ADAPTIVE DENSITY ESTIMATION WITH MASSIVE DATA SETS

David W. Scott, Rice University
Masahiko Sagae, Gifu University
David W. Scott, Dept. of Statistics, MS-138, Rice Univ, Houston, TX 77005-1892

Key Words: Kernel Methods, Method of Moments, Local Bandwidth, Splines, Histograms.

Abstract:

Locally adaptive estimation of densities requires optimization of either a local likelihood function or a global cross-validation criterion. For massive data sets, the promise of having sufficient data to do locally adaptive estimation well ironically runs into the computational limitations of numerical optimization. In this paper, we show how a method of moments algorithm can be used in place of local maximum likelihood to provide computationally and theoretically attractive local estimates. Algorithms and examples will be provided.

1. Challenge of Massive Data

Massive data sets (MDS) represent one of the grand challenge opportunities offered to Computational Statisticians in the next decade. While new statistical models may be required to successfully analyze and understand MDS, standard models and visualization tools will also be required. Some of the key tasks identified include clustering, dimension reduction, prediction, pattern recognition, density estimation, and data visualization. However, one already hears stories of logistic regression programs requiring days to fit on subsets of terabyte data sets. Our favorite exploratory data tools face a different set of challenges with such massive amounts of data. The core problem is the same for both — standard iterative algorithms and exploratory techniques require multiple passes through the data, and some require the data to be in core.

1.1 Reversing Efficiency Roles

What general solution can we propose? It is our position that massive data sets reverse our usual focus on statistical efficiency to a focus on numerical efficiency. Given the massive amounts of data, it is reasonable to assume that the parameter estimates will be virtually noiseless. (However, that is not to say the estimates will be without bias. Model verification and residual diagnostics represent another aspect of the grand challenge. Furthermore, most massive data sets are collected over time, and may not be stationary.) MDS provides the raw information to give us exactly what we want, if we can devise estimation strategies to get us past the numerical difficulties.

We suggest that the general solution will be a reversal of the fortunes of estimation by maximum likelihood versus the method of moments. Except in trivial cases, maximum likelihood always involves iterative solution, sweeping repeatedly through the data to estimate local curvature and gradient information in the parameter space. The method of moments, while not as fully developed, often involves solving equations of much smaller dimension without iteration.

1.2 Density Estimation Setting

In the case of density estimation, the first author has long advocated the use of binning methodology as an excellent computational tool for low-dimensional data of moderate size (what used to be thought of as huge). From the histogram to the frequency polygon to the averaged shifted histogram to the WARP class of estimators, a series of algorithms and theoretical results have been successfully adopted for implementing nonparametric methodology. A key observation is that all of these estimates are derived from bin counts, which should be viewed as zero-order local moments. It is the purpose of this paper to show how the next several higher-order local moments provide the information necessary to construct locally adaptive density estimators that are computationally attractive even with massive data sets.

We believe that an effort similar to that mounted by numerical analysts first faced with parallel com-

This research was supported in part by NSF grant DMS-9626187 and was performed while the second author was a Visiting Scholar at Rice University.

puter architectures should be mounted to revisited how we fit and estimate statistical models with massive data sets. We believe that the principles used in the special case of density estimation can be generalized and applied to other statistical models.

2. Polynomial Histograms

Intuitively, bin frequency counts may be used to determine such basic features of the unknown density as its slope by examining the relative heights of the histogram in adjacent bins, taking into account the amount of noise in those heights. Sagae suggested that that information was available wholly inside the bin itself! Specifically, if the true density is increasing over a bin interval, then there should be more data in the right half of the bin than in the left. Sagae suggested that computing the average of all the points in a bin and comparing that value to the location of the bin center should give a measure of the slope of the unknown density in that bin. Sagae and Scott (1997) provide explicit theoretical results of this phenomena, and extend such results by relating the local curvature of the unknown density to the variance of the points in an individual bin.

2.1 Definition and Construction

The result is a new estimator called the **polynomial histogram**, which is constructed bin by bin in the following manner. The density estimate, $\hat{f}$ is defined locally in each bin as a polynomial of order p that matches the local sample moments in the bin of orders $0, 1, \ldots, p$. Since a polynomial of order p has $p+1$ coefficients, these $p+1$ linear constraints determine the polynomial histogram.

As an example, we have plotted the polynomial histograms of order 0–3 in Figure 1. The 10,000 data come from a normal mixture. The $0th$-order polynomial histogram is of course the ordinary histogram. The first order polynomial histogram is a piecewise linear function that matches the histogram area and the individual bin means. If you look closely, you can see that it is different from the frequency polygon (Scott, 1985), which is also a piecewise linear function, but interpolates the midpoints of the histogram. In particular, the frequency polygon is by construction a continuous function, while the linear polynomial histogram uses its second degree of freedom to match the bin mean.

The second and third order polynomial histograms are also displayed in Figure 1. Again, since each piece of the polynomial histogram is constructed entirely by information within the bin, the global estimate is not smooth. It is also apparent that the third-order polynomial histogram is showing the usual signs of being undersmoothed.

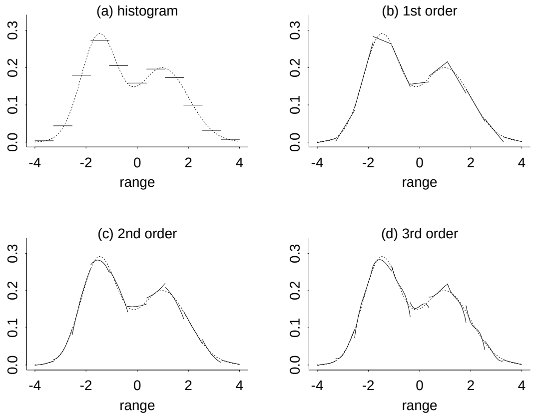

Figure 1: Polynomial histograms of order 0–3.

2.2 Theoretical Properties

Sagae and Scott (1997) prove that if a polynomial histogram of order p is constructed on a bin of width h, then the bias is of order $O(h^{p+1})$. The global bias properties of the polynomial histogram are of the same order. The variance of the polynomial histogram beyond the zeroth-order are not as attractive as the frequency polygon (and higher-order kernel) estimators. In fact, Sagae and Scott show that the global integrated variance is $(p+1)/nh$. While Scott has shown that this can be substantially reduced by allowing the definition of local moments to cover more than one bin, we simply note that such statistical inefficiency is entirely overwhelmed by the massive amount of data available. Thus the computational advantages of the polynomial histogram (one-pass to find the set of local bin moments, followed by solution of a small set of linear equations) far outweigh the increase in variance. Specifically, the global error for a first-order polynomial histogram has the form $c_1 n^{-4/5}$; therefore, the fact that the constant c_1 may be twice as large as for some other method is completely unimportant if $n = 10^{12}$.

2.3 Relation to Kernel Methods

A kernel estimator is given by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

What is the relationship, if any, to higher-order kernel estimators? Recall that a kernel is defined to be of order $p+1$ if the following conditions are true:

$$\begin{aligned} \int K(x)\,dx &= 1 \\ \int x^i K(x)\,dx &= 0, \quad \text{for } i=1,\ldots,p \\ \int x^{p+1} K(x)\,dx &\neq 0, \end{aligned}$$

in which case the bias is of order $O(h^{p+1})$. Many authors have discovered the sample-moment-matching properties of higher-order kernel estimates. For example, if $p > 0$, $\int x\,\hat{f}(x)\,dx = \bar{x}$. If $p > 1$, $\int (x-\bar{x})^2 \hat{f}(x)\,dx = s^2$, and so on.

Notice that the polynomial histogram also satisfies these equations, since it matches the *local* sample moments bin-by-bin, and, therefore, globally as well. In fact, Terrell has shown that sample moments matching is a key ingredient for understanding the properties of a general kernel, K(x,y), estimator.

2.4 Locally Adaptive Estimation

While a theory of constructing adaptive histograms is well-known, it is not commonly used to construct frequency polygons, since the estimates do not integrate to one and are not well-defined at the boundaries with empty bins. Note that (adaptive) frequency polygons are of special interest because they share the better bias properties $O(h^2)$ of positive kernel estimators. However, since the polynomial histogram is defined entirely bin-by-bin, and by construction integrates to one, it does provide a natural extension of histogram binned estimates to higher-order binned estimators. We will show below how to "fix" the visual problem of the discontinuity of the current definition.

Locally adaptive estimation can be obtained by careful application of the well-known least-squares or unbiased cross-validation algorithm. For example, Sain and Scott (1996) demonstrated such an algorithm where the data were grouped into bins for the purpose of sharing the same smoothing parameter, h_k. This may be accomplished with the polynomial histogram by computing the desired order of local moments on a fairly fine mesh, and then searching over adaptive meshes that can be formed by combining the finer mesh intervals. Clearly, local moments for the final adaptive mesh intervals can be computed from the (usually equally-spaced) finer mesh.

3. Adding Continuity

Suppose we wish to construct the polynomial histogram of order p over m bins with a mesh that is not necessarily equally spaced and given by $\{t_0, t_1, \ldots, t_m\}$. We begin by computing the sample moments, $\{\hat{\mu}_{ij},\ i=0,\ldots p,\ j=1,\ldots,m\}$, that is, $p+1$ local moments for each of the m bins. Thus we have $m(p+1)$ linear constraints.

Many years of experience suggests that cubic splines are visually pleasing. Our polynomial histogram is a piecewise polynomial, but lacks the continuity properties of a cubic spline. We can impose these additional constraints upon the polynomial histogram. Specifically, we ask that certain continuity conditions hold at the interior knots:

$$\hat{f}^{(k)}(t_j^-) = \hat{f}^{(k)}(t_j^+), \quad k=0,1,2, \quad j=1,\ldots,m-1.$$

This adds 3(m-1) linear constraints.

Thus for a polynomial histogram of order p satisfying the same smoothness properties as a cubic spline, we have $m(p+1)+3(m-1)$ linear constraints, or $m(p+4)-3$ linear constraints in all. We see that a piecewise polynomial of order $p+3$ is almost completely determined by these constraints, save for a few (3) leftover coefficients in the polynomials. By contrast, a cubic spline has 4 leftover coefficients. These can be defined away with a "natural spline", which constrains the second and third derivatives at the endpoints to vanish. Here, we have just 3 leftover coefficients, and no obvious similar "natural" conditions present themselves.

We decided to specify the extra coefficients by choosing the entire set to satisfy the constraints while minimizing the roughness functional

$$\int_{t_0}^{t_m} \hat{f}''(t)^2 dt.$$

We have implemented this **smoothed polynomial histogram** for the particular value $p=1$, which is local mean-matching. This requires only an order-4 piecewise polynomial, but our implementation adds an additional term to each polynomial for a total of $6m$ coefficients with $5m-3$ linear constraints and $m+3$ coefficients given to roughness smoothing. The entire set of coefficients can be solved simultaneously by a linear set of equations.

4. Is Local Adaptivity Needed?

4.1 Mixture Example (I)

We illustrate the algorithm with a modest data set of size 10,000 from the mixture density $\frac{3}{4}N(0,1)+$

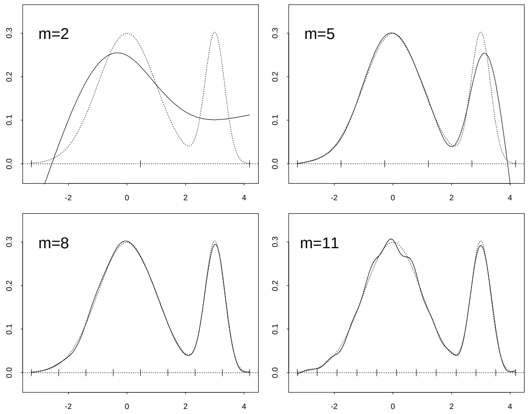

Figure 2: Smoothed PH's for several m's.

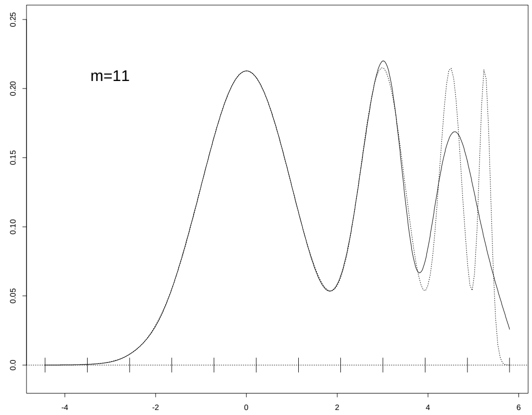

Figure 3: Smoothed polynomial hist for $m = 11$.

$\frac{1}{4}N(3, \frac{1}{3})$. In Figure 2, we show four smoothed polynomial histograms over the sample range with $m = 2, 5, 8, 11$ equally spaced bins. Clearly, the role of the smoothing parameter is played by the number of bins parameter, m. The sample information when $m = 2$ is not sufficient to reveal the simple structure, but 5 bins is sufficient. Eight bins gives a remarkably good estimate. Eleven bins shows the usual signs of a fixed bandwidth estimator, namely, when the smoothing is just right for the narrower mode, the smoothing is too rough for the wider mode.

4.2 Mixture Example (II)

Our second example is a mixture of four normals:

$$\frac{8}{15}N(0,1) + \frac{4}{15}N(3,0.5) + \frac{2}{15}N(4.5,0.25) + \frac{1}{15}N(5.25,0.125).$$

A large sample of size $n = 1{,}500{,}000$ was drawn, and the smoothed polynomial histogram computed with $m = 11$ and is displayed in Figure 3. Clearly, this oversmooths the high frequency components, and suggests the need for a locally adaptive estimate.

Next, we computed the smoothed polynomial histogram on a fixed mesh with $m = 160$ bins. This choice is about optimal for the narrowest component. The result is displayed in Figure 4. This estimate is excellent except for the widest component.

Finally, we computed the smoothed polynomial histogram on a nonregular mesh with $m = 23$ bins. The mesh was chosen so that each component had approximately six bins; see Figure 5. This estimate is excellent for all four components.

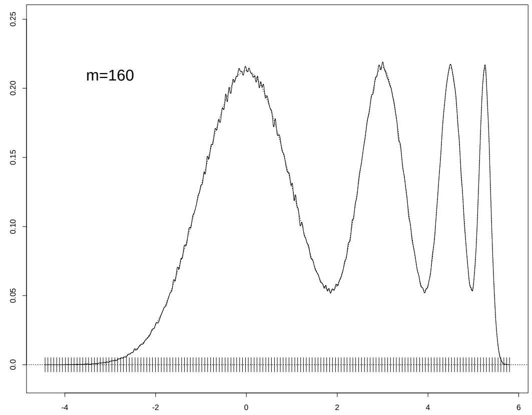

Figure 4: Smoothed polynomial hist for $m = 160$.

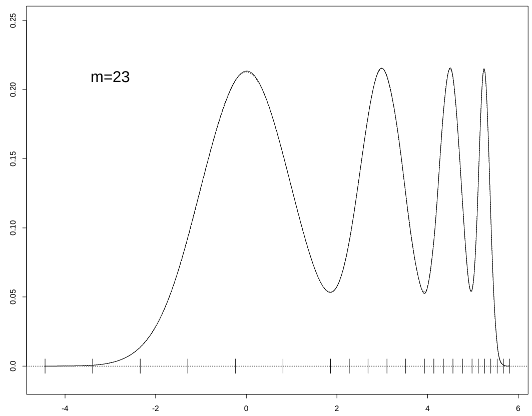

Figure 5: Smoothed polynomial hist for $m = 23$ with adaptive mesh.

5. Discussion

We now can attempt to answer the question: "Is Local Adapting Needed for MDS?" The answer may actually be, "No", if the sample size is large enough, and the dimension is not more than two or three. For sufficiently large sample sizes, while a ten-fold decrease in L_2 error may be possible with an adaptive mesh, the absolute magnitude of the L_2 error may already be so small that the extra reduction may not be even visually detected.

The actual implementation of least-squares cross-validation that would produce something as in Figure 5 has not yet been completed. Clearly, we would hope to use the binned data and to avoid any degenerate bin widths. This could be accomplished by using integer programming, since the set of bin widths must be a multiple of the fine mesh chosen initially.

6. References

Bowman, A.W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, **71**, 353-360.

Breiman, L., Meisel, W., and Purcell, E. (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, **19**, 135-144.

Fan, J., Hall, P., Martin, M., and Patil, P. (1993), "On Local Smoothing of Nonparametric Curve Estimators," Technical Report No. 440, Department of Statistics, Stanford University.

Hall, P., Hu, T.C., and Marron, J.S. (1994), "Improved Variable Window Kernel Estimates of Probability Densities," *Annals of Statistics*, To appear.

Hall, P. and Marron, J.S. (1988), "Variable Window Width Kernel Estimates," *Probability Theory and Related Fields*, **80**, 37-49.

Hall, P. and Schucany, W.R. (1989), "A Local Cross-Validation Algorithm," *Statistics & Probability Letters*, **8**, 109-117.

Hazelton, M. (1995), "Bandwidth Selection for Local Density Estimation," *Scandinavian Journal of Statistics*, To appear.

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

Jones, M.C. (1990), "Variable Kernel Density Estimates and Variable Kernel Density Estimates," *Australian Journal of Statistics*, **32**, 361-371.

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

Mielniczuk, J., Sarda, P., and Vieu, P. (1989), "Local Data-Driven Bandwidth Choice for Density Estimation," *Journal of Statistical Planning and Inference*, **23**, 53-69.

Rudemo, M. (1982), "Empirical Choice of Histogram and Kernel Density Estimators," *Scandinavian Journal of Statistics*, **9**, 65-78.

Sagae, M. and Scott, D.W. (1997). "Bin Interval Method of Locally Adaptive Nonparametric Density Estimation," submitted.

Sain, S.R. (1994), "Adaptive Kernel Density Estimation," Unpublished Ph.D. Dissertation, Department of Statistics, Rice University.

Sain, S.R. and Scott, D.W. (1995), "On Locally Adaptive Density Estimation," *Journal of the American Statistical Association*, **91**, 1525-1534.

Scott, D.W. (1985a). "Frequency Polygons." *J. Amer. Statist. Assoc.*, **80**, 348-354.

Scott, D.W. (1985b). "Average Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions." *Ann. Statist.*, **13**, 1024-1040.

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

Scott, D.W. and Sheather, S.J. (1985). "Kernel Density Estimation With Binned Data." *Comm. Statist.*, **14**, 1353-1359.

Scott, D.W. and Terrell, G.R. (1987), "Biased and Unbiased Cross-Validation in Density Estimation," *Journal of the American Statistical Association*, **82**, 1131-1146.

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

Terrell, G.R. (1984). "Efficiency of Nonparametric Density Estimators." Technical Report, Department of Statistics, Rice University.

Terrell, G.R. and Scott, D.W. (1992), "Variable Kernel Density Estimation," *The Annals of Statistics*, **20**, 1236-1265.

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