

# Nonparametric Estimation Methods

Ian Crawford

Department of Economics

### III. Nonparametric Regression Estimation (Part Deux)

i. Derivative Estimation

ii. Alternative methods (nearest neighbours series)

iii. Multivariate kernel regression

iv. The Curse revisited

v. Dimension Reduction - Overview

vi. Additively separable Models

## *Derivative Estimation*

Very often in applied work we care about *slopes*.

Typically these are heavily restricted by our choice of functional form - not with nonparametric regression.

Derivatives can be estimated by differentiating the estimate of  $m(x)$  the required number of times.

This works provided the estimate of  $m(x)$  is itself smooth enough. This can be achieved by, for example, using a kernel function  $K(\cdot)$  which is smooth (like the Gaussian density function). Unfortunately this is rather complicated for the Nadaraya-Watson estimator.

A further advantage of the local polynomial approach is that it provides an easy way of estimating derivatives of the function  $m(x)$ . The natural approach is to estimate  $m(x)$  by  $\widehat{m}_p(x)$  and compute the derivative analytically (hard) or numerically (slow). But a more efficient way is based on comparing

$$m(t) = m(x) + m'(x)(t-x) + \dots + \frac{m^{(p)}(x)}{p!}(t-x)^p$$

and

$$\min_{\beta_0 \dots \beta_p} \sum \left[ Y_i - \beta_0 - \beta_1(X_i - x) - \dots - \beta_p(X_i - x)^p \right]^2 \frac{1}{h} K\left(\frac{X_i - x}{h}\right)$$

gives by Taylor's Theorem the *local polynomial derivative estimator* for the  $v$ th derivative of  $m(x)$

$$\widehat{m}_p^v(x) = v! \beta_v(x)$$

## *Alternative Methods*

1. Nearest neighbour methods
2. Series Estimators

## *Nearest Neighbour Methods*

Kernel regression estimation can be viewed as a method of computing weighted averages of the response variables in a fixed neighbourhood around  $x$

The width of this neighbourhood is governed by the bandwidth  $h$ .

The  $k$ -nearest-neighbour estimator can also be viewed as a weighted average of the response variables in a neighbourhood around  $x$ , with the important difference that the neighbourhood width is not fixed but variable.

## *Nearest Neighbour Methods*

The values of  $Y$  used in computing the average, are those which belong to the  $k$  observed values of  $X$  that are nearest to the point  $x$

The NN estimator can be written as

$$\widehat{m}(x) = \frac{1}{n} \sum_{i=1}^n w_{ki}(x) Y_i$$

where the weights are

$$\begin{aligned} w_{ki}(x) &= n/k \text{ if } i \in \{i : X_i \text{ is one of the } k \text{ nearest obs to } x\} \\ &= 0 \text{ otherwise} \end{aligned}$$

## *Nearest Neighbour Methods*

Obviously if the data are sparse near  $x$  then the nearest neighbors are rather far away from (and each other) so we end up with a wide neighbourhood around  $x$ .

$k$  the smoothing parameter of this estimator. Increasing  $k$  makes the estimate smoother.

See Hardle (1990) for the variance of the NN estimator but note that it is proportional to  $1/k$  is does not depend on  $f(x)$ .

This is because the NN estimator always averages over  $k$  observations, regardless of how dense the data is in the neighborhood of the point where we estimate  $x$ .

## *Nearest Neighbour Methods*

In fact NN estimators can be viewed as kernels with rectangular weight functions, or indeed with more general weight functions.

$$\widehat{m}_{\widehat{h}_k}(x) = \frac{\frac{1}{n\widehat{h}_k} \sum_{i=1}^n K\left(\frac{X_i - x}{\widehat{h}_k}\right) Y_i}{\frac{1}{n\widehat{h}_k} \sum_{i=1}^n K\left(\frac{X_i - x}{\widehat{h}_k}\right)}$$

They suffer from the same drawback as NN density estimators (discontinuity where the bandwidth changes)

## *Series Estimators*

As we saw with density estimation, under mild regularity conditions, functions can be represented as a series of basis functions (e.g. a Fourier series).

That is

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x)$$

where  $\{\varphi_j(x)\}_{j=0}^{\infty}$  is a known basis of functions and  $\{\beta_j\}_{j=0}^{\infty}$  are the unknown Fourier coefficients. We've seen the cosine basis function already.

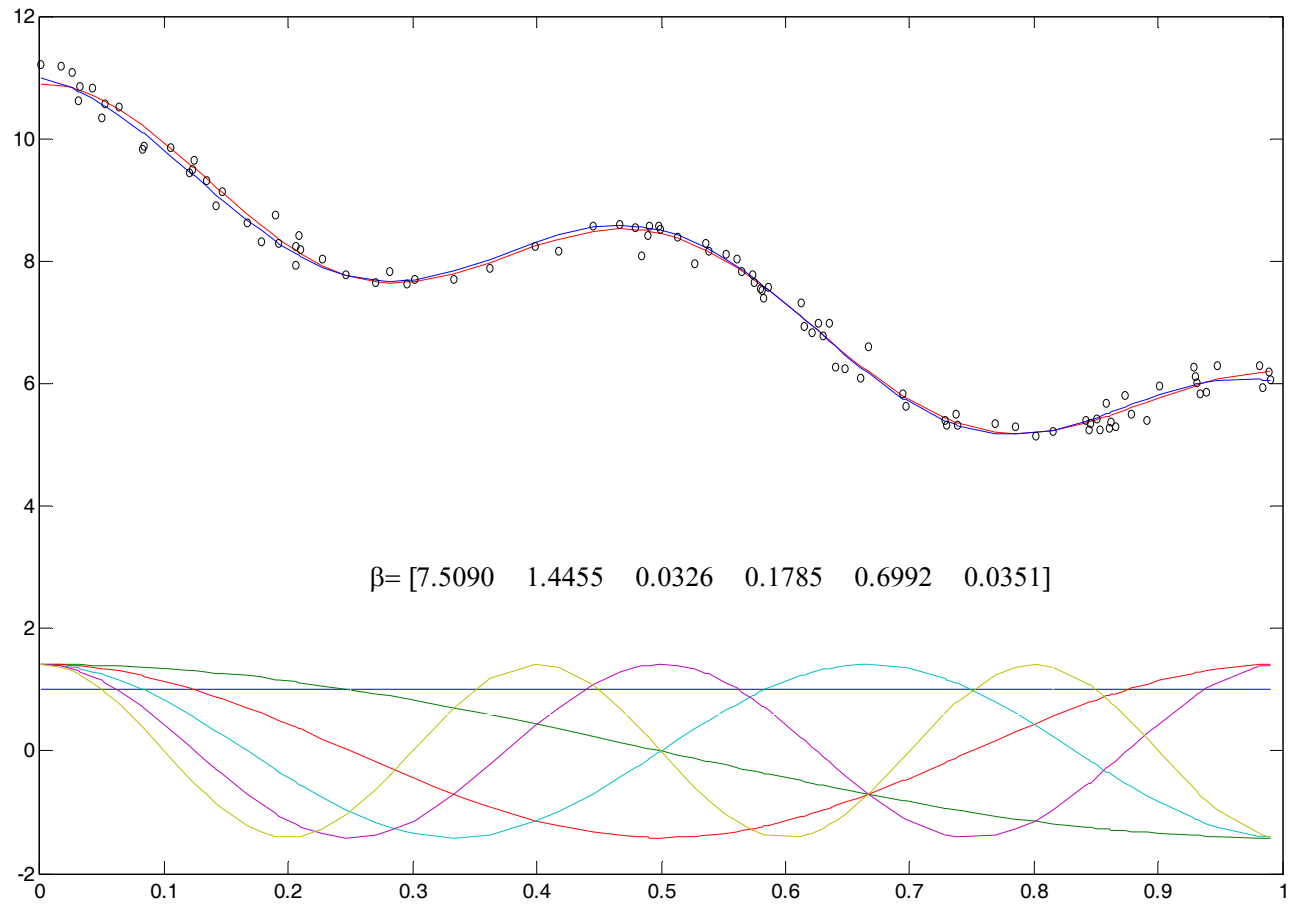
$$\varphi_0(x) = 1, \quad \varphi_j(x) = 2^{1/2} \cos(\pi j x), \quad \text{for } j = 1, \dots$$

Obviously, an infinite number of coefficients cannot be estimated from a finite number of observations. Hence, one has to choose the number of terms that will be included in the Fourier series representation. Series estimation proceeds in three steps:

(a) select a basis,

(b) select the cut-off  $J$ ,  $J < n$  ( $J$  is the smoothing parameter more terms  $\rightarrow$  interpolation. (see Efromovich (1997, p164)

(c) estimate the  $J$  unknown Fourier coefficients by a suitable method (OLS).



## *Series Estimators*

Given that series estimates are (or can be) constructed as sum of smooth curves derivative estimation is very easy.

The only hard bit is the choice of cut-off/smoothing parameter (see Efromovich for various plug-in and rule-of-thumb solutions)

Wavelets are examples of basis functions which can be used to model regression functions that feature varying frequencies and jumps.

## *Multivariate Regression*

So far we have looked at univariate regression methods.

In practice we are mainly interested in how the response variable depends on a *vector* of exogenous variables, denoted by  $\mathbf{x}$ .

This means we aim to estimate the conditional expectation

$$E(y|\mathbf{x}) = E(Y|x_1, \dots, x_d) = m(\mathbf{x})$$

$$E(y|\mathbf{x}) = \int y f(y|\mathbf{x}) dy = \frac{\int y f(y, \mathbf{x}) dy}{f(\mathbf{x})}$$

$$\hat{f}(y, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n h^{-1} K\left(\frac{Y_i - Y}{h}\right) \frac{1}{\det \mathbf{H}} \mathcal{K}\left(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right)$$

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\det \mathbf{H}} \mathcal{K}\left(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right)$$

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \frac{\frac{1}{n} \sum_{i=1}^n \frac{1}{\det \mathbf{H}} \mathcal{K}\left(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right) Y_i}{\frac{1}{n} \sum_{i=1}^n \frac{1}{\det \mathbf{H}} \mathcal{K}\left(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x})\right)}$$

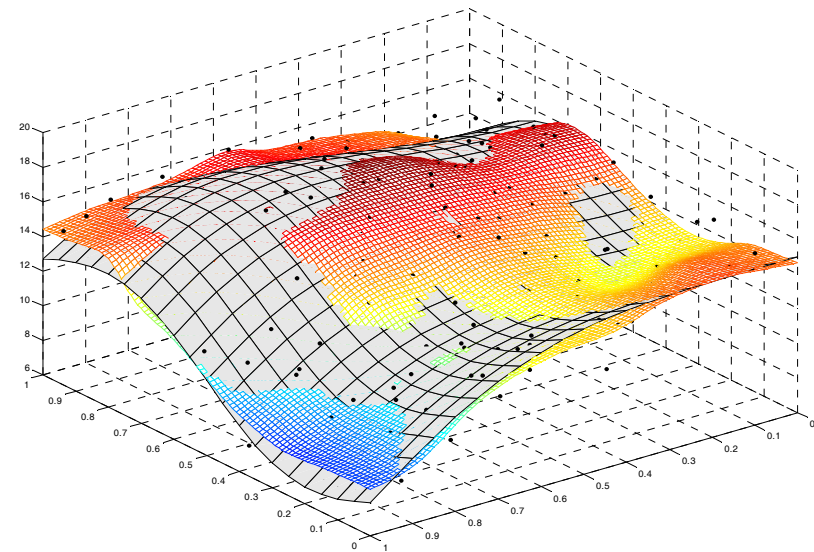
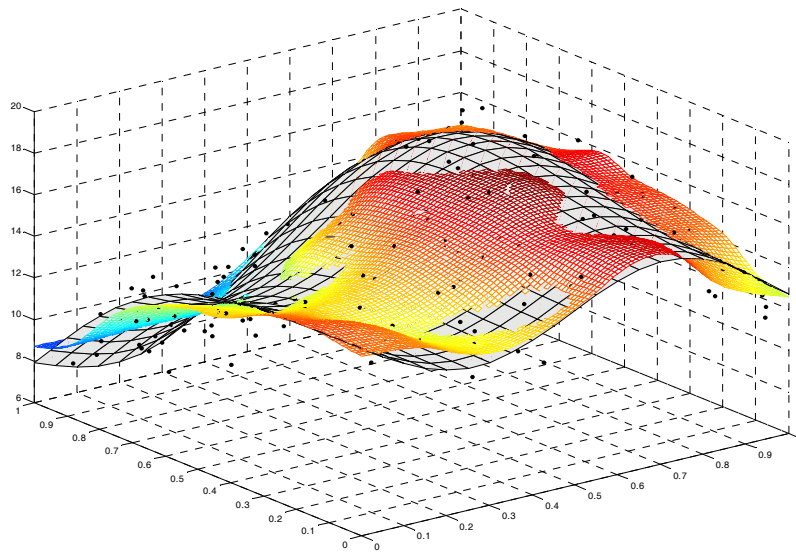
(a weighted sum of the observed responses in a ball or cube around  $\mathbf{x}$ ).

## *Multivariate Kernel Regression - Implementation*

```
x1grid = 0:0.01:1; x2grid = 0:0.01:1;
mx = NaN*ones(101,101);h1 = 0.07; h2 = 0.07;
for i = 1:101;
for j=1:101;
u = (ones(n,1)*[x1grid(i) x2grid(j)]-[X1 X2])./(ones(n,1)*[h1 h2]);
Kh = prod(pdf('Normal',u,0,1),2)/(h1*h2);
mx(i,j) = mean(Kh.*Y)/mean(Kh);
end;
end;
```

$$y = 10 - \sin(2\pi x_1) + 5 \cos(2x_2^2 - x_1) + \varepsilon$$

$$\varepsilon \sim N(0, 1.1), x_i \in [0, 1], n = 100$$



Just as we saw in the univariate the kernel, the multivariate Nadaraya-Watson estimator is a local constant estimator.

The definition of local polynomial kernel regression is a straightforward generalisation of the univariate case.

For example a local linear regression estimator solves the problem

$$\min_{\beta_0 \beta_1} \sum \left[ Y_i - \beta_0 - \beta_1' (\mathbf{X}_i - \mathbf{x}) \right]^2 \frac{1}{\det \mathbf{H}} \mathcal{K} \left( \mathbf{H}^{-1} (\mathbf{X}_i - \mathbf{x}) \right)$$

The solution can be written in LS form:

$$\hat{\boldsymbol{\beta}} = [\beta_0, \boldsymbol{\beta}'_1]' = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y}$$

where

$${}_{n \times (d+1)} \mathbf{X} = \begin{bmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})' \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})' \end{bmatrix}, \quad {}_{(n \times 1)} \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

$${}_{n \times n} \mathbf{W} = \text{diag} \left[ \frac{1}{\det \mathbf{H}} \mathcal{K} \left( \mathbf{H}^{-1} (\mathbf{X}_1 - \mathbf{x}) \right), \dots, \frac{1}{\det \mathbf{H}} \mathcal{K} \left( \mathbf{H}^{-1} (\mathbf{X}_n - \mathbf{x}) \right) \right]$$

$\beta_0$  estimates the regression function itself, whereas  $\boldsymbol{\beta}_1$  is proportional to the partial derivatives w.r.t. the components of  $\mathbf{x}$ .

The conditional asymptotic bias and variance terms and the pointwise confidence bounds are all intuitive generalisations of the univariate model and are discussed in Ruppert and Wand (1994).

Computation of multivariate (polynomial) kernel regression can be performed in any package which can do weighted least squares. But note that this has to be carried out at all observation points or on a grid of points in  $\mathbb{R}^d$ .

A practical problem is the graphical display for higher dimensional multivariate functions.

## The Curse - Revisited

The sparseness of observations in higher dimensions even for large sample sizes makes estimators based on local averaging perform unsatisfactorily.

Recall that asymptotically

$$MSE \approx \frac{1}{nh^d}C_1 + h^4C_2$$

where  $C_1$  and  $C_2$  are constants that neither depend on  $n$  nor  $h$ .

If we derive the optimal bandwidth we find that the optimal rate is  $\sim n^{-1/(4+d)}$

Increasing  $d$  dramatically slows convergence.

## *Dimension Reduction*

The problem of estimating fully nonparametric models (densities or regressions) in many dimensions is a serious one.

Especially since multivariate models are the norm in applied work (indeed univariate models are usually economically meaningless)

Because of this there has been a great deal of recent research aimed at mitigating the curse.

## *Dimension Reduction - Overview*

1. Variable selection.
2. Functional restrictions
3. Semi-parametric modelling.

## *Variable Selection*

The aim is to choose an appropriate subset of variables,  $\mathbf{X}_r \subset \mathbf{X}$ , from the set of all variables that could potentially enter the regression.

Of course this should be informed by the problem at hand and the relevant economic theory (which often includes dimension-reducing theoretical restrictions. Examples ...?).

This done, however, we need a statistical selection criterion.

## *Variable Selection*

Vieu (1994) has proposed to use the integrated square error to measure the quality of a given subset of variables. In theory, a subset of variables is defined to be an optimal subset if it minimizes the MISE

$$MISE(\mathbf{x}_r^*) = \min_{\mathbf{x}_r} MISE(\mathbf{x}_r)$$

where  $\mathbf{x}_r \subset \mathbf{x}$ .

In practice (and for the usual reason of unpenalised fit), the *MISE* is replaced by its sample analog, the multivariate analog of the cross validation function.

## *Functional Restrictions - additive models*

The additive model is a generalisation of the multiple linear regression model by introducing one-dimensional nonparametric functions in the place of the linear components.

Here, the conditional expectation of  $y$  given  $\mathbf{x} = [x_1, \dots, x_d]$  is assumed to be the sum of unknown functions of the explanatory variables:

$$E(y|\mathbf{x}) = \sum_{j=1}^d m_j(x_j) = m_1(x_1) + \dots + m_d(x_d)$$

Dimension reduction is achieved by estimating  $d$  functions of one-dimensional variables instead estimating one function of  $d$  variables.

## *Functional Restrictions - additive models*

Obviously lots of options are available, for example:

$$E(y|x_1, x_2, x_3) = m_1(x_1) + m_2(x_2, x_3)$$

Now there are two functions to estimate but the highest dimension is 2 rather than 3.

## *Semi-parametric models - the Partially Linear Model*

Sometimes, for theoretical reasons or for analytical reasons we might be happy to model part of the data linearly. For example the impact of a dummy variable might be adequately captured by an intercept/shifter.

Separate the explanatory variables into two groups  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . The regression of  $y$  on  $\mathbf{x}$  is assumed to have the form

$$E(y|\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1\boldsymbol{\beta} + m(\mathbf{x}_2)$$

where  $m(\mathbf{x}_2)$  is an unknown (multivariate) function of the data  $\mathbf{x}_2$ .

Estimating this involves a mixture of parametric and nonparametric techniques.

## *Semi-parametric models - Single Index Models*

Index models play an important role in econometrics.

Summarising information into one “single index” term greatly reduces the dimensionality of a problem.

Single Index Models have the following general form (parametric examples include, logit, probit, poisson etc)

$$E(y|\mathbf{x}) = g(\mathbf{x}\boldsymbol{\beta})$$

where  $g()$  is an unknown “link” function and  $\mathbf{x}\boldsymbol{\beta}$  is a linear index.

Estimation can be carried out in two steps. First, we estimate  $\boldsymbol{\beta}$ . Then, using the index values for our observations, we can estimate by nonparametric regression - only a one-dimensional regression problem.

## *Semi-parametric models - Hybrids*

The generalised additive model is

$$E(y|\mathbf{x}) = g\left(\sum_{j=1}^d m_j(x_j)\right)$$

The generalised partial linear model is

$$E(y|\mathbf{x}_1, \mathbf{x}_2) = g(\mathbf{x}_1\boldsymbol{\beta} + m(\mathbf{x}_2))$$

...and so on, and so on.

## *Summary*

1. Kernel weighted local polynomials provide a nice easy framework for estimating derivatives of nonparametric regressions.
2. There are various alternatives to kernel-based methods - nearest neighbours, splines, series. Series estimators are probably the easiest and the fastest to implement (they don't require anything more complicated than LS) and using wavelet basis functions is ideal for data which exhibit sharp changes in slopes.
3. All of those techniques generalise naturally to multivariate regression problems but beware the *Curse*

4. By slowing convergence rates dramatically the *Curse of Dimensionality* places a serious practical limitation on multivariate nonparametric estimation.
5. Solutions include - variable selection and functional restrictions.