Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation

Jeffrey S. Simonoff

International Statistical Review / Revue Internationale de Statistique, Vol. 66, No. 2. (Aug., 1998), pp. 137-156.

Stable URL:
http://links.jstor.org/sici?sici=0306-7734%28199808%2966%3A2%3C137%3ATSOSCD%3E2.0.CO%3B2-5

International Statistical Review / Revue Internationale de Statistique is currently published by International Statistical Institute (ISI).

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/isi.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Wed Apr 26 12:05:17 2006

International Statistical Review (1998), **66**, 2, 137–156, Printed in Mexico
© International Statistical Institute

Three Sides of Smoothing: Categorical Data Smoothing, Nonparametric Regression, and Density Estimation

Jeffrey S. Simonoff

Department of Statistics and Operations Research, Leonard N. Stern School of Business, New York University, New York, NY 10012–0258, USA

Summary

The past forty years have seen a great deal of research into the construction and properties of nonparametric estimates of smooth functions. This research has focused primarily on two sides of the smoothing problem: nonparametric regression and density estimation. Theoretical results for these two situations are similar, and multivariate density estimation was an early justification for the Nadaraya–Watson kernel regression estimator.

A third, less well-explored, strand of applications of smoothing is to the estimation of probabilities in categorical data. In this paper the position of categorical data smoothing as a bridge between nonparametric regression and density estimation is explored. Nonparametric regression provides a paradigm for the construction of effective categorical smoothing estimates, and use of an appropriate likelihood function yields cell probability estimates with many desirable properties. Such estimates can be used to construct regression estimates when one or more of the categorical variables are viewed as response variables. They also lead naturally to the construction of well-behaved density estimates using local or penalized likelihood estimation, which can then be used in a regression context. Several real data sets are used to illustrate these points.

Key words: Kernel estimator; Local likelihood estimator; Local polynomial estimator; Maximum penalized likelihood estimator; Poisson regression.

1 Introduction

Consider the following three data situations:

(1) Figure 1 is a scatter plot that refers to an exploration of the relationship between "objective" and subjective ratings of the difficulty of various school text passages. The horizontal axis gives the Flesch–Kincaid Grade Level, a common readability formula, while the vertical axis gives the subjective opinions of the appropriate grade level of the passage by members of the staff of an education program in New York City. These passages can be used to perform an Informal Reading Inventory, whereby the reading level of a student is assessed as being the grade level score of the most difficult passage that a student comprehends (based on their answers to questions after reading the passage). It is crucial to determine accurately the appropriate grade level of the passage. While it is clear that there is a relationship between the Flesch–Kincaid level and the opinions of the teachers, the nature of that relationship is not obvious. If teachers' ratings are very different from those of the Flesch–Kincaid score, that makes use of the easily available Flesch–Kincaid score problematic. The plot differentiates between fiction (circles)

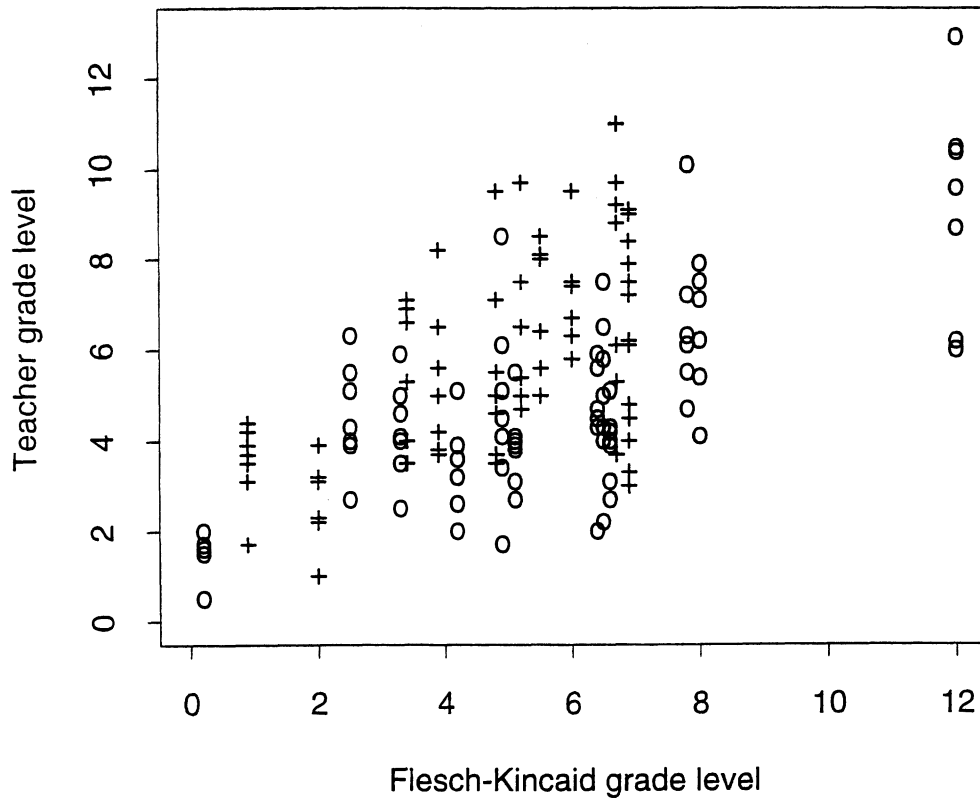

Figure 1. *Scatter plot of teacher assessments of grade level versus Flesch–Kincaid grade level scores for school text passages. Nonfiction books are given as pluses, while fiction books are given as circles.*

and nonfiction (pluses) books. Is the relationship between the two grade level scores different for the two types of books?

(2) Table 1 is a cross-tabulation of the fiscal health of 57 of the 75 largest U.S. cities, as measured by the Standard and Poor's 1994 rating for their general obligation bonds (the other 18 cities do not have general obligation bonds), by the number of not-for-profit organizations in that city among the city's top ten employers, according to Chamber of Commerce lists in 1995 (Abzug, Ahlstrom & Simonoff, 1997). Bond rating, which is ordered from AAA (top rating) to BB (lowest rating for these cities), has been shown to be a good proxy for a city's general fiscal health (Przybylski, Littlepage & Rosentraub, 1996), and number of organizations among the top ten employers is a measure of the importance of the not-for-profit sector in the life of the city. What is the relationship between prominence of the nonprofit sector in a city and its fiscal health?

(3) Figure 2 refers to a bivariate data set that gives the percentage of students who were from the top 10% of their high school class, and the percentage of students who graduated within six years, for 201 research universities in the United States (U.S. News and World Report, 1996). It is important for college administrators to understand the relationship between these two variables, since the former is a measure of the quality of students admitted, while the latter is a measure of the ultimate success of enrolled students. The scatter plot reveals a positive association between the variables, but gives little information about the possibility of relatively low or high density regions in the joint distribution.

Table 1

Standard and Poor's ratings of general obligation bonds by the number of not-for-profits among the top ten employers for U.S. cities.

Bond rating	Number of not-for-profits among top 10 employers					
	0	*1*	*2*	*3*	*4*	*5*
AAA	0	4	2	0	0	0
AA+	0	2	2	2	0	0
AA	4	9	5	2	1	0
AA−	1	1	4	0	0	0
A+	0	0	0	1	1	0
A	0	1	1	4	1	0
A−	1	0	2	0	0	0
BBB+	0	0	0	0	0	0
BBB	1	0	1	1	0	0
BBB−	0	0	0	0	0	2
BB+	0	0	0	0	0	0
BB	0	0	0	1	0	0

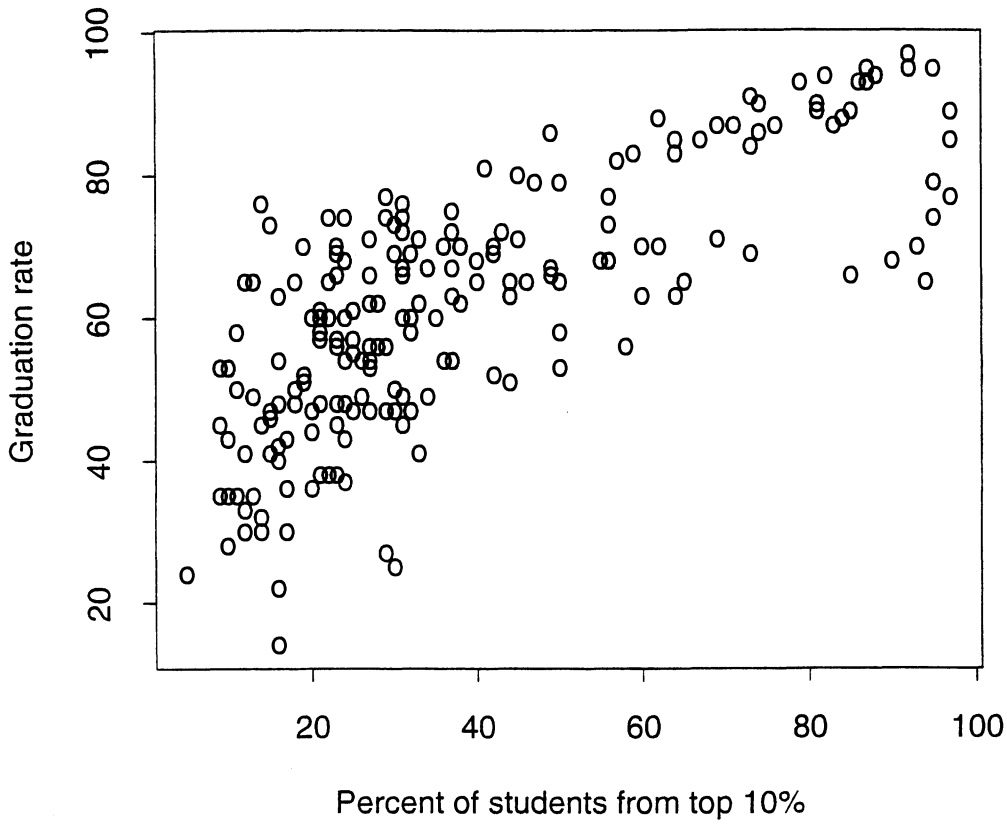

Figure 2. *Scatter plot of graduation rate of students versus percent of students from the top 10% of their high school classes for U.S. research universities.*

Despite the different natures of these three situations, all are closely related to each other. The idea that ties all of them together is the notion that whatever the form of the functional of interest (a regression curve in (1), a cell probability matrix in (2), and a bivariate density function in (3)), it is likely to be smooth, implying that data values that are "close" to a given point can be used to improve estimation at that location.

There is a vast literature on smoothing methods, with recent book-length treatments including those of Scott (1992), Green & Silverman (1994), Wand & Jones (1995), Fan & Gijbels (1996), and Simonoff (1996). A standard approach to describing smoothing methods is to start with the estimation of density functions, and then justify nonparametric regression through the relation

$$E(y|x) = \int \frac{yf(x, y)}{f_X(x)} dy \tag{1.1}$$

(Nadaraya, 1964; Watson, 1964).

The purpose of this paper is to describe a different way of thinking about smoothing methods. Section 2 starts with regression estimation (problem (1) above), and shows how this leads to natural estimation procedures for cell probabilities in ordered categorical data (problem (2) above). This is particularly true when estimation is based on the appropriate (Poisson) likelihood function. Categorical data smoothing methods provide a natural bridge to highly effective density estimators (problem (3) above), as discussed in Section 3. The paper concludes with discussion of extensions and open problems. Much of the material discussed here is based on Simonoff (1996, chapters 5 and 6). In particular, asymptotic results given there are not repeated here, the focus being on the application of smoothing methods rather than the theory.

2 From Regression to Categorical Data Smoothing and Back Again

Nonparametric regression estimation is based on the model

$$y_i = m(x_i) + \epsilon_i,$$

where the regression function $m(x)$ is the conditional expectation $m(x) = E(Y|X = x)$, and the errors satisfy $E(\epsilon|X = x) = 0$ and $V(\epsilon|X = x) = \sigma^2(x)$ not necessarily constant. If it is assumed that $m(x)$ is a smooth function (whatever its precise form), estimation of $m(x)$ at any point x_0 can proceed by "borrowing" information about $m(x_0)$ from observations "close to" x_0.

Many different methods for estimating $m(x)$ have been proposed, including kernel, regression spline, and wavelet estimators. This paper will focus on the two classes of methods known as local polynomial estimators and penalty function estimators. Local polynomial estimators estimate $m(x_0)$ as a polynomial using weighted least squares, where the weights decrease smoothly with increasing distance from x_0. The t^{th} order local polynomial regression estimator at x_0 is the constant term of the minimizer of

$$\sum_{i=1}^{n} [y_i - \beta_0 - \cdots - \beta_t(x_0 - x_i)^t]^2 W\left(\frac{x_0 - x_i}{h}\right),$$

where W is a kernel function, typically taken to be a continuous symmetric density function with zero mean and finite variance. In all of the examples used here, the tricube kernel

$$W(u) = \begin{cases} (1 - |u|^3)^3, & \text{if } u \in (-1, 1) \\ 0, & \text{otherwise,} \end{cases}$$

will be used. The order t is chosen to be $t = 1, 2,$ or 3. The bandwidth h determines how quickly the weights descend to zero with distance from x_0, and hence controls the smoothness of the resultant estimate (larger values of h lead to a smoother estimate).

The appeal of local polynomial estimators comes from their intuitive nature, amenability to asymptotic analysis, and good properties near the boundaries. Unlike the local averaging (kernel) estimator (which corresponds to a local polynomial estimator with $t = 0$), local polynomial estimators are based on the natural idea that $m(x)$ can be approximated locally using a polynomial, thereby accounting for local slope and curvature, leading to better estimates compared with using a local constant. See Fan & Gijbels (1996) or Simonoff (1996, chapter 5) for further details.

Penalty function estimators operate based on the principle of explicitly trading off fidelity to the data with smoothness of the estimate. The estimator is the minimizer over the class of functions with square integrable ℓ^{th} derivative and absolutely continuous derivatives up through order $\ell - 1$ of

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - m(x_i)]^2 + \alpha \int m^{(\ell)}(u)^2\, du.$$

Larger values of α penalize roughness of the estimate more, and hence lead to a smoother estimate. The most common version takes $\ell = 2$, resulting in a cubic smoothing spline. The asymptotic properties of the cubic smoothing spline are roughly similar to those of the local quadratic estimator. See Eubank (1988), Green & Silverman (1994), or Simonoff (1996, chapter 5) for further details.

The smoothness of a nonparametric regression estimate is controlled by a smoothing parameter (h for local polynomial estimators and α for penalty function estimators, respectively). There is a large literature on how to choose the smoothing parameter for various regression estimators; see Simonoff (1996, sections 5.3 and 5.6.3) for a discussion of some of the proposed methods. Hurvich, Simonoff & Tsai (1998) proposed a criterion based on the Akaike Information Criterion, where the smoothing parameter is chosen to minimize

$$AIC_C = \log\left\{\sum_i [y_i - \hat{m}(x_i)]^2/n\right\} + \frac{1 + \mathrm{tr}(H)/n}{1 - \mathrm{tr}(H)/n - 2/n}.$$

Here H is the so-called smoother matrix, which satisfies $\hat{\mathbf{y}} = H\mathbf{y}$. The AIC_C criterion has the advantage of being applicable to any linear estimator, including local polynomial estimators of any order and smoothing spline estimators, and was shown in Hurvich, Simonoff & Tsai (1998) to generally work well in practice. Unless stated otherwise, AIC_C is used as the criterion to choose the level of smoothing in all of the examples that follow. It is important to remember, however, that any automatic smoothing parameter selector should be viewed as only a guideline (or benchmark), and can be adjusted based on subjective impressions.

Figure 3 illustrates the way nonparametric regression methods can highlight structure. The data are those of Figure 1, and local quadratic estimates for the fiction (solid line) and nonfiction (dotted line) are given, along with a dashed line representing equality of the two grade level scores. It is clear that the objective and subjective scores do not match up, casting doubt on the use of the Flesch–Kincaid score from the point of view of the teachers. The two scores are reasonably linearly related for the nonfiction books, although the teachers consistently rate the passages roughly two grades higher than Flesch–Kincaid does. The pattern for the fiction books is very different, characterized by a wide range of book scores where the Flesch–Kincaid level is apparently unrelated to the teacher scores. This pattern of a leveling-off effect of readability scores has been noted before in the readability literature (see, e.g., Zakaluk & Samuels, 1988, p. 19, or Chall & Dale, 1995, pp. 74–75), although it is interesting to note that the Flesch–Kincaid score is designed to address this effect. The necessity of using different scoring systems for different types of texts was discussed in Chall *et al.* (1996, p. 3).

Consider now categorical data with ordered categories. The data consist of counts $\{n_j\}$, $j = 1, \ldots, K$, where K is the number of cells in the table and n_j is the number of observations that fall in the j^{th} cell. The goal is to estimate $\mathbf{p} = \{p_j\}$, the set of probabilities of an observation falling in a given cell. The standard model for this random variable is a multinomial distribution with sample

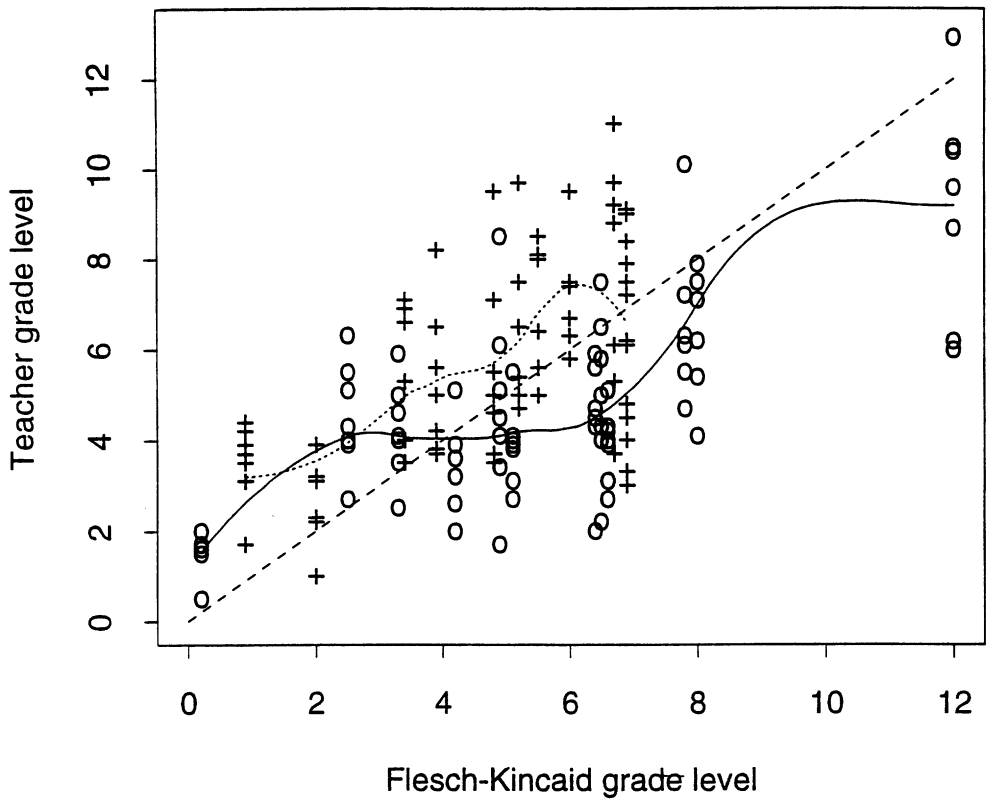

Figure 3. *Local quadratic estimates for reading grade level data. The solid line refers to fiction books, while the dotted line refers to nonfiction books. The dashed line corresponds to equality of the two grade level scores.*

size n and probability vector $\mathbf{p}$, with log-likelihood

$$\sum_{j=1}^{K} n_j \log p_j, \quad \text{subject to the constraint } \sum_{j=1}^{K} p_j = 1 \tag{2.1}$$

(ignoring constants).

The usual estimate of the cell probabilities is the set of observed cell frequencies $\bar{p}_j = n_j/n$. These estimates are accurate if the number of observations in each cell is large, but are not useful for sparse tables, where the table has small (or zero) counts. In this situation, borrowing information from nearby cells makes sense, and can lead to improved estimation.

For the multinomial distribution $E(n_j) = np_j$, and hence $E(\bar{p}_j) = p_j$. It is useful to model $\mathbf{p}$ as being generated from an underlying smooth density f on [0, 1] through the relation

$$p_j = \int_{(j-1)/K}^{j/K} f(u)du,$$

which is reasonable for categories with a natural ordering. This implies that the existence of deriva-

tives of f corresponds to smoothness of $\mathbf{p}$. Note that the Mean Value Theorem implies that

$$p_j = f(x_j)/K, \quad \text{for some } x_j \in [(j-1)/K, j/K]. \tag{2.2}$$

A natural way to define a smooth estimator $\hat{\mathbf{p}}$ is by analogy with a regression of response values $\bar{p}_j$ on the equispaced design j/K, $j = 1, \ldots, K$. So, for example, the local polynomial estimator $\hat{p}_i$ is the constant term of the minimizer $\hat{\beta}$ of

$$\sum_{j=1}^{K}\left[\bar{p}_j - \beta_0 - \cdots - \beta_t\left(\frac{i}{K} - \frac{j}{K}\right)^t\right]^2 W\left(\frac{i/K - j/K}{h}\right).$$

Aerts, Augustyns & Janssen (1997a,b) described the asymptotic properties of this estimator, showing that they were similar to those for local polynomial regression estimators. In particular, unlike the frequency estimator, the estimates are consistent under sparse asymptotics, where n and K become infinite at the same rate (these asymptotics model large sparse tables).

A difficulty with local polynomial probability estimates is that while an arbitrary regression function m can take on positive or negative values, a probability vector $\mathbf{p}$ cannot take on negative values. Negative local polynomial estimates are reflecting that when f is relatively small, it does not behave locally like a polynomial, since it must be nonnegative. Further, since $V(\bar{p}_j) = p_j(1 - p_j)/n$, sample frequencies that correspond to higher probability cells are more variable than those that correspond to low probability cells (that is, from a regression point of view, the data exhibit heteroscedasticity).

The problem is that the estimator is based on the minimization of a local least squares criterion, which is appropriate for regression data, but not for categorical data. The correct likelihood function is the multinomial likelihood (2.1). It is helpful to formulate this problem in terms of a Poisson regression model instead, with log-likelihood

$$\sum_{j=1}^{K}(n_j \log p_j - np_j)$$

(omitting constants). The canonical link for the Poisson distribution is the log link (McCullagh & Nelder, 1989), so the appropriate version of a local polynomial analysis is to use the local log-likelihood (Tibshirani & Hastie, 1987, 1990; Firth, Glosup & Hinkley, 1991; Fan, Heckman & Wand, 1995),

$$\sum_{j=1}^{K}\left\{n_j\left[\beta_0 + \cdots + \beta_t\left(\frac{i}{K} - \frac{j}{K}\right)^t\right] - \exp\left[\beta_0 + \cdots + \beta_t\left(\frac{i}{K} - \frac{j}{K}\right)^t\right]\right\} W\left(\frac{i/K - j/K}{h}\right) \tag{2.3}$$

for cell i. That is, it is the logarithm of $\mathbf{p}$ that is modeled locally as a polynomial, rather than $\mathbf{p}$ itself. The local polynomial likelihood estimator is then $\exp(\hat{\beta}_0)$, where $\hat{\beta}$ is the maximizer of (2.3). The estimates are thus guaranteed to be nonnegative, and better reflect the behavior of probabilities (especially in the tails).

Roughness penalty methods also can be adapted to this situation, by working in the $\log(\mathbf{p})$ scale, and penalizing the log-likelihood. The maximum penalized likelihood estimator is the maximizer of

$$\sum_j n_j \log p_j - \alpha \sum_j (\log \mathbf{p})^{(\ell)}$$

subject to $\sum_j p_j = 1$, where the "derivative" operation $(\log \mathbf{p})^{(\ell)}$ is actually a differencing operator. So, for example, for $\ell = 1$

$$(\log \mathbf{p})' = \log p_j - \log p_{j-1}$$

(Simonoff, 1983), while for $\ell = 2$

$$(\log \mathbf{p})'' = \log p_{j+1} - 2\log p_j + \log p_{j-1}.$$

The choice $\ell = 2$ has asymptotic advantages over $\ell = 1$, and will be the one used here. See Green & Silverman (1994, chapter 5) for general discussion of this estimation method.

The AIC_C smoothing parameter selector can be adapted to these likelihood-based estimators using a linearization approximation. The smoothing parameter is chosen to minimize

$$AIC_C = \log\left\{2\sum_{j=1}^{K} n_j \log[n_j/(np_j)]\right\} + \frac{1 + \mathrm{tr}(H)/K}{1 - \mathrm{tr}(H)/K - 2/K},$$

where H is the implied smoother matrix from the last iteration of the iteratively reweighted least squares fit of the model.

Figure 4 illustrates the use of these categorical smoothing methods. The data are from Eubank (1997), and refer to the probability of each of 10 balls (numbered 0 through 9) being selected in 150 draws from a machine used in the Pick 3 game of the Texas Lottery. The balls are loaded into the machine in sequential order (low numbers on the bottom) before mixing, so there is a natural ordering to the categories. The probability distribution of the number being chosen should be uniform, of course, but the machine that generated the data had been removed from use because of unusual selection patterns in the daily drawings. The observed proportions of each number being chosen out of 150 draws are given by the bars in the figure. Local linear likelihood (dotted lines connecting pluses) and penalized likelihood (solid lines connecting circles) probability estimates are superimposed on the bars, each with smoothing parameter chosen based on AIC_C.

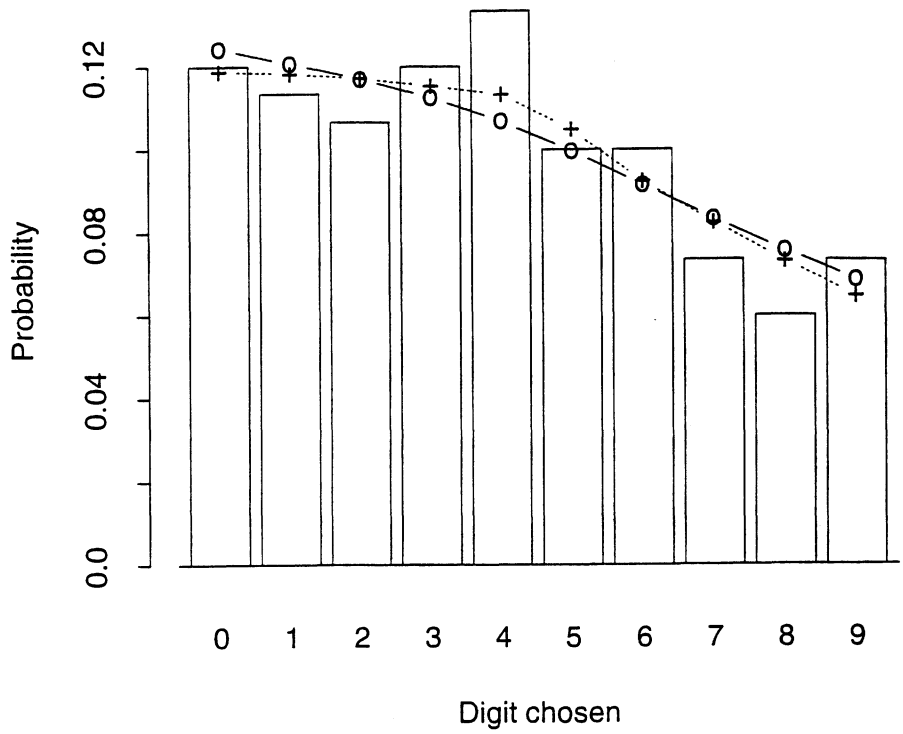

Figure 4. *Plot of probability estimates for lottery data. The bars are the frequency estimates, pluses connected by dotted lines are local quadratic likelihood estimates, and circles connected by solid lines are penalized likelihood estimates.*

As Eubank (1997) noted, even though χ^2 tests of uniformity are not close to statistical significance, there appears to be a pattern where the balls that are loaded into the machine last have a lower probability of being chosen. The local linear likelihood estimate suggests a sharp drop in probability of digits 5–9, while the penalized likelihood estimate fits a smoother trend, but both give the same qualitative results, which are consistent with those of Eubank (1997).

The problems associated with sparse multinomials are magnified when moving to higher dimensions (contingency tables). Since the number of cells in a table increases multiplicatively with the dimension of the table, higher dimensional tables are more likely to be sparse. Fortunately, smoothing methods extend directly to higher dimensions as well. Unfortunately, however, so does the so-called *curse of dimensionality*, which refers to the need for progressively larger sample sizes in higher dimensions to achieve comparable accuracy.

Local linear estimation illustrates how the methods generalize to higher dimensions. The least squares estimator $\hat{p}_{ij}$ for the probability of falling in the $(i, j)^{th}$ cell of an $R \times C$ two-dimensional table $\{n_{ij}\}$ is $\hat{\beta}_0$, where $\hat{\beta}$ is the minimizer of

$$\sum_{k=1}^{R}\sum_{\ell=1}^{C}\left[\bar{p}_{k\ell} - \beta_0 - \beta_1\left(\frac{i}{R} - \frac{k}{R}\right) - \beta_2\left(\frac{j}{C} - \frac{\ell}{C}\right)\right]^2 W_{h_R,h_C}(i, j, k, \ell, R, C),$$

where $W_{h_R,h_C}(i, j, k, \ell, R, C)$ is a two-dimensional kernel function with h_R and h_C the smoothing parameters for rows and columns, respectively (Aerts, Augustyns & Janssen, 1997c). A better version of this estimator is based on local likelihood, rather than local least squares. Equation (2.3) generalizes in the obvious way. For example, the local linear likelihood estimator for a two-dimensional table is $\exp(\hat{\beta}_0)$, where $\hat{\beta}_0$ is the constant term of the maximizer of

$$\sum_{k=1}^{R}\sum_{\ell=1}^{C}\left\{n_{k\ell}\left[\beta_0 + \beta_1\left(\frac{i}{R} - \frac{k}{R}\right) + \beta_2\left(\frac{j}{C} - \frac{\ell}{C}\right)\right]\right.$$

$$\left. - \exp\left[\beta_0 + \beta_1\left(\frac{i}{R} - \frac{k}{R}\right) + \beta_2\left(\frac{j}{C} - \frac{\ell}{C}\right)\right]\right\} W_{h_R,h_C}(i, j, k, \ell, R, C).$$

Table 2 gives local quadratic likelihood estimates for the data of Table 1. The estimator uses a product tricube kernel covering 35% of the cells, which is smaller than the amount suggested by AIC_C (the expected cell counts are given so that direct comparison with Table 1 is possible; the estimated cell probabilities are the expected cell counts divided by 57). The expected counts are generally concentrated in the upper left of the table, corresponding to higher bond ratings and fewer not-for-profits among the top ten employers. The table also shows a clear relationship between bond rating and the number of not-for-profits, with a higher bond rating being associated with a lower number of not-for-profits among the top ten employers. The "ideal" number of not-for-profits (in terms of higher probability of better bond rating) is apparently one, with zero or two not-for-profits being roughly similar. As the number of not-for-profits among the top ten employers gets larger, the probability of lower bond rating increases. Abzug *et al.* (1997) and Simonoff & Tutz (1998) discussed possible reasons for this observed pattern.

These cell probability estimates reflect an application of nonparametric regression to categorical data smoothing, but the direction of application also can be reversed. It is often the case that one (or more) ordered categorical variables can be considered response variable(s) (see Simonoff & Tutz, 1998, and the references therein). For example, the bond rating data of Table 1 are naturally viewed this way, with the number of not-for-profits being a potential predictor of the fiscal health of a city (as measured by the bond rating).

The goal is to estimate $p_i(j) = P(Y = i|X = j)$, where Y is the target variable and X is the

Table 2

Local quadratic likelihood estimated cell counts for U.S. city bond rating table.

Bond rating	Number of not-for-profits among top 10 employers					
	0	*1*	*2*	*3*	*4*	*5*
AAA	0.51	2.08	1.82	0.46	0.05	0.00
AA+	1.27	4.85	3.64	1.05	0.17	0.00
AA	1.72	5.60	3.97	1.45	0.33	0.00
AA−	1.25	2.61	2.91	1.70	0.55	0.00
A+	0.60	0.74	1.79	1.74	0.79	0.01
A	0.35	0.33	1.13	1.48	0.71	0.01
A−	0.36	0.18	0.84	0.97	0.36	0.02
BBB+	0.44	0.07	0.64	0.47	0.14	0.14
BBB	0.37	0.03	0.41	0.22	0.06	0.46
BBB−	0.19	0.05	0.20	0.20	0.09	0.51
BB+	0.05	0.03	0.14	0.22	0.14	0.42
BB	0.01	0.01	0.16	0.28	0.18	0.24

predicting variable. By the definition of conditional probability,

$$p_i(j) = \frac{P(Y = i \text{ and } X = j)}{P(X = j)} = \frac{P(Y = i \text{ and } X = j)}{\sum_i P(Y = i \text{ and } X = j)}. \tag{2.4}$$

A natural estimate of $p_i(j)$ is thus to simply substitute the local polynomial likelihood estimates of $P(Y = i \text{ and } X = j)$ into (2.4).

Table 3 shows how this works for the bond rating data. Each column represents a $\hat{p}_i(j)$, where $j = 0, \ldots, 5$ corresponds to the number of not-for-profits in the top ten and i corresponds to bond rating. The estimates confirm that the most favorable number of not-for-profits is one, with zero and two not-for-profits implying similar estimated probability distributions for bond rating. This is consistent with the good fit of a proportional odds model using $(\text{Nonprofits}-1)^2$ as the predictor that is discussed in Abzug *et al.* (1997).

Table 3

Local quadratic likelihood conditional probability estimates of bond rating given number of not-for-profits.

Bond rating	Number of not-for-profits among top 10 employers					
	0	*1*	*2*	*3*	*4*	*5*
AAA	.072	.125	.103	.045	.014	.000
AA+	.178	.293	.206	.103	.048	.000
AA	.241	.338	.225	.142	.092	.000
AA−	.176	.157	.165	.166	.154	.000
A+	.084	.045	.101	.170	.221	.006
A	.049	.020	.064	.145	.199	.006
A−	.051	.011	.048	.095	.101	.011
BBB+	.062	.004	.036	.046	.039	.077
BBB	.052	.002	.023	.021	.017	.254
BBB−	.027	.003	.011	.020	.025	.282
BB+	.007	.002	.008	.021	.039	.232
BB	.001	.001	.009	.027	.050	.133

These regression estimates also can be used to model higher dimensional probability matrices as the target distribution, using a generalization of (2.4) to multidimensional tables and summing in the denominator over the margin(s) corresponding to the predictor(s). Table 4 summarizes data of this type. The table is based on a survey of 341 undergraduate students in the Leonard N. Stern School of Business of New York University (LaBarbera & Simonoff, 1999). The response table is the joint

Table 4

Importance of number of anticipated employment opportunities by importance of anticipated salary size for undergraduate business students, separated by when the students chose their major. Ratings range from Not at all important (1) to Extremely important (5).

		Importance of employment opportunities				
		1	*2*	*3*	*4*	*5*
		Chose major as a high school student				
Importance	*1*	2	0	0	0	1
of	*2*	0	0	1	0	0
size	*3*	0	3	3	1	2
of	*4*	0	0	0	0	10
salary	*5*	1	1	0	8	25
		Chose major as a freshman				
Importance	*1*	0	0	0	0	0
of	*2*	0	0	0	1	0
size	*3*	0	0	2	4	2
of	*4*	0	0	1	6	15
salary	*5*	0	0	1	6	21
		Chose major as a sophomore				
Importance	*1*	2	0	0	0	0
of	*2*	0	1	1	0	0
size	*3*	0	0	4	7	5
of	*4*	0	1	3	19	22
salary	*5*	0	1	1	10	38
		Chose major as a junior				
Importance	*1*	0	0	0	0	0
of	*2*	0	1	0	1	1
size	*3*	0	0	1	4	2
of	*4*	0	1	1	15	14
salary	*5*	0	0	3	3	24
		Chose major as a senior				
Importance	*1*	2	0	0	0	0
of	*2*	0	1	0	0	0
size	*3*	1	0	1	0	0
of	*4*	0	0	2	8	6
salary	*5*	1	0	2	2	4

distribution of the importance to the student in choosing a major field of the number of anticipated employment opportunities by the anticipated size of the salary, each measured on a five point scale from not at all important to extremely important. The predicting variable is when the student chose their major field (as a high school student, freshman, sophomore, junior, or senior).

Table 5 gives the smoothing-based regression estimates of the joint distribution of importance of employment opportunities by importance of salary size given when the student chose their major field. The estimates are based on a local quadratic likelihood smoothing of the underlying three-dimensional table using a product tricube kernel that covers half of the cells. Several interesting patterns emerge from the estimated bivariate regression function.

It seems that in many ways students who choose their major field either very early (in high school) or very late (as seniors) are different from those who choose it in their first three years of college:

(1) The estimated probability of a student putting great importance (rating 4 or 5) on the anticipated number of employment opportunities is reasonably stable for students who chose their majors as freshmen (.899), sophomores (.893), or juniors (.863). It is noticeably lower, however, for students who choose their majors in high school (.818), and much lower for students who choose as seniors (.704).

(2) Similarly, the estimated probabilities of a student putting great importance on the anticipated

Table 5

Local quadratic regression estimates of joint probability distribution of importance of number of anticipated employment opportunities by importance of anticipated salary size for undergraduate business students, given when the students chose their major.

		Importance of employment opportunities				
		1	*2*	*3*	*4*	*5*
		Chose major as a high school student				
Importance	*1*	.018	.006	.002	.002	.002
of	*2*	.009	.009	.010	.008	.007
size	*3*	.005	.013	.030	.040	.033
of	*4*	.004	.010	.032	.094	.165
salary	*5*	.005	.007	.021	.102	.365
		Chose major as a freshman				
Importance	*1*	.007	.002	.001	.001	.001
of	*2*	.002	.004	.006	.006	.005
size	*3*	.001	.004	.023	.046	.038
of	*4*	.001	.004	.026	.107	.209
salary	*5*	.002	.003	.015	.095	.391
		Chose major as a sophomore				
Importance	*1*	.006	.001	.000	.000	.000
of	*2*	.001	.004	.007	.007	.004
size	*3*	.000	.004	.027	.067	.038
of	*4*	.001	.004	.032	.147	.218
salary	*5*	.001	.003	.015	.086	.326
		Chose major as a junior				
Importance	*1*	.010	.002	.000	.000	.000
of	*2*	.003	.005	.006	.005	.002
size	*3*	.001	.006	.027	.060	.034
of	*4*	.001	.006	.042	.164	.210
salary	*5*	.002	.005	.020	.096	.292
		Chose major as a senior				
Importance	*1*	.040	.005	.001	.000	.000
of	*2*	.021	.012	.006	.003	.001
size	*3*	.010	.017	.034	.046	.023
of	*4*	.008	.020	.068	.165	.160
salary	*5*	.009	.013	.033	.098	.208

salary size is higher for students who choose their majors as freshmen (.853), sophomores (.833), or juniors (.838) than it is for students who choose in high school (.805) or as seniors (.782).

(3) The strength of the association between importance of anticipated employment opportunities and importance of anticipated salary size also varies with time of choosing a major. The association is stronger when the choice is made as a freshmen, sophomore, or junior. For example, the estimated probabilities of the importance ratings differing by more than one level are .078, .073, and .076, respectively, for those years, but are .106 and .121, respectively, when the choice is made as a high school student or college senior.

Each of these results suggests a difference in the view of the purpose of a college education as a preparation for a job. Students who choose their majors as freshmen, sophomores, or juniors apparently view the vocational aspect of a college education as very important, but students who choose their majors early or late view that as less important. This could reflect that students who choose early do so out of a basic love for the field, while those who choose late have a tendency to procrastinate, and are ultimately choosing a major so that they can graduate.

3 From Categorical Data Smoothing to Density Estimation

Categorical data smoothing occupies a central position between nonparametric regression and nonparametric density estimation, providing a natural way to apply regression to the density estimation problem. The key link is through the idea of binning continuous data.

Let $f(x)$ be the density function for a continuous random variable X. Treating the data as (ordered) categorical by binning it and then smoothing the resultant histogram form has long been proposed on the basis of computational efficiency; see Fan & Marron (1994), and the references therein. Equation (2.2) shows that an estimate of a cell probability p_i gives a density estimate for $f(x_i)$ (it is easiest to take f to be supported on [0, 1], so that $x_i = i/K$, but any arbitrary interval can be treated by a simple translation, with the estimated density on $[a, b]$ being the estimated density on [0, 1] divided by $b - a$).

The local polynomial least squares density estimator is K times the constant term of the minimizer of

$$\sum_{j=1}^{K}\left[\bar{p}_j - \beta_0 - \cdots - \beta_t\left(x - \frac{j}{K}\right)^t\right]^2 W\left(\frac{x - j/K}{h}\right), \quad x \in [0, 1].$$

As the bins narrow (or equivalently, as $K \to \infty$), this is equivalent to the constant term of the minimizer of

$$\int W\left(\frac{x-u}{h}\right)\left[n^{-1}\sum_{i=1}^{n}\delta(u - x_i) - \beta_0 - \cdots - \beta_t(x-u)^t\right]^2 du,$$

where δ is the Dirac delta function. This estimator has been justified through both local polynomial arguments and as a generalized jackknifing boundary kernel estimator; see Sarda (1991), Lejeune & Sarda (1992), Jones (1993), Fan, Gijbels, Hu & Huang (1996), Cheng (1997a,b), and Cheng, Fan & Marron (1997).

This estimator inherits all of the properties of the local polynomial least squares cell probability estimator. These include the favorable ones (higher order convergence when using higher order polynomials, automatic boundary correction), but also the unfavorable ones (including possible negativity). A better approach is to base the estimate on the correct likelihood (Poisson regression, rather than Gaussian [least squares] regression), as was done in Section 2 for categorical data smoothing (see Efron & Tibshirani, 1996, Eilers & Marx, 1996, and Jones, 1996, for other Poisson regression-based proposals). The estimator is the (exponentiation of the) maximizer of

$$\sum_{j=1}^{K}\left\{n_j\left[\beta_0 + \cdots + \beta_t\left(x - \frac{j}{K}\right)^t\right] - \exp\left[\beta_0 + \cdots + \beta_t\left(x - \frac{j}{K}\right)^t\right]\right\} W\left(\frac{x - j/K}{h}\right).$$

As the bins narrow, this becomes equivalent to the maximizer of

$$\sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right)\log[f(x_i)] - n\int W\left(\frac{x-u}{h}\right) f(u)du$$

(ignoring constants). This is the local likelihood for density estimation proposed by Loader (1996), and is a special case of that proposed by Hjort & Jones (1996). Besides nonnegativity and automatic boundary bias correction, the local likelihood density estimator has other useful characteristics. Hjort & Jones (1996) showed that for the local linear likelihood estimator,

$$E[\hat{f}(x) - f(x)] \approx \frac{1}{2}\sigma_W^2 h^2[f''(x) - f_0''(x)],$$

where f_0 is the closest exponential density to the true $f(x)$. So, for example, if the true density has exponential-like tails, the bias will be small for a wide range of h (since $f''(x) \approx f_0''(x)$), allowing

the data analyst to choose h to highlight other structure or yield estimates with less bumpiness in the tails. Loader (1996) showed that for densities with exponential tails the kernel estimator has asymptotic relative efficiency equal to zero compared with the local linear likelihood estimator. A local quadratic likelihood estimator allows the ability to account for local curvature.

The maximum penalized likelihood probability estimator also has a direct analogue for density estimation. As the bins narrow, the estimator is the maximizer of

$$n^{-1}\sum_{i=1}^{n}\log f(x_i) - \alpha\int[\log f(u)]''\,du$$

subject to $\int f = 1$, which is the penalized likelihood estimator proposed by Good & Gaskins (1971) and Silverman (1982).

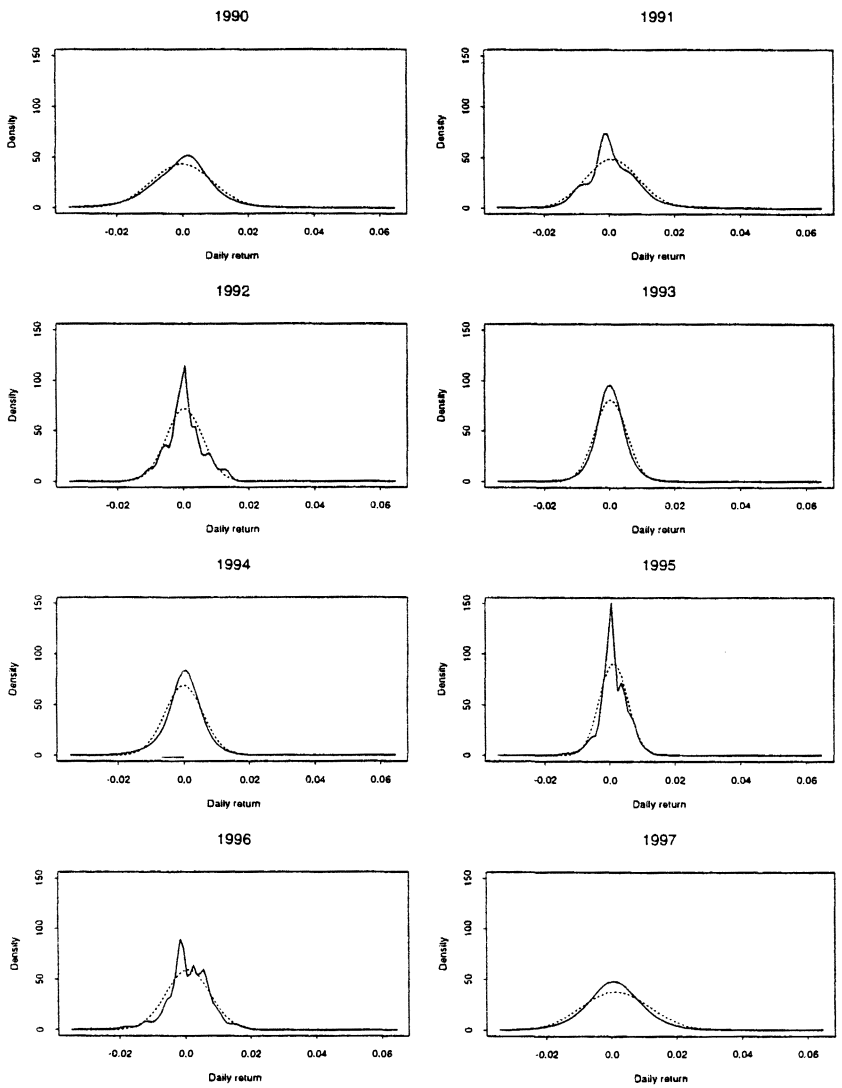

Figure 5. *Density estimates for daily returns of the New York Stock Exchange Composite Index for January 1990 through September 1997. Solid lines are penalized likelihood density estimates, while dotted lines are estimated Gaussian densities.*

Figure 5 demonstrates the practical benefits of this estimation scheme. The data are the daily returns of the New York Stock Exchange Composite Index for January 1990 through September 1997, separated by year. A common representation of stock prices is that they follow a geometric random walk with lognormal innovations; that is, stock returns are independent and identically distributed Gaussian random variables. Each plot in Figure 5 gives the AIC_C-based penalized likelihood density estimates (solid line) and the Gaussian density with observed mean and standard deviation (dotted line). All plots are on the same horizontal and vertical scale.

The random walk hypothesis is clearly not supported by the data. Returns are leptokurtotic, being more peaked and having fatter tails than the normal. Further, the returns are not identically distributed, as there are apparently periods of higher volatility (such as 1990–1991 and 1997) followed by periods of lower volatility (1992–1993 and 1995 particularly). These plots support the possibility of an autoregressive conditional heteroscedasticity (ARCH) model for returns; see Bollerslev, Chou, & Kroner (1992) for a discussion of applications of this model to financial data.

The advantages of estimating in the log f scale are seen in Figure 6. The figure gives the penalized likelihood density estimate for 1997 returns (dotted line), along with a kernel density estimate with smoothing parameter chosen to give a similar representation of the mode of the density (solid line). The kernel density estimate suffers from the common problem of bumpiness in the tails. This is avoided by basing the estimate on the likelihood, since in the tails the density behaves in a roughly exponential way.

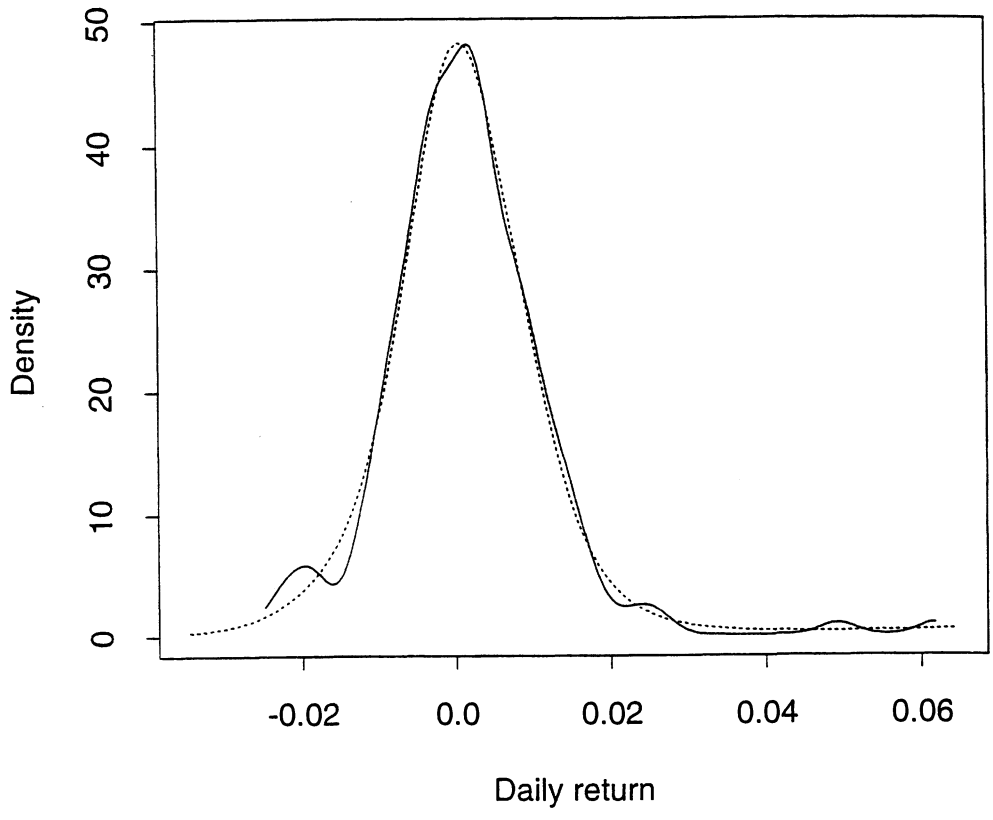

Figure 6. *Penalized likelihood density estimate (dotted line) and kernel density estimate (solid line) for 1997 stock return data.*

One advantage of the local polynomial likelihood density estimate over the penalized likelihood density estimate is that its structure can be manipulated to allow local variation in the amount of smoothing. Figure 7 gives local quadratic likelihood density estimates of the number of lifetime sexual partners for a sample of 1850 U.S. women (Chatterjee, Handcock & Simonoff, 1995, pp. 113–122). The top plot uses a constant bandwidth $h = 5$ based on a tricube kernel (the AIC_C bandwidth is $h = 44$, which smooths over the initial spike in the density). The bumpiness in the long tail is not spurious, as (not surprisingly) respondents tend to round off their answers, leading to multiple responses at 15, 20, 25, 30, and 35.

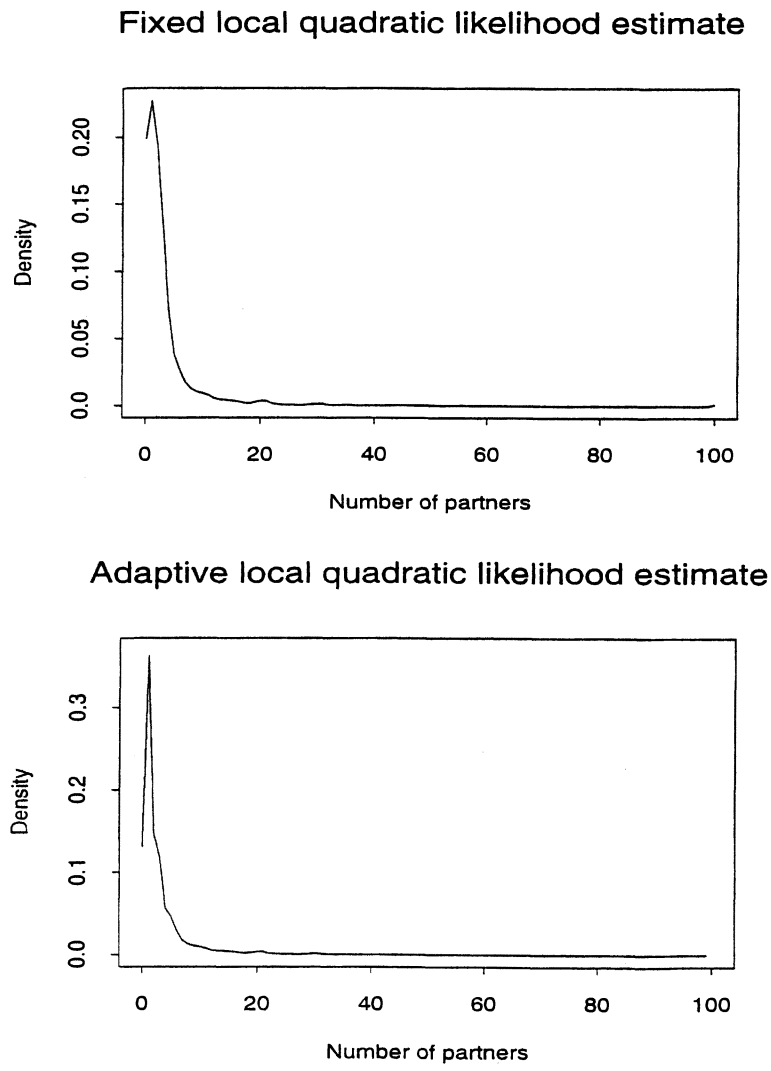

Figure 7. *Local quadratic likelihood density estimates for female lifetime sexual partners data. Top plot: fixed bandwidth local quadratic likelihood estimate. Bottom plot: local quadratic likelihood density estimate using smaller bandwidth for small number of partners.*

Examination of the original data suggests that the spike at the left of the density is not sharp enough, since three times as many respondents reported one lifetime partner than any other value.

This sharp increase and subsequent decrease can be incorporated into the local quadratic likelihood density estimate by using a smaller bandwidth in that region. The bottom plot in Figure 7 is identical to the top plot, except that the density estimate is based on $h = 2.5$ for 0–6 partners (this value was chosen subjectively here, although there are methods to do this automatically for local polynomial regression estimators; see, e.g., Fan & Gijbels, 1995).

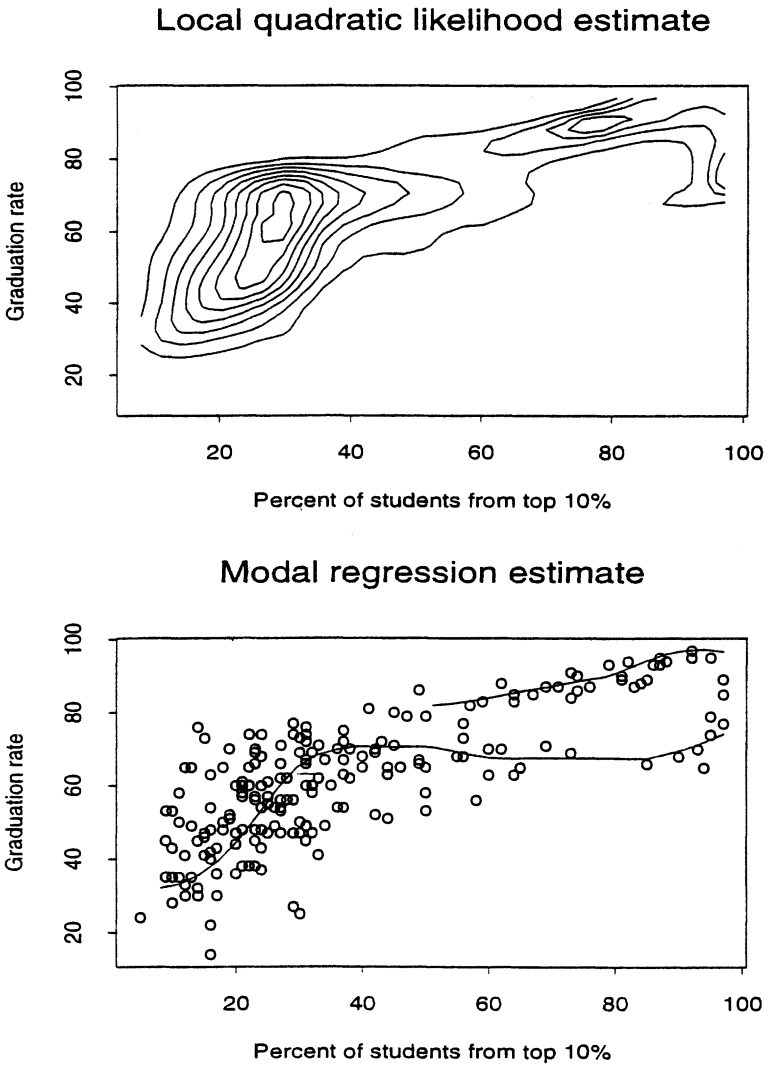

Figure 8. *Local quadratic likelihood estimates of research university data. Top plot: contour plot of bivariate density estimate. Bottom plot: modal regression estimate.*

Figure 8 concludes this section with a bivariate example, the college data of Figure 2. The estimate is a local quadratic likelihood estimate based on a 30×30 table of counts, with a product tricube kernel covering 10% of the cells (the AIC_C choice is 7% of the cells, which seems slightly undersmoothed). The estimate shows clear bimodality in the density that was not evident in the scatter plot. Most schools center at around 30% of students from the top 10% of their high school classes and a 60–70% graduation rate. The smaller mode corresponds to the top 25 universities in the U.S., where around

80% of the students come from the top 10% of their high school classes, and the graduation rate is around 90%. There is also indication of a third mode that corresponds to a high percentage of students from the top 10% of their high school classes, but lower graduation rates. These are schools in the University of California system, which effectively guarantees a spot in the system to all California high school students graduating in the top 10% of their class.

Just as probability estimates were used in Section 2 to create regression estimates, so too can multivariate density estimation be used to create nonparametric regression estimates based on (1.1). The bottom plot in Figure 8 is a modal regression plot (Scott, 1992, pp. 233–235), where the lines given correspond to modes in the underlying estimated conditional density $\hat{f}(x, y)/\hat{f}_X(x)$ (see Hyndman, Bashtannyk & Grunwald, 1996, for discussion of estimators of this type). The modal regression plot shows that for most universities, graduation rate is roughly linearly associated with percent of students from the top 10% of their high school classes until that percent is around 35%. Above that, there is virtually no relationship between the two variables, which might reflect the increasingly challenging nature of the school itself. The top 25 schools form a separate group shifted up from the rest of the schools (with graduation rates roughly 25 percentage points higher), exhibiting some evidence of a weak relationship between the two variables.

4 Conclusion

The theme of this paper is that likelihood-based regression methods provide a unified approach to constructing effective regression and density estimators, with categorical data smoothing occupying a position as a bridge between the two (in addition to its importance in the analysis of contingency tables). It would be hoped that this unified approach could lead to useful cross-fertilization between these areas. Further, given the increasing availability of Poisson nonparametric regression software (as part of generalized additive modeling software, for example), it seems that kernel density estimates should be replaced in routine application with likelihood-based density estimates as a matter of course.

Simonoff (1996, chapter 7) discussed several fruitful areas of application where better understanding of categorical data smoothing would be beneficial, including goodness-of-fit and smoothing-based parametric estimation and testing. The discussion here suggests that the nonparametric regression literature is a good place to look for useful ideas and results.

All of the models discussed here are purely nonparametric, but a major area of regression research in recent years has been that of semiparametric models. These models include aspects of both parametric and nonparametric models, and can thereby avoid many of the problems of the curse of dimensionality. Examples of such models include additive models, where m is represented as a sum of several smooth functions, and partially linear models, where m includes some variables entering linearly and others entering as smooth functions (see Hastie & Tibshirani, 1990, for a discussion of such models, and Simonoff & Tsai, 1999, for discussion of the application of AIC_C to them).

These models could also be adapted to contingency tables (and, by extension, multivariate density estimation) in the same way as is discussed here. So, for example, log-linear models for contingency tables could be replaced by log-additive or log-partially linear models, with potential gains in simplicity or interpretability. Density estimates based on such modeling, assuming they were appropriate for the given data, could also avoid the curse of dimensionality found in multivariate density estimators.

S-PLUS functions and the data sets used in this paper are available in the form of an S-PLUS dump file via the World Wide Web at the location
http://www.stern.nyu.edu/~jsimonof/three.dmp.

Acknowledgment

I would like to thank Chris Jones for helpful discussion of this material. I would also like to thank the referees for many suggestions that improved the presentation in this paper. John Gargani provided the student reading score data.

References

Abzug, R., Ahlstrom, D. & Simonoff, J.S. (1997). The organizational landscape of cities: a comparative geography. Unpublished manuscript.

Aerts, M., Augustyns, I. & Janssen, P. (1997a). Sparse consistency and smoothing for multinomial data. *Statistics and Probability Letters*, **33**, 41–48.

Aerts, M., Augustyns, I. & Janssen, P. (1997b). Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics*, **8**, 127–147.

Aerts, M., Augustyns, I. & Janssen, P. (1997c). Local polynomial estimation of contingency table cell probabilities. *Statistics*, **30**, 127–148.

Bollerslev, T., Chou, R.Y. & Kroner, K.F. (1992). ARCH modeling in finance. A review of the theory and empirical evidence. *Journal of Econometrics*, **52**, 5–59.

Chall, J.S., Bissex, G.L., Conard, S.S. & Harris-Sharples, S.H. (1996). *Qualitative Assessment of Text Difficulty: A Practical Guide for Teachers and Writers*. Cambridge, MA: Brookline Books.

Chall, J.S. & Dale, E. (1995). *Readability Revisited: The New Dale–Chall Readability Formula*. Cambridge, MA: Brookline Books.

Chatterjee, S., Handcock, M.S. & Simonoff, J.S. (1995). *A Casebook for a First Course in Statistics and Data Analysis*. New York: Wiley.

Cheng, M.-Y. (1997a). Boundary aware estimators of integrated density derivative products. *Journal of the Royal Statistical Society, Ser. B*, **59**, 191–203.

Cheng, M.-Y. (1997b). A bandwidth selector for local linear density estimators. *Annals of Statistics*, **25**, 1001–1013.

Cheng, M.-Y., Fan, J. & Marron, J.S. (1997). On automatic boundary corrections. *Annals of Statistics*, **25**, 1691–1708.

Efron, B. & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Annals of Statistics*, **24**, 2431–2461.

Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

Eubank, R.L. (1997). Testing goodness-of-fit with multinomial data. *Journal of the American Statistical Association*, **92**, 1084–1093.

Fan, J. & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Ser. B*, **57**, 371–394.

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

Fan, J., Gijbels, I., Hu, T.-C. & Huang, L.-S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, **6**, 113–127.

Fan, J., Heckman, N.E. & Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, **90**, 141–150.

Fan, J. & Marron, J.S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**, 35–56.

Firth, D., Glosup, J. & Hinkley, D.V. (1991). Model checking with nonparametric curves. *Biometrika*, **78**, 245–252.

Good, I.J. & Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Hjort, N.L. & Jones, M.C. (1996). Locally parametric density estimation. *Annals of Statistics*, **24**, 1619–1647.

Hurvich, C.M., Simonoff, J.S. & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Ser. B*, **60**, 271–293.

Hyndman, R.J., Bashtannyk, D.M. & Grunwald, G.K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, **5**, 315–336.

Jones, M.C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing*, **3**, 135–146.

Jones, M.C. (1996). On close relations of local likelihood density estimation. *Test*, **5**, 345–356.

LaBarbera, P. & Simonoff, J.S. (1999). Toward enhancing the quality and quantity of marketing majors: An empirical study. *Journal of Marketing Education*, **21**, to appear.

Lejeune, M. & Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, **14**, 457–471.

Loader, C.R. (1996). Local likelihood density estimation. *Annals of Statistics*, **24**, 1602–1618.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd. ed. New York: Chapman and Hall.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and Its Applications*, **9**, 141–142.

Przybylski, M., Littlepage, L. & Rosentraub, M.S. (1996). Philanthropy, nonprofits, and the fiscal health of cities. *Nonprofit Voluntary Sector Quarterly*, **25**, 14–39.

Sarda, P. (1991). Estimating smooth distribution functions. In *Nonparametric Functional Estimation and Related Topics*, ed. G.G. Roussas. Dordrecht: Kluwer Academic Publishers, 261–270.
Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.
Silverman, B.W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, **10**, 795–810.
Simonoff, J.S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Annals of Statistics*, **11**, 208–218.
Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.
Simonoff, J.S. & Tsai, C.-L. (1999). Semiparametric and additive model selection using an improved *AIC* criterion. *Journal of Computational and Graphical Statistics*, **8**, to appear.
Simonoff, J.S. & Tutz, G. (1998). Smoothing methods for discrete data. In *Smoothing and Regression. Approaches, Computation and Application*, ed. M.G. Schimek. New York: Wiley, to appear.
Tibshirani, R. & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559–568.
Tibshirani, R. & Hastie, T. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
U.S. News and World Report (1996). *1997 U.S. News and World Report America's Best Colleges*. Washington, D.C.
Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.
Watson, G.S. (1964). Smooth regression analysis. *Sankhyā, Ser. A*, **26**, 359–372.
Zakaluk, B.L. & Samuels, S.J., eds. (1988). *Readability: Its Past, Present, and Future*. Newark, DE: International Reading Association.

Résumé

Durant les quarantes dernières années, l'estimation fonctionnelle nonparamétrique a connu un développement considérable. Ce travail présente, un bilan des recherches portant sur l'estimation des fonctions de densités et de régression. Les résultats théoriques associés à ces deux problèmes d'estimation sont très similaires. De plus, l'estimateur de Nadaraya–Watson d'une fonction de régression trouve ses racines dans l'estimation de densités multivariées.

Un troisième volet de l'estimation fonctionnelle, moins exploité, est celui de l'estimation par lissage de lois de probabilité de données catégoriques. Ce travail explore le fait que ce type d'estimation constitue un pont entre l'estimation nonparamétrique de densités et de fonctions de régression. La régression nonparamétrique fournit un paradigme pour construire de manière efficace des estimateurs de lois de probabilités de données catégoriques. Un choix adéquat de la fonction de vraisemblance permet de construire des estimateurs possédant de nombreuses propriétés intéressantes. Les estimateurs ainsi obtenus peuvent être utilisés en estimation de régression aussi bien dans le cas de variables réponses catégoriques ou dans le cas d'une estimation préalable de densités par le biais de la vraisemblance locale ou pénalisée. Les divers problèmes abordés dans ce travail sont illustrés par l'entremise de plusieurs jeux de données réelles.

[*Received April 1997, accepted January 1998*]