Faculdade de Economia
da Universidade de Coimbra

Grupo de Estudos Monetários e Financeiros
(GEMF)
Av. Dias da Silva, 165 – 3004-512 COIMBRA,
PORTUGAL

gemf@fe.uc.pt
http://gemf.fe.uc.pt

RALPH W. BAILEY & JOHN T. ADDISON

A Smoothed-Distribution Form of Nadaraya-Watson Estimation

ESTUDOS DO GEMF
N.º 01 2011

PUBLICAÇÃO CO-FINANCIADA PELA
FUNDAÇÃO PARA A CIÊNCIA E TECNOLOGIA

Impresso na Secção de Textos da FEUC
COIMBRA 2011

A Smoothed-Distribution Form of Nadaraya-Watson Estimation

Ralph W. Bailey
University of Birmingham

John T. Addison
University of South Carolina,
Queen’s University Belfast, and University of Coimbra

ABSTRACT

A Smoothed-Distribution Form of Nadaraya-Watson Estimation

Given observation-pairs (x_i, y_i), $i = 1, \dots, n$, taken to be independent observations of the random pair (X, Y), we sometimes want to form a nonparametric estimate of $m(x) \equiv E(Y|X = x)$. Let Y^E have the empirical distribution of the y_i, and let (X^S, Y^S) have the kernel-smoothed distribution of the (x_i, y_i). Then the standard estimator, the Nadaraya-Watson form $\widehat{m}_{NW}(x)$, can be interpreted as $E(Y^E|X^S = x)$. The smoothed-distribution estimator $\widehat{m}_S(x) \equiv E(Y^S|X^S = x)$ is a more general form than $\widehat{m}_{NW}(x)$ and often has better properties. Similar considerations apply to estimating $Var(Y|X = x)$, and to local polynomial estimation. The discussion generalizes to vector $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

JEL Classification: C140

Keywords: nonparametric regression, Nadaraya-Watson, kernel density, conditional expectation estimator, conditional variance estimator, local polynomial estimator

Corresponding author:

Ralph W. Bailey
Department of Economics
Business School
University of Birmingham
Birmingham B15 2TT
U.K.
E-mail: r.w.bailey@bham.ac.uk

1 Introduction

Suppose that we have a sample of real observation-pairs (x_i, y_i) drawn independently from some joint probability distribution $f_{X,Y}$. (In fact, we aim to discuss observation-pairs of vectors, although the main ideas can be expounded in the scalar special case.) The problem we discuss is that of providing nonparametric estimates of the conditional mean and variance $m(x) \equiv E(Y|X=x)$, $V(x) \equiv Var(Y|X=x)$. Kernel-density-based methods for doing so originate in the work of Nadaraya (1964) and Watson (1964). However, study of the probabilistic nature of the Nadaraya-Watson estimators $\hat{m}_{NW}(x)$ and $\hat{V}_{NW}(x)$ suggests that they are to some extent inconsistent with the belief that a kernel-density-based function $\hat{f}_{X,Y}$ is our best available approximation to $f_{X,Y}$; inconsistent in the sense that they form estimates using the empirical distribution of the y_i, rather than the kernel-smoothed distribution. This consideration suggests that it would be more natural to use 'smoothed-distribution' estimators $\hat{m}_S(x)$ and $\hat{V}_S(x)$. Such estimators are derived and studied in the next section. The remainder of the present section sketches motives for obtaining estimates of $m(x)$ and $V(x)$.

Watson (1964, p. 359) introduced $\hat{m}_{NW}(x)$ as providing 'a simple computer method for obtaining a "graph" from a large number of observations', by which is meant drawing a curved regression-like line through a scatter plot, to reveal a pattern of relationship obscured by the number and variability of the points on the graph. Simonoff (1996, pp. 134-6) accordingly locates his useful introductory discussion of Nadaraya-Watson estimation within the subject of nonparametric regression. Applications appear in, for instance, Barrett and Dorosh's (1996) study of farmer welfare and rice prices in Madagascar. They provide Nadaraya-Watson regressions of such combinations as per-capita income against land holdings, and household activity against land holdings.

One motive for the study of conditional variance, in the scalar context, is the wish to know the dispersion of the (x_i, y_i) about the line $m(x)$. An estimate of $V(x) = E\left[(Y - m(x))^2 | X = x\right]$ allows us to draw such lines as $\hat{m}(x) \pm 2\sqrt{\hat{V}(x)}$, indicative of variability about $m(x)$. However, in order to construct confidence intervals that take into account the in-general non-normal shape of $\hat{f}_{Y|X}$, a bootstrapping method is usually applied, based on repeated resampling from the observations (x_i, y_i). (See Simonoff, 1996, p. 48; Barrett and Dorosh, 1996, p. 661, n. 10; Fiorio, 2004.) As a practical matter, we do not pursue the subject here.

Conditional variance is also the subject of study in its own right, especially as regards the heteroscedasticity of economic time series, inaugurated by Engle (1982) and Engle and Bollerslev (1986). (For a review of nonparametric methods in this context see Linton and Yan, 2011.) Giannopoulos (2008), for instance, models a situation in which it is known that the correlation between time series innovations increases at times of high volatility. To do so he uses an estimator of the Nadaraya-Watson type, obtaining the estimated variance-covariance matrix as a function of the innovation levels. The study illustrates that Nadaraya-Watson methods can be applied to variances as well as means, and vectors as well as scalars.

Nadaraya-Watson estimates of the conditional mean and variance have continued to be the focus of theoretical investigation. Parzen (1963) provided an early discussion of the asymptotic properties of kernel-density estimates. Simonoff (1996) discusses alternatives such as local polynomial regression and spline smoothing; also the characteristic kernel-density problems of boundary bias, bandwidth selection, and the presence of autocorrelation. And for a recent overview of non-parametric regression, see Wasserman (2006, particularly chapters 4 and 5).

2 Main Results

Let $\mathbf{w}_i = (\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, n$, $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{y}_i \in \mathbb{R}^q$ be a sample of observation-pairs drawn independently from the distribution with density $f_{\mathbf{X},\mathbf{Y}}$. We consider estimation of the vector $\mathbf{m}(\mathbf{x}) \equiv E(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ and of the matrix $\mathbf{V}(\mathbf{x}) \equiv Var(\mathbf{Y}|\mathbf{X} = \mathbf{x})$.

Our discussion is initially confined to the scalar case $p = q = 1$, and to estimators of the scalar value $m(x)$. We begin by reviewing the approach independently inaugurated by Nadaraya (1964) and Watson (1964), in which an estimate of $m(x)$ is formed by substituting kernel-density estimates of $f_{X,Y}$ and f_X into the exact expression $m(x) = \int y f_{X,Y}(x, y)\, dy / f_X(x)$. An estimate $\hat{f}_{X,Y}$ is of the form

$$\hat{f}_{X,Y}(x, y) = \frac{1}{n} \sum \frac{1}{|\mathbf{H}_i|} K_i \left[\mathbf{H}_i^{-1} (\mathbf{w} - \mathbf{w}_i)\right]$$

where the kernel function K_i satisfies $\int K_i(\mathbf{w})\, d\mathbf{w} = 1$, $\int \mathbf{w} K_i(\mathbf{w})\, d\mathbf{w} = \mathbf{0}_2$, $\int\int \mathbf{w}\mathbf{w}' K_i(\mathbf{w})\, d\mathbf{w} = I_2$, and the $\mathbf{H}_i$ are nonsingular 2×2 bandwidth matrices. (This and all subsequent summations are over i and run from 1 to n.) The corresponding estimate of the marginal distribution f_X is $\hat{f}_X(x) = \int \hat{f}_{X,Y}(x,y)\,dy$. In order to obtain a closed-form expression for $\int y\hat{f}_{X,Y}(x,y)\,dy$, the Nadaraya-Watson approach imposes on the kernel functions the product form $K_i(x,y) = K_{X,i}(x)\,K_{Y,i}(y)$ and on the bandwidth matrices the diagonal form $\mathbf{H}_i = \begin{pmatrix} h_{X,i} & 0 \\ 0 & h_{Y,i} \end{pmatrix}$. It is shown in section 3 below that the resulting estimator of $m(x)$ is

$$\hat{m}_{NW}(x) = \frac{\sum \frac{y_i}{h_{X,i}} K_{X,i}\left(\frac{x-x_i}{h_{X,i}}\right)}{\sum \frac{1}{h_{X,i}} K_{X,i}\left(\frac{x-x_i}{h_{X,i}}\right)}. \tag{1}$$

Notice that this expression involves neither $h_{Y,i}$ nor $K_{Y,i}$. If we specify constant bandwidths $h_{X,i} = h_X$, and constant marginal kernel functions $K_{X,i} = K_X$, we obtain the simpler form

$$\hat{m}_{NW}(x) = \frac{\sum y_i K_X\left(\frac{x-x_i}{h_X}\right)}{\sum K_X\left(\frac{x-x_i}{h_X}\right)}. \tag{2}$$

Equation (2) is the usual form quoted for the Nadaraya-Watson estimator; (1) generalizes to the case of variable bandwidths and variable kernel functions.

Irrespective of the derivation of (1) and (2), and the special assumptions they involve, we can provide the two definitions with a direct probabilistic interpretation, provided we exclude the case in which the kernel function K_i is permitted to be negative. The point of doing so is that, in these circumstances, the function $\hat{f}_{X,Y}$ is itself an exact density, corresponding to a pair of continuously-distributed random variables we call $\left(X^S, Y^S\right)$. Then the kernel-density approximation $\hat{f}_X$ can alternatively be interpreted as the exact marginal distribution of X^S. Write

$$p_j(x) \equiv \frac{\frac{1}{h_{X,j}} K_{X,j}\left(\frac{x-x_j}{h_{X,j}}\right)}{\sum \frac{1}{h_{X,i}} K_{X,i}\left(\frac{x-x_i}{h_{X,i}}\right)}. \tag{3}$$

To interpret $p_j(x)$, let I be an integer-valued random variable whose probability distribution is defined by $P(I = i) = n^{-1}$, $i = 1, \ldots, n$. Let

$\left(X^E, Y^E\right) = (x_I, y_I)$, so I selects one of the observations at random and equiprobably. Then $\left(X^E, Y^E\right)$ has the empirical distribution of the (x_i, y_i) (which distribution fails to possess a density function). The pair $\left(X^S, Y^S\right)$ smooths out the empirical distribution by adding a random perturbation pair (ε_i, ζ_i) to whichever (x_i, y_i) is chosen by I. (The superscript S stands for 'smoothed-distribution'.) Let (ε_i, ζ_i) have means zero and variance matrix Ω_i. The joint distribution of (ε_i, ζ_i) and $\left(\varepsilon_j, \zeta_j\right)$ for $i \neq j$ plays no part in the analysis, and need not be specified. (In terms of the discussion above, $\Omega_i = H_i H_i'$, and the pair $H_i^{-1} (\varepsilon_i, \zeta_i)'$ has the density K_i. We discuss later what values to choose for Ω_i. Here we say only that a good choice would reflect the covariance structure of points near (x_i, y_i).) Then $\left(X^S, Y^S\right) = (x_I + \varepsilon_I, y_I + \zeta_I) = \left(X^E + \varepsilon_I, Y^E + \zeta_I\right)$.

The kernel-density approximations $\hat{f}_{X,Y}$ and $\hat{f}_X$ can be interpreted as the exact densities f_{X^S,Y^S} and f_{X^S} respectively; hence without assuming either the product form for $K_i(x, y)$, or a diagonal form for $\mathbf{H}_i$, we obtain $\frac{\int y \hat{f}_{X,Y}(x,y) dy}{\hat{f}_X(x)} = E\left(Y^S | X^S = x\right)$, a quantity we write as $\hat{m}_S(x)$. In general $\hat{m}_S(x)$ differs from the Nadaraya-Watson estimator given by equations (1) and (2). To interpret the right-hand-sides of these equations, note that $\varepsilon_i / h_{X,i}$ has the marginal density $K_{X,i} = \int K_i dy$. Then by Bayes' theorem

$$P\left(I = j | X^S = x\right) = \frac{\operatorname{lik}\left(X^S = x | I = j\right) P(I = j)}{\sum \operatorname{lik}\left(X^S = x | I = i\right) P(I = i)} = p_j(x) \qquad (4)$$

showing that $p_j(x)$ yields the conditional probability that $I = j$, given that $X^S = x$. From (1), (3) and (4) we obtain

$$\hat{m}_{NW}(x) = \sum y_i P\left(I = i | X^S = x\right) = E\left(Y^E | X^S = x\right). \qquad (5)$$

This exact probabilistic interpretation of $\hat{m}_{NW}(x)$ draws attention to three of its characteristics. First, the right-hand side of (5) explains the fact, noted earlier, that the form of kernel-smoothing adopted for the y_i does not affect $\hat{m}_{NW}(x)$. Second, it shows that $\hat{m}_{NW}(x)$ is defined by a slightly unbalanced expression, in that it conditions an empirical random variable on a smoothed one. Finally, the central expression of (5) shows that $\hat{m}_{NW}(x)$ is confined to the convex hull of the y_i. This property causes $\hat{m}_{NW}(x)$ to flatten out in the tails even when the points of the scatter-plot of the (x_i, y_i) lie nearly on a sloping straight line.

As we have seen, the estimator

$$\hat{m}_S(x) \equiv E\left(Y^S | X^S = x\right), \tag{6}$$

is the appropriate generalization of $\hat{m}_{NW}(x)$, to non-independent perturbations (or in other words, to non-separable kernel functions K_i). Comparison of (5) and (6) shows that $\hat{m}_{NW}(x)$ looks at the empirical distribution of the y_i, while $\hat{m}_S(x)$ looks at their smoothed distribution. Thus $\hat{m}_S(x)$ is less likely than $\hat{m}_{NW}(x)$ to be strongly influenced by particular observations.

To obtain a closed-form expression for $\hat{m}_S(x)$ without imposing independence on the perturbations ε_i and ζ_i, we instead impose bivariate normality, by specifying $(\varepsilon_i, \zeta_i)' \sim N_2(\mathbf{0}_2, \mathbf{\Omega}_i)$. Let $\mathbf{\Omega}_i \equiv \begin{pmatrix} r_i & s_i \\ s_i & t_i \end{pmatrix}$. Then the distribution of ζ_i conditional on $\varepsilon_i = x - x_i$ is $N\left(\frac{s_i}{r_i}(x - x_i), t_i - \frac{s_i^2}{r_i}\right)$. Now $E\left(Y^S | X^S = x\right) = \sum E\left(Y^S | I = i, X^S = x\right) p_i(x)$. Using Bayes' theorem, as above,

$$p_j(x) = \frac{r_j^{-1/2} \exp\left(-\frac{(x-x_j)^2}{2r_j}\right)}{\sum r_i^{-1/2} \exp\left(-\frac{(x-x_i)^2}{2r_i}\right)}.$$

It is convenient to define $\hat{Y}_i^S \equiv \hat{Y}_i^S(x) \equiv E\left(Y^S | I = i, X^S = x\right) = y_i + \frac{s_i}{r_i}(x - x_i)$. With this notation,

$$\hat{m}_S(x) = \sum \hat{Y}_i^S p_i(x) = \hat{m}_{NW}(x) + \sum \frac{s_i}{r_i}(x - x_i) p_i(x). \tag{7}$$

As expected, $\hat{m}_S(x)$ and $\hat{m}_{NW}(x)$ are identical if the local covariances s_i are zero.

The discussion above extends naturally to the estimation of the conditional variance $V(x) \equiv Var(Y | X = x)$. As we show in the next section, the natural counterparts of (5) and (6) are

$$\hat{V}_{NW}(x) = Var\left(Y^E | X^S = x\right) = \sum y_i^2 p_i(x) - \hat{m}_{NW}(x)^2 \tag{8}$$

$$\hat{V}_S(x) = Var\left(Y^S | X^S = x\right) = \sum \left[t_i - \frac{s_i^2}{r_i} + \left(\hat{Y}_i^S\right)^2\right] p_i(x) - \hat{m}_S(x)^2 \tag{9}$$

The local covariance matrices $\mathbf{\Omega}_i$ are under our control, and we now consider how they might be chosen. This is really a separate subject, and we provide only introductory comments. One possibility is to make the $\mathbf{\Omega}_i$ all equal, and all proportional to the sample variance-covariance matrix $\hat{\Sigma}_{XY}$ of the (x_i, y_i). If x_i and y_i have, overall, a high level of correlation, we might expect $\hat{m}_S(x)$ with a constant $\mathbf{\Omega}$ to outperform $\hat{m}_{NW}(x)$, since the latter works best when $\mathbf{\Omega}$ is diagonal. These comments hold true in particular if $f_{X,Y}$ has the bivariate normal shape. Figures 1a, 1b, 1c show different estimated $\hat{m}(x)$, and indicative error bands $\hat{m}(x) \pm 2\sqrt{\hat{V}(x)}$, for a sample of observations (x_i, y_i) drawn independently from the distribution $N\left(\mathbf{0}_2, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right)$. The matrices $\Omega_i = \Omega$ were all set equal to the sample variance matrix, scaled so that the bandwidth $h_X = \Omega_{11}^{1/2}$ equals $1.059\sigma_X n^{-1/5} \cong 0.58$. (Here we follow the optimizing bandwidth criterion for normal samples given by Simonoff 1996, p. 45, with $\sigma = 1$.) The particular sample chosen for graphical portrayal was obtained by generating 101 samples of 20 observations each, and selecting the median sample, on the criterion statistic $\sum(y_i - \hat{m}_S(x_i))^2 / \sum(y_i - \hat{m}_{NW}(x_i))^2$. Sample t was generated using Stata's random seed t. We designed this automatic procedure in order to guard against selection of a sample particularly favourable to $\hat{m}_S(x)$ over $\hat{m}_{NW}(x)$, though we admit to having chosen a general context - small sample, highly correlated observations, unproblematic choice of Ω_i - where we expected the advantages of smoothed-distribution estimation over empirical-distribution estimation to be most clearly demonstrable.

Figures 1a, 1b, 1c about here.

Figure 1a shows the results from linear regression of y_i on x_i, including an intercept term; Figure 1b shows the Nadaraya-Watson estimates (5) and (8); and Figure 1c shows the corresponding smoothed-distribution estimates (6) and (9). Both Figure 1b and Figure 1c show the tendency, that we noted above, of $\hat{m}_{NW}(x)$ to flatten out to the right and left of the scatter plot, even when the observations exhibit a strong linear trend. The indicative error bands tighten to the right of Figure 1b because of the isolated point near $(3, 2.5)$. What happens is that for $x > 3$, the probability $p_i(x)$ tends to 1 for the relevant i, thus $Var\left(Y^E | X^S = x\right)$ tends to 0. The same phenomenon is visible, though less marked, in the other tail. Other samples show different behaviour: if the points at the right and left of the sample are more numerous and more dispersed, the Nadaraya-Watson indicative dispersion band stays wider. For the smoothed-distribution estimator $\hat{m}_S(x)$ in Figure 1c, the estimated regression line and dispersion band are much closer to those in

Figure 1a. In 99 of the 101 samples, the sum of squared errors was lower under smoothed-distribution estimation than under Nadaraya-Watson estimation.

The case of a normal data-generating distribution provides just one initial testing-ground for a kernel-density method, though we do want such a method to perform well in this simple context. In general $f_{X,Y}$ may have a shape far from that of the normal density, for instance having markedly non-elliptical, even non-convex, contours or being multimodal. Such possibilities are of course a major motive for introducing kernel-density methods in the first place. Thus in general we wish to adapt the $\mathbf{\Omega}_i$ to local conditions. The problem we face is a generalization of the scalar kernel-density problem of choosing the local bandwidth h_i. For discussion of the scalar version of this problem, see Simonoff (1996, pp. 54-6). And for discussion of the multivariate version, see Simonoff (1996, pp. 105, 114). A natural approach is to let $\mathbf{\Omega}_i$ be proportional to the sample covariance matrix of the (x_i, y_i), weighted by $p_i(x_i)$. In more detail, we could define

$$\mathbf{\Omega}(x) \propto \begin{pmatrix} Var\left(X^E|X^P = x\right) & Cov\left(X^E, Y^E|X^P = x\right) \\ Cov\left(Y^E, X^E|X^P = x\right) & Var\left(Y^E|Y^P = x\right) \end{pmatrix}$$

where X^P has a 'pilot' kernel-density function (Simonoff, 1996, p. 55), used to obtain an initial approximation to f_X. We could impose a suitable bandwidth on X^P by the criterion of Simonoff already alluded to. Then set $\mathbf{\Omega}_i \equiv \mathbf{\Omega}(x_i)$. A natural way to choose the scaling constant is to impose the condition $Var\left(X^S\right) = Var\left(X^P\right)$. But here we do not pursue this large topic further.

We now generalize our four estimators to the case where the observations are vector-pairs $(\mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^q)$. Assume that the perturbation vector $\left(\boldsymbol{\varepsilon}_i' \vdots \boldsymbol{\zeta}_i'\right)'$ has a multivariate normal distribution with zero means and $(p+q)\times(p+q)$ covariance matrix $\mathbf{\Omega}_i$, partitioned as $\begin{pmatrix} \mathbf{R}_i & \mathbf{S}_i' \\ \mathbf{S}_i & \mathbf{T}_i \end{pmatrix}$, where $\mathbf{R}_i$ is $p \times p$, $\mathbf{S}_i$ is $q \times p$ and $\mathbf{T}_i$ is $q \times q$. Bayes' theorem tells us that the posterior probability that $\mathbf{X}^S$ arises as a perturbation of the particular observation $\mathbf{x}_j$ is

$$p_j(\mathbf{x}) = P\left(I = j|\mathbf{X}^S = \mathbf{x}\right) = \frac{|\mathbf{R}_j|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_j)' \mathbf{R}_j^{-1} (\mathbf{x} - \mathbf{x}_j)\right)}{\sum |\mathbf{R}_i|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)' \mathbf{R}_i^{-1} (\mathbf{x} - \mathbf{x}_i)\right)}. \quad (10)$$

It is convenient to define

$$\hat{\mathbf{Y}}_i^S \equiv \hat{\mathbf{Y}}_i^S(\mathbf{x}) \equiv E\left(\mathbf{Y}^S|\mathbf{X}^S = \mathbf{x}, I = i\right) = \mathbf{y}_i + \mathbf{S}_i\mathbf{R}_i^{-1}(\mathbf{x} - \mathbf{x}_i).$$

The resulting conditional mean and conditional variance estimators, derived in the next section, are

$$\begin{aligned}
\hat{\mathbf{m}}_{NW}(\mathbf{x}) &\equiv E\left(\mathbf{Y}^E|\mathbf{X}^S = \mathbf{x}\right) = \sum \mathbf{y}_i p_i(\mathbf{x}) \\
\hat{\mathbf{m}}_S(\mathbf{x}) &\equiv E\left(\mathbf{Y}^S|\mathbf{X}^S = \mathbf{x}\right) = \sum \hat{\mathbf{Y}}_i^S p_i(\mathbf{x}) \\
&= \hat{\mathbf{m}}_{NW}(\mathbf{x}) + \sum \mathbf{S}_i\mathbf{R}_i^{-1}(\mathbf{x} - \mathbf{x}_i) p_i(\mathbf{x}) \\
\hat{\mathbf{V}}_{NW}(\mathbf{x}) &\equiv Var\left(\mathbf{Y}^E|\mathbf{X}^S = \mathbf{x}\right) = \sum \mathbf{y}_i\mathbf{y}_i' p_i(\mathbf{x}) - \hat{\mathbf{m}}_{NW}(\mathbf{x})\hat{\mathbf{m}}_{NW}(\mathbf{x})' \\
\hat{\mathbf{V}}_S(\mathbf{x}) &\equiv Var\left(\mathbf{Y}^S|\mathbf{X}^S = \mathbf{x}\right) \\
&= \sum \left[\mathbf{T}_i - \mathbf{S}_i\mathbf{R}_i^{-1}\mathbf{S}_i' + \hat{\mathbf{Y}}_i^S\hat{\mathbf{Y}}_i^{S\prime}\right] p_i(\mathbf{x}) - \hat{\mathbf{m}}_S(\mathbf{x})\hat{\mathbf{m}}_S(\mathbf{x})'. \quad (11)
\end{aligned}$$

Finally we comment briefly on local polynomial estimation, a generalization of Nadaraya-Watson estimation. Such estimation involves minimization of the function $\sum (y_i - g(x - x_i))^2 K\left(\frac{x - x_i}{h}\right)$, where g is a polynomial. Once the parameters of g have been estimated, $m(x)$ is estimated by $\hat{g}(0)$. (Nadaraya-Watson estimation is the case when g is of degree 0.) The procedure amounts to fitting a polynomial curve through the (x_i, y_i) by weighted regression, the weights being high near x. Since the weights $K\left(\frac{x - x_i}{h}\right)$ are proportional to $p_i(x)$, we could equivalently minimize the objective function $\sum (y_i - g(x - x_i))^2 p_i(x) = E\left[\left(Y^E - g\left(x - X^E\right)\right)^2 |X^S = x\right]$, a formulation that still holds when we generalize to variable bandwidths. We might ask what happens if the empirical variables X^E and Y^E are replaced by the smoothed variables X^S and Y^S. If both changes are made, the objective function reduces to $E\left[\left(Y^S - g(0)\right)^2 |X^S = x\right]$, which is minimized by $\hat{g}(0) = E\left(Y^S|X^S = x\right) = \hat{m}_S(x)$, the smoothed-distribution estimator already discussed. To avoid this simplification we could replace Y^E by Y^S but leave the first occurrence of X^E unchanged. If we do so, we obtain the objective function $E\left[\left(Y^S - g\left(x - X^E\right)\right)^2 |X^S = x\right]$. Estimation of $m(x)$ by minimization of this expression merits study as a promising variant of local polynomial estimation.

3 Proofs

To show how equation (1) follows from the assumption that K_i has the product form, note that in these circumstances

$$\hat{f}_{X,Y}(x,y) = \frac{1}{n}\sum \frac{1}{h_{X,i}h_{Y,i}} K_{X,i}\left(\frac{x-x_i}{h_{X,i}}\right) K_{Y,i}\left(\frac{y-y_i}{h_{Y,i}}\right)$$

and integrating out y yields

$$\hat{f}_X(x) = \frac{1}{n}\sum \frac{1}{h_{X,i}} K_{X,i}\left(\frac{x-x_i}{h_{X,i}}\right).$$

Since $\frac{1}{h_{Y,i}}\int yK_{Y,i}\left(\frac{y-y_i}{h_{Y,i}}\right)dy = y_i$, the final integral required is

$$\int y\hat{f}_{X,Y}(x,y)\,dy = \frac{1}{n}\sum \frac{y_i}{h_{X,i}} K_{X,i}\left(\frac{x-x_i}{h_{X,i}}\right).$$

Replacing $f_{X,Y}$ and $f_X(x)$ by $\hat{f}_{X,Y}$ and $\hat{f}_X$ in the exact expression $m(x) = \int yf_{X,Y}(x,y)\,dy/f_X(x)$ yields (1).

To prove the equations (11), we use a form of the law of total variance. If a random variable $\mathbf{W}$ is of exactly one of n types, and type i occurs with probability p_i and has mean $\boldsymbol{\mu}_i$ and variance $\mathbf{V}_i$, then $E(\mathbf{W}) = \sum p_i\boldsymbol{\mu}_i$ and $Var(\mathbf{W}) = \sum p_i(\mathbf{V}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i') - (\sum p_i\boldsymbol{\mu}_i)(\sum p_i\boldsymbol{\mu}_i)'$.

In the present context, the type is the value of I. Conditional on $\mathbf{X}^S = \mathbf{x}$, type i occurs with probability $p_i(\mathbf{x})$. In these circumstances, the unconditional distribution of $(\boldsymbol{\varepsilon}_i', \boldsymbol{\zeta}_i')'$ is by assumption $N\left(\mathbf{0}_{p+q}, \begin{pmatrix} \mathbf{R}_i & \mathbf{S}_i' \\ \mathbf{S}_i & \mathbf{T}_i \end{pmatrix}\right)$. Conditional on $I = i$ and $\mathbf{X}^S = \mathbf{x}$, implying $\boldsymbol{\varepsilon}_i = \mathbf{x} - \mathbf{x}_i$, $\boldsymbol{\zeta}_I$ has mean $\mathbf{S}_i\mathbf{R}_i^{-1}(\mathbf{x}-\mathbf{x}_i)$ and variance $\mathbf{T}_i - \mathbf{S}_i\mathbf{R}_i^{-1}\mathbf{S}_i'$; thus (under the same conditions) $\mathbf{Y}^E$ has mean $\mathbf{y}_i$ and variance $\mathbf{0}$; while $\mathbf{Y}^S$ has mean $\hat{\mathbf{Y}}_i^S$ and variance $\mathbf{T}_i - \mathbf{S}_i\mathbf{R}_i^{-1}\mathbf{S}_i'$. The results (11) follow immediately. The equations (7) and (9) are the scalar special case.

Acknowledgments

Both authors owe a debt of gratitude to their colleague Stan Siebert for his sustained encouragement and assistance. The first author thanks Jacqueline Smith for her invaluable support during the writing of this paper.

References

Barrett, C. B. and Dorosh P. A., 1996. Farmers' welfare and changing food prices: nonparametric evidence from rice in Madagascar. Amer. J. Agr. Econ. 78, 656-669.

Corzo Santamaría, T. and Gómez Biscarri, J., 2001. Nonparametric estimation of interest rate processes in Europe. Revista de Economía Aplicada 27, 83-101.

Engle, R. F., 1982. Autoregressive, conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987-1007.

Engle, R. F. and Bollerslev, T., 1986. Modeling the persistence of conditional variances. Econ. Rev. 5(1), 1-50.

Fiorio, C. V., 2004. Confidence intervals for kernel density estimation. Stata J. 4(2) 168-179.

Giannopoulos, K., 2008. Nonparametric, conditional pricing of higher order multivariate contingent claims. J. Bank. Fin. 32, 1907-1915.

Linton, O. B. and Yan, Y., 2011. Semi- and nonparametric ARCH processes. J Prob. Stat. Volume 2011, Article ID 906212. DOI:10.1155/906212.

Nadaraya, E. A., 1964. Some new estimates for distribution functions. Theory Probab. Appl. 15, 497-500.

Parzen, E., 1963. On estimation of a probability density function and mode. Ann. Math. Stat. 33(3), 1065-1076.

Simonoff, J. S., 1996. Smoothing methods in statistics. New York: Springer.

Wasserman, L., 2006. All of nonparametric statistics. DOI: 10.1007/0-387-30623-4.

Watson, G.S., 1964. Smooth regression analysis. Sankhya Ser. A 26, 359-372.

Fig. 1a. Linear Regression Estimates, Normal Sample (N=20, rho=0.8).

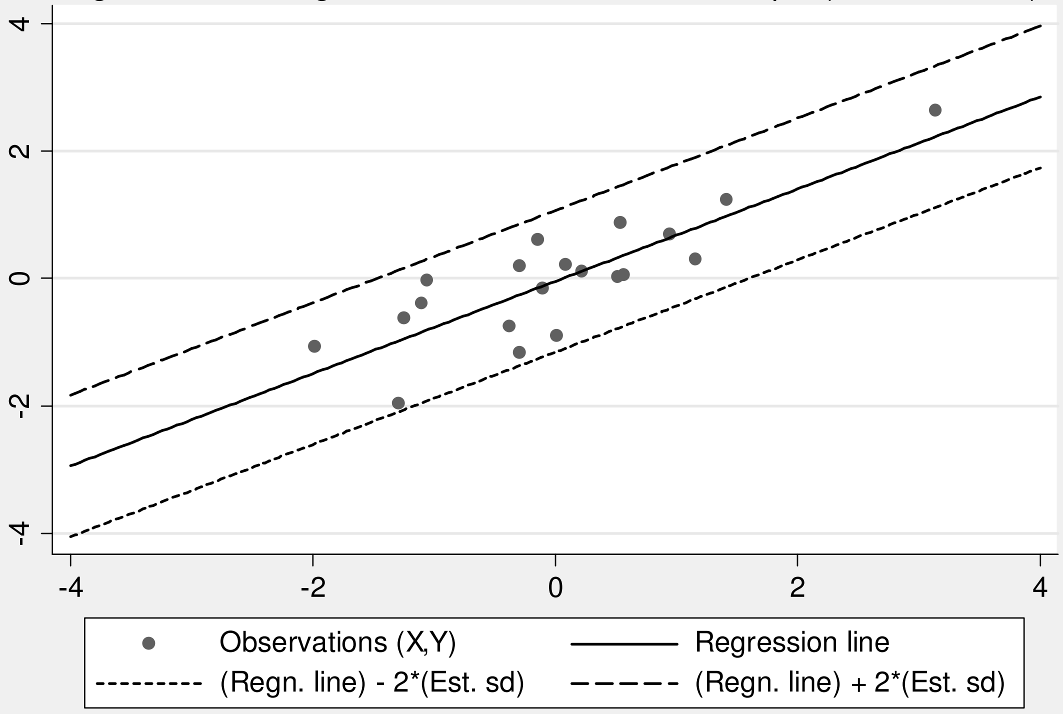
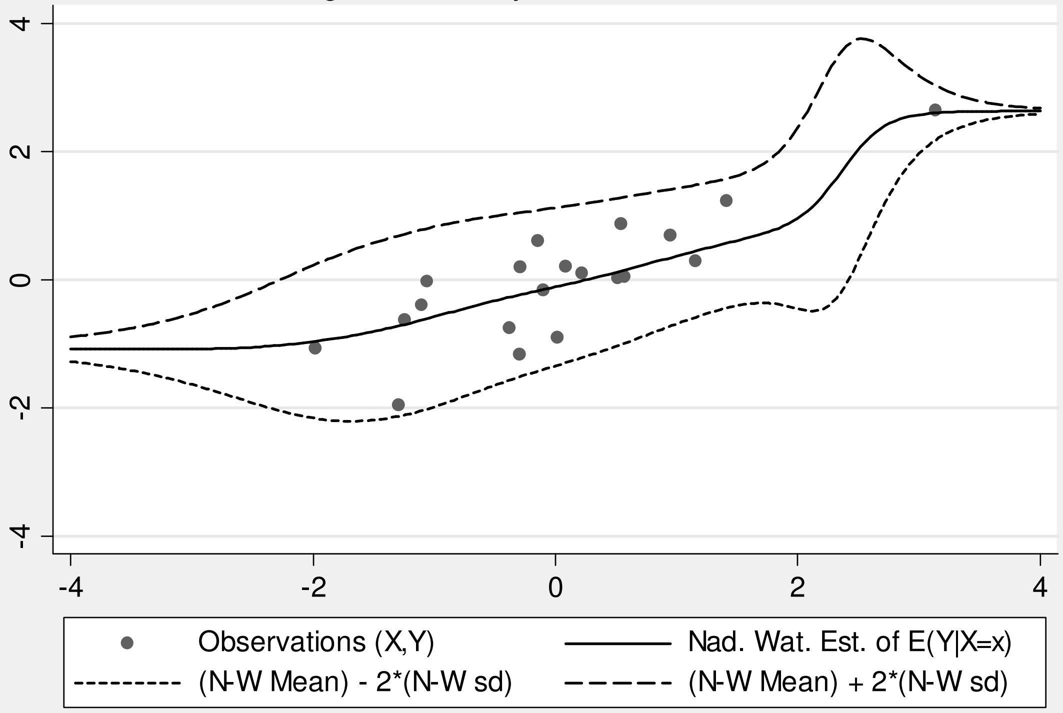


Fig. 1b. Nadaraya-Watson Estimates.



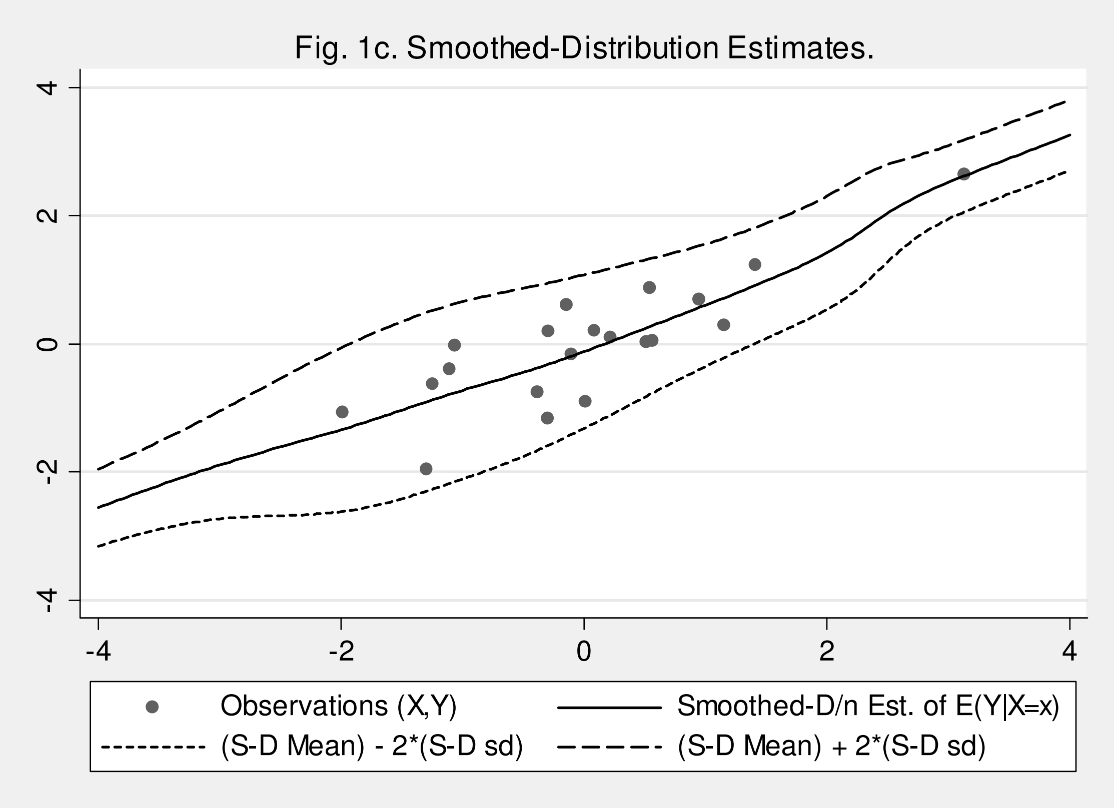


Fig. 1c. Smoothed-Distribution Estimates.
4
2
0
-2
-4
-4
-2
0
2
4
Observations (X,Y)
Smoothed-D/n Est. of E(Y|X=x)
(S-D Mean) - 2*(S-D sd)
(S-D Mean) + 2*(S-D sd)

ESTUDOS DO G.E.M.F.

(*Available on-line at* http://gemf.fe.uc.pt)

2011-01 *A Smoothed-Distribution Form of Nadaraya-Watson Estimation*
- Ralph W. Bailey & John T. Addison

2010-22 *Business Survival in Portuguese Regions*
- Alcina Nunes & Elsa de Morais Sarmento
2010-21 *A Closer Look at the World Business Cycle Synchronization*
- Pedro André Cerqueira
2010-20 *Does Schumpeterian Creative Destruction Lead to Higher Productivity? The effects of firms' entry*
- Carlos Carreira & Paulino Teixeira
2010-19 *How Do Central Banks React to Wealth Composition and Asset Prices?*
- Vítor Castro & Ricardo M. Sousa
2010-18 *The duration of business cycle expansions and contractions: Are there change-points in duration dependence?*
- Vítor Castro
2010-17 *Water Pricing and Social Equity in Portuguese Municipalities*
- Rita Martins, Carlota Quintal, Eduardo Barata & Luís Cruz
2010-16 *Financial constraints: Are there differences between manufacturing and services?*
- Filipe Silva & Carlos Carreira
2010-15 *Measuring firms' financial constraints: Evidence for Portugal through different approaches*
- Filipe Silva & Carlos Carreira
2010-14 *Exchange Rate Target Zones: A Survey of the Literature*
- António Portugal Duarte, João Sousa Andrade & Adelaide Duarte
2010-13 *Is foreign trade important for regional growth? Empirical evidence from Portugal*
- Elias Soukiazis & Micaela Antunes
2010-12 *MCMC, likelihood estimation and identifiability problems in DLM models*
- António Alberto Santos
2010-11 *Regional growth in Portugal: assessing the contribution of earnings and education inequality*
- Adelaide Duarte & Marta Simões
2010-10 *Business Demography Dynamics in Portugal: A Semi-Parametric Survival Analysis*
- Alcina Nunes & Elsa Sarmento
2010-09 *Business Demography Dynamics in Portugal: A Non-Parametric Survival Analysis*
- Alcina Nunes & Elsa Sarmento
2010-08 *The impact of EU integration on the Portuguese distribution of employees' earnings*
- João A. S. Andrade, Adelaide P. S. Duarte & Marta C. N. Simões
2010-07 *Fiscal sustainability and the accuracy of macroeconomic forecasts: do supranational forecasts rather than government forecasts make a difference?*
- Carlos Fonseca Marinheiro
2010-06 *Estimation of Risk-Neutral Density Surfaces*
- A. M. Monteiro, R. H. Tütüncü & L. N. Vicente
2010-05 *Productivity, wages, and the returns to firm-provided training: who is grabbing the biggest share?*
- Ana Sofia Lopes & Paulino Teixeira
2010-04 *Health Status Determinants in the OECD Countries. A Panel Data Approach with Endogenous Regressors*
- Ana Poças & Elias Soukiazis
2010-03 *Employment, exchange rates and labour market rigidity*
- Fernando Alexandre, Pedro Bação, João Cerejeira & Miguel Portela
2010-02 *Slip Sliding Away: Further Union Decline in Germany and Britain*
- John T. Addison, Alex Bryson, Paulino Teixeira & André Pahnke

2010-01 *The Demand for Excess Reserves in the Euro Area and the Impact of the Current Credit Crisis*
- Fátima Teresa Sol Murta & Ana Margarida Garcia

2009-16 *The performance of the European Stock Markets: a time-varying Sharpe ratio approach*
- José A. Soares da Fonseca
2009-15 *Exchange Rate Mean Reversion within a Target Zone: Evidence from a Country on the Periphery of the ERM*
- António Portugal Duarte, João Sousa Andrade & Adelaide Duarte
2009-14 *The Extent of Collective Bargaining and Workplace Representation: Transitions between States and their Determinants. A Comparative Analysis of Germany and Great Britain*
- John T. Addison, Alex Bryson, Paulino Teixeira, André Pahnke & Lutz Bellmann
2009-13 *How well the balance-of- payments constraint approach explains the Portuguese growth performance. Empirical evidence for the 1965-2008 period*
- Micaela Antunes & Elias Soukiazis
2009-12 *Atypical Work: Who Gets It, and Where Does It Lead? Some U.S. Evidence Using the NLSY79*
- John T. Addison, Chad Cotti & Christopher J. Surfield
2009-11 *The PIGS, does the Group Exist? An empirical macroeconomic analysis based on the Okun Law*
- João Sousa Andrade
2009-10 *A Política Monetária do BCE. Uma estratégia original para a estabilidade nominal*
- João Sousa Andrade
2009-09 *Wage Dispersion in a Partially Unionized Labor Force*
- John T. Addison, Ralph W. Bailey & W. Stanley Siebert
2009-08 *Employment and exchange rates: the role of openness and technology*
- Fernando Alexandre, Pedro Bação, João Cerejeira & Miguel Portela
2009-07 *Channels of transmission of inequality to growth: A survey of the theory and evidence from a Portuguese perspective*
- Adelaide Duarte & Marta Simões
2009-06 *No Deep Pockets: Some stylized results on firms' financial constraints*
- Filipe Silva & Carlos Carreira
2009-05 *Aggregate and sector-specific exchange rate indexes for the Portuguese economy*
- Fernando Alexandre, Pedro Bação, João Cerejeira & Miguel Portela
2009-04 *Rent Seeking at Plant Level: An Application of the Card-De La Rica Tenure Model to Workers in German Works Councils*
- John T. Addison, Paulino Teixeira & Thomas Zwick
2009-03 *Unobserved Worker Ability, Firm Heterogeneity, and the Returns to Schooling and Training*
- Ana Sofia Lopes & Paulino Teixeira
2009-02 *Worker Directors: A German Product that Didn't Export?*
- John T. Addison & Claus Schnabel
2009-01 *Fiscal and Monetary Policies in a Keynesian Stock-flow Consistent Model*
- Edwin Le Heron

A série Estudos do GEMF foi iniciada em 1996.