[27] Ullah, A. (1985) Specification analysis of econometric models, *Journal of Quantitative Economics*, **2**, 187-209.

[28] Wahba, G. (1990) *Spline Models for Observational Data*, SIAM, Philadelphia.

[29] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, London.

[30] Wand, M. P., Marron, J. S. and Ruppert, D. R. (1991) Transformations in density estimation, *Journal of the American Statistical Association*, **86**, 343-361.

[31] Weisberg, S. (1980) *Applied Linear Regression*, Wiley, New York.

[15] Minnotte, M. C. and Scott, D. W. (1993) The mode tree: a tool for visualization of nonparametric density features, *Journal of Computational and Graphical Statistics*, **2**, 51-86.

[16] Müller, H. G. (1988) *Nonparametric Analysis of Longitudinal Data*, Springer Lecture Notes in Statistics 46, Springer, New York.

[17] Park, B. U. and Marron, J. S. (1990) Comparison of data-driven bandwidth selectors, *Journal of the American Statistical Association*, **85**, 66-72.

[18] Park, B. U. and Turlach, B. A. (1992) Practical performance of several data driven bandwidth selectors (with discussion), *Computational Statistics* **7**, 251-285.

[19] Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *Journal of the American Staistical Association*, **85**, 617-624.

[20] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression, *Journal of the American Staistical Association*, **90**, 1257-1270.

[21] Schmitz, H. P. and Marron, J. S. (1992) Simultaneous estimation of several size distributions of income, *Econometric Theory*, **8**, 476-488.

[22] Scott, D. W. (1992) *Multivariate density estimation: theory, practice and visualization*, Wiley, New York.

[23] Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, **53**, 683-690.

[24] Silverman, B. W. (1978) Choosing the window width when estimating a density, *Biometrika*, **65**, 1-11.

[25] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[26] Terrell, G. R. (1990). The maximal smoothing principle in density estimation, *Journal of the American Statistical Association,* **85**, 470-477.

[3] Diggle, P. J. (1985) A kernel method for smoothing point process data, *Applied Statistics*, **34**, 138-147.

[4] Eubank, R. L. (1988) *Spline smoothing and Nonparametric Regression*, Dekker, New York.

[5] Fan, J. and Gijbels, I. (1996) *Local polynomial modelling and its applications*, Chapman and Hall, London.

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

[7] Green, P. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London.

[8] Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press.

[9] Izenman, I. and Sommer, C. J. (1988) Philatelic mixtures and multimodal densities, *Journal of the American Statistical Association*, **83**, 941-953.

[10] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a) A brief survey of modern bandwidth selection methods, *Journal of the American Statistical Association*, **91**, 401-407.

[11] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b) Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics*, **11**, 337-381.

[12] Kooperberg, C. and Stone, C. J. (1991) A study of log spline density estimation, *Computational Statistics and Data Analysis*, **12**, 327-347.

[13] Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error, *Annals of Statistics*, **20**, 712-736.

[14] Marron, J. S. and Udina, F. (1995) Interactive local bandwidth choice, Economics Working Paper 109, Universitat Pompeu Fabra.

Both families show the surprising bimodal structure known to be present in that data set.

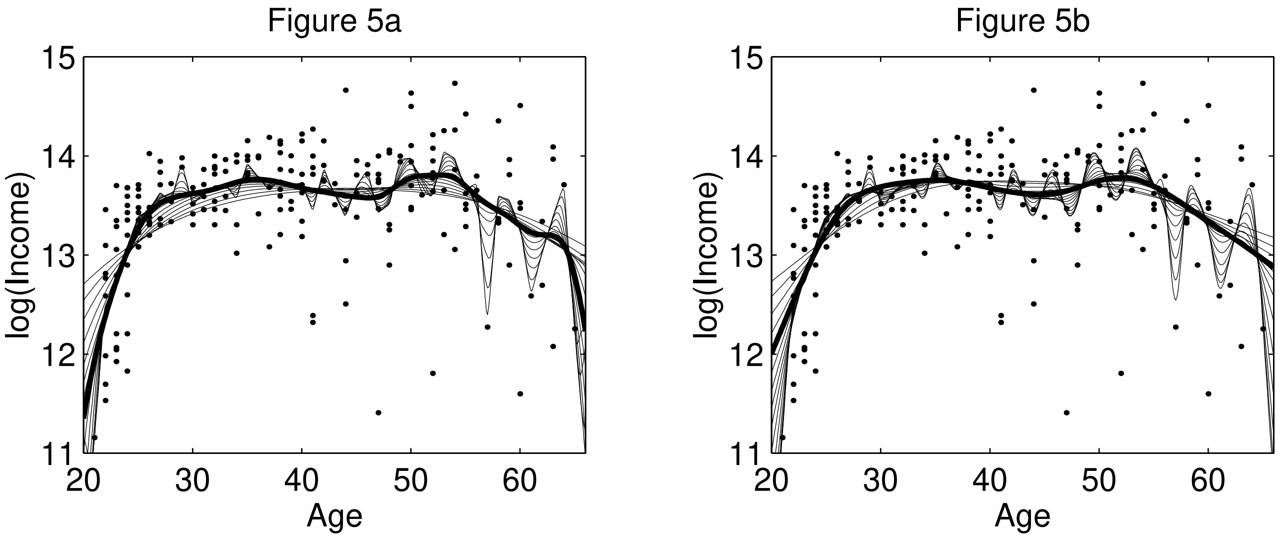

FIGURE 5: *Families of smooths, for Canadian Earning Power Data, overlaid with scatterplot of the raw data. Figure 5a is the local linear, centered at the Ruppert, Sheather and Wand data driven bandwidth. Figure 5b is smoothing splines, with smoothing range chosen to visually match Figure 5a.*

The family approach will not fit well with all types of smoothers. It is only effective when each member of the family is smooth enough to allow the eye to "visually connect" adjacent members. This is quite difficult with histograms and nearest neighbor based methods.

References

[1] Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Computational Statistics and Data Analysis,* **17** 153-176.

[2] Chu, C. K. and Marron, J. S. (1991) Choosing a kernel regression estimator, *Statistical Science*, **6**, 404-436.

2.3 Speed considerations

One drawback to construction of a family of smooths is that it requires more computational effort. Especially if the family approach is to be combined with interactive methods, it is important to use a fast computational technique. Fan and Marron [6] discussed two such methods. The binning approach suggested there is especially useful in the family context, because the data only need to be binned once. In many situations, the added overhead of finding $h_{\max}$ as in (1), and of construction of a family, is rather small with this approach.

2.4 Other smoothers

The family approach can be employed with a number of other different smoothing methods. For example, Figure 5 shows families of smooths for the Canadian Earning Data shown in Chu and Marron [2], from Ullah [27].

Figure 5a shows a local linear family. The above ideas are simply adapted to this context. In particular, we recommend using the Ruppert Sheather Wand bandwidth h_{RSW} as the center point, and again using 15 smooths. The scale of the family is now more complicated, since "peak height" is a more difficult concept. Since peaks play a role that is symmetric with the role of "valleys", we suggest studying "distances from the center". For the local linear smooth, the natural notion of center is the least squares fit line $\ell(x)$, which is the limit of the smooth $\widehat{m}_h(x)$ as $h \to \infty$. The one sided threshold of 60% for density estimation appears naturally as 30% in this two sided context. Hence define

$$h_{\max} = \max\left[\inf\left\{h > h_{RSW} : \left|\widehat{m}_h(x^*) - \ell(x^*)\right| \leq 0.3\left|\widehat{m}_{h_{RSW}}(x^*) - \ell(x^*)\right|\right\}, 3h_{RSW}\right],$$

where x^* is a location that maximizes $|\widehat{m}_{h_{SJPI}} - \ell|$, i.e.

$$\left|\widehat{m}_{h_{SJPI}}(x^*) - \ell(x^*)\right| = \max_x \left|\widehat{m}_{h_{SJPI}}(x) - \ell(x)\right|.$$

Using this $h_{\max}$, obtain $h_{\min}$ as in (2), and the corresponding grid as in (3).

A family based on smoothing splines is shown in Figure 5b. Since data based smoothing parameter selection is not so well understood for smoothing splines as for kernel / local linear methods, the range was chosen to be visually similar to Figure 5a.

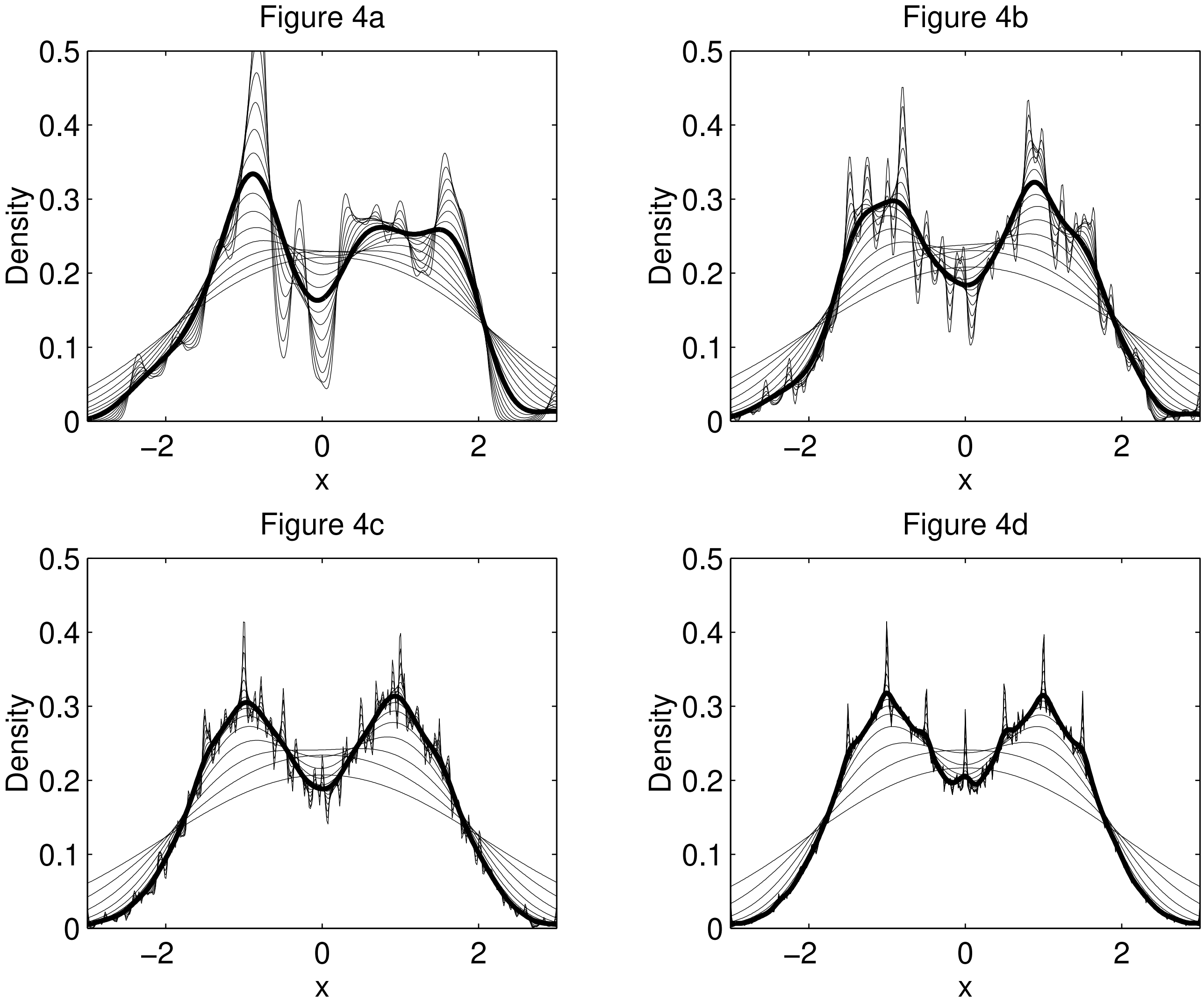

FIGURE 4: *Families of kernel density estimates for simulated data from the "double claw" distribution. Sample sizes are:* $n = 100$ *for Figure 4a,* $n = 1000$ *for Figure 4b,* $n = 10000$ *for Figure 4c and* $n = 100000$ *for Figure 4d.*

show that in Figure 4. The family approach gives visual insight as to how the variability in kernel smoothers decreases with n.(large scale for smaller samples, with both finer scale and more fine details for larger samples). Note that the even the crude bimodal structure is hard to discern for $n = 100$ (and in fact was not even present in the data for a test data set that is not shown here), but is quite clear for larger sample sizes. The finer spikes (which are a total of only 2% of the data) are not resolvable for sample sizes smaller than $n = 10^5$, and even there show up only in the family approach, not in the Sheather Jones estimate. This is another example where the family approach shows features that are not easily resolvable by conventional kernel density estimation.

15. Differences with nearby values are small, but much fewer than 15 makes it hard to "visually connect" the curves for nearby amounts of smoothing (which is essential for easy visual interpretation), as shown in Figure 3c. Many more than 15 seems to add distracting clutter, as shown in Figure 3d. An odd number is used to ensure that the centerpoint is included in the resulting grid.

Finally we consider the "breadth" of the family. A difficult aspect of this is that it depends on the sample size. In particular, since estimates with bandwidths that are any multiple of $n^{-1/5}$ converge to f as n grows, there is no universally acceptable multiple of the centerpoint, that can be used (because the difference is arbitrarily small for n large enough. We approach this problem by noting that we can maintain a "fixed range of smooths" by using an upper endpoint, which does not tend to 0 as n increases. For kernel density estimation, we recommend the bandwidth that smooths the height of the tallest peak in the central smooth down to 60% of its height. Mathematically, define

$$h_{\max} = \max\left[\inf\left\{h > h_{SJPI} : \widehat{f}_h(x^*) \leq 0.6\widehat{f}_{h_{SJPI}}(x^*)\right\}, 3h_{SJPI}\right], \qquad (1)$$

where x^* is a location that maximizes $\widehat{f}_{h_{SJPI}}$, i.e. $\widehat{f}_{h_{SJPI}}(x^*) = \max_x \widehat{f}_{h_{SJPI}}(x)$. The threshold of 60% is a personal choice, based on consideration of a number of examples. The lower bound of $3h_{SJPI}$ is sometimes useful for small sample sizes, to keep the family from being too narrow. For the lower endpoint, we suggest "going the same amount in the opposite direction". In particular, define

$$h_{\min} = \frac{h_{SJPI}}{(h_{\max}/h_{SJPI})} = \frac{h_{SJPI}^2}{h_{\max}}. \qquad (2)$$

The above recommendations may be summarized as the grid of bandwidths

$$\left\{h_{\min}\left(\frac{h_{\max}}{h_{\min}}\right)^{\frac{i-1}{14}} : i = 1, ..., 15\right\}. \qquad (3)$$

This grid was used at all points in Figures 1 - 4, except where noted otherwise.

To check that this range of bandwidths is appropriate over a wide range of settings, we tested it on simulated data sets of sample sizes $n = 10^2, 10^3, 10^4, 10^5$, for each of the family of test densities proposed by Marron and Wand [13]. The case of their density #11, the "Double Claw" was interesting, so we

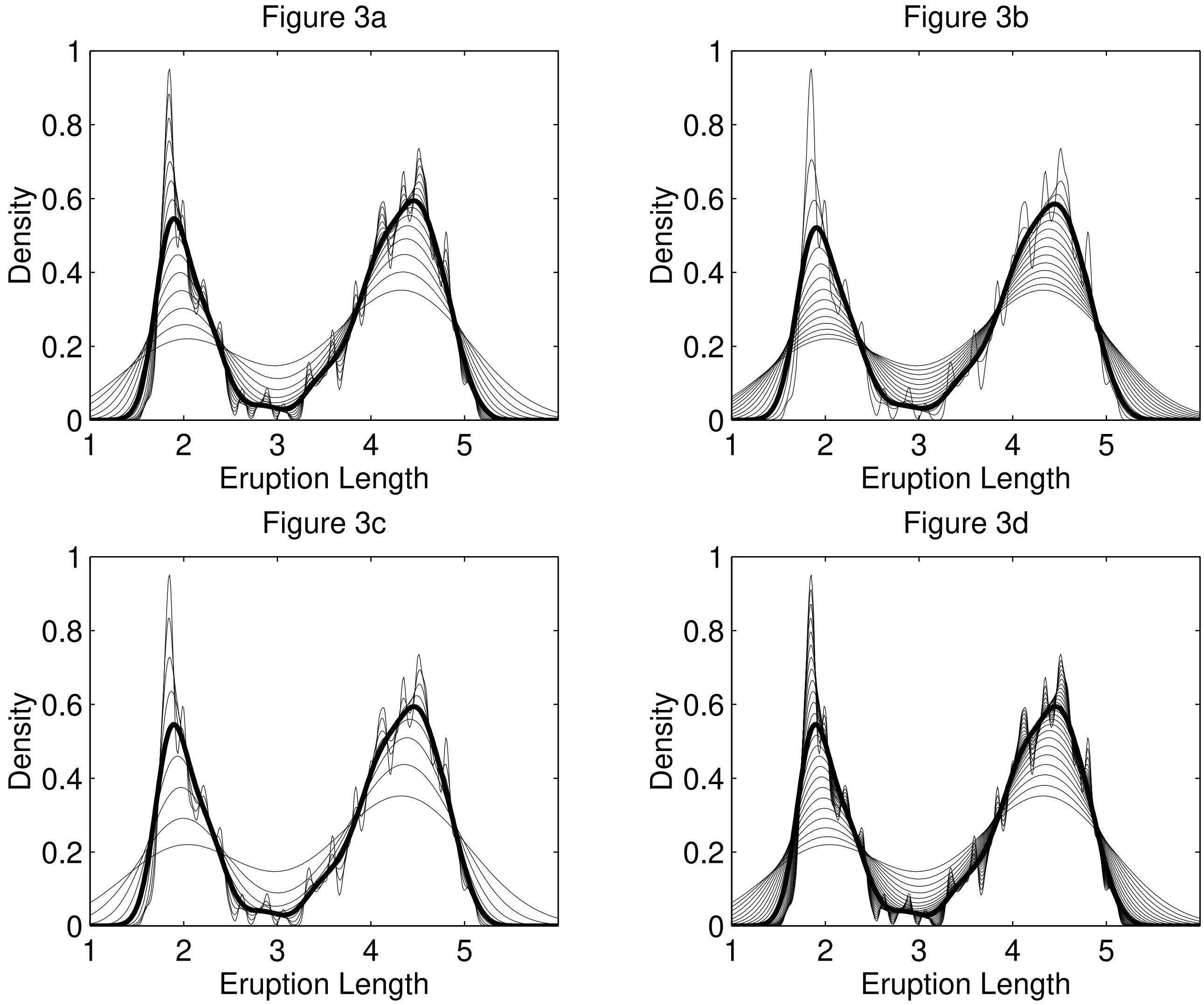

FIGURE 3: *Families of kernel density estimates for the Geyser data. The families in Figures 3a, c and d are logarithmically equally spaced, while linear spacing is used in Figure 3b. The number of members in the families are 15 for Figures 3a and b, 9 for Figure 3c and 25 for Figure 3d.*

Next we consider number of curves in the family, which involves some personal choice. After consideration of a number of examples, we suggest

and Scott, but not others because it is hard to separate from the two nearby larger modes, is clearly apparent.

For some types of smoothers, such as splines, the issue of good centerpoint is more difficult, because effective methods of smoothing parameter selection are not so readily available. Cross-validation type methods have the advantage that they are widely applicable, but they are not recommended here because they are too "noisy", as made clear in Jones, Marron and Sheather [10] [11]. However, even in the absence of a reliable center point, a nice feature of the family approach is that when one looks at a range, it is not crucial to get the amount of smoothing exactly right.

2.2 Scale of the family

Specification of the "scale" of the family of smooths involves two related issues. First there is the "breadth" of the family. Second is the "spacing" within the family. Both require a sensible quantification of "amount of smoothing". A simple approach is to recognize that the bandwidth works multiplicatively (i.e. it is more intuitive to double the bandwidth than to add 10 to it), so the log of the bandwidth is more natural.

Another approach to spacing, suggested by George Terrell, is to consider power transformations, and choose the power which makes the "sensitivity curve" of Scott [22], Section 3.2.5, as "symmetric as possible". One way of doing this is to choose the power so that the Taylor expansion at the minimum has a zero quadratic term. For dimension $d = 1$ this results in using the original linear bandwidth scale (but the scales $h^{2/3}$ for $d = 2$, $h^{1/3}$ for $d = 3$, and $\log(h)$ for $d = 4$).

A visual comparison of the linear versus logarithmic scale is available from Figures 3a and b. The data are the Geyser data from Table 2.2 of Silverman [25], which came from Weisberg [31]. The linear scale gives a family which is visually "more compressed" at the oversmoothed end, and "more sparse" at the undersmoothed end. The logarithmic scale is recommended, because the curves in Figure 3a look closer to "equally spaced in a visual sense".

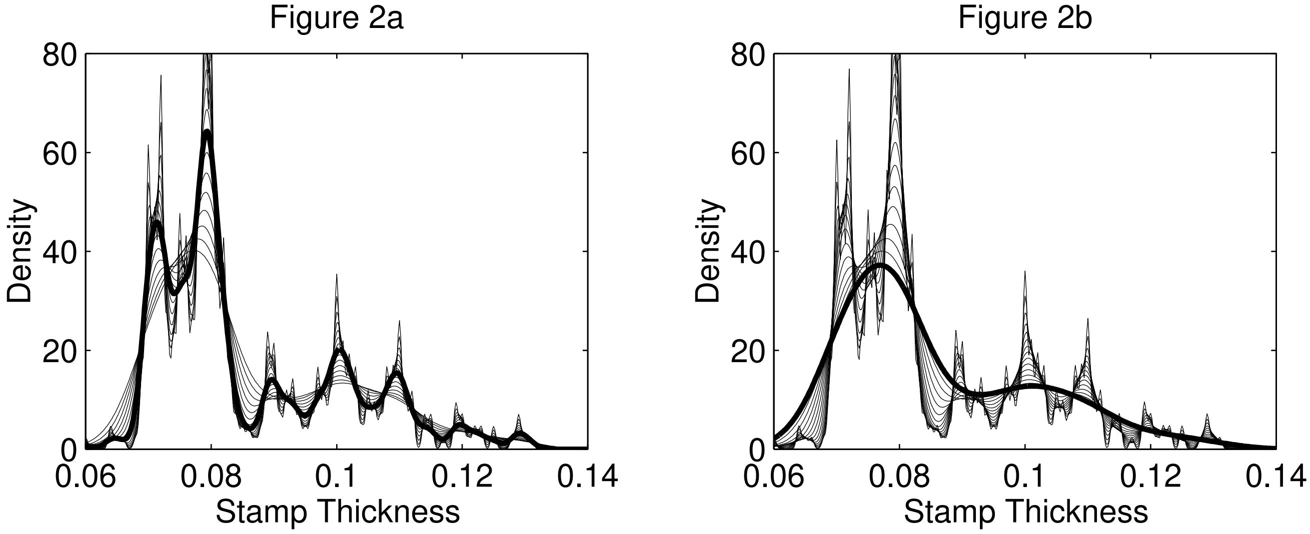

FIGURE 2: *Families of kernel density estimates for the Hidalgo Stamp data. Figure 2a shows the location of the family chosen centered at the Sheather Jones Plug In data driven bandwidth (shown as the thick curve). Figure 2b shows location chosen so the upper endpoint is at the over smoothed bandwidth (shown as the thick curve).*

One situation where centering at the Sheather Jones bandwidth is different from putting the upper end at the oversmoother, is when the data are unimodal and not too far from the normal distribution. Then the oversmoother is close to an appropriate amount of smoothing, so the family is skewed towards undersmoothing. The opposite extreme is when an appropriate amount of smoothing is quite small, which results in this family being skewed towards oversmoothing. Centering at h_{SJPI}, the Sheather Jones Plug In data driven bandwidth, is recommended because it more often gives a reasonable center point, and because it is more appealing to have the thick curve in the center of the family.

Note that the heavy curve in the middle of Figure 2a has the 7 modes found by Izenman and Sommer [9]. Additional structure in the data is shown by the other members of the family. For example, the locations of the 10 modes found by Minnotte and Scott, see their Figures 11 and 12, are now easy to see. Also the interesting mode near .76 that was found by Minnotte

Methods for specifying "location" are considered in Section 2.1. Choices of "scale" and related issues are discussed in Section 2.2.

2.1 Location of the family

A natural approach to controlling the location of the family to be presented, among the set of all smooths, is to use a "centerpoint" which is an effective choice of the global smoothing parameter. This is straightforward in kernel type smoothing, where good methods are known. See the review papers by Park and Turlach [18], Jones, Marron and Sheather [10], [11] and Cao, Cuevas, and González-Manteiga [1] about kernel density estimation, and Ruppert, Sheather and Wand [20] about local polynomial regression. In the kernel density estimation examples in this paper, the Sheather Jones Plug In method is used, unless indicated otherwise. The Ruppert Sheather Wand "Direct Plug In" bandwidth is recommended for local linear nonparametric regression.

An interesting alternative is the "oversmoothed" bandwidth, see Terrell [26]. This gives a bandwidth which attempts to be greater than or equal to an appropriate amount of smoothing. Hence, when this is used as the measure of location, it is natural to use it as the upper endpoint on the bandwidth scale. In many cases this gives a reasonably similar family to that centered as suggested above. An example is the Hidalgo Stamp data, from Izenman and Sommer [9], as shown in Figure 2.

The insights about the data that could be displayed through some type of location adaptive bandwidth are also all present in the plot. See Wand, Marron and Ruppert [30] and Kooperberg and Stone [12] for location adaptive methods which have been applied to this particular set of data. The advantage of the family approach is that it is simpler and more "interpretable" than a location varying smoothing method. In particular, it is much harder to understand at an intuitive level what most location varying smoothers (e.g. nearest neighbor based) are indicating about the data, except in the abstract sense of being a "curve estimate". See Marron and Udina [14] for explicit evidence of this.

An additional advantage of the family approach is that it easily adapts to situations where there is useful information in the data at several different levels of smoothing. See Diggle [3] for examples where this is an issue.

Some specific recommendations for choice of families are made in section 2. The ideas are developd for kernel density estimation, but apply also to other types of smoother, as shown in section 2.4.

Figure 1, and also each of the other figures in this paper, shows a phenomenon that was perhaps first discussed in print in Silverman [24]. In particular, when a bandwidth is adjusted from oversmoothing towards undersmoothing, changes tend to be relatively slow until a particular point where many small features suddenly begin to appear. Silverman's claim that the point where the big change occurs is a reasonable amount of smoothing is quite apparent from the figures here.

2 Specific Recommendations

Some specific suggestions for construction of a family are made and discussed in this section. The main discussion is in terms of kernel estimators, because they are very simple, and have excellent interpretability properties. However most of the essential ideas apply also to other types of smoothers. Discussion of adaption to some other smoothers, including local polynomials and smoothing splines, is given in Section 2.4. The recommendations are based on personal preference, gained through experience in visual presentation of smooths. They are intended only as rough guidelines.

A useful way to parameterize a family of smooths is in terms of "location" and "scale", among the set of all possible smooths for the given data set.

[17]. See Schmitz and Marron [21] for detailed discussion of these data. The "bandwidth" (i.e. smoothing parameter) in Figure 1a is chosen by a "plug-in" method, proposed by Sheather and Jones [23]. This method has been shown to be a good all-purpose smoothing parameter selector by Jones, Marron and Sheather [10] [11], and provides a good compromise for "best" choice of global bandwidth for these data. The single curve shows the surprising bimodality, which is the most important feature of the data. However, the family of curves reveals much deeper structure. In particular, it is immediately clear from Figure 1b that the thin peak on the left should be taller and thinner than is shown in Figure 1a, and the tail on the right should be smoother.

These facts about this data set could also be obtained from two other approaches:

- An interactive trial and error visual search through several bandwidths.
- A good location adaptive smoothing method (which would use a smaller bandwidth on the left, and a larger one on the right).

Another method of presenting a family of smooths, see e.g. Figure 2 of Roeder [19] and Figure 3 of Minnotte and Scott [15] (whose Figure 2 is also a precursor of our Figure 1b), is the "perspective surface" shown in Figure 1c. The surface approach was seen to be very effective for studying modality issues by Minnotte and Scott [15]. A reviewer has pointed out that this surface approach can be preferable when it is important to view many curves, such as when larger dynamic bandwidth range is important. However for exploratory data analysis, we view the overlay in Figure 1b as being preferable, because important features of the data are more immediately visible, while they are too often hidden in 3-d surface views. In particular, it is easier to miss both failings of the smooth in Figure 1a from studying only Figure 1c.

An advantage of the family approach over the usual interactive methods is that more information is summarized in one plot. This means that the user can spend less effort in keeping previous pictures in mind, and thus can concentrate more fully on the data. Furthermore one can learn about the data more rapidly. However the family approach should be viewed as more than just a competitor of interactive smoothing choice, indeed it can be an effective partner. In particular, the trial and error approach can be more efficient if one looks at a range of smooths on each trial.

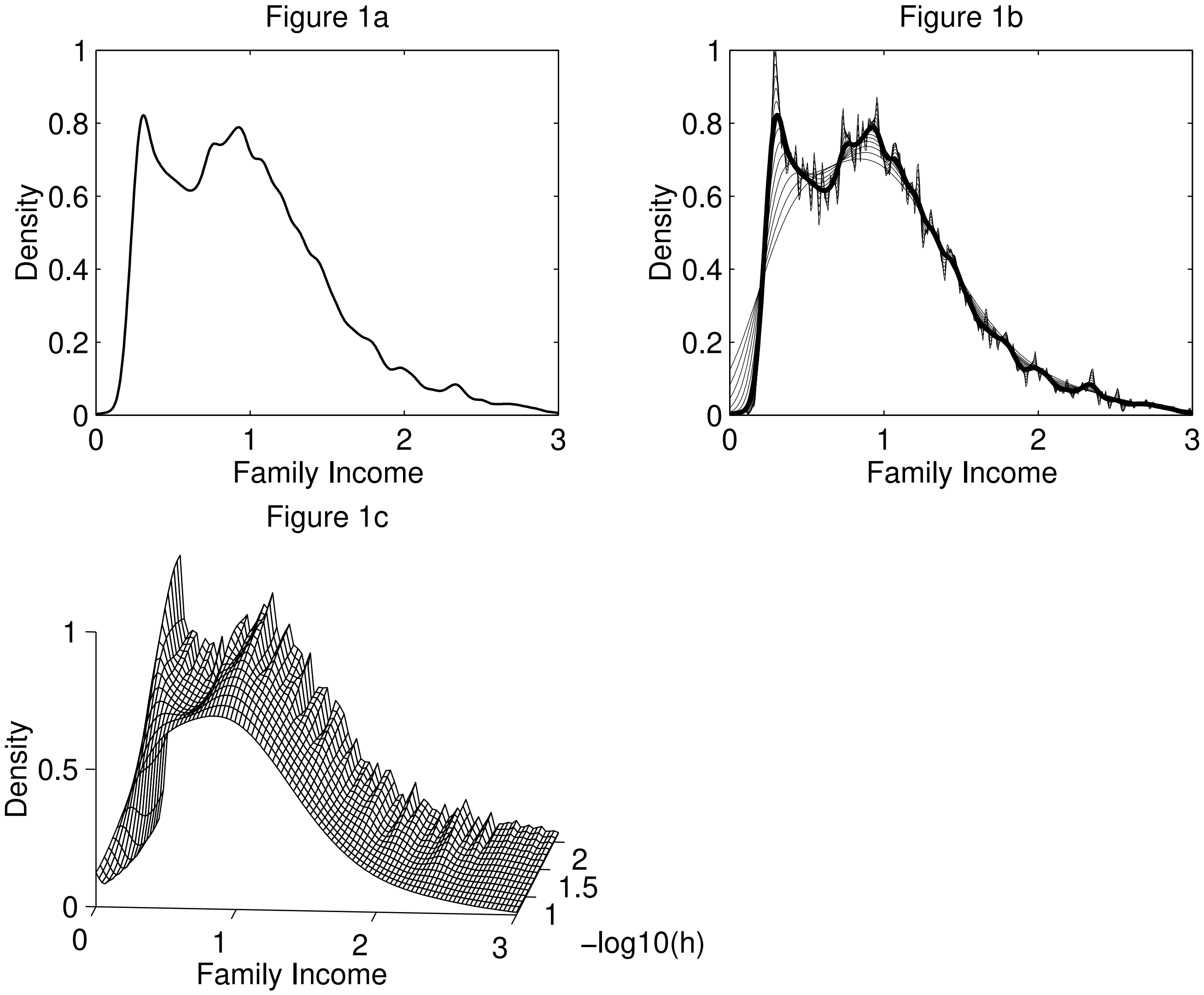

FIGURE 1: *Kernel density estimates for the Income Data, 1975. Figure 1a shows a single estimate using the Sheather Jones automatic bandwidth. Figure 1b shows a family of estimates centered at the estimate in Figure 1a. Figure 1c shows the same family, expanded as a surface, with bandwidth as the new dimension.*

This shows kernel density estimates using the Gaussian kernel, for the "1975 Income Data", which is part of a larger set shown in Park and Marron

usefulness for data analysis comes from an ability to easily see structure in data that can be very difficult to ascertain otherwise, see for example the monographs of Eubank [4], Härdle [8], Müller [16], Silverman [25], Scott [22], Wahba [28], Green and Silverman [7], Wand and Jones [29], and Fan and Gijbels [5]. The theoretical interest in smoothing is perhaps mostly due to the fact that it is fun to think about, but also presents challenging problems. Many important issues, such as smoothing parameter selection and assessing variability, are easy to understand, but have been difficult to address in a satisfactory way.

While parts of the available theory are deep and elegant, there is a tendency for the theoretical work to be distracting from points that are important to the analysis of data. For example, there is wide spread acceptance of the idea that the goal of smoothing should be a *single* curve estimate (or perhaps an estimate together with confidence bands). This idea comes very naturally from parametric statistics, where all important information is summarized in a single parameter. It is also theoretically convenient, as issues such as optimality are then straightforward to formulate.

The main premise of this paper is that smoothing becomes a more powerful graphical device, if the usual notion of a single curve estimate is replaced by a *family of smooths*, indexed by the smoothing parameter. This reveals structure in the data more quickly and easily than is possible from any single estimate, as illustrated in Figure 1.

Presentation of smoothers: The family approach

J. S. Marron*
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
S. S. Chung
Department of Statistics
Chonbuk National University
Chonju, Chonbuk 561-756
Korea

February 5, 1997

Abstract

The product of most statistical smoothing methods is a single curve estimate. A drawback of such methods is that what is learned varies with choice of the smoothing parameter. This paper proposes simultaneous display of all important features, through presentation of a family of smooths. Some suggestions are given as to how this should be done.

1 Introduction

Smoothing, or nonparametric curve estimation, is a useful data analytic tool, and also has been the subject of a great deal of theoretical research. It's

*Research Partially Supported by NSF Grant DMS-9203135, and by Division of Mathematics and Statistics, CSIRO.