodate a large range of regression settings.

Spline-based regression methods are extensively described in the statistical literature. While the theoretical properties of (unpenalized) regression splines and smoothing splines are well established, results for penalized regression splines have only recently become available. The monographs by Wahba (1990), Eubank (1999) and Ruppert et al. (2003) are good sources of information on spline-based methods.

## Acknowledgements

Based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC.

## References

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). New York: Marcel Dekker.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Wahba, G. (1990). *Spline models for observational data*. SIAM [Society for Industrial and Applied Mathematics].

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall.