The ed Method for Nonparametric Density Estimation and Diagnostic Checking

Ryan P. Hafen and William S. Cleveland

Abstract

The ed method of density estimation for a univariate $x$ takes a model building approach: an estimation method that can accurately fit many density patterns in data, and leads to diagnostic visual displays for checking fits for different values of the tuning parameters. The two-step estimator begins with a small-bandwidth balloon density estimate, which is the inverse of the distance of $x$ to the $\kappa$-th nearest neighbor. Next, the nonparametric regression method loess is used to smooth the log of the balloon estimate as a function of $x$. This turns nonparametric density estimation into nonparametric regression estimation, allowing the full power of regression diagnostics, model selection, statistical inference, and computational methods. The theory of gaps of order statistics provides a tentative regression model with a known variance. For data with regions of $x$ very sparse in data, pseudo-data can be added over the whole domain of $x$ in the form of a grid, and the estimate corrected at the end.

1 Introduction

This chapter introduces a new univariate nonparametric density estimation, ed which is based on local regression methods.

1.1 Background

Nonparametric density estimation has received substantial attention in the research literature of statistics and machine learning. A search of “density estimation” the ISI Web of Knowledge from 1950 to 2008 revealed over 22,000 papers. The principal reason is that the density of a distribution, univariate or multivariate, is such a fundamental statistical property.

Density estimation got its start as a major research area with the work of Rosenblatt et al. (1956) and Parzen (1962) on fixed bandwidth kernel density estimation. A kernel, which is a probability density such as the normal, is centered at each point in the data. The unnormalized density estimate at any point is the average of the kernels at that point. There are many excellent reviews centering on kernel methods (Izenman, 1991; Jones, Marron, and Sheather, 1996; Scott, 1992; Sheather, 2004; Silverman, 1986).
The kernel approach is simple and allows very fast computational algorithms (Gray and Moore, 2003a,b; Silverman, 1982). However, as discussed widely in the research literature, performance can be poor for a number of density patterns occurring in applications. For example, peaks can be chopped and valleys filled in, and significant bias can occur at a boundary for \( x \) when there is a sharp cut-off of the density at the boundary. Improvements to kernel density estimates have focused on reducing the bias through higher-order kernels and allowing the bandwidth to be adjusted locally, but results for this hard problem have not been entirely successful (Loader, 2004).

Diagnostic methods for assessing whether a kernel estimate follows the density patterns in data have been put forward (Farman and Marron, 1999; Loader, 1999; Marron and Udina, 1999), but the form of the estimation limits the possibilities, so there does not exist a set of comprehensive diagnostic tools, as there are, for example, for regression diagnostics (Daniel and Wood, 1980). The problem is that there is not a notion of local goodness-of-fit that can be exploited in a visual diagnostic tool, such as viewing scatterplots with fitted lines and viewing residuals in the regression setting. Automated model selection criteria for kernel estimates can help in finding an estimate that fits the data (Sheather, 2004), but the criteria are global and average out local places where lack of fit or overfitting occur.

1.2 Summary of Ed

The ed method has been designed to provide a model building approach to density estimation. The ed estimation method can readily fit a wide range of density patterns in data, and it enables the use of a rich set of visual displays for diagnostic checking of different ed fits and assumptions accompanying the fits. Ed does this by transforming density estimation into regression estimation, making available the models, methods, diagnostic checks, and computational methods of regression. The ed approach also includes a theoretical framework based on the theory of gaps for order statistics. The framework is useful for the fitting process and making statistical inferences based on the ed estimator.

The two-step estimator begins with a very small bandwidth balloon density estimate at \( x \), which is the inverse of distance to the \( \kappa \)-th nearest neighbor of \( x \) (Loftsgaarden and Quesenberry, 1965; Tukey and Tukey, 1981). The balloon density estimate is computed at certain values of \( x \). The log of the balloon estimate and the values of \( x \) provide a response and a factor that are treated as a univariate regression that theory suggests can be treated as having normal errors with constant known variance. A \( \kappa \) of about 10 typically works well to ensure this error distribution. Next, a regression curve is fitted using loess (Section 4). Polynomial fitting results in good bias and variance properties (Fan, 1993). The normal approximation can be checked in any application, and if it is warranted, the loess mechanism for statistical inference with normal errors provides characterizations of uncertainty.

Fitting polynomials to the log of the density is not new. Loa (1996) introduced a maximum likelihood method used directly on the raw data that has excellent variance-bias properties. But such direct maximum likelihood estimation does not facilitate powerful regression diagnostics. The ed approach works without sacrificing efficiency because just the small amount of smoothing of the balloon estimate followed by the log transformation tends to create a normal-error regression without introducing more than negligible bias.
The log transformation can perform poorly in certain applications when the data are sparse over a significant portion of the interval within which the data lie. One method to overcome this is gridding: pseudo-data are added over the interval in the form of a grid, and the estimate corrected at the end.

1.3 Overview of the Paper

Section 2 introduces three data sets that will be used throughout to illustrate ed and problems that the traditional kernel density estimator can encounter. Section 3 deals with the raw estimates, based on the $\kappa$-th nearest neighbor balloon estimate. Approximate distributional properties of the estimate are derived and the choice of $\kappa$ is discussed. Section 4 addresses smoothing the raw estimates using loess, and Section 5 applies methods of model selection, model diagnostic checking, and inference. Section 6 illustrates difficulties the traditional kernel density estimate can encounter and compares the kernel estimates to the ed estimates for the example data sets. Section 7 outlines a grid augmentation approach for estimating densities with sparse regions.

2 Examples

In this section, three data sets — named “Packet-Delay”, “Family-Income”, and “Normal-Mixture” — are introduced to illustrate the ed method as well as the well-known problems that the traditional kernel density estimator can encounter.

Packet-Delay is from a simulation study of quality of service for voice over the Internet (VoIP). The data are the 264,583 positive queueing delays for packets arriving for transmission on a router output link. The simulation recreates the router queueing mechanism, and arrivals at the queue are generated by a validated model for VoIP packet arrival times (Xi, Chen, Cleveland, and Telkamp, 2009). Some interesting features in the density of this data are possible discontinuities and a sharp cutoff of the density at 0 ms. A quantile plot for this data is shown in figure 1.

Another example that has had appearances in the density estimation literature is the family income data from the Family Expenditure Survey in the United Kingdom. This data was studied in detail in Marron and Wand (1992). The data consists of 7,201 family incomes for the year 1975, scaled to have a mean of 1. A preliminary analysis revealed large outliers, which were removed, reducing the size of the data to 7,162 observations. It is justifiable to remove these observations since it is not realistic to suppose that good estimates can be obtained in the extremities of the tails of the distribution by nonparametric methods. A quantile plot of the data with the outliers removed is shown in figure 2. Of interest in this data is the existence of a class system, indicated by multiple modes in the density. Different smoothing parameter choices lead to different conclusions.

The final example is a simulated mixture of normal random variables

$$f(x) = \frac{1}{4} (\phi(x;0,0.25) + \phi(x;3.5,0.25) + \phi(x;5,0.75) + \phi(x;10,1.5)),$$

where $\phi(x;\mu,\sigma)$ is the probability density function for a normal variable with mean $\mu$ and standard deviation $\sigma$. This density exhibits multiple modes of varying magnitude. It is used to illustrate
problems with fixed bandwidth kernel density estimation methods, and also for simulation comparisons in which knowing the true density is desirable. A quantile plot of a sample of size 3,000 from this density is shown in figure 3.

![Quantile plot for Packet-Delay data.](image)

**Figure 1:** Quantile plot for Packet-Delay data.

### 3 Estimation Step 1: Raw Estimates

The first step in the ed procedure is calculating the raw estimates. Consider a random sample of independent observations $x_1, \ldots, x_m$ from a random variable $X$ with unknown density $f$ and order statistics denoted $x_{(1)}, \ldots, x_{(m)}$. The raw estimates are based on non-overlapping gaps of groups of $\kappa$ order statistics,
Figure 2: Quantile plot for Family-Income data with 39 outliers removed.
Figure 3: Quantile plot for a sample of 3,000 observations from the Normal-Mixture example.
\[ g_i^{(\kappa)} = x_{(i\kappa+1)} - x_{((i-1)\kappa+1)}, \quad i = 1, \ldots, n, \]  
where \( n = \lfloor (m-1)/\kappa \rfloor \). For example, if \( \kappa = 10 \), then

\[
\begin{align*}
g_1^{(10)} & = x_{(11)} - x_{(1)}, \\
g_2^{(10)} & = x_{(21)} - x_{(11)}, \\
\vdots
\end{align*}
\]

Using the gaps, a quantity is computed that Tukey and Tukey (1981) call the balloon density estimate,

\[
\hat{b}_i^{(\kappa)} = \frac{\kappa}{mg_i^{(\kappa)}}, \quad i = 1, \ldots, n.
\]

The balloon density estimate will serve as the basis of the response variable in a regression problem. Each estimate, \( \hat{b}_i^{(\kappa)} \), is positioned at the midpoints of the gap intervals,

\[
x_i^{(\kappa)} = \frac{x_{(i\kappa+1)} + x_{((i-1)\kappa+1)}}{2}.
\]

Note that when positioned at \( x_i^{(\kappa)} \), the estimates \( \hat{b}_i^{(\kappa)} \) are equivalent to the uniform-weight nearest neighbor density estimation estimates proposed by Loftsgaarden and Quesenberry (1965) evaluated at \( x_i^{(\kappa)} \). Figure 4 illustrates a simple case of calculating the raw estimates \( (x_i^{(\kappa)}, \hat{b}_i^{(\kappa)}) \) for a sample \( x = (1, 3, 7, 2, 15, 17, 19) \) with \( \kappa = 2 \).

Favorable distributional properties result when working with the raw estimates on the log scale. It turns out that \( \log \hat{b}_i^{(\kappa)} \) are approximately independent and follow a non-standard log-gamma distribution, with

\[
\begin{align*}
\mathbb{E}[\log \hat{b}_i^{(\kappa)}] & = \log f(x_i^{(\kappa)}) + \log \kappa - \psi_0(\kappa), \\
\text{Var}(\log \hat{b}_i^{(\kappa)}) & = \psi_1(\kappa),
\end{align*}
\]

where \( \psi_0 \) and \( \psi_1 \) are the digamma and trigamma functions. A derivation of this is provided in section 3.2.

With this result, the raw estimate is computed as

\[
\hat{y}_i^{(\kappa)} = \log \hat{b}_i^{(\kappa)} - \log \kappa + \psi(\kappa)
\]

so that

\[
\mathbb{E}[\hat{y}_i^{(\kappa)}] = \log f(x_i^{(\kappa)}).
\]

Figures 5–7 show plots of \( (x_i^{(\kappa)}, \hat{y}_i^{(\kappa)}) \) for the three examples described in section 2. For the Packet-Delay data, there was one waiting time with as many as 67 ties, so a good choice is \( \kappa = 75 \). For this data, a value of \( \kappa = 75 \) is quite small relative to the sample size of 264,583.
Figure 4: Illustration of raw estimate calculation for a sample $x = (1, 3, 7, 2, 15, 17, 19)$ with $\kappa = 2$. The order statistics $x_1, \ldots, x_7$ are shown on a number line below the grid and the location points for the raw estimates $x_1^{(2)}, \ldots, x_3^{(2)}$ are indicated by the dots on the $x$-axis. The “×” plotting characters denote the raw estimates $(x_i^{(2)}, \hat{b}_i^{(2)})$. 
Figure 5: Raw estimates for Packet-Delay data with \( \kappa = 75 \). The plot pronounces the discontinuities in the density, at 0.16 and 0.32 (vertical lines).
Figure 6: Raw estimates for Family-Income data with $\kappa = 10$. 
Figure 7: Raw estimates for Normal-Mixture with $\kappa = 10$. The solid line is the log theoretical density.
Normal-Mixture data and the Family-Income data, \( \kappa = 10 \) was chosen. More on the selection of \( \kappa \) will be discussed in section 3.1.

These figures demonstrate the merit of viewing the raw estimates as a visualization technique, especially in the case of the Packet-Delay data. A discontinuity at 0.16 ms is prominent in the raw estimate plot, and another appears to be present at 0.32 ms. It would be much more difficult to unequivocally declare or even notice these points as discontinuities through some other technique like a histogram or a kernel density estimate. Further investigation revealed a simple explanation for these discontinuities – the service time of a packet is a constant time of 0.16, so packets that arrive when there is another packet being served will wait in the queue at most 0.16 time units, while packets that arrive when there is already one packet in the queue and another being served will wait in the queue at least 0.16 but at most 0.32 time units, etc. While this phenomenon should have been obvious from the start, visualizing the density via the raw estimates gave a quick understanding of what should have already been known about the data.

Figure 6 reveals the bimodal nature of the income distribution. The two peaks will become more evident after smoothing the raw estimates.

Figure 7 shows that the theoretical log density fits through the raw estimates for the Normal-Mixture sample nicely, affirming that the raw estimates are leading in the right direction.

### 3.1 Choosing \( \kappa \)

There are two considerations for the choice of \( \kappa \). First, \( \kappa \) should be small enough that there is as little distortion of the density as possible by the averaging that occurs, since the assumption is that the density is approximately uniform in small neighborhoods. Second, \( \kappa \) should be large enough that \( \hat{y}_i^{(\kappa)} \) is approximately normal. Choosing \( \kappa = 10 \) works quite well for approximate normality, and if \( m \) is not small, is adequate for not distorting the density.

To illustrate some of the consequences of choice of \( \kappa \), data was simulated from the Normal-Mixture data and raw estimates were computed with different \( \kappa \). Since \( f \) is known, the errors \( \hat{y}_i^{(\kappa)} - \log f(x_i^{(\kappa)}) \) can be studied. Figure 8 compares the sample order statistics of the errors with the expected order statistics of the normal with variance \( \psi_1(\kappa) \) and the non-standard log-gamma distribution for different values of \( \kappa \), for a random sample of size 3,000 from the Normal-Mixture data. The normal approximation seems to work well for \( \kappa = 10 \), but not for \( \kappa = 1 \) and \( \kappa = 4 \).

With proper choice of \( \kappa \), the raw estimates, \( (x_i^{(\kappa)}, \hat{y}_i^{(\kappa)}), i = 1, \ldots, n \), provide a regression problem, where the observations are independent, the error is approximately normal with constant known variance, and the expected value of \( \hat{y}_i^{(\kappa)} \) is the log density \( \log f(x_i^{(\kappa)}) \). Smoothing the raw estimates will result in a good estimate of the log density.

### 3.2 Approximate Distribution of the Raw Estimates

For a sequence of \( m \) independent identically distributed random variables \( X_1, \ldots, X_m \) with continuous density \( f \) and with order statistics denoted as \( X_{(1)}, \ldots, X_{(m)} \), consider the non-overlapping gaps
Figure 8: Normal and log-gamma probability plots of the sample errors $\hat{y}_i^{(k)} - \log f(x_i^{(k)})$ for the Normal-Mixture sample, where the normal distribution has variance $\psi_1(\kappa)$. The dashed line is the mean of the errors.
of groups of $\kappa$ order statistics,

$$G_i^{(\kappa)} = X_{(i\kappa+1)} - X_{((i-1)\kappa+1)}, \quad i = 1, \ldots, n,$$

(11)

where $n = \lceil (m - 1) / \kappa \rceil$, which is a random quantity. An alternate way to view this is as the sum of distances to consecutive order statistics,

$$G_i^{(\kappa)} = \sum_{j=1}^{\kappa} \Delta_i^{(j)}.$$

(12)

Recall that $G_i^{(\kappa)}$ is used to calculate an estimate of the density positioned at the interval mid-points,

$$X_i^{(\kappa)} = \frac{X_{(i\kappa+1)} + X_{((i-1)\kappa+1)}}{2}.$$

(13)

When $\kappa$ is small, the density of $f$ around the neighborhood of $X_i^{(\kappa)}$ is approximately uniform, so that each of the $\kappa$ distances, $\Delta_i^{(1)}$, ..., $\Delta_i^{(\kappa)}$ behave approximately like an exponential distribution with rate $mf(X_i^{(\kappa)})$. With this distributional assumption for each $\Delta$,

$$E[G_i^{(\kappa)}] \approx \frac{\kappa}{mf(X_i^{(\kappa)})}.$$

(14)

For observed data, replacing the expected value with the observed $g_i^{(\kappa)}$ and solving the above for $f(X_i^{(\kappa)})$ yields the balloon estimate in equation 9.

With the approximate distribution of $G_i^{(\kappa)}$ being gamma, the balloon estimate

$$\hat{B}_i^{(\kappa)} = \frac{\kappa}{mG_i^{(\kappa)}}, \quad i = 1, \ldots, n$$

(15)

has an approximate inverse gamma distribution whose variance depends on $f(X_i^{(\kappa)})$.

Now, to simplify notation, use $B$ and $G$ instead of $\hat{B}_i^{(\kappa)}$, and $G_i^{(\kappa)}$ and consider the transformation $Y = \log B$. With this,

$$F_Y(y) = P(\log B \leq y) = P(\log \kappa - \log m - \log G \leq y) = 1 - F_G(\exp(\log \kappa - \log m - y)).$$

(16)

The gap, $G$, has an approximate gamma distribution shape parameter $\kappa$ and rate parameter $\lambda = mf(X_i^{(\kappa)})$, with probability density function

$$f_G(y; \kappa, \lambda) = \frac{\lambda^\kappa y^{\kappa-1} e^{-\lambda y}}{\Gamma(\kappa)}, \quad x > 0.$$  

(17)
Now let $a = \log \kappa - \log m$. Differentiating (16),

$$f_Y(y) = -f_G(e^{a-y})(-e^{a-y})$$

$$= \frac{1}{\Gamma(\kappa)} \exp \left\{ -\kappa(y - a - \log \lambda) - e^{-(y-a-\log \lambda)} \right\}$$

$$= \frac{1}{\Gamma(\kappa)} \exp \left\{ -\kappa(y - \log \kappa - \log f(X_i^{(\kappa)}) - e^{-(y-\log \kappa-\log f(X_i^{(\kappa)}))} \right\},$$  

(18)

where the last step was obtained by substituting back in $a = \log \kappa - \log m$ and $\lambda = mf(X_i^{(\kappa)})$. Equation 18 is the probability density function for the non-standard log-gamma distribution (Kotz and Nadarajah, 2000), with

$$E[Y] = \log f(x_i^{(\kappa)}) + \log \kappa - \psi_0(\kappa)$$

(19)

$$\text{Var}(Y) = \psi_1(\kappa),$$

(20)

where $\psi_0$ and $\psi_1$ are the digamma and trigamma functions, respectively.

Thus the log transformation yields an unbiased estimate, up to an additive known constant, of the log density at $x_i^{(\kappa)}$, and that the variance is now constant. Of course, these results are approximate, based on the assumption of near-uniformity of the density in small neighborhoods.

Figure 7 gives some empirical validation of the expectation property (19) of the raw estimates since the true log density fits nicely through the raw estimates. To give some empirical support to the variance property (20) of the raw estimates, figure 9 plots the variance of the errors $y_i^{(\kappa)} - \log f(x_i^{(\kappa)})$ for several samples of size 3,000 from the Normal-Mixture data at different values of $\kappa$. The black line joining through the circles represents the trigamma function, and it passes nicely through the median variance value for each $\kappa$. Also, figure 8 provides empirical validation of the normality property of the raw estimates for the Normal-Mixture data.

### 4 Estimation Step 2: Loess Smoothing

Local regression methods are used to smooth the raw estimates, $(x_i^{(\kappa)}, y_i^{(\kappa)})$, $i = 1, \ldots, n$. Since the approximate distribution of the response variable is known to be a non-standard log-gamma, local likelihood methods could be used (Tibshirani and Hastie, 1987) to obtain a log-density estimate. However, for large enough $\kappa$, the distribution of $y_i^{(\kappa)}$ tends to the normal distribution, enabling the use of loess smoothing. Throughout the presentation of smoothing the raw estimates, the simplified notation of $(x_i, y_i)$ is used in place of $(x_i^{(\kappa)}, y_i^{(\kappa)})$.

From section 3, for large enough $\kappa$, $y_i$ is approximately normal with expected value equal to the log of the density at $x_i$. Thus, the model is

$$y_i = \ell(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

(21)

where $\ell(x_i) = \log f(x_i)$ and $\varepsilon_i$ is approximately normal with

$$E[\varepsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \psi_1(\kappa),$$

(22)
Figure 9: Variance of sample errors $\hat{y}_i^{(\kappa)} - \log f(x_i^{(\kappa)})$ for several Normal-Mixture samples of size 3,000 and for $\kappa$ ranging from 5 to 20. The boxes represent the interquartile range for the variances at each $\kappa$ with the horizontal line in each box representing the median value. The points with the dashed line connecting them represent the $\psi_1(\kappa)$, the theoretical variance for the raw estimates.
where $\psi_1$ is the trigamma function. Since the raw estimates are computed at disjoint intervals, the $\varepsilon_i$ are also independent.

Once the loess fit $\hat{\ell}(x_0)$ is obtained, the ed density estimate is calculated by exponentiating,

$$\hat{f}(x_0) = \exp(\hat{\ell}(x_0)).$$

(23)

Although the ed estimate is not guaranteed to integrate to one, experience has shown that a good fit results in an estimate whose numeric integral is extremely close to one. If one desires to have an estimate that integrates exactly to one, it is possible to renormalize.

Figure 10: Loess fit to the Packet-Delay data on the log scale. Smoothing parameters used were: $\lambda = 1$ and $\alpha = 0.75$ for the lowest interval, $\lambda = 2$ and $\alpha = 0.5$ for the middle interval, and $\lambda = 1$ and $\alpha = 1.5$ for the highest interval.
Figure 11: Loess fit with $\lambda = 3$, $\alpha = 0.24$ to the Normal-Mixture raw estimates. The dashed line represents the true log density.
Figure 10 and 11 show loess fits to the Packet-Delay and Normal-Mixture raw estimates, and figures 12 and 13 show the exponentiated fits. The Family-Income data will be studied separately in the next section, where it receives more attention in terms of parameter selection and model diagnostics.

The Packet-Delay data was fitted separately on each interval of continuity. Although in general the assumption is that the density is continuous, if discontinuities are noticed in the raw estimates as in this case, each interval of continuity can be fit separately. Loess smoothing was applied to the raw estimates on each interval, using $\lambda = 1, \alpha = 0.75$, for the lowest interval, $\lambda = 2, \alpha = 0.5$, for the middle interval, and $\lambda = 1, \alpha = 1.5$ for the highest interval.

The Normal-Mixture data was fit using $\lambda = 3$ and $\alpha = 0.24$. As can be seen in the figure, the resulting fit is quite good. Selection of $\lambda$ and $\alpha$ was done with the help of Mallows’ $C_p$, discussed
5 Diagnostic Checking, Model Selection, and Inference

With density estimation set as a regression problem, a multitude of visual and numeric diagnostic tools are available to evaluate the goodness-of-fit of the ed estimate, aid in parameter selection, and check the validity of model assumptions.

5.1 Evaluating Goodness-of-Fit

In the density estimation setting, typically peaks and valleys are present so one can quite often expect $\lambda$ to be 2 (local quadratic fitting) or 3 (local cubic fitting). Diagnostics are carried out on the fits and residuals, where the residuals are calculated as

$$\hat{e}_i = y_i - \hat{\ell}(x_i).$$  \hspace{1cm} (24)

Figures 14 and 15 show the residuals for the Packet-Delay and Normal-Mixture fits shown in figures 10 and 11. The Packet-Delay residuals indicate that perhaps a smaller bandwidth should be used in the 0–0.16 ms range.

Throughout the remainder of this section, attention will be given to the Family-Income data. Figures 16 and 17 show the fitted density and residuals for different choices of $\lambda$ by varying

---

Figure 13: Exponentiated loess fit to the Normal-Mixture data. The dashed line represents the true density.
Figure 14: Residuals from loess fit for the Packet-Delay data.
Figure 15: Residuals from loess fit for the Normal-Mixture data.
smoothing degrees of freedom, \( \nu \) to the Family-Income raw estimates. The value \( \nu \) corresponds to different choices of \( \alpha \). The fit is very rough for large \( \nu \) (small \( \alpha \)) and gets smoother as \( \nu \) decreases (\( \alpha \) increases). However, as \( \nu \) gets too small, bumps begin to be apparent in the residuals, as emphasized by a loess smooth, indicating that the larger bandwidths are cutting off the peaks. With the goal of keeping the fit as smooth as possible while not introducing bias, choosing a fit with \( \lambda = 3 \) and \( \nu \) at 16 or slightly larger looks like a good option. The \( \nu = 16 \) fit with \( \lambda = 3 \) corresponds to a smoothing bandwidth of \( \alpha = 0.24 \).

![Family-Income data log density fits for different choices of \( \lambda \) and varying degrees of freedom \( \nu \). The vertical lines roughly indicate the two peaks in the density.](image)

Figure 16: Family-Income data log density fits for different choices of \( \lambda \) and varying degrees of freedom \( \nu \). The vertical lines roughly indicate the two peaks in the density.

### 5.2 Mallows \( C_p \) Model Selection

Another useful visualization tool for parameter selection is looking at a plot of the Mallows \( C_p \) model selection criterion (Mallows, 1966), which shows the trade-off of bias and variance. Mal-
Figure 17: Family-Income data log density residuals for different choices of $\lambda$ and varying degrees of freedom $\nu$. A loess smooth is added to highlight deviations from 0. The vertical lines roughly correspond to the two peaks in the density.
lows $C_p$ was extended to the local regression setting by Cle (1988). Here, an estimate of the scaled mean-squared error, $M$, is plotted vs. the equivalent number of parameters, $v$, which is a measure of the amount of smoothing done and is analogous to the number of parameters in a parametric regression setting. $M$ is calculated as

$$M = \sum_{i=1}^{n} \frac{\hat{\epsilon}^2_i}{\hat{\sigma}^2(\epsilon)} - n + 2v,$$

where $\hat{\sigma}^2(\epsilon)$ is estimated from a loess fit with a small bandwidth so that the bias of the fit will be negligible, which results in a nearly unbiased estimate of the error variance. The quantity $v$ is a function of the fitting parameters, and details of its calculation can be found in Cle (1988). An approximate relation is $v \approx 1.2\lambda/\alpha$, indicating that smaller $v$ corresponds to smoother fits.

The idea of the $C_p$ plot is that when the fit follows the pattern in the data, asdf should see $E(M) \approx v$. The amount by which $M$ exceeds $v$ is an estimate of the bias. asdf stress here that this criterion is far more useful as a visualization tool as opposed to just a criterion that one minimizes. The $C_p$ plot nicely complements the fitted value and residual diagnostic plots.

A $C_p$ plot for the Family-Income data can be seen in figure 18. The $C_p$ plot reinforces the parameter choices $\lambda = 3$, $v = 16$ arrived at from visual diagnostics. Plots of the final loess fit and residuals for these parameters can be seen in figure 19.

5.3 Validating Model Assumptions

To validate the assumptions made in modeling, several diagnostic plots were made. Figures 21–23 show plots checking the assumptions of constant variance, independence variance, and approximate normality for the residuals of the Family-Income ed fit with $\lambda = 2$ and $\alpha = 0.16$. It appears that the assumptions hold, making inference possible.

6 Comparison with Kernel Density Estimates

The traditional fixed-bandwidth kernel density estimate (KDE) was introduced over a half-century ago by Rosenblatt et al. (1956) and Parzen (1962) and is arguably still the most commonly used density estimation procedure, next to the histogram (Silverman, 1986). This is most-likely due to the simplicity of the method and its widespread availability in software implementations.

Given data $x_1, \ldots, x_m$ from a density $f$, the KDE $\hat{f}(x)$ is calculated as

$$\hat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x-x_i}{h}\right),$$

where $K$ is a kernel function that integrates to 1 and is typically a symmetric probability density function. The parameter $h$ is the bandwidth smoothing parameter that controls the variance of the kernel and hence the amount of smoothing.

The quality of a KDE is much more dependent on choice of $h$ than choice of $K$. The estimated density curve inherits all smoothness and differentiability properties of the kernel function, but
Figure 18: $C_p$ plot for the Family-Income data. The plotting characters correspond to the degree of the fit.
Figure 19: Final loess fit with $\lambda = 3, \alpha = 0.24$ for the Family-Income data.
Figure 20: Final loess fit residuals with $\lambda = 3$, $\alpha = 0.24$ for the Family-Income data.
Figure 21: Absolute residuals vs. income for Family-Income fit.
Figure 22: Lag 1 correlation plot for Family-Income fit residuals.
Figure 23: Normal quantile plot for Family-Income fit residuals.
the efficiency of the estimator quite similar across many kernel choices (Silverman, 1986). In this section, the gaussian kernel is used with mean 0 and variance 1.

Choosing $h$ has been a major area of research and there are many automatic selection criteria, four of which are utilized in this section: Silverman’s rule (Silverman, 1986), least squares (unbiased) and biased cross-validation (Scott and Terrell, 1987), and the Sheather-Jones method (Sheather and Jones, 1991). Typically the Sheather-Jones bandwidth is recommended. An excellent discussion of these bandwidth selection methods can be found in Sheather (2004). A problem with these and other bandwidth selection methods is that they rely on global and/or asymptotic properties.

The three examples are used in this section to highlight some problems with the traditional kernel approach, namely problems with: fixed bandwidths, boundaries and discontinuities, bias, and choosing optimal smoothing parameters. These difficulties are contrasted with the simplicity with which ed deals with these problems. There have been several improvements to kernel density estimation methods to address these problems, such as variable bandwidths, boundary kernels, and higher-order kernels (Karunamuni and Alberts, 2005; Silverman, 1986; Wand and Jones, 1995), but the ed approach can manage all of these issues within a simple framework, while providing insight to the analyst as to whether these problems are present or not.

### 6.1 The Family-Income data

Figure 24 shows kernel density estimates for 6 different bandwidths for the income data, ranging from $h = e^{-4}$ to $h = e^{-1}$. The largest bandwidth yields a unimodal estimate, while the smallest bandwidth exhibits a multi-modal estimate. Thus different choice of bandwidth can lead to different conclusions on whether the income distribution indicates the presence of an income class system. Here one would definitely want to know the “correct” bandwidth with confidence.

Another possible problem that is brought to attention is that if a small bandwidth is desired to adequately capture the first peak, the smoothness in the tail is sacrificed. Also, positive probability is estimated for negative incomes for some of the bandwidth choices. This could be corrected with a boundary kernel, or with a power transformation of income as is done in Wand and Jones (1995). Finally, sharp peaks in densities are usually chopped with KDEs. This can be explained by the equivalence of the KDE to local constant fitting of the log density (Loader, 1999), which cannot adequately capture curvature. Thus considerable bias can be expected in this and all densities with peaks.

The ed fit of this data as studied in section 5 has given confidence in the bimodal nature of this density and a satisfactory fit has been obtained. Figure 25 shows the exponentiated final ed fit compared to the fit obtained from kernel density estimation using the Sheather-Jones bandwidth selection criterion. The ed fit has eliminated the wiggles in the upper tail, and the first peak is more pronounced.
6.2 The Packet-Delay Data

Figure 26 shows the KDE for the Packet-Delay data, using a bandwidth chosen by Silverman’s rule. Due to the large size of the data, the other selection criteria took a very long time to compute and did not find a suitable parameter choice. This data illustrates the problem kernel density estimates can have with discontinuities and boundaries. The KDE trails off at 0 ms and gives positive density for negative waiting times. From figure 5, it is clear that the density cuts off abruptly at 0 ms. The problem with the KDE lies in the fact that at the boundary, there is no data on the left side of the kernel.

The KDE hints at the discontinuity at 0.16 ms. One may try fitting a KDE separately on each interval, as was done with the ed fit, but doing so would lead to the boundary problem at the endpoints of each interval. Of course a boundary kernel method could be used, but these problems are more easily exploited and modeled with ed.

6.3 The Normal-Mixture Data

Kernel density estimate plots for the Normal-Mixture data are shown in figure 27. The solid gray line in the density plots is the true density. The bandwidths used in the density plots were calculated based on the four criteria discussed above. The density plot shows how extremely different results can be obtained from the different bandwidth selection criteria. Silverman’s rule and biased cross-validation indicate a tri-modal density. Relying on one of these methods could
Figure 25: Final loess fit to Family-Income data on the identity scale, compared to the kernel density estimate with the Sheather-Jones bandwidth. The dashed lines indicate the differences in the peaks for the two estimates.
Figure 26: Kernel density estimate plot of waiting time in queue for Packet-Delay data.
lead to confusion over which features are really present. Also note that the Sheather-Jones and unbiased cross-validation bandwidths are more faithful to the true density, especially capturing the narrowest peak, but exhibit too many wiggles for the first and fourth peak. This is a case where a variable bandwidth would be beneficial.

Figure 27: Kernel density estimate plots of 3,000 simulated values from the Normal-Mixture density. The gray line is the true density.

While a multitude of density estimation methods exist that may be able to deal with many of the issues that have been identified in these examples, or while extra care may avoid some of the problems discussed, the ed method lends itself as a simple comprehensive approach to visualizing and modeling these densities.

6.4 Inference

The approximate normality of the residuals and the capabilities of loess enable construction of confidence intervals for the ed density estimate. Doing so assumes that the estimate is unbiased. The diagnostic methods presented in section 5 assist in determining whether there is bias present in the loess modeling, but this does not account for bias that may have been introduced when computing the raw estimates. If the diagnostics look good, it is safe to assume the bias is negligible.

Pointwise confidence limits can be computed. Figure 28 shows the ed fit to the British income data density with 99% pointwise confidence limits.
Figure 28: Pointwise 99% confidence limits for Family-Income log density estimate. The gray line is the fitted log density and the vertical lines indicate the interval over a grid of points. The bottom panel shows the interval with the fit removed.
7 Grid Augmentation

One problem with fitting on the log scale is the fact that it is difficult to estimate densities when they are close to 0. For example, in the Normal-Mixture fit shown in figure 13, the difficulty with estimating the troughs is evident, although for this example it is not as evident when viewing the exponentiated fit.

To deal with low density regions, the data can be augmented with a grid of uniform points, creating a mixture of the form

$$g(x) = (1 - \omega)f(x) + \omega u(x),$$

where $f$ is the true density, $u$ is a uniform density over the range of the data plus a buffer if the support of the density is not finite on either end, and $\omega$ is the desired proportion of mixing. The ed procedure is then applied to the augmented data and the estimate of $\log g(x)$ is used to obtain the estimate of $f(x)$.

The density $u(x)$ is uniform over an interval $(a, b)$. Choice of $a$ and $b$ depends on the data at hand. In the case of a density that does not trail off at one or both tails, one can choose $a$ and/or $b$ to be the minimum or maximum of the observed data, or a value that makes sense in the context of the data (such as $a=0$ for non-negative data). If the tails of the density trail on either end, the range of the uniform should extend beyond the range of raw estimate design points to capture the transition of the $f$ into $u$. Furthermore, the range of the uniform should extend even further so that loess estimation at the tails of $f$ incorporates data on both sides of the estimation point. The uniform limits are computed as

$$a = 2x(1) - x(\lceil \alpha m \rceil)$$

and

$$b = 2x(m) - x(m - \lceil \alpha m \rceil + 1).$$

(28)

To obtain a grid-augmented fit, the user specifies a desired proportion of blending $\omega$. A grid of equally-spaced points is generated (not a random uniform sample) over $(a, b)$. The number of points to generate is

$$m_g = \left\lfloor \frac{\omega m}{1 - \omega} \right\rfloor,$$

so that the original data $x_i, i = 1, \ldots, m$ is augmented with values

$$u_i = a + (i - 1) \frac{b - a}{m_g - 1}, \quad i = 1, \ldots, m_g.$$

(30)

Putting both data sets together yields a new sample, $x^*_i, i = 1, \ldots, m^*$, where $m^* = m + m_g$, from which the raw estimates can be calculated.

With the $u_i$ defined as above, a slight adjustment needs to be made to $\omega$ to ensure that in in regions where the uniform density dominates, the raw estimates obtained from the augmented data agree exactly with the log of the uniform density $u(x; a, b)$. For a value $x$ a region where all $\kappa$ neighbors are from the uniform grid, the raw estimate will be

$$\hat{\ell}(x) = \log \frac{k}{m(u_{\kappa+1} - u_1)} - \log \kappa + \psi(\kappa).$$

(31)
Since \( x \) lies in a region surrounded by only uniform points, assume that \( f(x) = 0 \), so that \( \log g(x) = \log(\omega u(x)) \). Then the raw estimates in this region need to match \( \log g \), which gives

\[
\omega^* = \frac{b - a}{e^{\ell(x)}}. \tag{32}
\]

So for a user-specified proportion of grid augmentation \( \omega \), the \( x_i^* \) follow a distribution of the form

\[
g(x) = (1 - \omega^*)f(x) + \omega^*u(x). \tag{33}
\]

Now the ed estimate, \( \hat{g}(x) \), is calculated using loess. While \( \log g(x) \geq \ell(x) \) should always hold, loess may give estimates at certain points where this inequality does not hold. At any such point, \( x \), simply set \( \log \hat{g}(x) = \hat{\ell}(x) \). Now, since \( u(x) \) is known, the estimate for \( f \) can be calculated

\[
\hat{f}(x) = \frac{\hat{g}(x) - \omega^*u(x)}{1 - \omega^*}. \tag{34}
\]

Figure 29 shows a plot of a grid augmentation fit to the mixture of normals with \( \kappa = 10, \lambda = 3, \alpha = 0.16, \) and \( \omega = 0.4 \) (resulting in \( \omega^* = 0.38 \)). The first panel shows the raw estimates with \( \log \hat{g} \), and the second panel shows \( \hat{f} \). The last panel shows the residuals on the log scale for the mixture.

Note that when using grid augmentation, the numeric integral has, from experience, consistently come up slightly less than 1. When using grid augmentation, normalizing is recommended.

## 8 Computational Methods

There are two major computational steps in obtaining the ed density estimate. The first is the computation of the raw estimates, and the second is smoothing the raw estimates. Each is discussed below.

The main computational chore in calculating the raw estimates is obtaining the order statistics. A simplistic way to do this is to sort the data and extract the order statistics as needed. This is the current approach and the quicksort algorithm is used, which on average takes \( O(m \log m) \) time although its worst case is \( O(m^2) \). Since only every \( \kappa \)th order statistic is needed, however, computation time is wasted sorting the values in between, and a better approach might be to recursively apply a selection algorithm. There are selection algorithms that operate in worst-case linear time (Blum, Floyd, Pratt, Rivest, and Tarjan, 1973).

After the raw estimates are computed, there are \( n \) values to smooth, where \( n = \lfloor (m - 1)/\kappa \rfloor \). Loess is a computationally efficient method, where direct evaluation of the local fit is evaluated at a fixed and sufficiently dense set of points from a \( k \)-d tree, and interpolated elsewhere. Typically, a user would like to compute the final density estimate at a dense grid of equally-spaced points, enabling plotting of the density or numerical integration for normalization. Once \( k \)-d tree has been constructed, interpolation on the grid is performed in linear time. Overall runtime of loess is proportional to \( n \) (Cleveland and Grosse, 1991).
Figure 29: Plots of ed fit using grid augmentation. The blending parameter used was \( \omega^* = 0.38 \), and the mixture \( g = 0.62f + 0.38u \) was fit with local cubic fitting and \( \alpha = 0.16 \). The top panel is the fit, \( \log \hat{g} \), and the bottom panel is \( \hat{f} \).
Another computational requirement in loess smoothing is estimating the standard error of the estimates, as discussed in section 6.4. Direct calculation of the standard error estimate and corresponding degrees of freedom is computationally burdensome. In the loess implementation, there is a clever approach to estimating these quantities efficiently. In this application, the theoretical variance of the residuals is known, so that this calculation is unnecessary.

9 Discussion

Similar work and ideas include Marron and Udina (1999), where the idea is briefly mentioned of using residuals to indicate when performance of the density estimate is poor, by binning the data to a fine grid and using the difference between the kernel estimate and the bin frequencies as residuals. Also, Farmen and Marron (1999) discuss binning the data and then fitting locally using a smoothing method that works in the presence of heteroscedastic noise. Cleveland, Mallows, and McRae (1993) briefly discuss a notion similar to ed.

There are some existing visualization tools for density estimation. SiZer (Cha, 1999) addresses the question of what features of a density are significant, examining a family of different fixed bandwidth kernel density estimates. It provides analysts with a visual tool to quantitatively decide which features are really there. Mode trees (Min, 1993) is a similar tool. These tools are very useful for exploring the features of a density, but are not helpful in obtaining a final density estimate, as they only consider families of fixed-bandwidth estimates, and most densities cannot be accurately estimated using fixed bandwidths.

Ed shows promise as a useful tool for exploring densities with moderate or large samples. Simply viewing the raw estimates alone can be extremely insightful. The procedure is very straightforward, and problems such as discontinuities, boundary cut-offs, and mis-fitting peaks and valleys can be detected and dealt with very easily all within the same framework. In the traditional kernel approach, on the other hand, one would need to resort to boundary kernels, variable bandwidths, higher-order kernels, or other means to deal with these problems. The low computational complexity of ed is also very attractive from a practical standpoint.

Future work includes extending ed to higher dimensions, more work on grid augmentation, and determining more rigorously the theoretical properties of the bias and variance. Another possible research direction would be work on automating parameter selection for ed. Although the method has been presented here as a modeling tool, which implies an iterative human-guided approach, the favorable bias and variance properties along with the sufficiently fast computation time may make this method a useful tool in machine learning algorithms.
References


