On Fitting and Adapting of Density Estimates

David W. Scott
Department of Statistics, MS-138
Rice University
Houston, TX 77005-1892

Abstract

We discuss two uses of polynomials for constructing locally adaptive density estimates. These ideas build upon the well-established literature on spline approximation and local polynomial techniques. The first builds a polynomial approximation by computing low-order sample-moments of data in bins. The second approach considers "polynomial patches" over kernel density estimates. The first technique is especially promising for large data sets, while the second provides an interesting approach for testing features in a density Each attempts to provide a density estimate that retains only significant features.

1 Introduction

Producing reliable density estimates is our goal. Our discussion focuses on univariate data, as multivariate problems introduce a host of additional practical problems. A good density estimate should not only be close to the true density function, but should also contain important features of the density. Features of interest include the proper number and location of modes and bumps (Good and Gaskins, 1980) in the density. A histogram may be close to the true density, yet simultaneously have too many modes. An incorrectly specified parametric model may miss important features. If we could assure that $\hat{f}$, $\hat{f}'$, and $\hat{f}''$ are all close to truth, then we have done our work. This task is not easy, and is complicated by the fact that in nonparametric estimation, many more samples are required to estimate the second derivative than the function itself. Proper estimation requires a number of compromises, which may be difficult to articulate for a particular data set.

There are different strategies one can employee for this task. A parametric model may be posed followed by testing. A mixture model may be fit using the EM algorithm, with some iterative or testing procedure for determining the number of components. A nonparametric estimate fitted by cross-validation is generally indicated, but has been criticized for being too variable. In my opinion, much of the variability can be traced to the lack of scale-adaptivity inherent in the estimate when applied to data that are not scale homogeneous. However, testing of features in the (say, kernel) estimate is a conceptually valuable post-estimation procedure. Finally, many locally adaptive algorithms have come into existence and attempt to go directly to a proper estimate.

We consider several of these in some detail. Imagine a density estimate with a bump (not a mode) adjacent to the primary mode. If the density estimate is a mixture, testing for that secondary feature is not always straightforward (i.e., a null hypothesis of one versus two components). If the main portion of the density requires several components for modeling, then there may not be a clean way of isolating other features.

Testing extra modes in kernel estimates by excess mass and bootstrap simulation is a good example of the third strategy (Müller and Sawitzki, 1991; Minnotte, 1997). The size of the extra mode is estimated and then compared to the size of the extra mode in bootstrap estimates. Testing the more subtle feature of an extra bump might be feasible by defining an excess mass criterion on the derivative function.

However, the cleanest approach to the testing problem in my mind is the original idea proposed by Good and Gaskins (1980). They proposed an iterative procedure that modified the data in order to eventually produce an estimate that did not contain that feature. Since their estimation procedure used a penalized maximum likelihood criterion, they proposed evaluating the two estimates by comparing their log penalized likelihood on the sample data. A primary advantage of their technique is the production of an estimate with a "false feature" removed. However, they did not provide a mechanism for doing this globally (if 10 tests revealed that 4 features were false). Yet, it seems a promising path to try to follow. We do so below.

Techniques that try to directly estimate an optimally locally adaptive density have great intuitive appeal. Examples are local polynomial (Hjort and Jones, 1996;

Loader, 1996), adaptive mixture (Marchette, et al., 1996; Priebe, 1994; Kooperberg and Stone, 1991; for example), and kernel (Sain and Scott 1996). These algorithms typically involve decisions such as knot placement, cross-validation, and other procedures which ultimately determine features present in the estimate. The reliability of these algorithms is still under study. Local testing may still be appropriate, if only to guage the strength of evidence. Such algorithms have greater practical limitations of course.

2 A Polynomial Algorithm for Large Data Sets

With massive data sets, many algorithms are severely tested. A computationally efficient algorithm such as the averaged shifted histogram (Scott, 1992) does not directly provide derivative and curvature estimates. Any algorithm requiring multiple passes in order to perform optimization tasks is automatically eliminated. Thus likelihood-based algorithms are next eliminated. The binned kernel estimate (Silverman, 1982, Scott and Sheather, 1985) is still feasible. However, studies of appropriate binning by Holmstrom (1998, unpublished) suggest that the practical granualarity may be far from optimal for feature estimation.

Thus a simple algorithm which still uses all the data would be most interesting. In particular, a locally adaptive estimator may not be required since the massive sample size makes even a fixed bandwidth algorithm sufficiently accurate in the sense described above. Such a simple algorithm is the binned polynomial histogram algorithm of Scott and Sagae (1997). They showed (Sagae and Scott, 1998) that if additional information was computed on a bin by bin basis, that higher order density estimators would result. Specifically, if the first $p+1$ sample moments are computed for all the data in each bin, then that is sufficient information in order to construct an order-p estimator by matching the "theoretical" bin moments of the estimator (a polynomial, in our implementation). For example, the ordinary histogram corresponds to $p=0$ in which the $0th$ moment or bin count is matched. A better estimator corresponding to usual kernel methods results if the bin count and local bin mean are matched, and so on.

In Figure 1, the bin-by-bin polynomial estimates for $0 \leq p \leq 3$ are displayed for a sample of 1,000 $N(0,1)$ points. Continuity is not imposed, but it is interesting to see how the estimate behaves near the bin edges. Scott and Sagae (1997) describe an implementation that does impose continuity as additional constraints. Specifically, suppose we consider $p=1$ over a mesh with m bins. Then we begin with $2m$ linear constraints on the polynomials, defined bin-by-bin. We have much experience in the subjective aspects of polynomial continuity visually, and cubic splines are the most widely accepted choice. This means at each of the $m-1$ interior bin edges, the polynomials must join in a smooth fashion (i.e., level, derivative, and curvature) for an additional $3(m-1)$ linear constraints, or $5m-3$ in all. Thus a fourth degree polynomial would have only 3 degrees of freedom unspecified. We choose instead to implement a fifth degree polynomial and to use the $m+3$ extra degrees of freedom to minimize $\int \hat{f}''(x)^2\,dx$. An example with 9 bins and 100,000 points from a mixture density is shown in Figure 2.

The polynomial histogram algorithm is essentially a one-pass algorithm through the massive data set. Starting with a large number of bins, the sample moments are computed in one pass. Now any bin adjustments are easily accomplished without the original data. The estimation complexity is controlled by the size of the mesh rather than the length of the data set. However, in a real sense, all of the data are reflected in the local bin moments. Experiments with complex mixtures with millions of data points gave almost perfect estimates. Since real massive data sets are often a million times larger, we are satisfied that we can accurately and computationally efficiently represent such data.

3 Testing Local Features in a Kernel Density Estimate

3.1 Previous Work

Fixed bandwidth kernel estimates cannot generally represent all density features equally well. If the bandwidth is chosen appropriate to the narrowest feature, the other features will be undersmoothed (with false modes and bumps the result). On the other hand, if the bandwidth is chosen appropriate to the widest feature, then narrower features may be completely lost or severely biased downwards. Thus the critical bandwidth test of Silverman (1978) which attempts to bootstrap a fixed bandwidth kernel estimate is generally unsatisfactory in any realistic situation. Excess mass ideas are local in nature, and thus can be employed at multiple scales. The original implementation of Müller and Sawitzki (1991) does not, but Minnotte's (1997) clever use of the mode tree does. Neither provides a "final estimate" reflecting the results of the testing.

3.2 The Density Patch

The bump surgery technique of Good and Gaskins (1980) has much to offer. However, a key step in the data manipulation has a renormalization that affects the data globally. Since log-likelihoods are heavily influenced by changes in the tails, the odds they compute are sensitive to the location of data far from the feature under consideration. Furthermore, their estimate was comprised of some 4,000 Fourier series terms. It was an amazing computational feat for the time.

Our idea is similar in that we explicitly attempt to perform "surgery" to remove the feature in question in order to provide a density which can be used in a null hypothesis concerning the veracity of that feature. Furthermore, we want the surgery to be entirely local to the feature. Specifically, given a fixed kernel estimator at an appropriate critical bandwidth for the feature of interest, we examine closely the location of zeros of $\hat{f}'$ and $\hat{f}''$ to define a pair of intervals that surround the feature and over which the patch should be inserted. For example, suppose an estimate has two adjacent modes (see Figure 3). Then if we wish to test if the pair could be replaced by a single mode, then we are doubting the entire estimate extending to critical points in $\hat{f}''$ on either side of the original modes. Thus potential locations for the patch extend beyond those points to the next set of critical points (or possibly beyond). These are indicated in Figure 3 as well.

Now at this point in Good and Gaskin's algorithm, they would start reallocating the data locally in an attempt to smooth out the two bumps and then renormalize repeatedly. We propose instead to define the patch directly. In Figure 3, the patch should begin in the interval AB and end somewhere in the interval CD. Let B' and C' denote the endpoints of the patch in those respective intervals. At each potential knot B' and C', we compute the following: $\hat{f}(B')$, $\hat{f}'(B')$, $\hat{f}''(B')$, and also at $x = C'$. The patch is defined to be a single polynomial over (B', C') that interpolates these values (in the cubic spline sense as with our smoothed polynomial histogram example) and is as smooth as possible (in the $\int p''(x)^2 dx$ sense over (B', C')). Such as procedure would be biased towards narrower intervals, so our exact criterion is

$$\int_A^{B'} \hat{f}''(x)^2 dx + \int_{B'}^{C'} p''(x)^2 dx + \int_{C'}^{D} \hat{f}''(x)^2 dx.$$

This includes the roughness of the kernel estimate and the polynomial patch together. Our patch is also constrained to match the area of the kernel estimate over (B', C'). Thus we have seven constraints plus a smoothness criterion. Our implementation uses a sixth degree polynomial and a grid search to find the smoothest patch.

An example of a patch on the $n = 63$ Buffalo snowfall data is shown in Figure 4. Here we are testing whether the rightmost mode is real. The patch is highlighted, and minimized the roughness criterion, whose contours are shown in Figure 5. Notice that the criterion surface is not strictly convex, which is why we have chosen a simple grid search algorithm. However, a more refined optimization would be appropriate as the patch contains a few very small zeros in its second derivative that could be removed by slight changes in the parameters. Figure 6 displays the optimal patch for testing the left mode.

Following Good and Gaskins, we use a likelihood ratio statistic to guide our evaluation of the weight of evidence in favor of the patch or the feature. The p-value indicated assumes the statistics follows a χ_1^2 distribution, which has not been established at this time. Because the two estimates are identical outside the interval (B', C'), the likelihood ratio statistic is entirely determined by data points falling in the interval containing the feature. This local behavior is appropriate for a nonparametric test.

For the snowfall data, the weight of evidence in favor of the right mode is nearly 3 times greater than that in favor of the left mode. (A p-value less than 0.05 is signally that the patch does not adequately represent the data in the interval of the feature.) Similarly a test of the small bump on the left in the tail (a single data point) also fails to support that mode ($p = 0.26$). Thus our final estimate would include two patches in the kernel estimate, eliminating the modes at 22 and 50, and leaving a single bump in their place. Other examples involving galaxy data will be presented in the talk.

We have conducted a number of simulations and tests using the p-value indicated, and it seems to be reasonable. We have considered using a polynomial density estimator in place of the kernel estimator so that the patch would form a more natural alternative, but have not pursued this so far.

4 Summary and Conclusions

Polynomial algorithms have been shown to be useful in two extreme settings. A pure polynomial spline-like estimator highly appropriate to the massive data set setting, and a local polynomial patch for testing features in a kernel estimator. Since density estimation is demanded in a very large spectrum of cases, we believe these and other algorithms will prove valuable in practice.

Acknowledgements

This research was supported in part by the National Science Foundation under grant DMS-9306658 and the Department of Defense under contract MDA 904-95-C-2203.

References

Good, I.J. and Gaskins, R.A. (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by the Scattering and Meteorite Data," (with Discussion) *J. Amer. Statist. Assoc.*, 75, pp. 42–73.

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

Kooperberg, C. and Stone, C.J. (1991), "A study of logspline density estimation," *Comput. Stat. Data Analysis*, 12, 327–347.

Loader, C.R. (1996), "Local Likelihood Density Estimation," *Annals of Statistics*, 24, 1602–1618.

Marchette, D.J., Priebe, C.E., Rogers, G.W. and Solka, J.L. (1996), "The Filtered Kernel Estimator," *Computational Statistics*, 11, 95–112.

Minnotte, M.C. (1997), "Nonparametric Testing of the Existence of Modes," *Annals of Statistics*, 25, 1646–1660.

Müller, D.W. and Sawitzki, G. (1991), "Excess Mass Estimates and Tests for Multimodality," *J. Amer. Statist. Assoc.*, 86, pp. 738–746.

Priebe, C.E. (1994), "Adaptive Mixtures," *J. Amer. Statist. Assoc.*, 89, 796–806.

Sagae, M. and Scott, D.W. (1998), "Bin Interval Method of Locally Adaptive Nonparametric Density Estimations," Rice Technical Report.

Sain, S.R. and Scott, D.W. (1996), "On Locally Adaptive Density Estimation," *J. Amer. Stat. Assoc.*, 91, 1525-1534.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley, New York.

Scott, D.W. and Sagae, M. (1997), "Adaptive Density Estimation With Massive Data Sets," Proceedings of the Statistical Computing Section, ASA, pp. 104–108.

Scott, D.W. and S.J. Sheather (1985), "Kernel Density Estimation with Binned Data." *Communications in Statistics*, 14, 1353–1359.

Silverman, B.W. (1978), "Choosing the Window Width When Estimating a Density," *Biometrika*, 65, 1–11.

Silverman, B.W. (1982), "Algorithm AS176. Kernel Density Estimation Using the Fast Fourier Transform," *Appl. Statist.*, 31, 93–99.

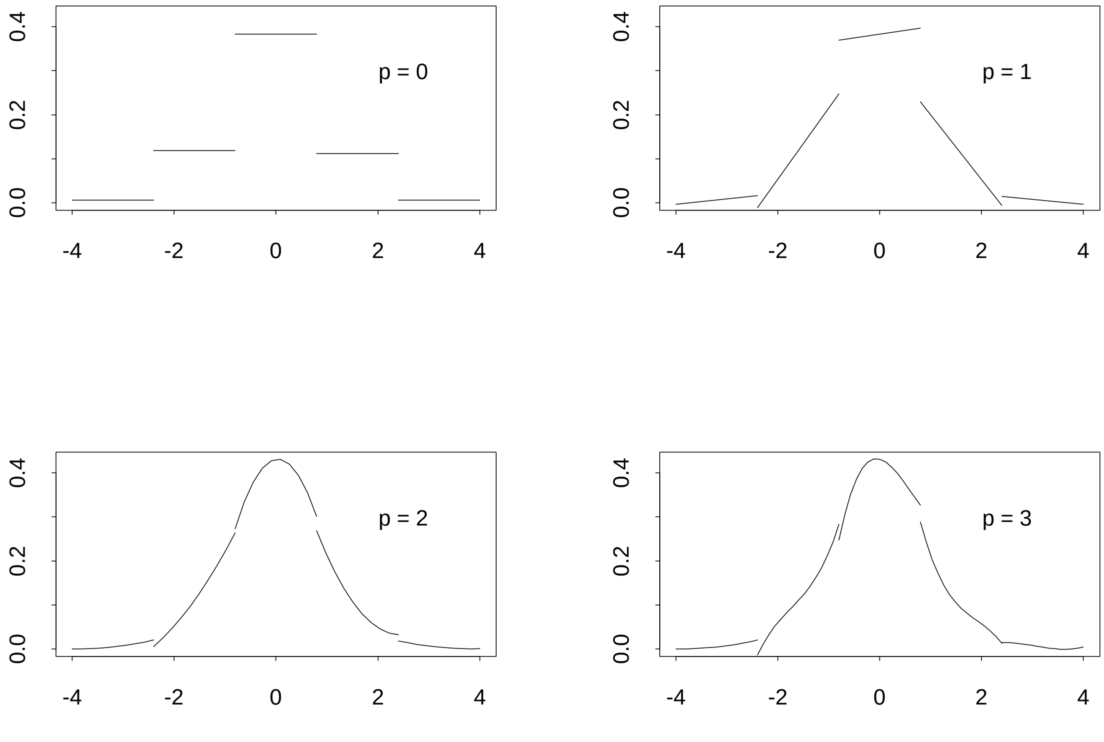

Figure 1: Polynomial histograms of order 0 to 3 for a $N(0,1)$ sample of size 1,000. The algorithm is applied bin by bin, and continuity is not imposed.

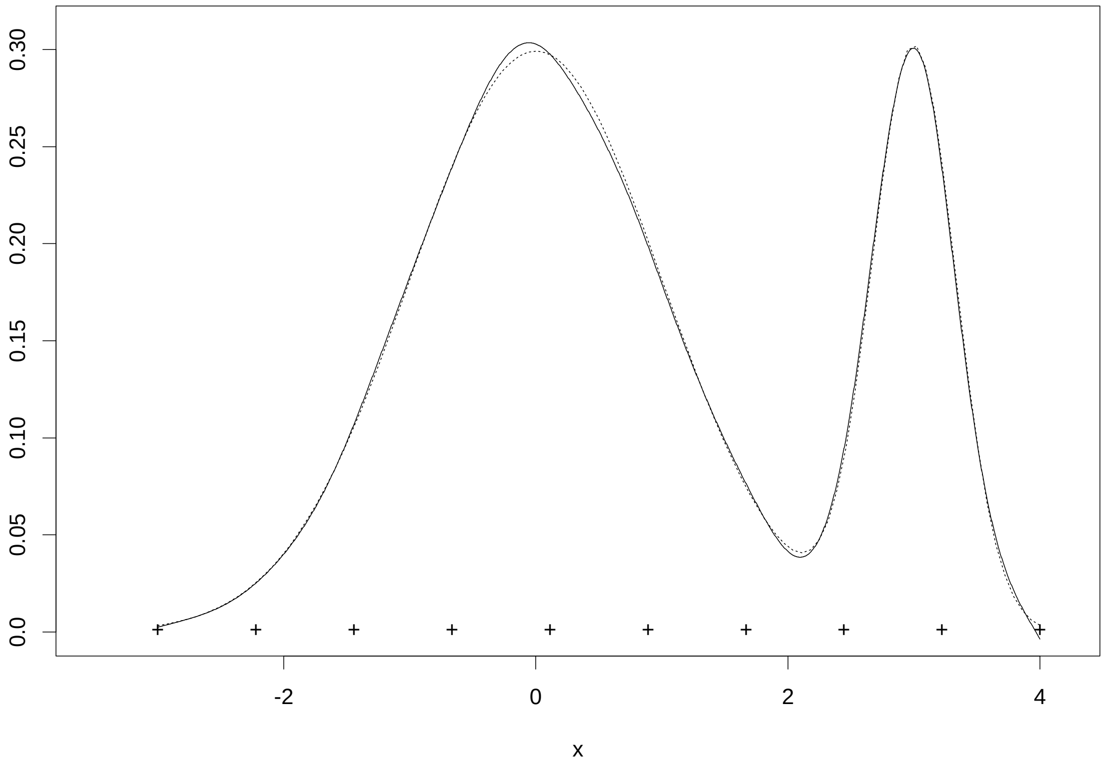

Figure 2: A continuous version of the polynomial histogram with 9 bins from normal mixture and 100,000 points. The estimate and true functions are nearly identical.

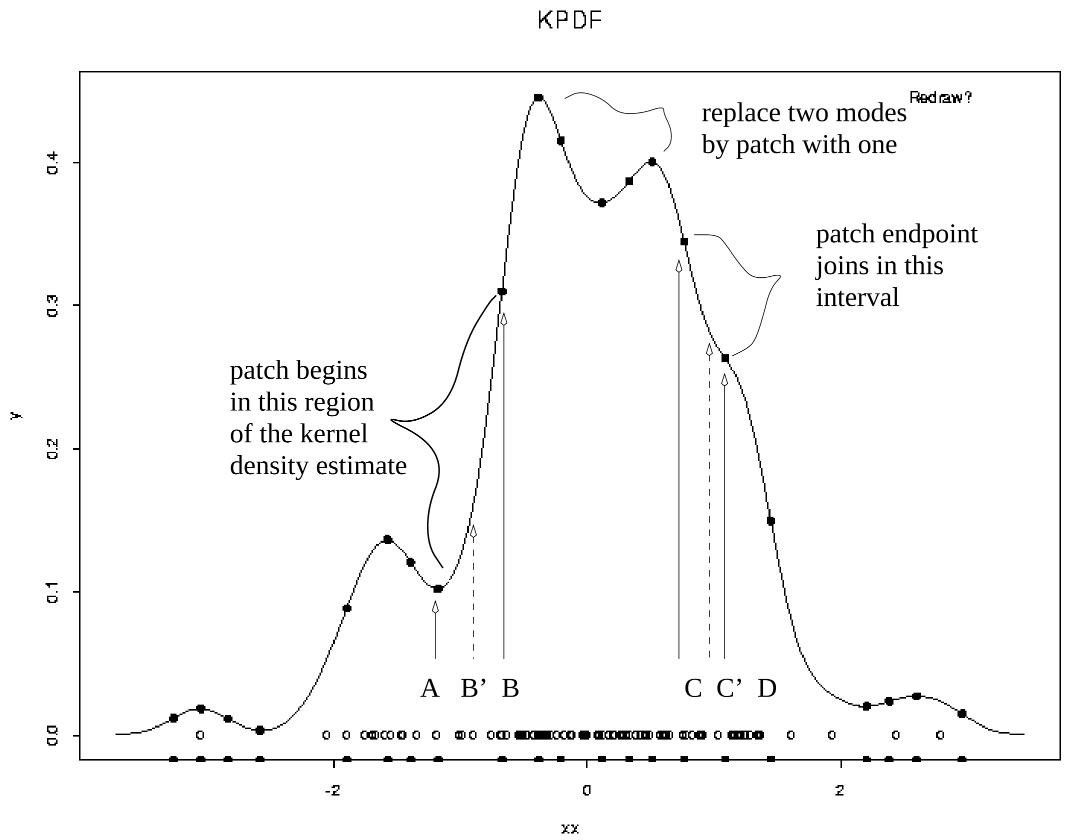

Figure 3: Graphic illustrating the determination of the polynomial patch to find a unimodal alternative. The search for the patch will begin on the kernel estimate above the interval AB and end above the interval CD. The optimal endpoints are denoted by B'C'. The dots on the kernel estimate indicate inflection points and points where the derivative is zero. Such points determine the two intervals AB and CD.

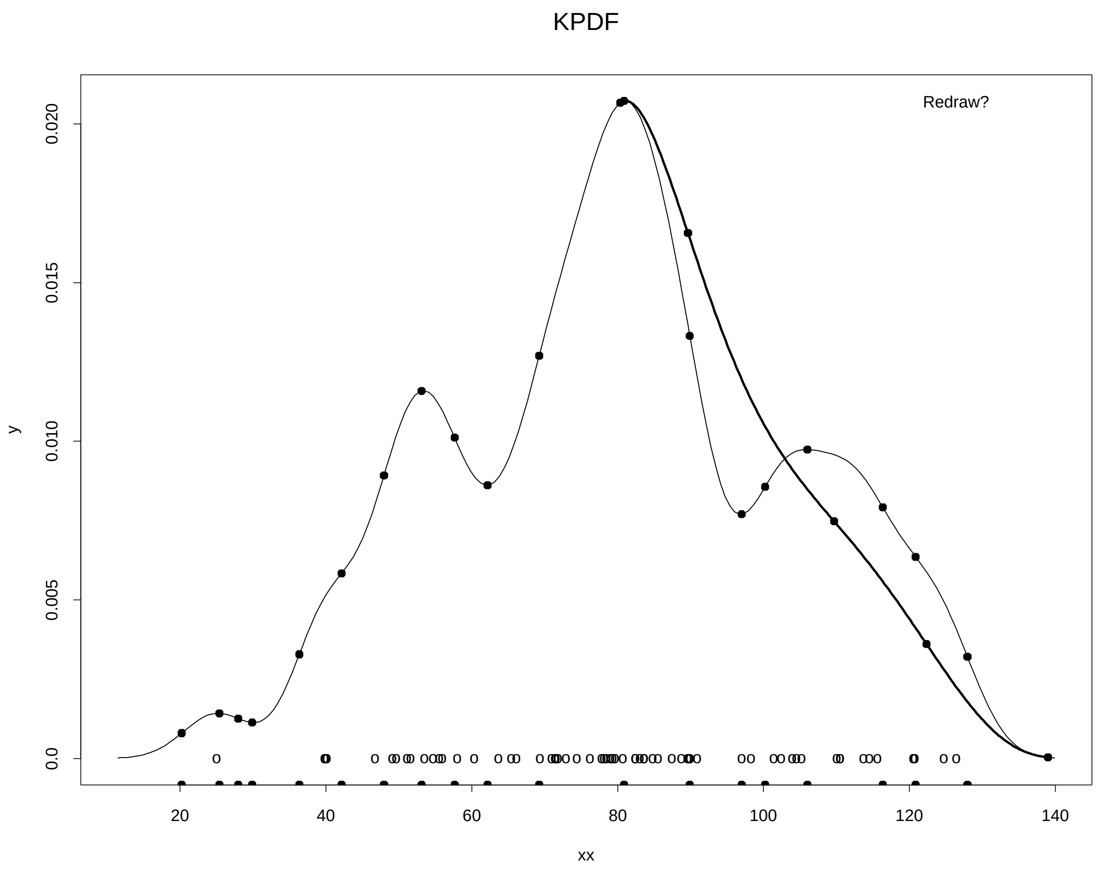

Figure 4: This graphic depicts the testing the rightmost mode in the Buffalo snowfall data with the bandwidth $h = 4.5$. The optimal patch is the thicker line extending from the highest mode to the right end of the graph. The test statistic computed from the likelihood ratio corresponds to a p-value of 0.046. Thus the patch is rejected as too simple a representation of the data in that region compared to the full kernel estimate. Note that the best patch has three rather than one inflection point.

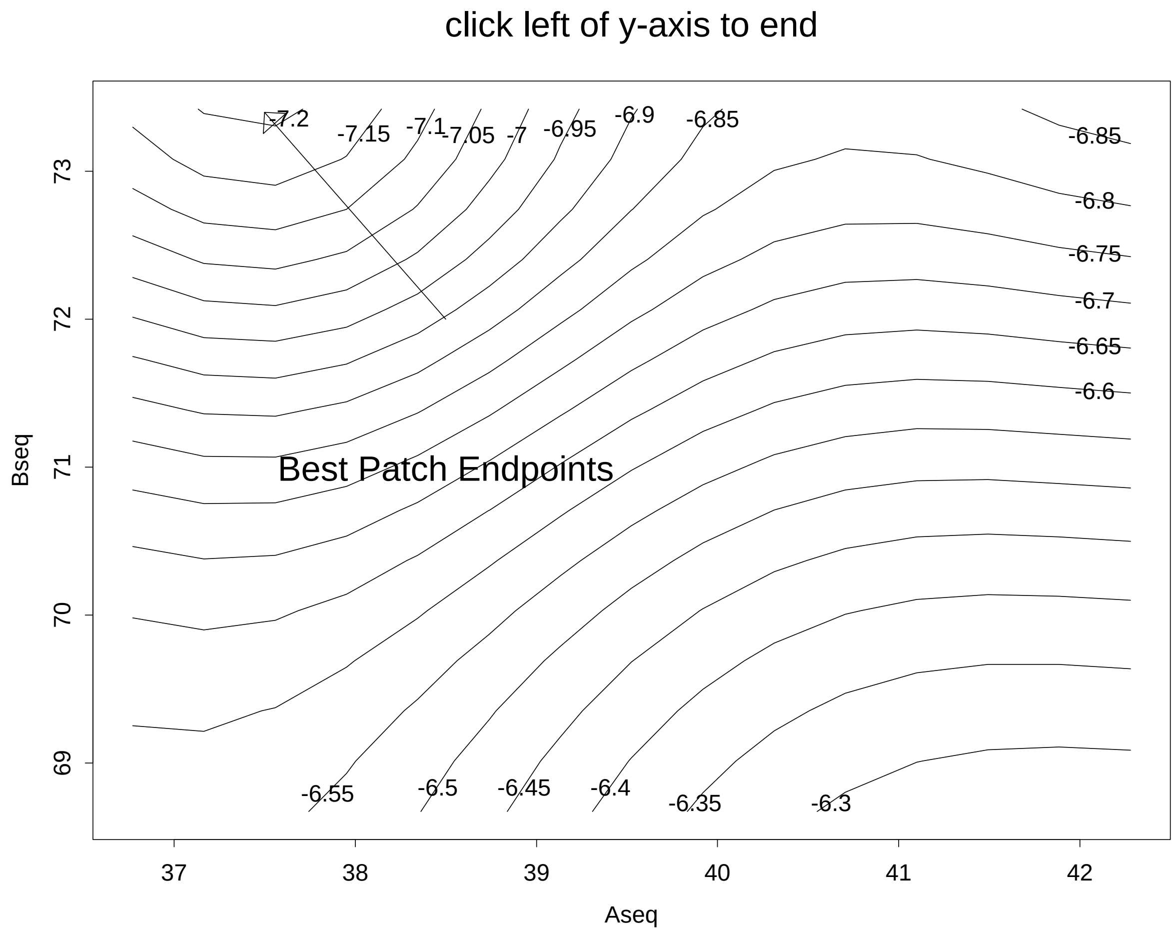

Figure 5: For testing the leftmost mode in the Buffalo snowfall data, these contours depict the log10 of the density roughness over the entire interval AD for over 200 hundred patches extending from B' to C' (labeled Aseq to Bseq in this figure). The optimal patch endpoints is indicated by the arrow.

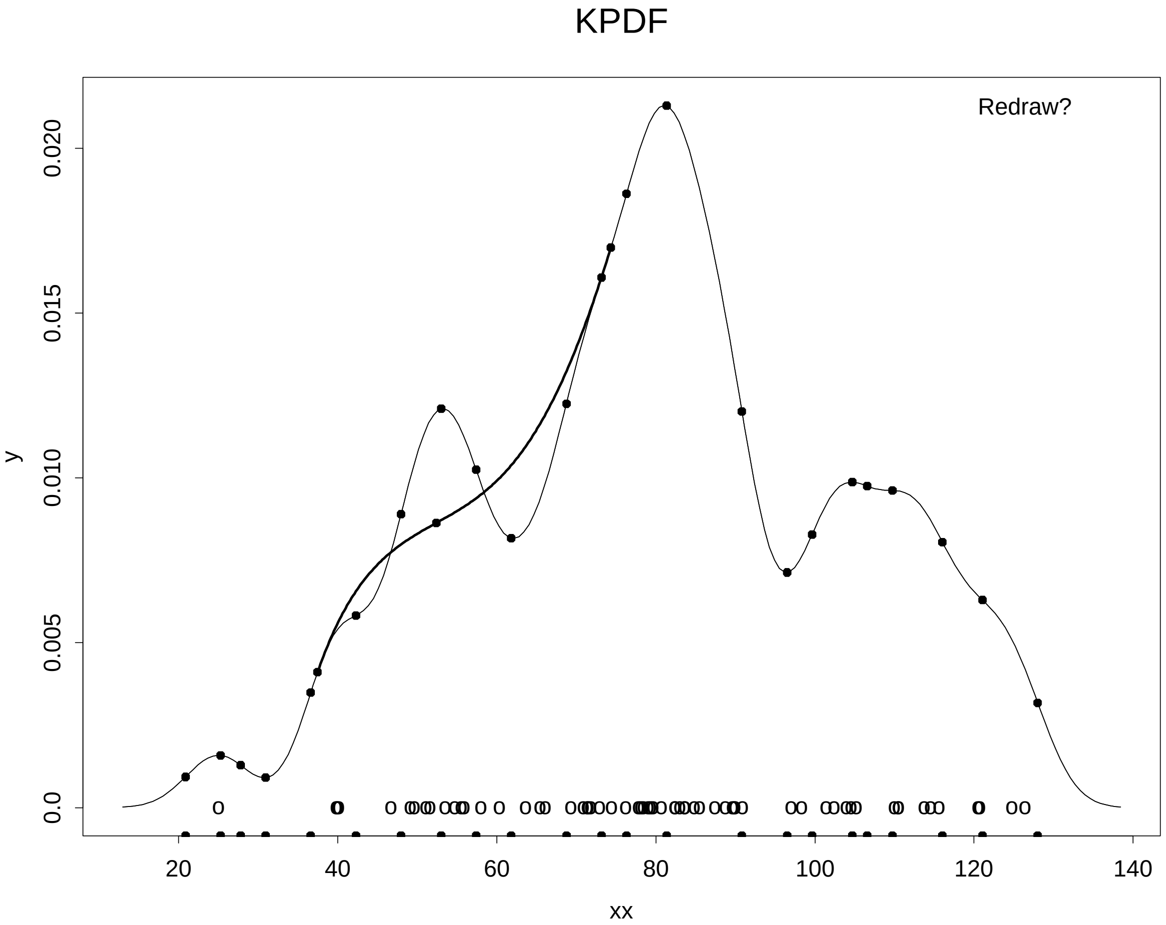

Figure 6: Best patch corresponding to the selection in Figure 5. Here the p-value is approximately 0.14, which suggests that the patch is an adequate representation of the data in that region. In other words, the mode in the kernel estimate is now strongly supported.