NONPARAMETRIC REGRESSION OF POSSIBLY SIMILAR CURVES

by

Abdoul Gadiry Sam

A Thesis Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

In the Graduate College

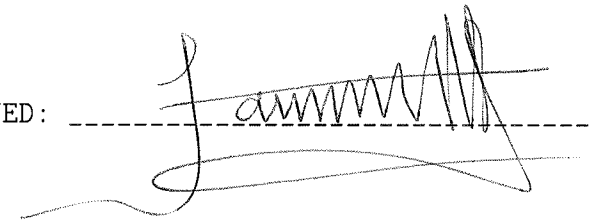
THE UNIVERSITY OF ARIZONA

2002

STATEMENT BY AUTHOR

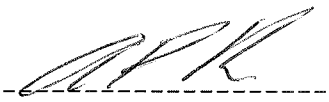
This thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

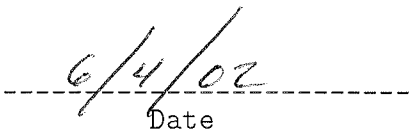
SIGNED: 

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:



Alan P. Ker
Associate Professor of
Agricultural and Resource Economics



Date

Acknowledgments

My heartfelt gratitude goes to my supervisor Alan P. Ker whose commitment, assistance, guidance and expertise made possible the accomplishment of this thesis. I also thank the other committee members Satheesh Aradhyula and Gary Thompson for their helpful input. I am particularly grateful to my family back in Senegal whose inexorable support and affection made it possible for me to complete a graduate program.

Table of Contents

TABLE OF CONTENTS—*Continued*

List of Figures

List of Tables

Abstract

In many situations it is needed to estimate a set of curves that are believed to be similar in structure. In such case, Ker (2000) suggests the use of external information from the other curves in order to reduce the bias of the standard nonparametric estimator for an individual regression function. In the density case, Ker showed that the inclusion of external data in the estimation of a given density generates sizeable efficiency gains when the different underlying densities are similar. While Ker focuses on bias reduction, Racine and Li (2000) and Altman and Casella (1995) devised estimators that can be used to reduce the variance of the standard nonparametric methods by smoothing across possibly similar curves. All of these techniques have however the same objective: improve on the standard nonparametric estimators. This thesis undertakes Monte Carlo simulations and two empirical applications to evaluate potential gains obtained by using nonparametric techniques that integrate external information. The simulations undertaken show that when the curves are similar in shape, the gains can be enormous: some of these methods outperform the standard nonparametric estimator significantly by reducing its mean integrated squared error by as much as 55 %. The replications also show that if the curves are dissimilar, some of the methods incorporating external data remain competitive to the standard nonparametric estimator.

1. Introduction

Economic data are often presented as a set of curves measured on different experimental units:

$$y_i = m_i(x_i) + \epsilon_i, i = 1...n \tag{1.1}$$

where y_i is the response variable, x_i is the matrix of explanatory variables, $m_i(.)$ is the true unknown curve and ϵ_i an error term for experimental unit i . These curves belong to the same population in which a common phenomenon is observed. In some circumstances, the individual curves are believed to have some structural similarities that can be utilized to produce improved estimates. Examples of populations of curves include crop-reporting districts (CRD) where the individual curves are the county yields. When rating cop insurance policies for example, it is necessary to detrend the yields for each county in a given CRD, which amounts to estimating a regression function for each county. When estimating the curve for county j, the standard approach is to use only the yield data of that county.[1] An alternative to the standard approach is to include the yield data of the other counties in the estimation process.[2] In a same CRD, weather patterns, soil type and technology use among others factors could be considered similar across counties. These "similarities" provide a solid basis for assuming that the curves of the different counties are related even though the magnitude of such relationships is unknown. Hence when estimating

[1]Henceforth such estimation methods will be referred to as standard methods or techniques.

[2]Methods making use of external information will hereafter be referred to as advanced methods or techniques.

the curve of county j, it seems reasonable to use external information (data) from the remaining counties in the same CRD for potential efficiency gains. In that case, the individual curve estimate can be described mathematically by the equation:

$$\hat{m}_i(x_i) = f(x_1, y_1, x_2, y_2 ... x_n, y_n). \tag{1.2}$$

The concept of similarity is loosely used in this thesis given the difficulty to delineate it explicitly: the extent to which the different curves are similar is unknown. One can only make an assumption of similarity based on certain elements such as weather patterns, technology use or soil type in the case of a CRD.

In this thesis three advanced methods will be explored: the Altman and Casella Nonparametric Empirical Bayes estimator (NEB); the Racine and Li estimator for combined continuous and discrete data; and the Ker nonparametric estimator with a pooled start. For each one of these methods, the goal is to estimate the true regression function in a way that will not only significantly outperform the standard techniques when the hypothesis of similarity is correct but also produce reliable estimates when the curves are dissimilar. All three techniques employ nonparametric theory.

If indeed the different curves were similar in shape, the efficient estimator would pool the data to estimate a single curve. To do so, a Nadaraya-Watson pooled estimator can be used as the estimate for all the different individuals. However when the curves are not similar as hypothesized, the pooled estimator is inconsistent.

A natural question is how do the advanced methods perform when the individual curves are dissimilar; could they produce less efficient estimates or worse misleading results? As already stated, the objective of the advanced estimators investigated in this thesis is not to lose much efficiency to the standards methods when the curves are dissimilar. One of the goals of this thesis is to carry simulations and an empirical

application in order to answer that question.

1.1. Objective of the Study

The contribution of this thesis is to evaluate the performance of each of the advanced methods compared to the standard nonparametric estimator, which is also called the Nadaraya-Watson estimator. To do so, Monte Carlo simulations as well as two empirical applications are used. Comparison of the different techniques is based on the mean integrated squared error (MISE) for the simulations since the true regressions functions are known. The MISE is an error metric that captures both the bias and the variance of an estimator. The first empirical application consists of wage modeling using a cross-section dataset of men and women. Comparison of the estimators is based on the mean integrated prediction error (MIPE), which is the average of the squared deviations between the data points and their fitted values. The second empirical application relates to the rating of crop insurance polices; performance comparison in that case is judged by the magnitude of the loss ratio of the insurance companies. Both the MISE and the MIPE have been extensively used in econometric applications to assess the efficiency and the predictive ability of competing estimators.

1.2. Plan of the Study

The remainder of the thesis is organized as follows. The second chapter introduces nonparametric methods for regression and density estimation in order to familiarize the reader with the basics of nonparametric econometrics. The third chapter elaborates on three regression techniques earlier mentioned: (1) the Altman and Casella

estimator, (2) the Racine and Li estimator and (3) the Ker estimator. The fourth chapter presents the results of the simulations about the finite sample performances of each of these three techniques. Data is generated from four different curves and the estimates for different sample sizes are computed. The empirical applications are presenting in the fifth chapter. Finally, the sixth chapter provides concluding remarks.

2. NONPARAMETRIC METHODS IN ECONOMETRICS

Before elaborating on nonparametric regression techniques, it is important to introduce some aspects of nonparametric methods to familiarize the reader with the basics on these methods. There are different classes of nonparametric estimators but all the estimators considered in this thesis are kernel based. Kernel methods are the most investigated nonparametric methods from a statistical standpoint. The main reason for the use of nonparametric methods is their ability to display a "better" structure of the true function as a result of less rigid assumptions. Parametric-type estimators generally assume that either the data or the error term follows a certain distribution (Maximum likelihood estimator in regression case or density estimation methods) constraining the estimated function in shape. Nonparametric methods in contrast do not assign any functional form to neither the data nor the disturbance term over the whole data set, which leads to more flexibility in the analysis and exploration of the data. The idea is to let the data "speak for themselves"; that is why nonparametric methods are referred to as "density-free" techniques. However, this can be a little misleading in the sense that nonparametric methods do assume some parametric form in the neighborhood of an observation. It should be mentioned that nonparametric and parametric methods are not incompatible to each other. In fact several authors have combined the two methods to produce improved estimators (see Hjort and Glad(1995), Fan and Ullah (1997)).

2.1. Nonparametric Density Estimation

Regression analysis is driven by density estimation; it can be reduced to estimating a ratio of two densities:

$$E[Y|x] = m(x) = \frac{\int yf(x,y)dy}{\int f(x,y)dy} \tag{2.1}$$

A probability density function can be estimated parametrically by assuming that the data generating process is known. If one believes for example that the data is normally distributed, then the estimator will be:

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\bar{x}}{s^2})} \tag{2.2}$$

where $\bar{x}= \frac{1}{T}\sum_{t=1}^{T} x_t$ and $s^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})$ are respectively maximum likelihood estimates of and μ and σ^2. Again imposing such restriction on the data can be too strong and costly. In fact if the data is not normally distributed but has a bimodal distribution for example, such restriction of normality would smooth away one of the peaks resulting in a loss of possibly valuable information.

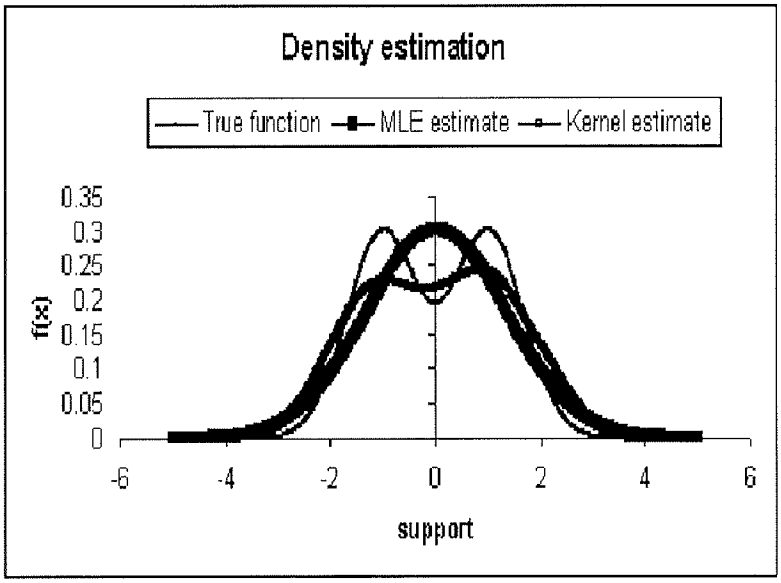


Figure 2.1. Graphic illustration of a misspecified model

The nonparametric estimate displays the economically-important bimodal structure of the true density while the parametric estimate does not.

2.1.1. The Univariate Kernel Estimator

There are numerous nonparametric density estimators such as the histogram, the naive estimator or series estimators. The simplicity of the Kernel estimator makes it one of the most appealing nonparametric density estimators for both practical and theoretical purposes. The formula of the univariate kernel estimator is:

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} K(\frac{x - X_t}{h}) \tag{2.3}$$

or $\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} K_h(x - X_t)$ where $K(.)$ is chosen to be a symmetric probability density function centered at zero.[3] The scalar h is called the smoothing parameter; it controls the spread of the estimator. The following assumptions are made on the smoothing parameter:

$(i) \lim_{T\to\infty} h = 0 \qquad (ii) \lim_{T\to\infty} Th = \infty$

Assumption (ii) says that the smoothing parameter approaches zero at a slower rate than T^{-1}. $K(.)$ is also called the "individual" kernel. In most cases it is chosen to be the normal distribution although other distributions such as the uniform or the Epanechnikov may be used. The individual kernel being a probability density function guarantees that the Kernel estimate itself is a density. Notice that $\hat{f(x)}$ is simply the average of the T individual kernels centered at x. A great benefit of such an estimator is that it tends to show a better structure of the true density. Intuitively, one can see that in regions where there are a lot of observations, the value of the estimator will be large because the closeness between the data points raises their weight. Similarly in regions with relatively distant observations the estimator is expected to have a low value because the spread of the data points decreases their weight.

[3] If $K(.)$ is not symmetric around zero it would not be possible to have an individual kernel function be centered at an observation.

The mean square error (MSE) is a commonly used error metric in density estimation because it measures closeness between the estimated function and the true function for a given observation by encapsulating both the variance and the bias of the estimate. Capturing the bias in the loss function is fundamental when dealing with nonparametric estimators since they are typically biased.

$$\mathrm{M}SE = E(\hat{f} - f)^2 = var(\hat{f}) + bias(\hat{f})^2. \tag{2.4}$$

It can be shown by using the Taylor series expansion that

$$\mathrm{E}\hat{f}(x) = f(x) + \frac{1}{2}h^2 f''(x) \int z^2 K(z)dz + o(h^2) \tag{2.5}$$

$$\mathrm{E}\hat{f}(x) = f(x) + \frac{1}{2}h^2 \mu_2(K) f''(x) + o(h^2) \tag{2.6}$$

having $z = \frac{x-y}{h}$. The bias is therefore $\mathrm{E}(\hat{f}) - f = \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2)$ This is characteristic of nonparametric estimators: they are biased. However, assumption (i) guarantees that Kernel estimator is asymptotically unbiased. In the same way, the variance of the estimator can be shown to be:

$$\mathrm{var}\hat{f}(x) = (Th)^{-1}R(K)f(x) + o(Th)^{-1} \tag{2.7}$$

where $\mathrm{R}(K) = \int K(z)^2 dz$. As with the bias, the variance of the kernel estimator goes to zero as the sample size goes up because of assumption (ii). The bias and the variance known, one can evaluate the mean squared error:

$$\mathrm{M}SE = (Th)^{-1}R(K)f(x) + \frac{1}{4}h^2\mu_2(K)f''(x)^2 + o(h^4 + (Th)^{-1}) \tag{2.8}$$

Since the MSE is a point wise error metric, it is necessary to integrate it over the entire support in order assess the performance of the estimator. Integrating the MSE

gives the mean integrated squared error (MISE):

$$MISE = \{(Th)^{-1}R(K)f(x) + \frac{1}{4}h^2\mu_2(K)^2 \int f''(x)^2\}\hat{f}(.,h) + o(h^4 + (Th)^{-1} \quad (2.9)$$

$$\mathrm{M}ISE = AMISE(\hat{f}) + o(h^4 + (Th)^{-1}) \quad (2.10)$$

where AMISE stands for asymptotic mean integrated squared error. Wand and Jones (1995) proposed that AMISE be used to approximate the MISE as the sample size T gets large. It is easy to see that AMISE is solely a function of the unknown smoothing parameter h. This tells how important the choice of the smoothing parameter is to kernel density estimation. Asymptotically, minimizing the MISE amounts to finding the h that minimizes the asymptotic mean integrated squared error.

The smoothing parameter controls the spread of the kernel estimator $\hat{f}(x,h)$. If the normal distribution is used as $K(.)$ than h^2 is equal to the variance σ^2. A large value of h tends to over smooth the density estimate whereas a small h will make it spiky. This is quite understandable since when h is large, it means that more observations will get a higher weight. When h is small the opposite is observed, that is few observations will be given large weight. The accuracy of the estimates depends crucially on the selection of the smoothing parameter. Then the question that arises is how to choose the right smoothing parameter? Unfortunately the optimal h (i.e. the one that minimize AMISE) is not computable practically because the true function $f(x)$ is unknown:

$$h_{AMISE} = [\frac{R(K)}{\mu_2(K)^2R(f'')T}]^{1/5} = cT^{-1/5}$$

It needs to be estimated using the data at hand. One can see easily how reasonable

assumptions (i) and (ii) are: since $h = cT^{-1/5}$ its limit as T goes to infinity is zero. Similarly $Th = cT^{4/5}$ therefore the limit of Th as T goes to ∞ is infinity.

There is a rich literature about the estimation of the smoothing parameter. Terrell (1990, Theorem I), showed that the smoothing parameter that minimizes the mean integrated squared error is bounded:

$$h_{AMISE} \leq [\frac{243R(K)}{35\mu_2(K)^2T}]^{1/5}\sigma$$

where σ denotes the population standard deviation. This led to what is called the oversmoothing parameter:

$$\hat{h}_{os} = [\frac{243R(K)}{35\mu_2(K)^2T}]^{1/5}s$$

where s is the sample standard deviation. The oversmoothing parameter $\hat{h}_{os}$ provides a good start when using cross-validation methods to search for h. Many authors have written about the selection process of the smoothing parameter. Techniques generally used include the Silverman's "rule of thumb", the least-squares cross validation method and the biased cross-validation method.[4]

2.2. **Nonparametric regression**

Regression analysis whether parametric or nonparametric has two major goals:

- the first one is to investigate a relationship between the explanatory variable(s) and the explained variable,
- the second objective is to provide a tool for forecasting, if necessary, the values of the explained variable.

[4] For an extensive coverage of the smoothing parameter selection techniques the reader is directed to Kernel Smoothing by Wand and Jones (1995)

Regression analysis is however not limited to these two goals only. For a growth curve analysis, one might be interested in the direction of the growth as well as the speed of it. In such case, another objective is to estimate the first and the second derivatives of the regression function. In either case, the accuracy of the results depends critically on appropriateness of the estimator used.

Nonparametric methods are recommended generally when the data cannot be related to a parametric family. If the data belong to a given parametric family then any nonparametric estimator will be less efficient than the parametric one.

Nonparametric regression methods find their usefulness in a lot of empirical applications among which the crop insurance program. Setting accurate premiums is vital to the continuation of the program given the enormous amounts of money involved. To set accurate rates, it is crucial among other things to estimate the yield curves efficiently. Given the huge number of curves to estimate, it would be too rigid to assign a parametric form to the yield curves for the data might belong to a different parametric family from a county to another. This makes nonparametric estimators good candidates to provide a reliable alternative in the estimation process because of their flexibility.

Let (x_i, y_i) be a sequence of n vectors sampled from bivariates densities f_i. A relationship between the response variable and the design variable(s) can be summarized in the regression function:

$$y_i = m(x_i) + \epsilon_i, i = 1.....n \tag{2.11}$$

where $m(x) = E[y|x]$ is the individual regression function after dropping the subscript. As we mentioned previously, estimating $m(x)$ can be done by assuming what the data generating process is. In such case, $m(x)$ would be estimated using a para-

metric estimator. If the assumed functional form is incorrect, the resulting estimates will be biased and inefficient. This summarizes in general the risk of assuming any parametric model. An alternative way of estimating a regression function is to use nonparametric techniques; which means that $m(x)$ is estimated without any assumption about the functional form of the data.

2.2.1. The Nadaraya-Watson Estimator

This estimator is widely used in the literature. It was introduced by Nadaraya(1965) and Watson (1964). The Nadaraya-Watson (NW) estimator is given by:

$$\dot{m}(x) = \sum_{t=1}^{T} W_h y_t \tag{2.12}$$

where $W_h = \frac{K_h(x-X_t)}{\sum_{t=1}^{T} K_h(x-X_t)}$ is a weight matrix and $K()$ the individual kernel. The NW estimator is therefore a weighted average of the response variable. As for the local weighted least squares, closer observations are assigned more weight than farther ones. A big h tends to flatten the Kernel density making the NW estimator behave like OLS while a small h tends to make the estimator spiky.

As for nonparametric methods in general, the Nadaraya-Watson estimator is biased. It can be proved that the leading terms of the bias are:

$$E(\dot{m}(x)) - m(x) = \frac{1}{2}\mu_2 h^2 m''(x) + 2m'(x)\frac{f'(x)}{f(x)} \tag{2.13}$$

after dropping the subscript on f and $m(x)$.

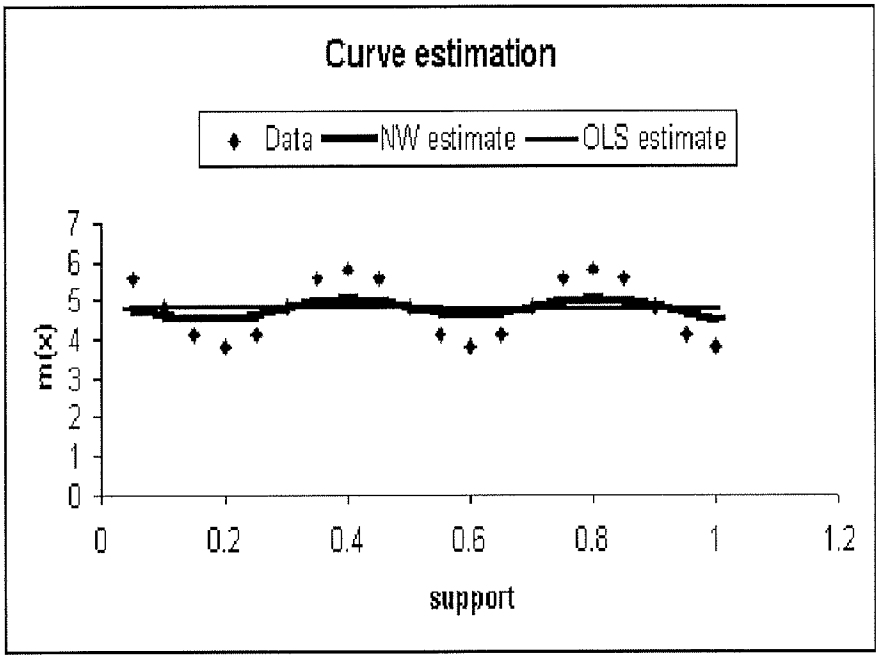


Figure 2.2. Nonparamtric vs ols

The nonparametric estimate follows the patterns of the true curve ($y = cos(5\pi)+2$) while the parametric estimate cuts across the observed data unable to pick any drop or increase; h =0.11.

2.2.2. **Locally Weighted Least Squares**

This method consists of fitting a line in the neighborhood of an observation using the least squares method. It is a weighted local least squares; observations closer to the one of interest are assigned significant weight while those farther are given a smaller weight. The estimator can be written as follows:

$$\hat{m}(x) = x\hat{\beta} \tag{2.14}$$

where

$$\hat{\beta} = (X'W(h)X)^{-1}X'W(h)y \tag{2.15}$$

and

$$W(h)_{ii} = \frac{K_h(x - X_i)}{\sum_{j=1}^{T} K_h(x - X_j)}, W(h)_{ij} = 0 \ \ \forall i \neq j \tag{2.16}$$

is the weight matrix; X is the matrix of explanatory variables. This method differs from OLS in the way it localizes the linear relationship between the response variable and the explanatory variable(s). By fitting a line only locally, the objective is to provide better estimates by capturing a drop or increase in the true function if any. Like any other kernel method the estimates are greatly dictated by the value of the smoothing parameter.

2.2.3. **Selecting a Smoothing Parameter in a Regression Context**

Whether it is for density estimation or regression estimation, the selection of the smoothing parameter (or scaling factor as it is referred by some authors) is the main issue to address; its value affects the estimates more than anything else. In a regression context as in density case, the choice of h is a tedious task. One widely

used technique is the cross-validation smoothing method or "leave-one out" method. Silverman (1985, page 5) describes the intuition of the method succinctly:

> "The basic principle of cross-validation is to leave the data points out one at a time and to choose the value of h under which the missing data points are best predicted by the remainder of the data The cross validation choice of h is then the value of h that minimizes the cross validations score CV(h)".

The cross-validation method consists of finding the smoothing parameter h that minimizes the function CV(h)=$\sum_{t=1}^{T}(y - \hat{m}(x)_{-t})^2$ where $\hat{m}(x)_{-t}$ is a leave-one out estimator.

When the smoothing parameter is big, it tends to flatten the estimated curve making it look like OLS. In contrast, a small smoothing parameter makes the curves "noisy" with a lot of peaks. This is understandable since h controls the amout of weight given to farther and closer observations. In summary the size of the smoothing parameter gives some insight about how informative farther observations are in the estimation of the observation of interest. If h is small it suggest that y_t is better explained by the observations closer to it while a big h would suggest that both closer and farther observations have a predctive information on y_t.

The next two graphs show the effect of a large and small smoothing parameter on the estimates.

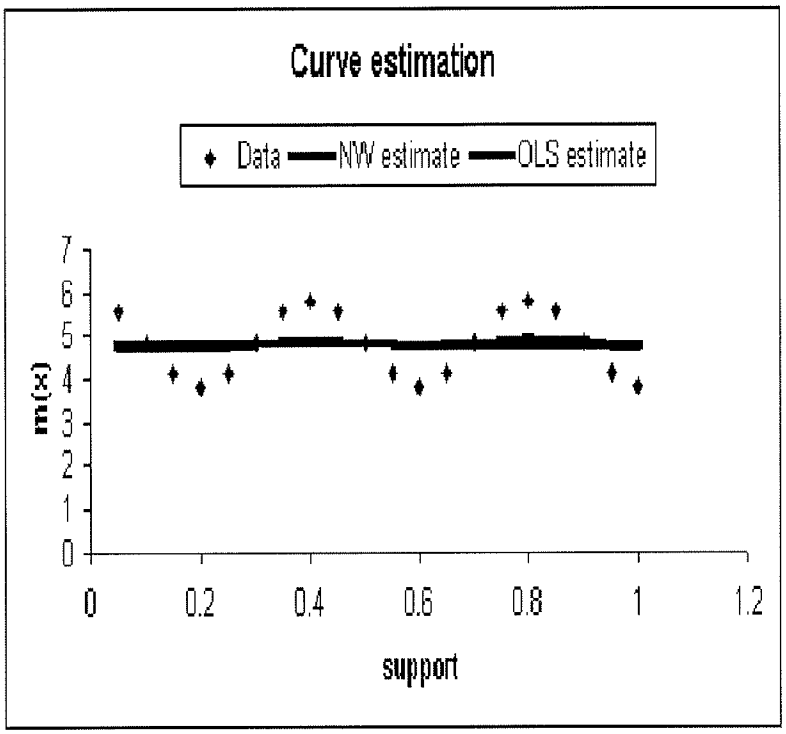


Figure 2.3. Impact of a large smoothing parameter

When the bandwidth is large($h = 0.15$), it flattens the nonparametric estimates to a point where it tends to behave like OLS.

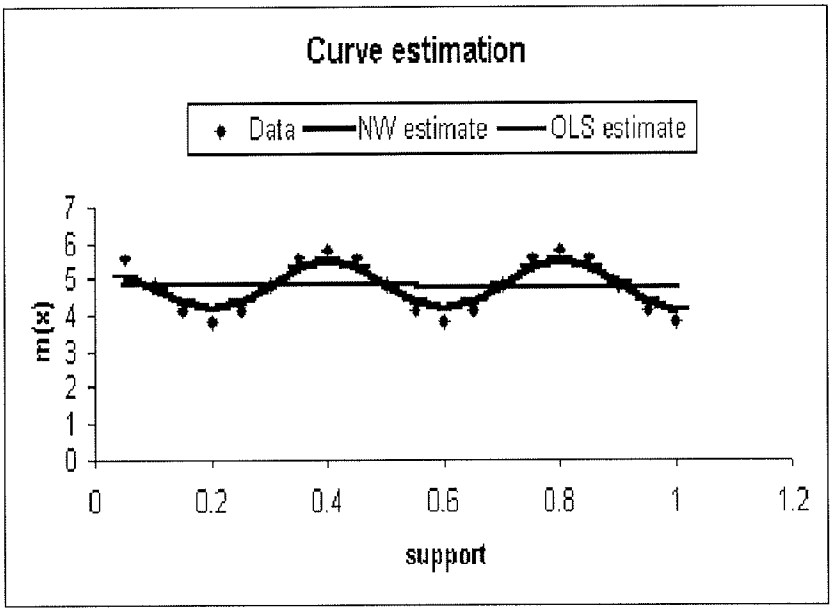


Figure 2.4. Impact of a small smoothing parameter

When the bandwidth is small($h = 0.06$), the nonparametric estimates are spiky and tend to fit the data without getting rid of the noise.

2.2.4. Critiques of Nonparametric Methods

Although nonparametric methods provide a suitable alternative to parametric techniques because of their flexibility, they have some drawbacks that need to be addressed. The choice of smoothing parameter in density estimation as well as in regression estimation is crucial to the accuracy of the nonparametric estimates. There are numerous methods available, each with their proponents. Unfortunately sometimes they produce different parameters, leading to different estimates.

The most popular method for the selection of h is cross-validation, which can be very computer-intensive when dealing with big datasets or a lot of individual curves as in a multiple curve estimation.

The biggest disadvantage of nonparametric methods is certainly the "curse of dimensionality"[5]. The "curse of dimensionality" refers to the poor performance of nonparametric methods when the number of regressors (denote p the number of regressors) increases. The poor performance is attributed to the slow rate of convergence which gets worse as the dimension p rises. The rate of convergence is the rate at which the estimated regression function converges to the unknown true function.[6] It is determined theoretically. Typically nonparametric estimators have a rate of convergence of order $O(T^{-4/(4+p)})$. When p is equal to 1, parametric estimators have a rate of convergence of order $O(T^{-1})$ under the null whereas nonparametric methods have a slower convergence rate of order $O(T^{-4/5})$[7]. The rate of convergence being of

[5] A detailed explanation of the curse of dimensionality can be found in Yatchew (1998)

[6] A more extensive discussion of the parametric versus nonpararametric debate and the rate of convergence can be found in Ker and Coble (2000)

[7] For more on the order notation and theory the reader is directed to Kernel smoothing by Wand and Jones (1995)

order $O(T^{-1})$ means that

$$\lim_{T\to\infty} T \times MSE < \infty.$$

Hence for nonparametric estimators with p=1,

$$\lim_{T\to\infty} T^{5/4} \times MSE < \infty.$$

Clearly the rate of convergence is worse than that of a parametric estimator and it gets worse as p rises. Estimators with a fast rate of convergence tend to be preferred to those with a slow rate of convergence. One of the reasons why nonparametric methods are not used overwhelmingly is that they have a typically slower rate of convergence than their parametric counterparts under the null that is when the assumed parametric form is the correct one. The "curse of dimensionality" has been used a lot by proponents of parametric methods to advocate for the use of parametric models. Some proponents of nonparametric methods propose the use of additive models as a way to reduce the "curse of dimensionality". It is worthwhile to note that parametric methods have a faster rate only when the assumed functional form is correct. Under the alternative, the rate of convergence of nonparametric methods is still $O(T^{-4/5})$ while parametric methods do not converge to the true; which gives the nonparametric methods the benefit of the doubt since it is unknown what the true functional form is.

Both proponents of parametric and nonparametric methods have valid arguments and legitimate concerns about the use of either method. That is why some authors have combined the two methods in order to develop improved estimators that would make use of the strengths of each technique. In the density estimation case, Hjort and Glad (1995) developed a combined parametric-nonparametric estimator to reduce the

bias of the Kernel density estimator. The method performs well in practice. In the regression case, Olkin and Spiegleman (1987) developed a semiparametric estimator that is a convex combination of the parametric and the nonparametric estimators.

Ullah and Vinod (1993, page 87) summarize the strengths and weaknesses of nonparametric methods as follows:

> "Why study nonparametric smoothing methods? Because they offer versatility and flexibility in estimating and forecasting -one does not need to specify functional form (...) Why do they work? Because they use local smoothing and because the theory (...) demonstrates so. What does not work? They fail when there are too many regressors and/or too few observations."

3. ADVANCED NONPARAMETRIC METHODS

As mentioned in chapter I, it is often necessary to estimate a set of curves representing different experimental units. Generally, these curves are structurally related to each other even though the degree of this relationship may be unknown. In this case, when estimating an individual curve, it would be inefficient not to use the data on the other individuals. Standard regression techniques ignore external information by estimating each curve separately. In this chapter, three advanced techniques allowing for the use of external information are studied. [8]

3.1. The Altman and Casella Estimator

This estimator is also known as the nonparametric empirical Bayes estimator (NEB). It mixes both parametric and nonparametric theory in an attempt to improve fixed-design regression performance by making use of the possible similarities between the individual curves[9].

3.1.1. The Method

Let $y_{it} = m_i(t) + \epsilon_{it}$ be the regression equation of individual i where t represents the fixed design points; this type of model is known as a fixed-design regression model. Let m_i be estimated nonparametrically using a linear smoother in the form:

[8] The term "advanced " has more to do with the fact that these techniques allow the use of external data than their mathematical complexity.

[9] This estimator was introduced by Altman and Casella (1995)

$\hat{m}_i(t) = \sum_{t=1}^{T} W_h y_{i,t}$; if the data are normally distributed, so is the nonparametric estimator. It is easy to see that $\hat{m}_i$ is biased:

$$E(\hat{m}_i(t)|h = \dot{h}) = \sum_{t=1}^{T} W_h m_i(t) \neq m_i(t)$$

Given the biasness of the nonparametric estimator, it can be rewritten as: $\hat{m}_i = \psi_i + \upsilon_i$ where υ_i is an error term such that $E(\upsilon_i(t)) = 0$ and $var(\upsilon_i(t)) = \frac{\sigma^2}{T}$.

Using the Bayesian hierarchical model, the authors derived the posterior mean of $m_I(t)$ to be:

$$\tilde{m}_i(t) = \bar{m}(t) + \alpha(t)[\hat{m}_i(t) - \psi(t)]$$

where $\alpha(t)$ is called the shrinkage parameter. In practice, the unknown parameters would be replaced by sample estimates, which lead to the final NEB estimator:

$$\tilde{m}_i(t) = \bar{y}_{.t} + \hat{\alpha}(t)[\hat{m}_i(t) - \overline{\hat{m}(t)}] \tag{3.1}$$

where: $\bar{y}_{.t} = \frac{1}{n}\sum_{i=1}^{n} y_{it}$ is the cross-individual sample mean of the data at point t and $\hat{\alpha}(t) = \frac{\hat{\sigma}_{y(t)m(t)}}{\hat{\sigma}^2_{\hat{m}(t)}}$ (the estimator of the shrinkage factor) is the ratio of the covariance between the data and the nonparametric estimates and the variance of the nonparametric estimates, $\overline{\hat{m}(t)} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_{it}$. Notice that this estimator uses the data from the other individuals in the population in the regression of the curve of interest. This is known as the Stein paradox in the sense their may not be any similarities at all between the different individuals in the population. Then the question is how can one gain efficiency? To answer this question, Altman and Casella showed that minimizing the common squared error loss function (SE) could also derive the estimator:

$$SE = \sum_{i=1}^{n} (\hat{m}_i - m_i)^2. \tag{3.2}$$

Given $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and $\overline{\hat{m}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i$, it is can be proved that there is an α such that: $\tilde{m}_i = \bar{m} + \alpha[\hat{m}_i - \overline{\hat{m}}]$ has a lower squared error (SE) than the nonparametric estimator. Evaluating the estimator at all t points yields to the NEB estimator. Then the efficiency gains come from the minimization of a common loss function.

3.1.2. **What is the Role of $\alpha(t)$?**

The scalar $\alpha(t)$ is known as the shrinkage parameter. It's role is to "shrink" the individual estimator to the mean curve. Three scenarios can arise; the first one is $0 < \alpha(t) < 1$. In that case $\alpha(t)$ "shrinks" the individual curve to the population average since $\hat{\alpha}(t)[\hat{m}_i(t) - \overline{\hat{m}(t)}]$ tends to zero. If $\hat{m}_i$ is unbiased for m_i then $\alpha(t)$ will be between 0 and 1.

If $\hat{m}_i$ is biased for m_i, $\alpha(t)$ could be less than 0 or greater than 1; $\alpha(t)$ is less than 0 if and only if the data are negatively correlated with the nonparametric estimates since the denominator is positive. In such case, $\hat{\alpha}(t)$ will make the data and the nonparametric estimates positively correlated: $\hat{\alpha}(t)[\hat{m}_i(t) - \overline{\hat{m}(t)}]$ will transform the smallest $\hat{m_i}(t)$ value into the largest $\tilde{m_i}(t)$ value and vice versa. By changing the sign of the estimates the "shrinkage" parameter will ensure that the data and the final estimates are correlated positively.

The last scenario is $\alpha(t)$ greater than 1, which happens when the covariance between the data and the nonparametric estimates is larger than the variance of the latter. This situation arises when the data points are more disperse than the estimates. In this case, the term $\hat{\alpha}(t)[\hat{m}_i(t) - \overline{\hat{m}(t)}]$will augment the spread of the NEB estimates:$\hat{\alpha}(t)[\hat{m}_i(t) - \overline{\hat{m}(t)}]$will turn the small $\hat{m_i}(t)$ value into a smaller $\tilde{m_i}(t)$ value and the big $\hat{m_i}(t)$ value into a bigger $\tilde{m_i}(t)$ value thereby making the NEB

estimates more scattered. Clearly in the last two scenarios $\alpha(t)$ does not have a shrinkage role. The reader is directed to Altman and Casella (1995) for a complete derivation of the model.

3.1.3. **The Disadvantages of the Method**

This estimator is confined to a fixed and balanced design type of regression only; the inability to handle random design regression makes the method very limited. Another major limitation of the method is its inability to forecast, which is quite important in a lot of econometric studies.

Besides, when the data points are similar from an individual to another (it happens often when error terms are correlated), the estimator breaks down. To see that, let's focus on the second expression of the estimator:

$$\lim_{y_{i,t} \to y_{j,t}} [\hat{m}_i(t) - \overline{\hat{m}(t)}] = 0, \qquad \forall i, j. \tag{3.3}$$

This equation says that the similar the data points are, the closer their estimates also are hence the difference between the estimate of an observation for curve i at a given time t and the average of the estimated values at that time is very close to 0. The implication of that result is that the Altman and Casella estimator will be dominated by population average $\bar{y}_{.t} = \frac{1}{n}\sum_{i=1}^{n} y_{it}$. But for a given period of time t, the population average $\bar{y}_{.t}$ and the y_{it}'s would have almost the same values making the estimator fit the raw data. This problem will be further discussed in chapter IV.

Last, it is worthwhile to note that the estimator is constructed on the assumption that the data (response variable) is normally distributed. This assumption could be wrong for small samples; in which case the estimator will be mispecified.

3.2. The Racine and Li Estimator

This estimator was motivated by the failure of nonparametric methods to estimate categorical data satisfactorily. The goal is to adequately estimate regression functions with many discrete variables without having to split the data into subsets, the number of which depends on the values that the categorical variables take on. The estimator can be however adapted in a context of multiple curve estimation.

Let $X^c \in R^q$ be a matrix of random continuous variables and X^d denote the discrete regressors. Instead of smoothing the discrete variables with a kernel function, the technique treats them separately of the continuous variables. They are smoothed as follows:

$$L(X^d(t), x^d(t), \lambda) = \begin{cases} 1 & \text{if } X^d(t) \in x^d(t) \\ \lambda & \text{otherwise} \end{cases} \tag{3.4}$$

In the context of multiple curve estimation, the discrete variable represents the different experimental units considered. If there are n individual curves to estimate the discrete variable is then $X^d = 1, 2, 3, ...n$ where the numbers $1, 2, ...n$ indicate the number of each curve. Consider the example of yield data in two counties from the same crop-reporting district. When estimating the curve of county one, the discrete variable smoother is:

$$L(X^d(t), x^d(t), \lambda) = \begin{cases} 1 & \text{for data from county one} \\ \lambda & \text{for data from county two} \end{cases} \tag{3.5}$$

When estimating the curve of county two it is the other way around:

$$L(X^d(t), x^d(t), \lambda) = \begin{cases} 1 & \text{for county two} \\ \lambda & \text{for county one} \end{cases} \tag{3.6}$$

The scalar λ is the smoothing parameter for the discrete variable. When the curves are not similar, λ should be as small as possible reverting the estimator back to the

ordinary Nadaraya-Watson estimator. When the curves are similar, the weight on the external observations should be high to reflect the similarities, hence boosting the performance of the estimates. It should be pointed out that for the smoothing of the discrete variable(s) the weights do not add up to 1. However that is a not a problem considering that the weights appear on both the numerator and the denominator of the estimator. As for the continuous variables, they will be smoothed by the standard c-variate kernel: $K_h = h^{-c}\prod_{i=1}^{c} k_h(x_i)$ where $k_h(x_i$ is the individual kernel estimate for each curve. Then the Racine and Li estimator can be written as:

$$\breve{m}(x) = \frac{\sum_{t=1}^{nT} y_t W_{h,x}}{\sum_{t=1}^{nT} W_{h,x}}$$

where $W_{h,x} = K_h(x)L(X_t^d, x_t^d, \lambda)$ after dropping the subsript indicating the different curves for simplicity. The smoothing parameters λ and h are both chosen by minimizing the cross-validation function:

$$CV(h,\lambda) = \sum_{t=1}^{nT}[y_t - \tilde{m}_t^{-t}]^2;$$

$\hat{m}_t^{-t}$ is a leave-one out estimator.

3.3. The Ker Estimator

The objective of this estimator is to reduce the bias of the standard Nadaraya-Watson estimator (NW) by starting with a nonparametric-pooled estimator and then using a correction function to adjust for individual effects. Recall that the leading terms of the bias of the Nadaraya-Watson estimator are a function of the curvature of the true function:

$$E(\dot{m}(x)) - m(x) = \frac{1}{2}\mu_2 h^2 m''(x) + 2m'(x)\frac{f'(x)}{f(x)} \tag{3.7}$$

Clearly one can decrease the bias by reducing the curvature of the $m(x)$.

3.3.1. The Method

The idea is to pool the data and estimate a single regression function $\hat{m}_p(x)$ using the Nadaraya-Watson estimator. Then when estimating the individual curve, the pooled estimator is multiplied by a correction factor $r(x) = \frac{m(x)}{m_p(x)}$ in order to adjust for individual effects. The correction factor is itself estimated nonparametrically by:

$$\hat{r}(x) = \sum_{t=1}^{T} \frac{y_t}{\hat{m}_p(X_t)} K_h(x - X_t). \tag{3.8}$$

The Ker estimator is:

$$\hat{m}(x_t) = \frac{\sum_{t=1}^{T} y_t [\frac{\hat{m}_p(x)}{\hat{m}_p(X_t)}] K_h(x - X_t)}{\sum_{t=1}^{T} K(\frac{x - X_t}{h})} \tag{3.9}$$

Again the subscript indicating individual i's curve has been drop for notational convenience. Ker(2000)showed that $\hat{m}(x)$ is biased and the leading terms of the bias are:

$$E(\hat{m}(x)) - m(x) = \frac{1}{2}\mu_2 h^2 m_p(x) r''(x) + 2m(x) r' \frac{f'(x)}{f(x)} \tag{3.10}$$

It can be shown that the variance converges at the same order than that of the Nadaraya-Watson estimator.

It is easy to see that the bias of the Ker estimator is not a function of the curvature of the true regression function as it is for the standard Nadaraya-Watson estimator. Rather it is function of the second derivative of the correction factor $r(x) = \frac{m(x)}{m_p(x)}$ which represents the "residual" curvature after starting with the pooled estimator. If the pooled estimator coincides with the true function, then $r(x)$will be a straight line hence $r'' = 0$. Statistically, this means that a significant part of the bias equal

to $\frac{1}{2}\mu_2 h^2 m_p(x) r''(x)$ will go away. But this is the ideal scenario that is not likely to happen in practice. The objective is to estimate $m_p(x)$ and $m(x)$ nonparametrically so that their ratio fluctuates around 1 making the "residual" curvature close to zero. If indeed there are structural similarities between the individual curves, the correction factor should be less variable than the individual true curves hence its curvature should be smaller than that of the individual curves.

Why use a pooled estimator as a start? Ker suggested that pooling the data of the different curves should provide more information when the curves are similar. Using a nonparametric-pooled estimator instead of a parametric one is driven by the possibility that the individual curves could belong to different parametric families.

3.3.2. **Advantages of the Ker Estimator**

The nonparametric estimator with a pooled start has some advantages that ought to be enumerated:

- when the curves are similar, it has a lower bias thanks to the correction factor and the "abundance" of data;
- unlike the Altman and Casella estimator, the Ker estimator does not assume any functional form of the data or the error term, circumventing the possibility of a wrong assumption about the underlying density;
- also this estimator does not require that the regressor be a fixed or balanced design. Random as well as fixed designs can be used as explanatory variables and they do not have to be of the same length for each individual.

3.4. The Major Limitation of The Advanced Methods

The main limitation of the competing estimators is the selection of the optimal number of individual curves. Since the objective is to gather enough "similar" data in order to produce more accurate estimates, it might be problematic if unnecessary or extraneous information is included in the estimation process. To limit such a possibility Ker proposes that cross validation be used for 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. The results of curve selection for the Ker estimator are presented in chapter IV. The procedure is applicable to both Ker and Racine and Li estimators.

4. Monte Carlo Simulations

4.1. **Efficiency Comparisons**

The main goal of this thesis is to undertake a performance comparison of the advanced methods based on the empirical and simulated data. This chapter focuses on Monte Carlo simulations. Monte Carlo methods are widely used in the applied econometric literature as a tool of exploration of finite sample properties of an estimator. The replications are undertaken for each estimator and for different sample sizes. Since the focus is on the advanced methods and their incorporation of similarities between different experimental units, four curves are generated with [0,1] interval as a support. A gaussian white noise error process is added to randomize the data. The four curves are[10]:

$$f(t) = sin(15\pi t) + \epsilon \tag{4.1}$$

$$g(t) = sin(5\pi t) + \epsilon \tag{4.2}$$

$$h(t) = 10e^{-10t} + \epsilon \tag{4.3}$$

$$z(t) = t + \epsilon \tag{4.4}$$

Estimates of the true functions are computed using the Nadaraya-Watson estimator; the Altman and Casella estimator; the Racine and Li estimator; the Ker estimator; and finally the pooled Nadaraya-Watson estimator. Two scenarios are

[10] The first three 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). The last curve was used by A.Yatchew in his article "Nonparametric Regression Techniques in Economics (1998).

investigated in the simulations. In the first scenario, all the individual curves are considered to be exactly the same and equal to $g(t)$. The choice of $g(t)$ does not bear any statistical reason and does not need to. The data from an individual to another will not be the same because of the random errors added. This is the ideal case obviously; the goal is to see how "better" the estimators making use of external information would be in a situation where all the individual curves were identical. Remember that the incorporation of external information is based upon the assumption that the different experimental units are similar to some extent.

In the second scenario, all of the four curves are dissimilar. This is another extreme case that would provide some insights about how the estimators making use of external information behave when the assumption of similarity is violated. Unfortunately the assumption of similarity could be wrong in some instances. If that happens, the use of external data could lead to an efficiency loss. The choice of these two scenarios is motivated by empirical econometric analysis. It is hard to believe that in empirical applications, the different experimental units in a set will be identical or totally unrelated; the truth lies somewhere between these two extreme cases.

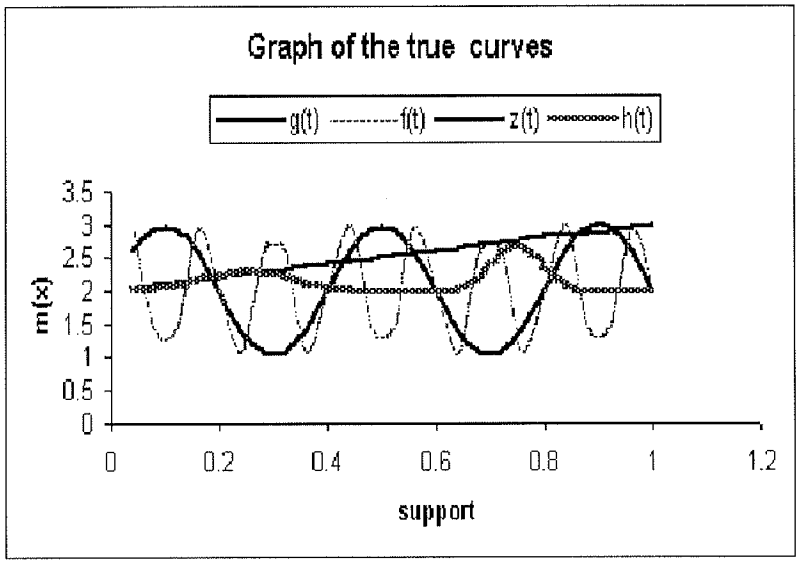


Figure 4.1. Graph of the true curves

4.1.1. Case of Identical Curves

As said above, only one curve $g(t)$ is generated and individual errors added to make the data vary from an individual to another. Advanced methods are expected to do much better (in terms of efficiency) than the standard methods because they have more information. The performance of each curve is assessed via its mean integrated squared error. The mean integrated squared error is computed twice; first using cross-validation to choose the smoothing parameter and second using squared error minimization to choose the smoothing parameter. The last method is not applicable in empirical studies because it requires the knowledge of the true regression function. The smoothing parameter was computed using two different methods: the least-squares cross-validation and the squared error minimization (best h). A summary of the mean integrated squared error results is reported in tables 4.1 and 4.2. More detailed results are posted as an appendix.

Overall the pooled estimator performed better than all the competing estimators; which was expected. Since all the four curves are exactly the same pooling the data (the pooled estimator has four times more data then the competing estimators) should result in a more precise estimation. However the advantage of the pooled estimator over the Racine and Li and the Ker estimators shrunk as the sample size grew up. The general tendency is that the advanced estimators did moderately to drastically better than the standard method for all the sample sizes. For the nonparametric estimator with a pooled start for example, the gains ranged between 50% and 66%[11]. Recall that the main objective of the nonparametric estimator with a pooled start (Ker estimator) is to reduce the bias of the Nadaraya-Watson estimator by replacing the

[11] Gains with respect to the Nadaraya-Watson estimator.

curvature of the true function-in the expression of the bias of $m(x)$- by the curvature of a correction factor. If the start is sufficiently close to the true function then the correction function should be close to 1 therefore having little curvature. Judging by the results of the simulations, it is clear that the goal was realized. When the curves are similar, the bias of the Ker estimator is considerably lower than that of the Nadaraya Watson as theoretically derived. The success of the Ker estimator in this case is highly attributable to a lessening in the bias of the Nadaraya-Watson estimator as shown in table 4.7.

The Racine and Li estimator also gained appreciably with respect to the Nadaraya-Watson estimator.

As for the NEB estimator, it remained less competitive than the other estimators as the sample size grew up. Altman and Casella (1995, page 513) stated that "the effectiveness of the NEB estimator is greater when there are more curves per set and when there are fewer design points per curve." In these simulations only four curves are used, which could account for the poor performance of the NEB.

4.1.2. **Case of Dissimilar Curves**

Before getting into the results of the replications, it should be mentioned that the study of this case is purely motivated by an investigation of the performance of the estimators when the curves are dissimilar. In an empirical analysis the advanced methods should be used only if a reasonable assumption of similarity between the individual curves can be made, otherwise their use could lead to an efficiency loss.

Contrary to the previous case, all four different curves are generated in order to assess the performance of the advanced methods when the assumption of similarity is

wrong. These simulations are very important; the odds of having a wrong assumption are not negligible at all. If the estimators lose too much with respect to the standards methods it may not be "safe" to use them. A summary of the mean integrated squared error results is reported in tables 4.3 and 4.4.

The results are encouraging for the Ker estimator and for the Racine and Li estimator. They did better than the NEB estimator and did not lose to the Nadaraya-Watson estimator as the sample size increased. It should be noted that the performance of the Ker estimator is greatly attributable to the correction factor $\hat{r}(x)$. Since the curves are not similar in this particular case, the correction factor adjusts the pooled estimator up or down depending on the individual curve that is being estimated. The possibility of accounting for individual effects makes the estimator perform well even when the start function is not suitable. As for the Racine and Li estimator, it lost slightly to the Nadaraya Watson due to the fact that irrelevant data are given a weight (ideally negligible since the data is not similar) but its performance was very improved as the data set grew up.

The replications are quite informative: they have shown there is a lot to gain and not much to lose by making use of external data. When the similarity assumption is correct, the advanced methods have a noticeably lower mean integrated squared error with the exception of the NEB. Even with quite dissimilar curves, the Ker and the Racine and Li estimators remained highly competitive. As said above this does not imply that the Ker and Racine and Li estimators can be used whenever there are two curves or more. They should be used only if one is reasonably confident in the similarities between the curves. The pooled estimator was expected to do less well because it uses extraneous data to estimate the individual curve without accounting

for local corrections. The nonparametric estimator with a pooled start worked well basically because it allows for individual corrections to be made. Therefore even when the curves are quite dissimilar, starting with the pooled estimator is still acceptable thanks to the correction factor.

As previously said, the logic behind the choice of the two extreme cases is that the truth lies somewhere between the two when it comes to empirical applications. Judging from their performance in the simulations, there is a strong probability that the use of the Ker and Racine and Li estimators will generate a better estimation of the individual curves even when the similarity assumption is violated.

4.2. **Advanced Methods and Curve Selection**

Until now, the emphasis has been on the inclusion of external data from possibly similar curves for efficiency gains. However, finding the "optimal" curves to include in the estimation process is an important concern in the use of the advanced methods. It is clear that if the individual curves do not have the similarities they are assumed to have, using the advanced methods could lead to inefficient or worse misleading estimates. To select the individual curves for the advanced methods, Ker proposes that cross-validation be used. For the estimation, one would use the sample of curves that minimizes the cross-validation score. The method works the same way than the cross-validation for the smoothing parameter and is applicable to both the Ker and Racine and Li estimators. In this chapter only the curve selection results for the Ker estimator are presented for sample sizes T = 25, 50, 100 and 500. As in the previous section, two extreme cases are investigated: a case of identical curves and a case of

dissimilar curves. For simplicity only the results for the curve $g(t)$ are presented.

The results are posted in tables 4.5 and 4.6. They show the number of times the different curves $f(t)$, $g(t)$, $h(t)$ and $z(t)$ have been included in the pooled start of the Ker estimator for the estimation of $g(t)$.

The replications show that when the curves are all similar, the Ker estimator makes use of the data from the remaining curves in the population. Each of the four curves has been selected more than 70% of the time for all sample sizes; which is consistent with the expectations given the curves are identical. The percentage of inclusion grew up with the sample size. The number of times the Nadaraya-Watson estimator has been used in negligible.

When the curves are dissimilar, the opposite is observed: fewer curves are included in the pooled start. This also is consistent with expectations since the curves are dissimilar. Although the remaining curves have been included the percentage of inclusion decreased as the sample size grew up except for $g(t)$ which is the curve being estimated here.

TABLE 4.1. Average MISE*100 for the case of identical curves: method of least squares cross-validation.

T = 25				
NW	NEB	Ker	Pooled	Racine and Li
20.76907	23.856748	10.393694	4.6224098	7.4805452
T=50				
9.1796395	15.363084	4.7476092	2.7597298	3.3706197
T=100				
4.9576954	12.676808	2.4369758	1.5798329	1.73607
T=500				
1.3344219	9.9716733	0.6116169	0.4474773	0.4543125

TABLE 4.2. Average MISE*100 for the case of identical curves: method of squared error minimization.

T = 25				
NW	NEB	Ker	Pooled	Racine and Li
11.993081	21.491211	5.964608	4.110164	4.1446102
T =50				
7.2192163	16.024076	3.4170944	2.4731641	2.4938289
T =100				
4.2317549	12.200125	1.9229589	1.4583455	1.4704434
T =500				
0.9461135	8.721456	0.5456322	0.4007325	0.4272035

TABLE 4.3. Average MISE*100 for the case of dissimilar curves: method of least squares cross-validation.

T = 25				
NW	NEB	Ker	Pooled	Racine and Li
22.604844	29.590071	23.472613	31.511426	23.243858
T = 50				
9.6582812	20.557586	11.085589	29.579152	13.15461
T = 100				
4.9297477	17.785701	6.4515517	27.606241	7.0070324
T = 500				
1.3260667	15.722575	1.9691218	25.589078	2.0751343

TABLE 4.4. Average MISE*100 for the case of dissimilar curves: method of squared error minimization.

T = 25				
NW	NEB	Ker	Pooled	Racine and Li
11.558408	22.864163	12.823735	30.437596	6.7373031
T = 50				
7.0655439	19.550934	8.8510286	28.77789	10.953052
T = 100				
4.218905	17.488481	5.8698167	27.253604	16.843468
T = 500				
1.2245831	15.694512	1.8748517	25.554444	2.0525395

TABLE 4.5. Curve selection for the Ker estimator for the case of similar curves.

T = 25				
No start	g(t)	f(t)	z(t)	h(t)
0.2	81.8	79.8	73.8	72.6
T = 50				
0.4	78	81.8	78.8	78.6
T = 100				
0	75.2	81.6	80.8	79.2

Tables 4.5 and 4.6 exhibits the percentage each of the four curves has been selected to be part of the pooled start of the Ker estimate of $g(t)$. The column "No start" refers to the ordinary Nadaraya-Watson estimator; if no curve is included in the pooled start, the Ker estimator reverts back to the Nadaraya-Watson estimator. .

TABLE 4.6. Curve selection for the Ker estimator for the case of dissimilar curves.

T = 25				
No start	g(t)	f(t)	z(t)	h(t)
1.6	66.8	35.8	41.4	36.2
T = 50				
0.8	92	14	48.8	20.8
T = 100				
0.6	94	16	36.5	18.8

TABLE 4.7. Bias results for the case of similar curves.

T = 25		
NW	Ker	Pooled
1.5435512	0.642875	0.8548098
T = 50		
1.4156884	0.4390969	0.5059055
T = 100		
0.8851497	0.2987018	0.3167598
T = 500		
0.2686621	0.089205	0.0940388

TABLE 4.8. Bias results for the case of dissimilar curves.

T = 25		
NW	Ker	Pooled
12.833349	12.92856	28.440004
T = 50		
1.0290213	1.6715469	26.95288
T = 100		
0.920278	1.6132288	25.81984
T = 500		
0.2846808	0.4738972	25.060287

5. EMPIRICAL APPLICATIONS

5.1. Nonparametric Wage Modeling

To assess the predictive ability of the competing estimators it is important to carry out an out of sample prediction based on empirical data. To do so a cross-section dataset on male and female wages is used. More explicitly, the dataset considered has three explanatory variables:

- a dummy variable representing the sex of the worker
- a dummy variable representing the union status of the worker
- a continuous variable representing the experience of the worker.

The explained variable is the natural log of the wage of the worker. The data can be summarized in a regression function as follows:

$$y_t = m(sex_t, union_t, experience_t) + \epsilon_t, t = 1..534 \tag{5.1}$$

where $y_t = log(wage)_t$. The dataset has 534 observations. The first dummy variable (sex) was used to separate the data into two subsets representing male and female data. The data of each subset is used to estimate the male and female curves and also to predict out of sample. Out of the 534 observations, 106 (roughly 20%) were set aside as a hold-out sample for prediction purposes. The prediction performance of each estimator is measured by the mean integrated prediction error (MIPE). The formula for the MIPE is:

$$MIPE = \frac{1}{T_2}\sum_{t=1}^{T_2}(\hat{y}_t - y_t)^2 \tag{5.2}$$

where T_2 is the size of the hold out of sample upon which the prediction is based and $\hat{y}_t$ is the predicted value of y_t. For this study the Altman and Casella estimator is disqualified by since the dimension of the regression matrix exceeds 1.

Even though this is just an application on labor data, it is appropriate to explain the structure of wage equations to help the reader who is not familiar with labor economics. The dependant variable is the natural log of wage instead of wage, the reason being that the distribution of the wages is not believed to be normal but rather positively skewed. [12] The log of the wages looks normally distributed than that of wages hence the use of log(wage) as a dependant variable. Besides it is widely believed in labor economics that factors such as experience and education affect wages in a multiplicative manner. Combining these two elements produces a linear relationship between log(wage) and the explanatory variables; such relationship is needed to apply linear least squares methods. The normality assumption of the log of wages also guarantees that the error term is normally distributed.

Single equation estimation has been commonly used by labor economists to investigate important economic issues such as wage inequalities between male and female workers or whites and non-whites workers. It is also used to analyze the effect of unions on earnings. A pioneering work has been done by Oaxaca (1973) in wage differentials. In most cases however, the estimation of wage equations has been confined to parametric estimators such as the ordinary least squares method. In this application, OLS as well as advanced nonparametric methods have been used to estimate the wage equations of male and female workers and compare their predictive ability.

[12] See Schooling, Experience, and Earnings by Mincer, J (1974).

Let the following equations

$$y_f = g(x_{1f}, x_{2f}, ..., x_{Tf}) \tag{5.3}$$

$$y_m = h(x_{1m}, x_{2m}, ..., x_{Tm}) \tag{5.4}$$

$$y_p = p(x_{1f}, x_{1m}, x_{2f}, x_{2m}, ..., x_{Tf}, x_{Tm}) \tag{5.5}$$

be the female, male and pooled male-female wage equations respectively where $y_k = log(wage)_k$ and the x_{ij}'s are the explanatory variables such as experience, years of schooling, union status etc. Also let $\bar{y_f}$, $\bar{y_m}$ and $\bar{y_p}$ represent the mean relative wages for females, males and the pooled males and females respectively. $\hat{\bar{y}_f}$ and $\hat{\bar{y}_p}$ is interpreted as the estimated overpay or underpay (depending on the sign) of female workers. Similarly the difference between $\hat{\bar{y}_m}$ and $\hat{\bar{y}_p}$ measures the favoritism or discrimination of male workers. Given the sensibility of the wage discrimination issue it is important to estimate it accurately.

As said previously, the performance comparison between the estimators is based on the out of sample prediction error (MIPE). The results are displayed in table 5. The Wilcoxon nonparametric test suggests that the MIPE of the Ker estimator is significantly lower than that of the competing estimators at a 1% level. To reduce criticism of the results, the out of sample was randomly picked and the prediction error computed; the procedure was replicated 100 times. The reported values are the averages of the MIPE across the 100 replications.

TABLE 5.1. Out of sample prediction error of the different methods.

Out-of-sample=106 obs				
NW	Ker	Racine and Li	Pooled	OLS
397.97496	378.81303	386.07591	440.34925	544.26751
(8.588)	.	(8.276)	(8.544)	(8.626)

The numbers in brackets are the computed critical values of the Wilcoxon (Sign) test $H_0 : X - Ker = 0$ against $H_1 : X - Ker > 0$, X= {NW,Racine and Li, Pooled, OLS}. The critical values reveal that H_0 can be rejected in favor of H_1 at a 1% significance level.

5.2. Rating Crop Insurance Policies Using Advanced Methods

Crop insurance is a major component of the agricultural system in the U.S [13]. In his 1999 state of the union address to Congress, President Clinton reiterated the important role of the crop insurance program:

> " As this Congress knows very well, dropping prices and the loss of foreign markets have devastated too many family farmers. Last year, the Congress provided substantial assistance to help stave off disaster in American agriculture, and I am ready to work with lawmakers of both parties to create

[13] A detailed discussion of the U.S. crop insurance program can be found in Ker and McGowan (2000); McGowan (1999); and Goodwin and Ker (1998).

a farm safety net that will include crop insurance reform and farm income assistance."

In 2000, the Congress passed the "Agricultural Risk Protection Act (ARPA)", which is estimated to cost 8.2 billion over a five-year period. ARPA effectively raised the federal expenses on crop insurance to an estimated 16.1 billion. The key feature of the federal crop insurance is its regulation by the federal government via the Department of Agriculture' Risk Management Agency (RMA). The RMA is in charge of rating crop insurance policies. Given the rates set by the RMA, the private insurance companies either accept or decline the proposed contracts depending on their expected payoffs. For simplicity, we assume that insurance companies are risk-neutral and profit maximizers. Since information on yield data is freely available, insurance companies have the possibility to assess whether the proposed contracts are underrated or overrated by using different estimators than the RMA. If their estimator(s) offer a superior performance, they can weed out all the contracts deemed undesirable.

In this section, we undertake simulations designed to investigate the performance of the advanced methods relative to the one-knot linear spline used by the RMA to model the temporal process of yields:

$$y_t = \alpha + \beta_1(tI_{(0,\delta)}(t) + \delta(1 - I_{(0,\delta)}(t)) + \beta_2(1 - I_{(0,\delta)}(t)(t - \delta)) + \epsilon(t) \qquad (5.6)$$

where $I(.)$ is an indicator function.

To investigate the relative performance of the competing estimators, we use soybeans data for 87 counties in Iowa grouped in 9 crop-reporting districts. For each county, data is measured between 1956 and 1999. Out of the 44 observations available

for each county 15 were used as a hold out sample to evaluate the empirical rates. To evaluate the rate for say 1990, all the data up to and including 1989 was included.[14] Let Γ be the universe set of 1305 policies (87 counties x 15 years), then the loss ratio can be defined as:

$$LR = \frac{\sum_{i\in\Gamma} max(0, \alpha y^e - y_i)w_i}{\sum_{i\in\Gamma} \hat{\eta}_i w_i} \tag{5.7}$$

where w is a weight matrix such that w_i=1 if policy i is retained and 0 otherwise; $\hat{\eta}_i$ is the premium rate estimated by the RMA for policy i, y^e is the predicted yield for the year of coverage using the different estimators and α is the coverage level.

As outlined earlier, the temporal process is believed to be similar from a county to another within the same CRD or sate for reasons such as weather, soil type or technology among other factors. If indeed the similarity assumption is correct, using external data while rating policies for a given county should yield a smaller loss ratio for the insurance company. The simulations are based on the following basic assumption:

Insurance companies use both the Ker, the Racine and Li and the pooled Nadaraya-Watson estimators to evalutate their loss ratios given that the government is using the one-knot spline estimator to model the temporal process of yields.

[14]The data was obtained from the National Agricultural Statistics Service website: www.nass.gov.

Table 5.2 provides the results of the simulations for the 75% and 85% coverage levels. The results indicate that the advanced methods provide a much smaller loss ratio for the insurance companies if the governement is using the one-knot linear spline. As such, the use of the Racine and Li or Ker estimator enables insurance companies to save a lot of money buy effectively weeding out contratcs with big loss ratios: the overal loss ratio of the program is almost twice as big as that of the insurance companies if they were to use the Racine and Li or the Ker estimator for a 75% coverage level. However, only the Ker estimates are found statistically significant at a 5% level for the 85% coverage level. For the 75% coverage level both the Racine and the Ker are significant but the pooled Nadaraya-Watson is not.

TABLE 5.2. Crop Insurance Rating simulation for soybeans for all 87 counties in Iowa

75 Percent Coverage Level			
	Ker	Pooled-NW	Racine and Li
Insurance Company Loss Ratio	0.48943	0.8413287	0.49456
Percentage of Contracts Retained	0.4812261	0.4015326	0.06283
Loss Ratio of RMA	1.3630571	0.8206314	0.86809
Overall Loss Ratio	0.832077	0.832077	0.832077
85 Percent Coverage Level			
	Ker	Pooled-NW	Racine and Li
Insurance Company Loss Ratio	0.6273756	0.7408744	0.455656
Percentage of Contracts Retained	0.5111111	0.3478927	0.0666667
Loss Ratio of RMA	1.5783423	1.2031936	1.023927
Overall Loss Ratio	0.968261	0.968261	0.968261

6. Summary and Conclusion

A main objective of applied econometrics is to find the most efficient estimators so as to be closer to the true unknown function to estimate. Years of research have enriched the econometric literature with a multitude of estimation techniques, each of which has its own merits. Nonparametric methods are certainly younger than their parametric counterparts but they are not any less worthy. It is now proven that they can be either suitable substitutes or complementary (semiparametric methods) to parametric techniques for both regression and density estimation. The attractiveness of nonparametric methods lies in their flexibility. By not constraining the data to any parametric family, they generally succeed in displaying a good structure of the true relationship. Such flexibility however comes with the cost of a much more tedious statistical inference and a lower rate of convergence compared to parametric estimators when the assumed functional form is correct. The latter concern should not discourage anyone from using nonparametric methods since the very reason for their use is that the true function is unknown!

In this thesis, the goal was to evaluate some leading nonparametric regression methods, which were divided into two categories: the standard estimators and the advanced ones. In the quest of finding the most appropriate estimators for multiple curve regression, the mean squared error was used as the sole arbiter. From yield data to stock market data, economic data is often presented in a form of a set of curves. Advanced methods can then be used to boost the precision of the estimates. As pointed out earlier in the thesis, these methods have two goals: significantly out

perform the standard methods when the assumption of similarity is correct while not loosing much of efficiency when the curves are quite dissimilar. The results of the replications are enlightening about what are the "best" nonparametric methods for a multiple curve estimation. The simulations showed that when the curves are very similar, the advanced methods have a considerably lower mean integrated squared error than the standard ones except for the Altman and Casella estimator. It is not really surprising for the advanced methods to perform well when the curves are the same since they use more data than the standard ones do. However when the curves are quite dissimilar, the results are very disappointing for two of the four competing estimators: the NEB and the NW pooled estimators. The pooled estimator was not expected to do well because it uses extraneous data to estimate a single regression function without allowing for individual corrections. The Racine and Li estimator performed very well as the data set got bigger. The nonparametric estimator with a pooled start worked satisfactorily basically because it allows for individual corrections to be made. Therefore even when the curves are quite dissimilar, starting with the pooled estimator is still acceptable thanks to the correction factor.

The nonparametric with a pooled start and the Li and Racine estimator did not loose much to the standard method when the curves were all dissimilar and gained a lot when the curves were all the same; which means that these two methods are the most appropriate to use among the advanced methods studied in this thesis for they remain viable even when the curves are dissimilar.

A. APPENDIX: DERIVATION OF THE NW ESTIMATOR.

By definition a probability density function is:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h). \tag{A.1}$$

Therefore an estimator of the true density can be obtained by replacing $P(x - h < X < x + h)$ by the percentage of the data falling in the interval $(x - h < X < x + h)$, which leads to:

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} w \frac{(x - X_t)}{h}$$

where $w(u) = \frac{1}{2} I_{|u|<1}$. However $w(u)$ is generally replaced by a smooth twice differentiable function and symmetric around 0 called kernel function, hence

$$\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} K \frac{(x - X_t)}{h} \tag{A.2}$$

where $K(.)$ is the kernel function. By definition,

$$m(x) = E[Y|X] = \frac{\int y f(x, y) dy}{\int f(x, y) dy} \tag{A.3}$$

Analogically to the one-diemnsional estimator $\hat{f}(x)$, the bi-diemnsional estimator is:

$$\hat{f}(x, y) = \frac{1}{Th} \sum_{t=1}^{T} K \frac{(x - X_t)}{h} K \frac{(y - Y_t)}{h} \tag{A.4}$$

Using one of the properties of the kernel funtion : $\int K(u) du = 1$, yields
$\int \hat{f}(x, y) dy = \frac{1}{Th} \sum_{t=1}^{T} K \frac{(x - X_t)}{h}$
$\int y \hat{f}(x, y) dy = \frac{1}{Th} \sum_{t=1}^{T} K \frac{(x - X_t)}{h} \int y K_h(y - Y_t) dy$
$\int y \hat{f}(x, y) dy = \frac{1}{Th} \sum_{t=1}^{T} K \frac{(x - X_t)}{h} Y_t.$
Substituting these results in the $E[Y|X]$ gives the desired result.

B. APPENDIX: MISE RESULTS FOR EACH INDIVIDUAL CURVE.

B.1. Method Of Least Squares Cross-Validation

B.1.1. Case Of Identical Curves

T = 25					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	20.72418	21.243134	10.018895	4.6224098	7.4651668
f(t)	20.129997	21.09666	10.393866	4.6224098	7.5496414
z(t)	21.537767	27.186135	10.590098	4.6224098	7.4038587
h(t)	20.684336	25.901065	10.571917	4.6224098	7.503514
T = 50					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	9.2709134	15.534941	4.6577056	2.7597298	3.3291956
f(t)	9.1627238	15.568016	4.6879064	2.7597298	3.3969154
z(t)	9.1018502	15.142989	4.8303613	2.7597298	3.3296771
h(t)	9.1830705	15.206391	4.8144634	2.7597298	3.4266907
T = 100					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	4.9435386	12.610485	2.4422237	1.5798329	1.7234327
f(t)	5.0232013	12.852991	2.4192199	1.5798329	1.7572632
z(t)	4.9248814	12.737523	2.4049929	1.5798329	1.7151046
h(t)	4.9391602	12.506234	2.4814668	1.5798329	1.7484793
T = 500					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	1.3310111	9.9587196	0.6162842	0.4474773	0.453653
f(t)	1.3373609	9.9085542	0.5989836	0.4474773	0.454861
z(t)	1.3294284	10.005928	0.6288069	0.4474773	0.4532038
h(t)	1.3398871	10.013492	0.602393	0.4474773	0.4555322

B.1.2. **Case of Dissimilar Curves**

T = 25					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	21.673759	31.148529	21.24108	44.080997	26.619727
f(t)	51.468621	41.397166	51.671092	53.944872	50.330942
z(t)	7.7171983	22.396525	9.1187993	19.295811	8.3986675
h(t)	9.5597957	23.418062	11.859478	8.7240261	7.6260945
T = 50					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	9.2709134	20.301806	9.8094033	39.879477	11.582789
f(t)	20.628567	29.595526	20.14619	50.449841	20.507959
z(t)	3.4872522	15.278363	6.4695597	18.181921	10.233868
h(t)	5.2463925	17.054648	7.9172049	9.8053679	10.293824
T = 100					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	4.9435386	17.938383	5.6296867	35.989214	5.7132496
f(t)	9.9676379	22.250233	9.8606091	45.508545	11.772208
z(t)	1.6718378	14.679067	5.1230165	17.907822	5.1937916
h(t)	3.1359762	16.27512	5.1928944	11.019384	5.3488807
T = 500					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	1.3310111	16.070122	1.6200937	34.484919	1.6932427
f(t)	2.6901558	17.386571	2.6263977	38.987476	3.4192943
z(t)	0.4112157	14.514309	1.8710859	17.646661	1.5814448
h(t)	0.8718843	14.919299	1.7589098	11.237256	1.6065554

B.2. Method of Squared Error Minimization

B.2.1. Identical Curves

T = 25					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	12.07322	21.033366	5.8891287	4.110164	4.1370921
f(t)	11.956001	21.567046	5.9276014	4.110164	4.1499222
z(t)	11.868447	21.792663	6.0209777	4.110164	4.1406882
h(t)	12.074658	21.571768	6.0207242	4.110164	4.1507383
T = 50					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	7.1237467	17.613589	3.3718964	2.4731641	2.4875523
f(t)	7.29325	15.467719	3.3917286	2.4731641	2.4972436
z(t)	7.1826636	15.577487	3.4474006	2.4731641	2.4941105
h(t)	7.2772049	15.437507	3.457352	2.4731641	2.4964093
T = 100					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	4.1834606	12.012177	1.9036003	1.4583455	1.4663148
f(t)	4.2975123	12.048319	1.9323622	1.4583455	1.4754626
z(t)	4.1769992	12.399812	1.9412467	1.4583455	1.4688432
h(t)	4.2690475	12.340195	1.9146263	1.4583455	1.471153
T = 500					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	0.8489	2.9323	0.3648	0.3884	0.4032
f(t)	0.8158	2.9516	0.3517	0.3884	0.4028
z(t)	0.8141	2.9255	0.3516	0.3884	0.4041
h(t)	0.8484	2.9363	0.368	0.3884	0.4078

B.2.2. **DissimiLar Curves**

T = 25					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	12.07322	23.626352	12.368647	42.358604	14.354198
f(t)	26.805463	35.249329	26.783875	52.929029	31.093744
z(t)	2.8590875	15.92908	5.0556504	18.185358	10.97880
h(t)	4.4958631	16.651892	7.086767	8.2773916	10.947127
T = 50					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	7.1237467	19.605836	7.8859699	37.54812	8.953505
f(t)	16.018863	27.690728	15.897717	48.659762	19.467898
z(t)	1.6882511	14.580068	5.3427849	18.22484	7.676760
h(t)	3.4313147	16.327104	6.2776424	10.678838	7.7140464
T = 100					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	4.1834606	17.606796	4.9931094	35.341562	5.3857198
f(t)	9.3490553	22.016201	9.1886389	43.817845	11.608215
z(t)	0.9553908	14.307412	4.5827652	18.255705	4.8808498
h(t)	2.3877132	16.023513	4.7147534	11.599305	5.0744279
T = 500					
Curve	NW	NEB	Ker	Pooled	Racine and Li
g(t)	1.2152272	16.048592	1.5405973	34.544591	1.663336
f(t)	2.6655793	17.421745	2.5673953	38.519849	3.4182853
z(t)	0.2695979	14.44411	1.7407919	17.777637	1.5512695
h(t)	0.7479282	14.8636	1.6506221	11.3757	1.5772674

C. APPENDIX: BIAS SIMULATION RESULTS

C.1. **Case of Identical Curves**

T = 25			
Curve	NW	Ker	Pooled
g(t)	1.7086196	0.6599748	0.8548098
f(t)	1.5754048	0.6713257	0.8548098
z(t)	1.3577803	0.627912	0.8548098
h(t)	1.5324003	0.6122874	0.8548098
T = 50			
Curve	NW	Ker	Pooled
g(t)	1.3027837	0.4468328	0.5059055
f(t)	1.5293041	0.4380713	0.5059055
z(t)	1.4090016	0.4327291	0.5059055
h(t)	1.4216641	0.4387543	0.5059055
T = 100			
Curve	NW	Ker	Pooled
g(t)	0.8580819	0.3071596	0.3167598
f(t)	0.9220165	0.3044441	0.3167598
z(t)	0.8705085	0.3039829	0.3167598
h(t)	0.8899919	0.2792207	0.3167598
T = 500			
Curve	NW	Ker	Pooled
g(t)	0.2639614	0.088872	0.0940388
f(t)	0.2836919	0.0883134	0.0940388
z(t)	0.2627882	0.0937041	0.0940388
h(t)	0.2642067	0.0859305	0.0940388

C.2. Case of Dissimilar Curves

T = 25			
Curve	NW	Ker	Pooled
g(t)	1.3443316	1.58388	41.009574
f(t)	48.211876	46.67916	50.873449
z(t)	0.4336558	0.9883468	16.224388
h(t)	1.3435328	2.4628515	5.6526033
T = 50			
Curve	NW	Ker	Pooled
g(t)	1.3027837	1.4848416	37.253205
f(t)	1.5177201	1.7175819	47.823569
z(t)	0.2409215	1.4033739	15.555649
h(t)	1.0546599	2.0803902	7.1790961
T = 100			
Curve	NW	Ker	Pooled
g(t)	0.8580819	1.0383854	34.202813
f(t)	1.823169	1.8398797	43.722144
z(t)	0.1576054	1.828484	16.121421
h(t)	0.8422557	1.7461662	9.2329829
T = 500			
Curve	NW	Ker	Pooled
g(t)	0.2639614	0.3253766	33.956128
f(t)	0.6460724	0.5591815	38.458685
z(t)	0.0581047	0.5272765	17.11787
h(t)	0.1705845	0.4837541	10.708465

References

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

[2] Fan, Y., and A. Ullah. "Asymptotic Normality of a Combined Regression Estimator." *Journal of Multivariate Analysis* 71(1997): 191-240 *Annals of Statistics* 12(1984): 210-229.

[3] Goodwin, B., and A. P. Ker. "Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group-Risk Crop Insurance Contracts." *American Journal of Agricultural Economics* 80 (February 1998): 139-153.

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

[5] 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. *Journal of the American Statistical Association* 83(1988): 1084-1089.

[6] Ker,A.P. "Nonparametric Estimation of Possibly Similar Densities With Application to the U.S. Crop Insurance Program." Working paper, Department of Agricultural and Resource Economics, University of Arizona, 2000.

[7] Ker, A.P., and K. Coble. "Modeling Yields." Working paper, Department of Agricultural and Resource Economics, University of Arizona, 2000.

[8] Ker, A.P., and B.K. Goodwin. "Nonparametric Estimation of Crop Insurance Rates Revisited." *American Journal of Agricultural Economics* 82 (May 2000): 463-478.

[9] Ker, A.P., and P. McGowan. "Weather-Based Adverse Selection and the U.S. Crop Insurance Program: The Private Insurance Company Perspective." *American Journal of Agricultural Economics* 25 (December 2000): 386-410.

[10] McGowan, J. P. "El Nino/Southern Oscillation Based Adverse Selection in the U.S. Crop Insurance Program." MS thesis, Department of Agricultural and Resource Economics, University of Arizona, 1999.

[11] Mincer, J. *Schooling, Experience, And Earnings.* New York: Columbia University Press, 1974.

[12] Nadaraya, E. A. "On Nonparametric Estimates of Density Functions and Regression Curves." *Theory of Applied Probability* 10(1965): 186-190.

[13] Oaxaca, R."Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14 n3(1973): 693-709.

[14] Olkin,I., and Spiegelman, H."A Semiparametric Approach to Density Estimation." *Journal of the American Statistical Association* 82 n3(399): 858-865.

[15] Racine,J. S., and Qi Li. "Nonparametric Regression with Both Categorical and Continuous Data." Working paper, Department of Economics, University of South Florida , 2000.

[16] Silverman, B. W. "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting." *Journal of the Royal Statistical Society* Series B, 47(1985): 1-52.

[17] Terrell, G. R. "The maximal Smoothing Principle in Density Estimation." *Journal of the American Statistical Association* 85(1990): 470-477.

[18] Ullah, A., and Vinod, H.D. "General Nonparametric Regression Estimation and Testing in Econometrics." *Handbook of Statistics* 11(1993): 85-115.

[19] Yatchew, A. "Nonparametric Regression Techniques in Economics." *Journal of Economic Literature* 36(June 1998): 669-721.

[20] Wand, M.P., and M.C. Jones. *Kernel Smoothing.* London, New York: Chapman and Hall, 1995.

[21] Watson, G. S. "Smooth Regression Analysis." *Sankhya* 26:15 (1964): 175-184.