Nonparametric Regression under Alternative Data Environments

By
Abdoul G. Sam and Alan P. Ker

University of Arizona

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Denver, Colorado, July 1-4, 2004

Correspondence: Abdoul G. Sam, Department of Economics, University of Arizona, Tucson AZ, 85721. E-mail: abdoul@u.arizona.edu. Tel: (520) 621 6256.

Nonparametric Regression under Alternative Data Environments

Abdoul G. Sam and Alan P. Ker *

May 15, 2004

Abstract

This paper proposes a nonparametric bias-reduction regression estimator which can accommodate two empirically relevant data environments. The first data environment assumes that at least one of the predictor variables is discrete. In such an empirical framework, a "cell" approach, which consists of estimating a separate regression for each discrete cell has generally been employed. However, the "cell" estimator may be inefficient in that it does not include data from the other cells when estimating the regression function for a given cell. The second data environment assumes that the researcher is faced with a system of regression functions that belong to different experimental units. In each case, the new estimator attempts to reduce estimation error by incorporating extraneous data from the remaining experimental units (or cells) when estimating a given individual regression function. Consistency of the proposed estimator is established and Monte Carlo simulations demonstrate its strong finite sample performance.

Keywords: Extraneous information, bias reduction, correction factor, data environments.

*The authors are Ph.D. Student and Professor, respectively, in the Departments of Economics and Agricultural and Resource Economics, University of Arizona.

1 Introduction.

Consider $\{(X_i, Y_i)\}_{i=1}^n$ a sequence of independently and identically distributed R^{p+1}-valued random vectors where Y_i represents the response variable and X_i an R^p-valued vector of predictor variables. In this paper, we concern ourselves with the estimation of the conditional mean $E(Y|X = x)$. In doing so the following specification is assumed:

$$Y_i = m(X_i) + \sigma\epsilon_i \tag{1}$$

where ϵ_i is a zero-mean unit variance independently and identically distributed error process. For notational simplicity we will allow x_i and y_i to denote both the random variables X_i and Y_i and their sample counterpart. Parametric estimation of the conditional mean requires an assumption about the form of the data generating process (DGP). This assumption is the source of both the strength ($\sqrt{n}$-convergence rate under the null) and weakness (mispecification under the alternative) of parametric methods. On the other hand, nonparametric regression methods such as kernel estimators have become widespread because they eliminate the issue of mispecification and are consistent under mild regularity conditions.

Let x_i in (**??**) be R-valued, then the Nadaraya-Watson (NW) estimator of the conditional mean is given by:

$$\tilde{m}(x) = \frac{\sum_{i=1}^n y_i K_h(x_i - x)}{\sum_{i=1}^n K_h(x_i - x)} \tag{2}$$

where h is the smoothing parameter and $K_h(u) = \frac{1}{h}K(\frac{u}{h})$ with$K(u)$ being the Kernel function. Denoting $\mu_2 = \int z^2 K(z)dz$ and $\mathrm{R}(K) = \int K^2(z)dz$, the standard properties of the NW estimator are:

$$E(\tilde{m}(x)) - m(x) = \frac{1}{2}\mu_2 h^2 m''(x) + 2m'(x)\frac{f'(x)}{f(x)} + o(h^2)$$

$$Var(\tilde{m}(x)) = \sigma^2(nhf)^{-1}R(K) + O(h/n)$$

where f is the probability density function of x. Since the bias is $O(h^2)$ and $h = h(n)$ goes to 0 as n becomes large, it is follows that the NW is consistent.

However, the downside of the NW estimator is its finite sample bias which can be quite large. In empirical applications, the finite sample bias constitutes the main concern when using Kernel estimators in general. Several papers have proposed estimators which reduce (Hardle and Brownman, 1988; Hjort and Glad, 1995; Glad, 1998) or eliminate Racine (2001) the bias of Kernel methods.

In this paper we propose a new type of nonparametric estimator with two attractive features; the new estimator may substantially lower the bias of the Kernel estimators and accommodates alternative data environments. The first data environment assumes that at least one of the explanatory variables is discrete. While this situation is easily accommodated in a parametric framework, the continuity assumptions required for nonparametric regression are violated. As a result, a separate nonparametric regression estimation is required for each unique discrete value. That is if one of the explanatory variables is discrete and can only take values $\{1, 2, 3, 4\}$, then the sample data must be partitioned into four subsets corresponding to four discrete values. In the nonparametric literature this is termed the "cell" estimator as a separate estimation is required for each of the four cells. As a result, the nonparametric regression estimation for one subset of data does not utilize the data in the other subsets. Recently, Racine and Li (2003) have developed a nonparametric estimator that smoothes across the discrete values this reducing variance at a cost of an increased bias. Conversely, our proposed estimator attempts to reduce the bias by making use of the entire data set in estimating the shape of the regression curves. The second data environment assumes that one is required to estimate a set of regression

curves rather than a single regression curve. Empirically, this situation arises often and led to the famous Stein' s Paradox. Altman and Casella (1995) have developed a Stein-type Bayesian nonparametric estimator that uses empirical Bayes techniques pointwise across the function space to reduce estimation error. A requirement of the Altman and Casella estimator is that the design be fixed. In essence the latter data environment can be viewed as a generalization of the former with each of the discrete cells representing an experimental unit.

The remainder of the paper is organized as follows. The second section outlines the Racine and Li and the Altman and Casella estimators while the third section lays out the proposed estimator and investigates its asymptotic properties. The fourth section studies the finite sample performances of the competing estimators: our estimator, the Racine and Li, the Altman and Casella and the Locally linear Kernel estimators. Finally, the fifth section summarizes the findings.

2 Nonparametric Estimation of Multiple Curves

In many relevant empirical studies we must estimate a set of regression curves, say one for each experimental unit of interest. These regression curves can be arranged into a system of equations as follows:

$$y_{ij} = m_j(x_{ij}) + \sigma_j \epsilon_{ij}, i = 1...n_j, j = 1...Q \tag{3}$$

where j references the j^{th} experimental unit, y_{ij} is the scalar response variable, x_{ij} is the vector of explanatory variables, $m_j(x_{ij})$ is the true conditional mean function and ϵ_{ij} a zero-mean and unit variance independent and identically distributed error process. Note that the two data environments fall under this model with the latter being directly related while the former simply sets Q, the number of curves, equal to the

number of values the discrete explanatory variable can take. Standard Kernel regression techniques (NW, Locally linear Kernel, Gasser-Müller) estimate each individual conditional mean separately thus ignoring the data from the other experimental units. However, if the conditional means $m_1, m_2 ... m_Q$ are functionally similar, it would be inefficient not to exploit the abundance of information to overcome the paucity of data. A limited number of studies have proposed estimators built upon the idea of using of extraneous data to improve the efficiency of individual estimates. For example Hart and Wehrly (1986) used hourly measurements of plasma citrate for a sample of 10 human subjects to estimate a population mean plasma concentration as function daytime. Abaffy and al (2003) evaluate the position of eleven European Union members in the Euro bond market by assuming that their underlying yield curves can be nonparametrically modeled as a sum of an individual factor and a common factor. The common factor captures cross-country similarities, which resulted from the elimination of exchange rate risk (due to the launching of the Euro). Ker (2000) proposed a Stein-type empirical Bayes estimator for crop insurance rating which uses extraneous yield data from counties that belong to the same crop-reporting district than the county of interest. The rationale for the inclusion of extraneous yield data is the belief that these individual yield densities are functionally similar as they are sampled from the same population (crop-reporting district in the latter case) where factors such as weather pattern, soil type, technology, chemical products used etc. are comparable. Hence when estimating the density of county j, it seems reasonable to use information from the remaining counties in the same crop-reporting district for potential efficiency gains. This section outlines the Racine and Li (2003) nonparametric estimator and the Altman and Casella (1995) semiparametric estimator; which both accommodate multiple curve estimation.

2.1 The Racine and Li Estimator

This estimator is motivated by the failure of nonparametric methods to estimate categorical data satisfactorily. The goal is to adequately estimate regression functions with many discrete independent variables without having to split the data into subsets the number of which depends on the values of the categorical variables. Suppose we have data on one experimental unit: y_i a scalar response variable, x_i^c a vector of continuous variables and x_i^d an r-dimensional vector of discrete regressors. The continuous variables are smoothed using a c-variate kernel function while the discrete variables are smoothed as follows:

$$S(x_{it}^d, x_t^d, \lambda) = \begin{cases} 1 & \text{if } x_{it}^d = x_t^d \\ \lambda & \text{otherwise }, 0 \leq \lambda \leq 1 \end{cases} \tag{4}$$

where x_{it}^d is the t^{th} component of the vector x_i^d. The resulting estimator is a NW estimator with a product weight function:

$$\tilde{m}^{RL}(x^c, x^d) = \frac{\sum_{i=1}^n y_i W_{h,\lambda}(x_i^c, x^c, x_i^d, x^d)}{\sum_{i=1}^n W_{h,\lambda}(x_i^c, x^c, x_i^d, x^d)} \tag{5}$$

where $W_{h,\lambda}(x_i^c, x^c, x_i^d, x^d) = K_h(x_i^c - x^c) \prod_{t=1}^r S(x_{it}^d, x_t^d, \lambda)$.

The estimator can be easily adapted in a context of multiple equation estimation to allow the use of extraneous data. This is done first by vertically concatenating the observation pairs (y_{ij}, x_{ij}) in (**??**) and then generating a discrete "regressor" x_{ij}^d that references the experimental unit to which the pair (y_{ij}, x_{ij}) belongs, hence for a system of Q regression equations the domain of x_{ij}^d is $\{1, 2, ...Q\}$[1]. Then the Racine and Li estimator for curve j is:

$$\tilde{m}_j^{RL}(x) = \frac{\sum_{l=1}^N y_l W_{h_j,\lambda_j}(x_l, x, x_l^d, x^d)}{\sum_{l=1}^N W_{h_j,\lambda_j}(x_l, x, x_l^d, x^d)} \tag{6}$$

[1]It is implicitly assumed here that the vector of explanatory variables x_{ij} contains no discrete components; this assumption involves no loss of generality.

where $N = \sum_{j=1}^{Q} n_j$. When estimating the conditional mean for experimental unit j, the discrete smoother $S(x_{lt}^d, x_t^d, \lambda_j)$ controls the inclusion of extraneous information by assigning a weight λ_j $(0 \leq \lambda_j \leq 1)$ to the data belonging to the remaining experimental units. The boundedness of λ_j within the unit interval allows the Racine and Li estimator to nests both the pooled and the NW estimator. When the individual mean functions are dissimilar, λ_j should be as small as possible reverting the estimator back to the NW estimator. When the conditional means are functionally similar, the weight placed on the "external" observations should be close to 1 to reflect the similarities, hence boosting the performance of the estimates [2].

The smoothing parameters λ_j and h_j can be both chosen by minimizing the cross-validation function:

$$CV(h_j, \lambda_j) = \sum_{l=1}^{N} [y_l - \tilde{m}_{j(l)}^{RL}(x)]^2$$

where $\tilde{m}_{j(l)}^{RL}(x)$ is a leave-one out estimator.

2.2 The Altman and Casella Estimator

The Altman and Casella estimator is a nonparametric empirical Bayes estimator. It assumes that each experimental unit is measured at equispaced design points $x_i = i/n$ so that (3) can be rewritten as $y_{ij} = m_j(x_i) + \epsilon_{ij}$. It is also assumed that each curve can be written as $m_j(x_i) = m(x_i) + \eta_j(x_i)$; that is the curve for experimental unit j at design point x_i is the population mean curve plus a term which captures the deviation from the population mean curve. Underlying this last assumption is

[2] The pooled estimator vertically concatenates the entire data set to estimate a unique conditional mean without accounting for potential dissimilarities between the regression functions. At the contrary the ordinary Kernel estimators treat the different regression functions as separate and estimate individual conditional means. When λ_j goes to zero, it simply means that the extraneous data is getting a weight of zero: $S(x_{lt}^d, x_t^d, \lambda_j)$ goes to zero for the extraneous data while the observations belonging to the equation of interest get a weight of $S(x_{lt}^d, x_t^d, \lambda_j) = 1$ which amounts to using the NW estimator for each individual curve. When λ_j goes to 1, the estimator becomes the pooled estimator since $W_{h_j,\lambda_j}(x_l^c, x^c, x_l^d, x^d) = K_h(x_l^c - x^c)S(x_l^d, x^d, \lambda_j) = K_h(x_l^c - x^c)$.

the fact that the curves are all sampled from the same population hence share certain intrinsic characteristics. Denote $\tilde{m_j}$ the nonparametric estimator of m_j; since $\tilde{m_j}$ is typically biased, it can be expressed as $\tilde{m_j} = \phi_j + \upsilon_j$ where υ_j is an error term such that $E[\upsilon_j(i)] = 0$ and $var[\upsilon_j(i)] = \alpha^2/n$. Based on the asymptotic properties of the nonparametric estimator $\tilde{m}_j$, Altman and Casella form a hierarchical model ($\tilde{m}_j|\phi_j$ is normally distributed, and ϕ_j and m_j are jointly normally distributed) and derive the posterior mean of m_j:

$$\tilde{m}_j(x) = \bar{m}(x) + \alpha(x)[\tilde{m}_j(x) - \phi(x)] \tag{7}$$

In practice, the hyperparameters are replaced by sample estimates, which leads to the Altman and Casella estimator for experimental unit j:

$$\tilde{m}_j^{AC}(x) = \bar{y}_x + \tilde{\alpha}(x)[\tilde{m}_j(x) - \overline{\tilde{m}}(x)] \tag{8}$$

where $\bar{y}_x = \frac{1}{Q}\sum_{j=1}^{Q} y_{xj}$ is the cross-individual sample mean of the data at design point x, $\tilde{\alpha}(x) = \frac{\tilde{\sigma}_{y(x)m(x)}}{\tilde{\sigma}^2_{m(x)}}$ is the ratio of the covariance between the data and the nonparametric estimates and the variance of the nonparametric estimates, and $\overline{\tilde{m}}(x) = \frac{1}{Q}\sum_{j=1}^{Q}\tilde{m}_j(x)$. The reader is directed to Altman and Casella (1995) for a complete derivation of the model. Notice that this estimator uses the data from the other individuals in the population in the regression of the curve of interest through $\overline{\tilde{m}}(x)$and $\bar{y}_x$.

If the individual curves are similar then $[\hat{m}_i(t) - \overline{\hat{m}}(t)]$ goes to zero and the final estimates behave like $\bar{y}_x$ which is unbiased for the population mean curve. This estimator performs better when the number of experimental units is large enough so that $\bar{y}_x$ provides a good approximation of the population mean.

3 Nonparametric Estimator with a Pooled Start

Underlying our estimator is that we presume that the curves in a set are similar in shape without explicitly modeling the extent of similarity. If the curves were identical, that is if $m_1 = m_2 = ...m_Q = m$, the efficient estimator would pool the data and estimate one common curve for all the experimental units. Conversely, if the curves were not similar the efficient estimator would estimate a separate curve for each experimental unit. We are purposively vague with respect to the form or extent of similarity between the curves because in an empirical situation it is generally impossible to know if the curves are either identical, similar and to what extent and how, or completely dissimilar in structure. We have adapted the Hjort and Glad (1995) estimator to the situation of combining pooled and individual nonparametric estimators. As a result, the proposed estimator resembles the pooled estimate if the curves are identical and the individual estimate if the curves are dissimilar. Again, a major advantage of the proposed estimator is that the form or extent of similarity is not required to be known.

Without loss of generality, let the x_{ij} in (**??**) be R-valued, then we propose that the conditional mean for experimental unit j be estimated as follows:

$$\hat{m}_j(x) = \hat{m}_p(x)\hat{r}_j(x) = \frac{\sum_{i=1}^{n} y_{ij}[\frac{\hat{m}_p(x)}{\hat{m}_p(x_{ij})}]K_{h_j}(x_{ij} - x)}{\sum_{i=1}^{n} K_{h_j}(x_{ij} - x)} \tag{9}$$

The new estimator is implemented in two-steps. The first step pools the measurements from all experimental units to estimate a pooled estimator denoted $\hat{m}_p(x)$. The second step consists of multiplying the pooled estimator by a nonparametrically estimated correction factor $\hat{r}_j(x)$ to account for individual effects. The new estimator is designed so as to outperform standard Kernel methods when the hypothesis of similarity is tenable but also produce reliable estimates when the curves are dissimilar.

The motivation behind the construction of the estimator is the reduction of the

finite sample bias seen in the ordinary Kernel estimators such as the NW and the Locally linear Kernel estimators. Intuitively, if the curves are identical, that is if $m_1 = m_2 = ...m_Q = m_p$, $\hat{r}_j(x)$ is an estimate of unity and the efficiency gains in this case are substantial because our estimator behaves like the pooled start. In general when the guide is not distant from $m_j(x)$ in the sense that $\|m_p(x) - m_j(x)\|^2 = o(1)$, $\hat{m}_p(x)$ is more efficient than the ordinary Kernel estimator since it pools the measurements from all the experimental units and the correction factor will be less rough than $m_j(x)$ resulting in smaller bias than in the NW estimator for example. This claim will be more apparent in the next section. The asymptotic bias and variance properties and finite sample performance of the new estimator are explored next. The basic properties of the new estimator show that the asymptotic variance of the NW estimator and that of the new estimator differ only by $O(Nh_p)^{-1}))$ where $N = \sum_{j=1}^{Q} n_j$ while the bias of the new estimator can be substantially lower when the nonparametric guide is in the vicinity of the individual conditional mean.

3.1 Estimation with a Non-random Start

We first start by assuming that the guide belongs to class of fixed functions and expand our findings by allowing the feasible guide to be a nonparametric estimate. Suppose that the start m_p belongs to a class of non-random functions but the correction factor is itself nonparametrically estimated using the NW estimator. This leads to the following version of the proposed estimator:

$$\dot{m}(x) = m_p(x)\dot{r}(x) = \frac{\sum_{i=1}^{n} y_i[\frac{m_p(x)}{m_p(x_i)}]K_h(x_i - x)}{\sum_{i=1}^{n} K_h(x_i - x)} \tag{10}$$

after then dropping the subscript j for notational convenience. In deriving the asymptotic mean, variance, and ditribution of the new estimator we will assume the following regularity conditions generally made in the asymptotic theory of Kernel estimators (see Pagan and Ullah, 1999).

A1 The x_{ij}'s are i.i.d and independent of the error process ϵ_{ij}

A2 The density function $f(x)$ and the conditional mean $m(x) \in \mathcal{C}^2(\Theta)$ with finite second derivatives and $f(x) \neq 0$ in Θ, the neighborhood of point x.

A3 The density function $g(x)$ and the conditional mean $m_p(x)$ of the pooled data $\in \mathcal{C}^2(\Theta)$ with finite second derivatives and $g(x) \neq 0$ in Θ, the neighborhood of point x [3].

A4 The kernel function $K(z)$ is bounded, real-valued, with the following characteristics: (i) $\int K(z)dz = 1$, (ii) $K(z)$ is symmetric about 0, (iii) $\int z^2 K(z)dz < \infty$, (iv) $|z|K(|z|) \to 0$ as $|z| \to \infty$, (v) $\int K^2(z)dz \leq \infty$

A5 $h_j \to 0$ and $n_j h_j \to \infty \ \forall \ j = 1....Q$.

A6 $E|\epsilon_i|^{2+\delta}$, $\int |K(\omega)|^{2+\delta}$, and $\int |\frac{m_p(x)}{m_p(x_i)}|^{2+\delta}$ are finite for some $\delta \geq 0$

A7 $h_p \to 0$ and $n_j h_p \to \infty \ \forall \ j = 1....Q$; h_p being the smoothing parameter for the pooled estimator.

PROPOSITION I: Let $m_p \in \mathcal{C}^2(\Theta)$ be a non-random function such that $m = m_p r$ and $|m_p| > \eta > 0$. Then under the assumptions A1-A5, we have

- $E[\dot{m}(x)] - m(x) = \frac{1}{2}\mu_2 h^2 [m_p(x)r''(x) + 2m_p(x)r' \frac{f'(x)}{f(x)}] + o(h^2)$

[3] $g(x) = f(x)$ if the conditional means $m_1, m_2, ..., m_Q$ are identical. If the conditional means are different, $g(x)$ is a mixture density generated by the set of Q individual probability density functions: $g(x) = \sum_{j=1}^{Q} w_j f_j(x)$.

- $Var[\dot{m}(x)] = \sigma^2(nhf)^{-1}R(K) + o(h/n)$

Proof: For a non-random start the derivation of the asymptotic properties of our estimator are very similar to those of the Nadaraya-Watson estimator (see appendix A herein).

3.2 Estimation with a Nonparametric Start

Now we assume that both the start $\hat{m}_p$ and the correction factor are nonparametrically estimated using the NW estimator, that is:

$$\hat{m}(x) = \hat{m}_p(x)\hat{r}(x) = \frac{\sum_{i=1}^{n} y_i[\frac{\hat{m}_p(x)}{\hat{m}_p(x_i)}]K_h(x_i - x)}{\sum_{i=1}^{n} K_h(x_i - x)} \tag{11}$$

PROPOSITION II: Let $\hat{m}_p$ be the NW estimate of $m_p \in \mathcal{C}^2(\Theta)$ intended to best approximate the conditional mean m such that $m = m_p r$ and $|m_p| > \eta > 0$. Then under the assumptions A1-A5, we have

- $E[\hat{m}(x)] - m(x) = \frac{1}{2}\mu_2 h^2[m_p(x)r''(x) + 2m_p(x)r'\frac{f'(x)}{f(x)}] + o(h^2)$
- $Var(\hat{m}(x)) = \sigma^2(nhf)^{-1}R(K) + O(h/n + (Nh_p)^{-1})$

Proof: See appendix A.

PROPOSITION III: Under the assumptions A1-A7, $\hat{m}(x)$ has a limiting normal distribution:

$$\sqrt{nh}(\hat{m}(x) - m(x)) \to N(B(\hat{m}(x)), \Sigma) \tag{12}$$

where $B(\hat{m}(x)) = \frac{1}{2}\mu_2 h^2[m_p(x)r''(x) + 2m_p(x)r'\frac{f'(x)}{f(x)}]$ and $\Sigma = \frac{\sigma^2}{f(x)}R(K)$

Proof: See appendix B.

The bias of the new estimator is not a function of the slope and curvature of the true regression function as it is for the standard Nadaraya-Watson estimator (see section I), the Locally linear kernel or the Gasser-Müller estimator. Rather it is function of the slope and second derivative of the correction factor $r(x) = \frac{m(x)}{m_p(x)}$. If the pooled estimator coincides with the true function, then $r(x)$ will be a line hence both r' and $r'' = 0$. Statistically, this means that the leading terms of the bias will vanish. But this is the ideal scenario that is not likely to happen in practice. But if $m_p(x)$ and $m(x)$ are not too far apart so that their ratio fluctuates around unity, then the correction factor should be less variable than the individual conditional mean hence leading to bias reduction. However, as we will explain in the fourth section, the pooled start really does not have to be a good approximation of the $m(x)$ for our estimator to remain competitive to the the ordinary Kernel estimators even when the curves are dissimilar.

Two reasons motivate us to use a nonparametric start instead of a parametric one as in Hjort and Glad (1995) and Glad (1998). First, using a nonparametric prior frees us from issues of functional mispecification especially in the event that the individual curves belong to different parametric families. Second, the pooled estimator helps make up for the paucity of data when the measurements for each experimental units are limited thus reducing estimation error if the underlying curves are similar.

A potential limitation of our estimator is the selection of the optimal number of curves to be used when estimating a given individual curve so as to achieve bias reduction. We propose that cross validation be used to select the "optimal" curves to be included. Such procedure could certainly be computationally intensive and time consuming when dealing with a big data set and/or numerous individual curves but should reduce the likelihood of contamination bias. This problem is similar to the choice of instruments in Instrumental variable estimation when the number of

instruments is large. Following Donald and Newey (2001) we propose that the cross-validation procedure be applied to a subset (or subsets) of curves that are sought to be likely to satisfy the similarity hypothesis[4].

4 Simulations

The simulations were undertaken to the end of evaluating the finite sample performance of our estimator under both alternative data environments outlined in section I. The first set of simulations compares our estimator, the nonparametric estimator with a pooled start (NEPS) with the Locally linear Kernel (LLK) estimator and the Racine and Li (R&L) estimator (to accommodate the data environment when at least one the predictor variables is discrete and takes values from the set $\{1, 2...Q\}$). The R-valued continuous predictor variable is assumed to be uniformly distributed on the [0,1] interval. The second set of simulations compares the NEPS to the LLK and the Altman and Casella (A&C) estimator (to accommodate the data environment where the researcher is faced with a system of regression equations). As required by the Altman and Casella estimator, an equi-spaced design is used with $x_i = i/n$.

In each of the above cases, two scenarios are investigated in the simulations. In the first scenario, which we call the "case of identical curves", four individual curves were generated and constrained to have the same conditional mean equal to $m(x) = sin(5\pi x)$. The choice of the functional form of $m(x)$ does not bear any statistical reason. Individual-specific errors added to differentiate the data across experimental units. This is the ideal case obviously; the goal is to see how much "better" the

[4]The cross-validation procedure consists of alternating the pooled start from the set formed by the Q curves, for total of 2^Q possible pooled guides: (uniform start,$\{1\}, \{2\}...\{Q\}, \{1, 2\}, \{, 1, 3\}...\{1, 2, ...Q\}$). Then choose the one pooled guide whose loss function is the lowest. Cross-validation becomes increasingly cumbersome when the number of experimental units, Q is large. To mitigate the problem of choosing a start when Q is large, the researcher can set select a subset of k $\leq$ Q curves and apply the cross-validation procedure.

estimators making use of extraneous information would be in a situation where all the individual curves were identical. Recall that the incorporation of external information is based upon the assumption that the Q curves are similar to some extent.

In the second scenario, which we refer to as the "case of dissimilar curves", also four curves were generated, however, with dissimilar conditional means (see graph).[5] The four curves are:

$$m_1(x) = sin(15\pi x) + \epsilon_1 \tag{13}$$

$$m_2(x) = sin(5\pi x) + \epsilon_2 \tag{14}$$

$$m_3(x) = .3e^{(-64(x-.25)^2)} + .7e^{(-256(x-.75)^2)} + \epsilon_3 \tag{15}$$

$$m_4(x) = 10e^{-10x} + \epsilon_4 \tag{16}$$

This is another polar case that should provide some insights about how the advanced estimators perform compared to the standard methods which are unaffected by the dissimilarity of the conditional means.

The choice of these two scenarios is motivated by empirical analysis. It is not very likely that in empirical applications, the individual conditional means in set will

be identical or totally unrelated; the truth lies somewhere between these two polar cases. Throughout the simulations a Gaussian Kernel is used and the bandwidth is the one that minimizes the integrated squared error:

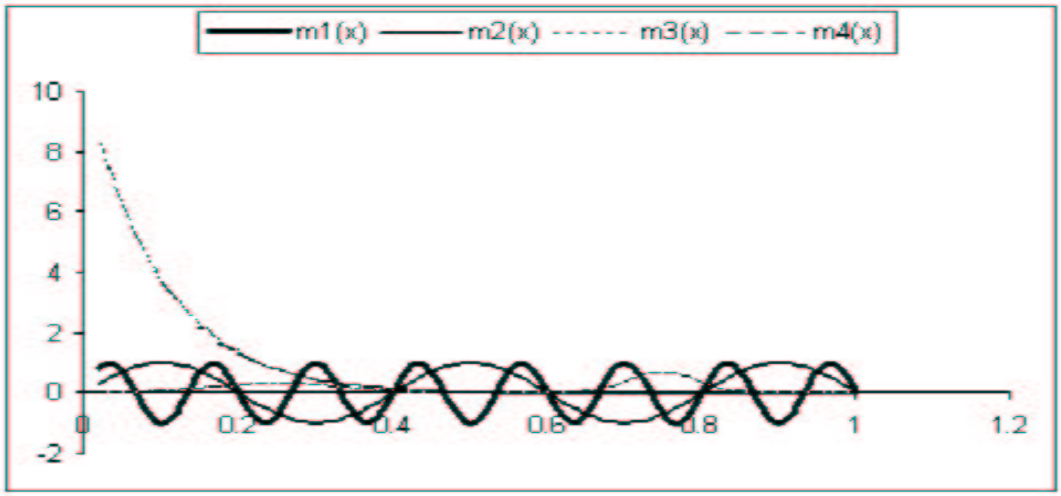
$$ISE[\hat{m}(x)] = \int [\hat{m}(x) - m(x)]^2 dx \tag{17}$$

The performance of each curve is assessed via its mean integrated squared error (MISE). Tables 1 and 2 (see appendix C) report the average MISE of the four curves

[5] These curves were used by C. Hurvitch and J. Simonoff in their article "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion" (1998).

along with the average mean integrated squared bias (MIB^2) for 500 replications

Figure 1: Graphic depiction of the four conditional means



respectively for the random and fixed designs.

The results in table 1 show that the new estimator (NEPS) significantly outperformed the Locally linear kernel, which was expected since all four conditional means are identical and that the Locally linear kernel does not make use of extraneous information. Interestingly, our estimator also beat the Racine and Li estimator for all samples size in the case of similar curves. The performance of our estimator is attributable to a lower bias than its competitors as reported in table 1. As said earlier, the Racine and Li estimator trades off variance for bias; so while it has a lower variance, its finite sample bias remains relatively high compared to that of our estimator.

Interestingly, our estimator outperformed the LLK while the Racine and Li estimator remained competitive in the case of dissimilar curves. The strong performance of our estimator although not anticipated given the dissimilarity of the conditional means is not surprising. This is because the standard Kernel estimator is a special case of our estimator with $m_p(x)$ being equal to a constant over the entire support

of x. Clearly a flat prior is quite conservative for most curves and therefore can be improved upon, hence $m_p(x)$ need not be a good approximation of $m(x)$. The performance of the Racine and Li estimator lies in its ability to revert back to the ordinary NW estimator by having $\hat{\lambda_j} \to 0$ when the curves are dissimilar.

Similar conclusions are drawn from the results in table 2. Our estimator outperforms its competitors, the Locally Linear Kernel and the Altman and Casella's nonparametric Empirical Bayes estimator, when the conditional means are identical. As in the random design case, our estimator remained competitive to the LLk even when the similarity hypothesis is clearly wrong. The performance of the A&C is in general disappointing even in the case of identical curves. This could be explained in part by the small number of experimental units ($Q = 4$) considered in our simulations.

5 Conclusion

In this paper, we have proposed a simple nonparametric regression method which admits two important empirical frameworks: multiple curve estimation and "cell" estimation. The method has been designed so as to achieve bias reduction by utilizing extraneous information from the set of available curves when estimating a a given conditional mean. Consistent with the expression of the asymptotic bias, the simulations conducted show that the new estimator significantly outperformed the ordinary Kernel estimator (LLK) when the conditional means were similar thanks to a lower bias. Perhaps more interesting is that our estimator did not lose much if at all to the ordinary Kernel estimator (which does not incorporate extraneous information) when the similarity hypothesis was untenable. The new estimator also performed admirably against the Racine and Li and the Altman and Casella estimators.

A potential shortcoming of the proposed estimator is the choice of the pooled start; that is the "optimal" extraneous information that could lead to bias reduction. When

the number of experimental units is large, such enterprise may not be trivial. On the other hand, cross-validation techniques could be used to select the pooled start when the number of of experimental units is limited. In the former case, the difficulty in choosing the pooled start can be mitigated by applying the cross-validation procedure to a subset of curves which are sought to be the best "candidates".

References

[1] Abaffy, J, and al, "A Nonparametric Model for Analysis of the Euro Bond Market". *Journal of Economic Dynamics and Control* 23 (2003), 1113-1131

[2] Altman, N.S, and G. Casella, "Nonparametric Empirical Bayes Growth Curve Analysis." *Journal of the American Statistical Association 90 (1995): 508-514.*

[3] Donald, S. G., and W. K. Newey, "Choosing The Number of Instruments". *Econometrica* 69 (2001), 1161-1191

[4] Fan, J. , "Design-adaptive Nonparametric Regression", *Journal of the American Statistical Association* 87 (1992), 998-1004

[5] Glad, I., "Parametrically Guided Nonparametric Regression". *Scandinavian Journal of Statistics* 25(1998), 649-668

[6] Hart, J. D., and T. E. Wehrly "Kernel Regression Estimation using Repeated Measurements Data." *Journal of the American Statistical Association* 81 (1986), 1080-1087

[7] Hjort, N.L., and I. Glad, "Nonparametric Density with a Parametric Start." *The Annals of Statistics 23 (1995), 882-904.*

[8] Hjort, N.L., and M.C. Jones, "Locally Parametric Nonparametric Density Estimation." *The Annals of Statistics 24 (1996), 1619-1647.*

[9] Hurvich, C., and J.S. Simonoff, "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of Royal Statistical Society Series B, 60 (1998), 271-293.*

[10] Ker, A.P., and B. K. Goodwin, "Nonparametric Estimation of crop insurance Rates Revisited. *American Journal of Agricultural Economics* 83 (2000), 463-478

[11] Pagan,A. R., and A. Ullah , *Nonparametric Econometrics.* New York, Cambridge University Press, 1999.

[12] Racine,J. S., and Qi Li, "Nonparametric Regression with Both Categorical and Continuous Data." *fothcoming, Journal of Econometrics.*

A Proof of Proposition II

We have $\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)[\frac{y_i}{\hat{f}(x)}][\frac{\hat{m}_p(x)}{\hat{m}_p(x_i)}]$. A Taylor series expansion of $\frac{\hat{m}_p(x)}{\hat{m}_p(x_i)}$ at $\frac{m_p(x)}{m_p(x_i)}$ yields

$$\begin{aligned}\hat{m}(x) &\simeq \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)\frac{y_i}{\hat{f}(x)}[\frac{m_p(x)}{m_p(x_i)} + \frac{\hat{m}_p(x) - m_p(x)}{m_p(x_i)} \\ &\quad -\frac{m_p(x)}{m_p(x_i)}(\frac{\hat{m}_p(x_i) - m_p(x_i)}{m_p(x_i)})]\end{aligned}$$

The expressions $\frac{1}{n}\sum_{i=1}^{n} K_h(x_i-x)\frac{\epsilon_i}{\hat{f}(x)}\frac{\hat{m}_p(x)-m_p(x)}{m_p(x_i)}$ and $\frac{1}{n}\sum_{i=1}^{n} K_h(x_i-x)\frac{\epsilon_i}{\hat{f}(x)}\frac{m_p(x)}{m_p(x_i)}\frac{\hat{m}_p(x_i)-m_p(x_i)}{m_p(x_i)}$ are of order $o_p(h_p^2)$ hence

$$\begin{aligned}\hat{m}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)\frac{y_i}{\hat{f}(x)}[\frac{m_p(x)}{m_p(x_i)} + \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)\frac{m(x_i)}{\hat{f}(x)}\frac{\hat{m}_p(x) - m_p(x)}{m_p(x_i)} \\ &\quad -\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)\frac{m(x_i)}{\hat{f}(x)}\frac{m_p(x)}{m_p(x_i)}(\frac{\hat{m}_p(x_i) - m_p(x_i)}{m_p(x_i)})] + o_p(h_p^2) \\ \hat{m}(x) - m(x) &= \frac{m_p(x)}{n\hat{f}(x)}\sum_{i=1}^{n} K_h(x_i - x)(r(x_i) + \epsilon_i^* - r(x)) + \frac{1}{n\hat{f}(x)}\sum_{i=1}^{n} K_h(x_i - x)r(x_i)(\hat{m}_p(x) - m_p(x) \\ &\quad -\frac{1}{n\hat{f}(x)}\sum_{i=1}^{n} K_h(x_i - x)\frac{m_p(x)}{m_p(x_i)}r(x_i)(\hat{m}_p(x_i) - m_p(x_i)) + o_p(h_p^2) \\ &= \frac{A_n}{\hat{f}(x)} + \frac{B_n}{\hat{f}(x)} + o_p(h_p^2)\end{aligned}$$

where $\epsilon_i^* = \frac{\epsilon_i}{m_p(x_i)}$, $A_n = \frac{m_p(x)}{n}\sum_{i=1}^{n} K_h(x_i - x)(r(x_i) + \epsilon_i^* - r(x))$ and $B_n = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i-x)r(x_i)(\hat{m}_p(x)-m_p(x)) - \frac{1}{n}\sum_{i=1}^{n} K_h(x_i-x)\frac{m_p(x)}{m_p(x_i)}r(x_i)(\hat{m}_p(x_i)-m_p(x_i))$.

$$\begin{aligned}E(A_n) &= m_p(x)E[n^{-1}\sum_{i=1}^{n} K_h(x_i - x)\{r(x_i) - r(x)\}] \\ &= m_p(x)\int K_h(x_1 - x)\{r(x_1) - r(x)\}f(x_1)dx_1 \\ &= m_p(x)\int K(\omega)\{r(x + h\omega) - r(x)\}f(x + h\omega)d\omega \text{ after a change of variable.} \\ &= \frac{h^2}{2}[m_p(x)f(x)r''(x) + 2m_p(x)f'(x)r'(x)]\mu_2(K) + o(h^2) \qquad (18)\end{aligned}$$

Denote B_n^1 and B_n^2 respectively the first and second terms of B_n.

$$\begin{aligned}E(B_n^1) &= E\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)r(x_i)E_{x_i}[\hat{m}_p(x) - m_p(x)] \\ &= \frac{1}{2}\mu_2 h_p^2[m_p''(x) + 2m_p'(x)\frac{g'(x)}{g(x)}]E(\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)r(x_i)) \\ &= r(x)f(x)Bias[\hat{m}_p(x)] + o(h^2).\end{aligned}$$

Similarly,

$$\begin{aligned}
E(B_n^2) &= E(\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)r(x_i)\frac{m(x)}{m_p(x_i)}E_{x_i}[\hat{m}_p(x_i) - m_p(x_i)] \\
&= \frac{1}{2}\mu_2 h_p^2 E([m_p''(x_i) + 2m_p'(x_i)\frac{g'(x_i)}{g(x_i)}](\frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)r(x_i)\frac{m_p(x)}{m_p(x_i)}) \\
E(B_n^2) &= \frac{m_p(x)}{2}\mu_2 h_p^2 \int K_h(x_i - x)\frac{r(x_i)}{m_p(x_i}(m_p''(x_i) + 2m_p'(x_i)\frac{f'(x_i)}{f(x_i)})f(x_i)dx_i \\
&= \frac{1}{2}\mu_2 h_p^2 \int K(\omega)(r(x) + o(1))(M(x) + o(1))(f(x) + o(1))d\omega \\
&\quad \text{where the definition of } M(x) \text{ should be apparent. Hence} \\
E(B_n^2) &= r(x)f(x)Bias[\hat{m}_p(x)] + o(h^2)
\end{aligned}$$

Since $plim\hat{f}(x) = f(x)$, it follows that $E(\hat{m}(x) - m(x)) \simeq f(x)^{-1}E(A_n + B_n)$ by Slutsky's theorem. This completes the first part of the proof.

$Var[A_n] = \sigma^2(nh)^{-1}R(K)f(x) + O(h/n)$. The computation of the variance of B_n and the covariance of A_n and B_n is also straightforward but significantly longer thus not provided in details. Both $Var[B_n]$ and $Cov(A_n, B_n)$ are found to be the order $O[(Nh_p)^{-1}]$. Again $Var(\hat{m}(x)) \simeq f(x)^{-2}[Var(A_n) + Var(B_n) + 2Cov(A_n, B_n)]$ by Slutsky's theorem, which completes the second part of the proof.

B Proof of Proposition III

Write $(\hat{m}(x) - m(x))\hat{f}(x) = C_n + D_n + o_p(h_p^2)$

where $C_n = \frac{m_p(x)}{n}\sum_{i=1}^{n} K_h(x_i - x)(r(x_i) - r(x)) + \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)r(x_i)(\hat{m}_p(x) - m_p(x)) - \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)\frac{m_p(x)}{m_p(x_i)}r(x_i)(\hat{m}_p(x_i) - m_p(x_i))$ and

$D_n = \frac{m_p(x)}{n}\sum_{i=1}^{n} \epsilon_i^* K_h(x_i - x)$. From proposition II, we know that $E(C_n) = \frac{h^2}{2}[m_p(x)f(x)r''(x) + 2m_p(x)f'(x)r'(x)]\mu_2(K) + o(h^2)$. Straightforward calculations show that $Var(C_n) = o(h^4) + O(\frac{1}{n_1 h_p + n_2 h_p + ...+ n_Q h_p})$. By assumption A7, $n_j h_p \to \infty \ \forall \ j = 1....Q$ hence the last term of the variance of C_n can be ignored for asymptotic results. Combining the expectation and variance of C_n, it follows that

$$\begin{aligned}
C_n &= E(C_n) + o_p(h^2) \\
&= \frac{h^2}{2}[m_p(x)f(x)r''(x) + 2m_p(x)f'(x)r'(x)]\mu_2(K) + o_p(h^2) \\
&= f(x)B(\hat{m}(x)) + o_p(h^2);
\end{aligned}$$

Similarly, $E(D_n) = 0$ and $Var(D_n) = (nh)^{-1}\{\sigma^2 R(K)f(x) + o(1)\}$. D_n is a triangular array of i.i.d random variables thus, under assumption A6, we can apply Liapounov's central limit theorem to obtain: $\sqrt{nh}(D_n) \to N(0, f^2(x)\Sigma)$.

Since $plim\hat{f}(x) = f(x)$, it also follows that

$$\sqrt{nh}(\hat{m}(x) - m(x) - B(\hat{m}(x))) = \sqrt{nh}\frac{D_n}{\hat{f}(x)} + o_p(1) = \sqrt{nh}\frac{D_n}{f(x)} + o_p(1) \to N(0, \Sigma) \tag{19}$$

C Simulations results

Table 1: Average error of the four curves: random design.

Case of similar curves

n	LLK		R&Li		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
25	21.7348	10.4943	6.5461	4.1736	5.1653	0.6823
50	10.644	5.6111	3.8884	2.6365	2.6987	0.3431
100	5.7383	3.5223	2.3816	1.7222	1.4638	0.1710

Case of dissimilar curves

n	LLK		R&Li		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
25	25.4681	16.8525	27.7545	20.8580	24.7851	15.5193
50	18.5100	14.3130	20.3510	16.2460	18.1560	12.7090
100	14.8060	12.3190	15.5110	12.9280	14.5970	11.0740

Table 2: Average error of the four curves: fixed design.

Case of similar curves

n	LLK		A&C		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
25	16.8787	11.9162	9.4340	0.01441	5.0308	0.5366
50	5.6786	1.5812	7.8547	0.0152	2.7299	0.3296
100	3.1335	0.6486	7.4178	0.0149	1.5254	0.1980

Case of dissimilar curves

n	LLK		A&C		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
25	20.9578	16.8456	43.3380	23.9114	20.7351	13.3161
50	12.4690	8.9958	18.3277	9.1848	11.9579	7.3078
100	4.4908	2.0037	11.1970	3.1989	5.1113	1.7449