Nonparametric Function Estimation
Chapters 4–5 Stat 550[1]

David W Scott[2]

Rice University

October 7
Fall 2023
Rice University

[1] A course based upon the 2nd edition of *Multivariate Density Estimation; Theory, Practice, and Visualization*, *John Wiley & Sons, 2015*

[2] www.stat.rice.edu/∼scottdw/

Chapter IV: Frequency Polygons

- The discontinuities in the histogram limit its usefulness as a graphical tool for multivariate data.
- The **frequency polygon** (FP) is a continuous density estimator based on the histogram, with linear interpolation.
- However, once again Fisher (1932) did not like the frequency polygon

  > *The advantage is illusory, for not only is the form of the curve thus indicated somewhat misleading, but the utmost care should always be taken to distinguish the infinitely large hypothetical population from which our sample of observations is drawn, from the actual sample of observations which we possess; the conception of a continuous frequency curve is applicable only to the former, and in illustrating the latter no attempt should be made to slur over this distinction.*

Frequency Polygon Thoughts

- Fisher could not know there was a theoretical reason to prefer the FP to the histogram
- In one dimension, the frequency polygon is the linear interpolant of the mid-points of an equally spaced histogram.
- The frequency polygon extends beyond the histogram into an empty bin on each end.
- The frequency polygon is easily verified to be a *bona fide* density function, that is, nonnegative with integral equal to 1.
- Note that the FP curve actually has slope, as opposed to the jumps of a histogram.

MISE Analysis of the Frequency Polygon

- We focus on a typical pair of adjacent histogram bins:

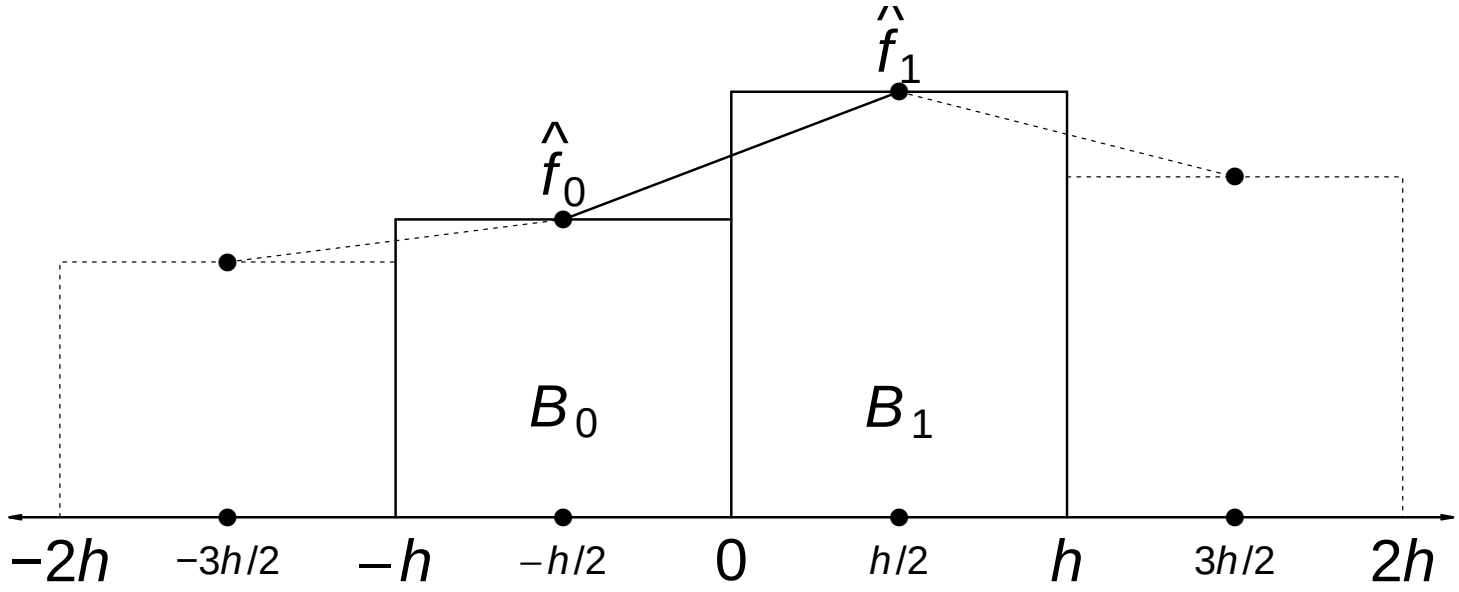

Figure: The frequency polygon in a typical bin, $(-h/2, h/2)$, which is derived from two adjacent histogram bins B_0 and B_1.

- The FP is defined by the formula

$$\hat{f}(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)\hat{f}_1, \qquad -\frac{h}{2} \le x < \frac{h}{2}$$

MISE Analysis of the FP (continued)

- Proceeding as with the histogram, we find

$$AMISE(h) = \frac{2}{3nh} + \frac{49}{2{,}880} h^4 R(f'');\quad \text{hence,}$$

$$h^* = 2[15/(49R(f''))]^{1/5} n^{-1/5}$$

$$AMISE^* = (5/12)[49R(f'')/15]^{1/5} n^{-4/5}$$

- Note the differences : h^4, $R(f'')$, $n^{-1/5}$, and $n^{-4/5}$.
- Recall the $AMISE(h)$ for the histogram:

$$AMISE(h) = \frac{1}{nh} + \frac{1}{12} h^2 R(f')$$

- For example, with 800 normal data points, the optimal bin width for the FP is 50% wider than the corresponding histogram bin width.

Tentative Conclusions from the FP *MISE* Analysis

- The FP can better approximate regions in the pdf where the derivative is large.
- The FP has problems where the second derivative is large (peaks and tails)
- Surprisingly, the improvement is not just to the constant in the *AMISE*, but a whole order due to better approximation in the squared bias (h^4 versus h^2)
- Unlike the histogram, a known discontinuity (e.g. negative exponential at $x = 0$) cannot be "fixed" because of interpolation across bins.

Sensitivity of FP to bin width $h = ch^*$

- Again, we may use the $AMISE(h)$ to see how close to h^* we want to be. For the FP

$$\frac{\text{AMISE}(ch^*)}{\text{AMISE}(h^*)} = \frac{c^5 + 4}{5c}$$

Table: Sensitivity of AMISE to Error in Bin Width Choice $h = ch^*$

c	Histogram $(c^3 + 2)/(3c)$	FP $(c^5 + 4)/(5c)$	Higher Order $(c^9 + 8)/(9c)$
1/2	1.42	1.61	1.78
3/4	1.08	1.13	1.20
1	1	1	1
4/3	1.09	1.23	1.78
2	1.67	3.60	28.89

AMISE for Histogram versus FP for Normal Data

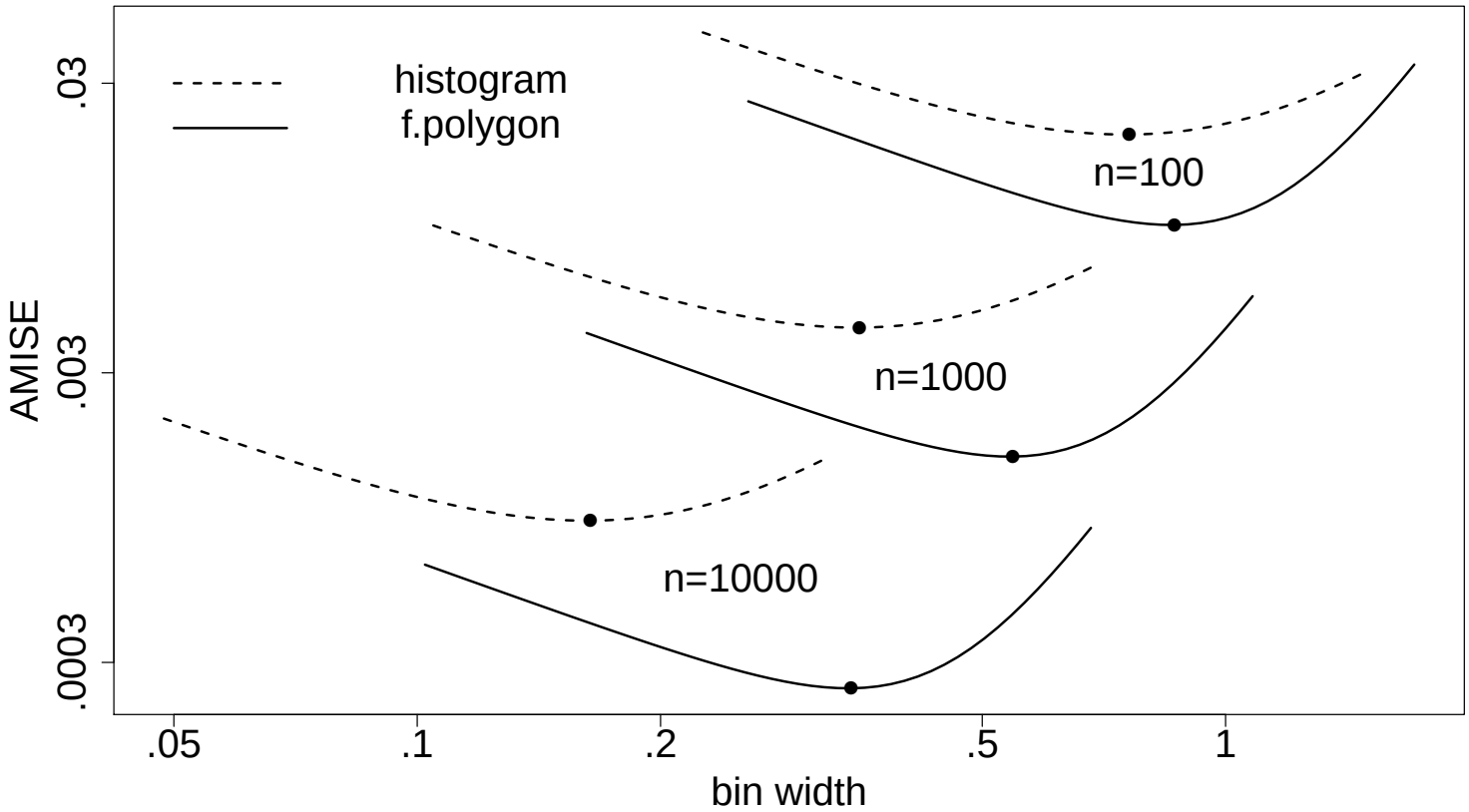

Figure: AMISE for histogram and frequency polygon for standard normal density.

Equivalent Sample Sizes, Updated

- Compare 4 estimators of $N(0,1)$ data

  Table: Sample Sizes Required for $N(0,1)$ Data So That AMISE$^* \approx 1/400$ and $1/4000$

Estimator	Equivalent Sample Sizes	
$N(\bar{x}, 1)$	57	571
$N(\bar{x}, s^2)$	100	1,000
Optimal FP	546	9,866
Optimal histogram	2,297	72,634

- Much improved!

Normal Reference Rule for the FP

- For the normal density $N(\mu, \sigma^2)$,

$$R(\phi'') = 3/(8\sqrt{\pi}\sigma^5)$$

- Plugging into the expressions for h^* and $AMISE(h^*)$,

$$h^* = 2.15\,\sigma\,n^{-1/5}$$

$$AMISE(h^*) = \frac{0.3870}{\sigma}\,n^{-4/5}$$

- Thus the data-based normal reference rule for the FP is

$$\boxed{h^* = 2.15\,\hat{\sigma}\,n^{-1/5}}$$

- If you look back at the 4 histograms of a million normal data point, the last one displayed $h = 4h^*$. This turns out to the histogram from which the optimal FP should be constructed.

Some More Practical FP Bin Width Rules

- In addition to the normal reference rule, we can easily implement *BCV* and *UCV*
- *BCV* follows from the estimate of $R(f'')$

$$\hat{R}(f'') = \frac{1}{n^2 h^5} \sum_k (\nu_{k+1} - 2\nu_k + \nu_{k-1})^2 - \frac{6}{nh^5} .$$

which when plugged into the AMISE expression results in

$$\mathrm{BCV}(h) = \frac{271}{480nh} + \frac{49}{2880n^2 h} \sum_k (\nu_{k+1} - 2\nu_k + \nu_{k-1})^2$$

- Since *UCV* uses $\hat{f}_{-i}(x_i)$, there is not such a simple formula

Comparison of $BCV(h)$ for the Histogram and FP

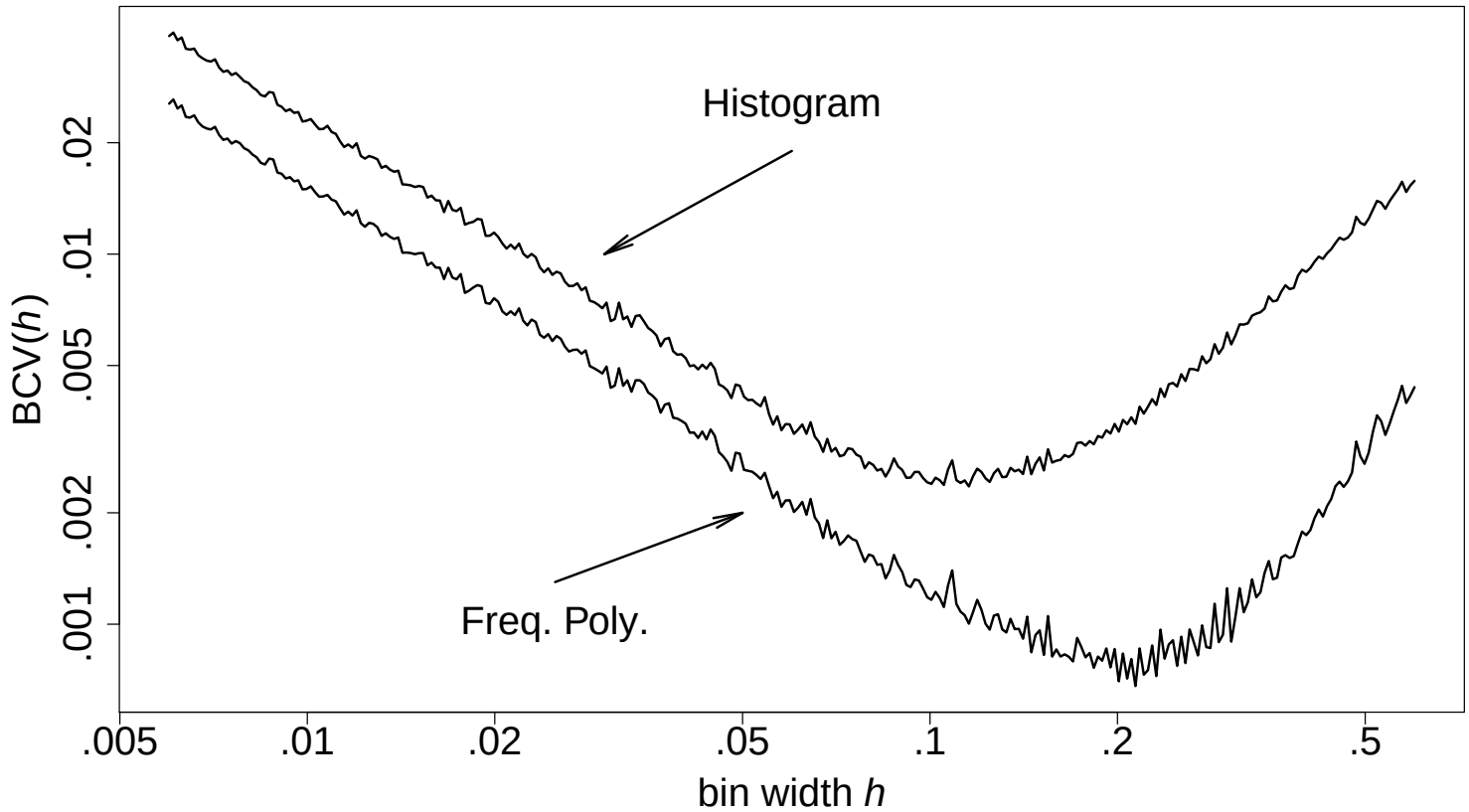

Figure: BCV for histogram and frequency polygon for German income data.

Oversmoothed Bin Widths for the Frequency Polygon

- The criterion is to minimize $R(f'')$ subject to various scale constraints.
- For example, among densities supported on the interval $[-0.5, 0.5]$, the smoothest density is

$$f_3(x) = \frac{15}{8}(1 - 4x^2)^2 \; I_{[-0.5,0.5]}(x)$$

- On a general interval $[a, b]$, this leads to the inequality

$$R(f'') \geq \frac{720}{(b-a)^5}$$

- Applying this inequality to h^* and re-arranging leads to

$$\boxed{\text{number of bins} = \frac{b-a}{h^*} \geq \left(\frac{147}{2}n\right)^{1/5}}$$

Oversmoothed Bin Widths (continued)

- As an example, with the large LRL dataset of 25,752 points, the optimal FP requires at least 18 bins, while the optimal histogram requires at least 37 bins.
- Next, among all densities with variance σ^2, the smoothest density is

$$f_4(x) = \frac{35}{96\sigma}\left(1 - \frac{x^2}{9\sigma^2}\right)^3 I_{[-3\sigma,3\sigma]}(x) \quad \text{so that}$$

$$R(f'') \geq \frac{35}{243\sigma^5}$$

- Substituting this inequality into the expression for h^* gives

$$\boxed{h \leq \left(\frac{23{,}328}{343}\right)^{1/5} \sigma\, n^{-1/5} = 2.33\,\sigma\, n^{-1/5} \equiv h_{\mathrm{OS}}}$$

- Recall that the constant to 2.15 for $N(\mu, \sigma^2)$ data.

Optimally Adaptive Meshes for the FP

- Using the following expression for the pointwise *MSE*

$$AMSE(x) = \frac{2f(x)}{3nh} + \frac{49}{2{,}880}h^4 f''(x)^2$$

- We have the results

$$h^*(x) = 2[15f(x)/49f''(x)^2]^{1/5} n^{-1/5}$$

$$AMSE^*(x) = (5/12)[49/15]^{1/5}[f''(x)^2 f(x)^4]^{1/5} n^{-4/5}$$

$$AAMISE^* = (5/12)[49/15]^{1/5} \left\{ \int [f''(x)^2 f(x)^4]^{1/5} dx \right\} n^{-1/5}$$

- Local adapting is always better by Jensen's inequality

$$AAMISE^* \leq AMISE^* \qquad \text{since}$$

$$\int \left[f''(x)^2 f(x)^4\right]^{1/5} dx \leq \left[\int f''(x)^2 dx\right]^{1/5}$$

Optimally Adaptive Meshes for the FP

- Jensen's inequality here is

$$\mathrm{E}\left[\frac{f''(x)^2}{f(x)}\right]^{1/5} \leq \left[\mathrm{E}\frac{f''(x)^2}{f(x)}\right]^{1/5}$$

- Thus, asymptotically, the *MISE* of an adaptive FP is only 91.5% and 76.7% of the *MISE* of a fixed-bin-width FP for normal and Cauchy data, respectively.

Optimal Adaptive Histogram for the FP

- Interesting, even though the FP interpolating non-equally-spaced bins does not generally integrate to 1.

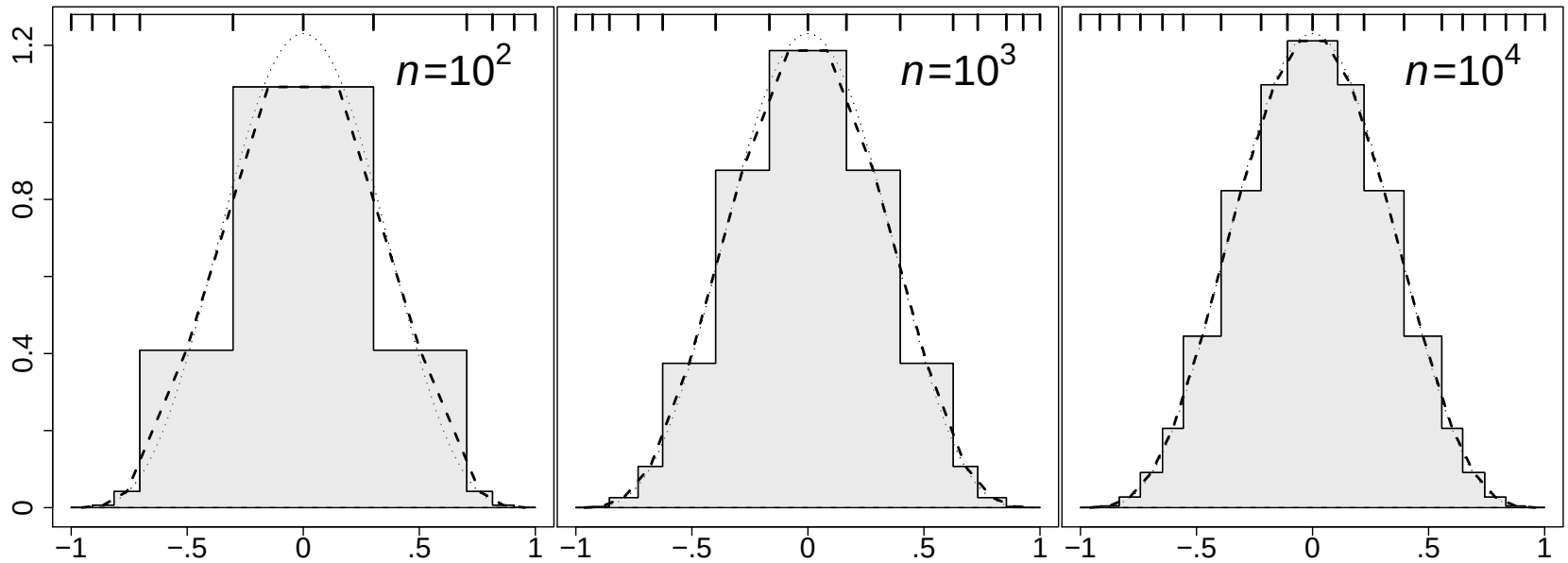

Figure: Optimal adaptive frequency polygon meshes for the scaled Beta(5,5) density. The histogram is drawn from which the FP (dotted line) is derived. The tick marks for the adaptive mesh are shown above each figure.

Modes and Bumps in a Frequency Polygon

- As we saw in the last chapter, an optimal histogram does not zero in on the true modal bin
- But for a FP applied to a density with mode at $x = 0$, write

$$h^* = cn^{-1/5} \quad \text{and define} \quad \beta \equiv -\frac{1}{2} c^{5/2} \frac{f''(0)}{\sqrt{f(0)}}$$

- Then

$$\lim_{n\to\infty} \Pr\left(\nu_0 = \arg\max_{|j|\le k} \nu_j\right) = \int_y \prod_{\substack{j=-k \\ j\neq 0}}^{k} \Phi(y + j^2\beta)\, \phi(y)\, dy$$

- Note that β is a measure of the "size" of the mode; a large value of β means the mode is more prominent

Where is the Sample Mode for a Frequency Polygon?

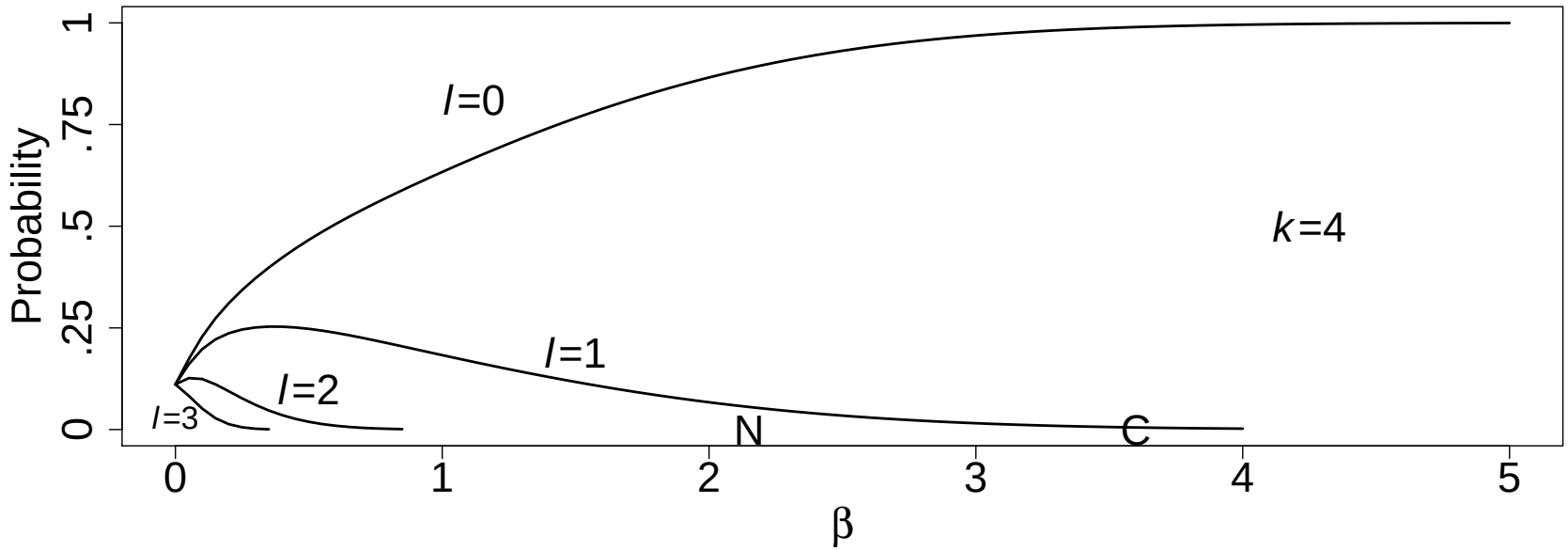

Figure: Probability distribution of the location of the sample mode as a function of β for the choice $k = 4$. The values of β for normal and Cauchy densities are 2.15 and 3.6, respectively, and are indicated by the letters N and C.

- Can use this diagram to obtain a rough confidence interval for the location of the true mode. Estimate β and compute the probability the mode might be off by a bin or two.

Multivariate Frequency Polygons

- How to interpolate a bivariate histogram? Mesh of triangles or use the linear blend.

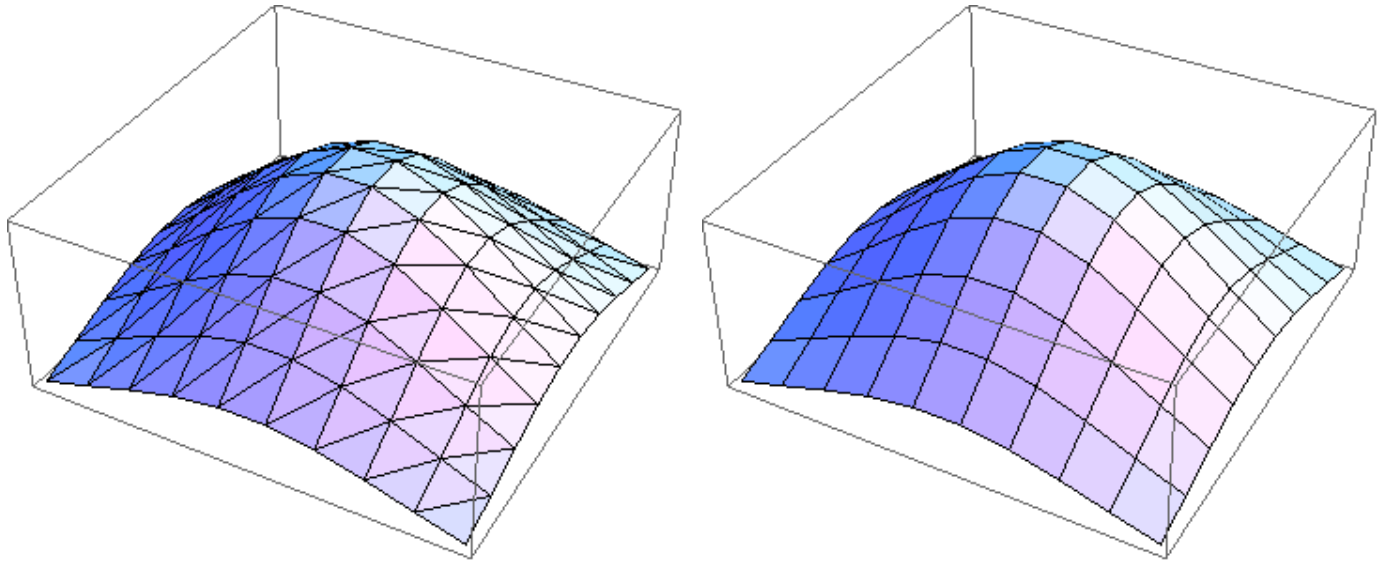

Figure: Construction of a bivariate frequency polygon using triangular meshes (left) and linear blend elements(right).

- The linear blend uses the four adjacenct values and the formula (linear parallel to axes; quadratic along diagonal)

$$f(x, y) = a + bx + cy + dxy$$

Asymptotics of the Linear Blend Frequency Polygon (LBFP)

- Hjort showed the error $AMISE(\mathbf{h})$ takes the form

$$\frac{2^d}{3^d n h_1 \cdots h_d} + \frac{49}{2{,}880} \sum_{i=1}^{d} h_i^4 R(f_{ii}) + \frac{1}{32} \sum_{i<j} h_i^2 h_j^2 R\left(\sqrt{f_{ii} f_{jj}}\right)$$

- This leads to the general optimal results

$$h_i^* = O(n^{-1/(4+d)}) \quad \text{and} \quad \text{AMISE}^* = O(n^{-4/(4+d)})$$

- The FP is also better than the Histogram in dimensions $d > 1$.

Rates in Higher Dimensions

d	Histogram	Frequency Polygon
1	$n^{-2/3}$	$n^{-4/5}$
2	$n^{-2/4}$	$n^{-4/6}$
3	$n^{-2/5}$	$n^{-4/7}$
4	$n^{-2/6}$	$n^{-4/8}$
5	$n^{-2/7}$	$n^{-4/9}$
6	$n^{-2/8}$	$n^{-4/10}$
7	$n^{-2/9}$	$n^{-4/11}$
8	$n^{-2/10}$	$n^{-4/12}$

Table: Asymptotic order of MISE for multivariate histogram and frequency polygon density estimators. The rate of convergence decreases as the dimension, d, increases. The arrows indicate identical rates for the histogram and frequency polygon.

Rates in Higher Dimensions: Conclusions

- This is very encouraging. Bivariate and trivariate histograms are widely used. This table suggests using a smoother density estimator should carry forward up to 5-6 dimensions!
- For graphical reasons, the first definition of a FP is simpler to work with because the resulting contours are comprised of piecewise polygonal sections. CAD-CAM systems always assume everything is made up of triangular meshes.

Application to Normal Data

- Using the triangular mesh with a bivariate normal data, that the optimal bin widths are approximately equal to

$$h_i^* = 2.105 \left(1 - \frac{107}{208}\rho^2 + \cdots\right) \sigma_i \, n^{-1/6}, \quad i = 1, 2$$

- For multivariate normal data with $\Sigma = I_d$, the optimal smoothing parameters in each dimension are equal with the constant close to 2.
- Normal reference rule proposed by Scott is

Approximate normal FP reference rule : $h_i = 2\,\hat{\sigma}_i \, n^{-1/(4+d)}$

Bin Edge Problems

- The asymptotic theory indicates that the choice of bin origin is asymptotically negligible.
- However, since FP bins are wider, the collection of possible FP's can vary quite a bit.
- We will consider Parzen's Buffalo Snowfall data ($n = 63$) on the next slides
- For certain choices of (h, t_0), the histogram looks unimodal or trimodal. What do you think?

Six Choices of h for the Buffalo Snowfall Data

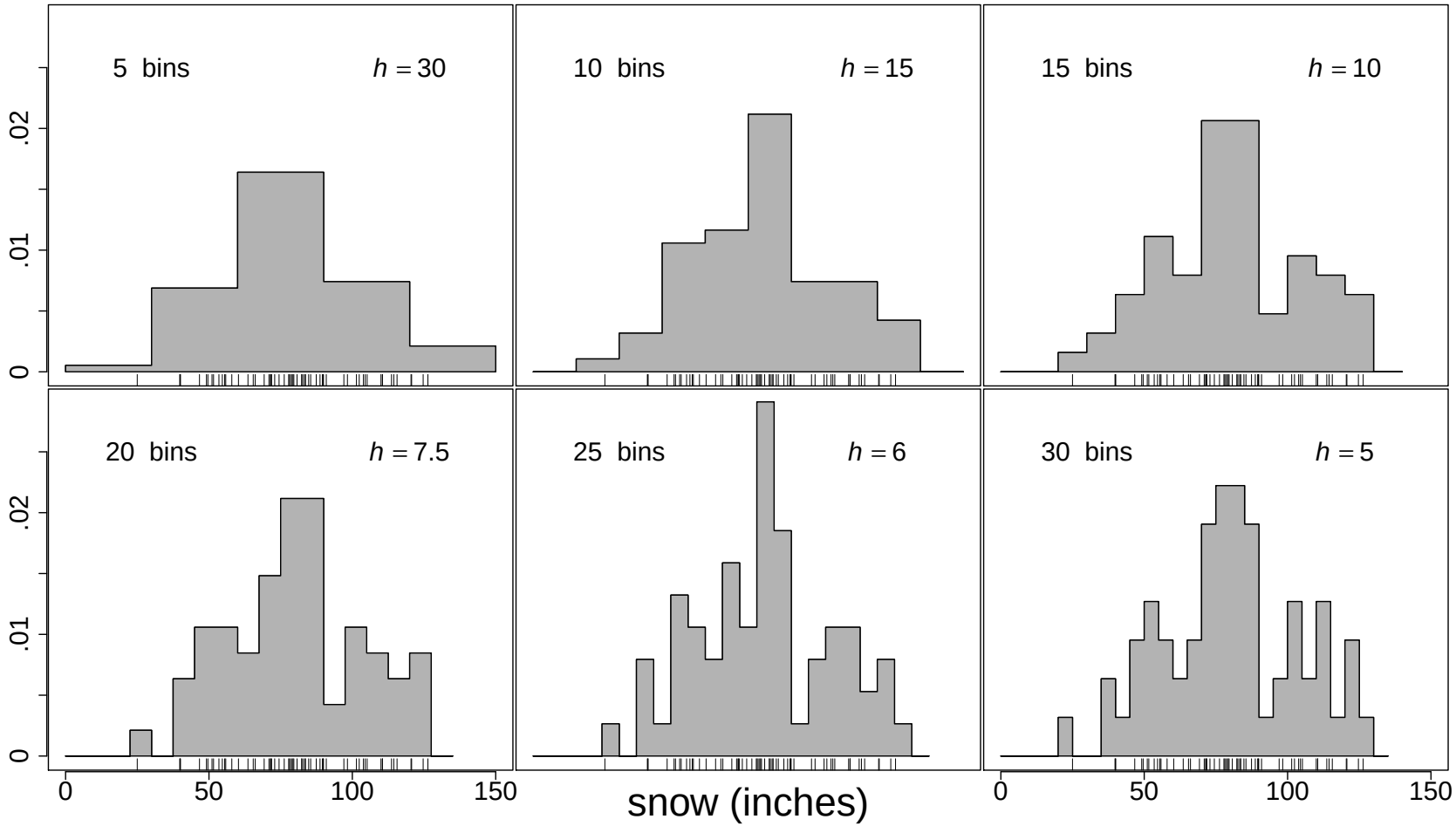

Figure: Histograms of the Buffalo snowfall data with bin origin $t_0 = 0$, and bin widths of 30, 15, 10, 7.5, 6, and 5 inches over the interval $(0, 150)$.

Six Choices of t_0 for the Buffalo Snowfall Data

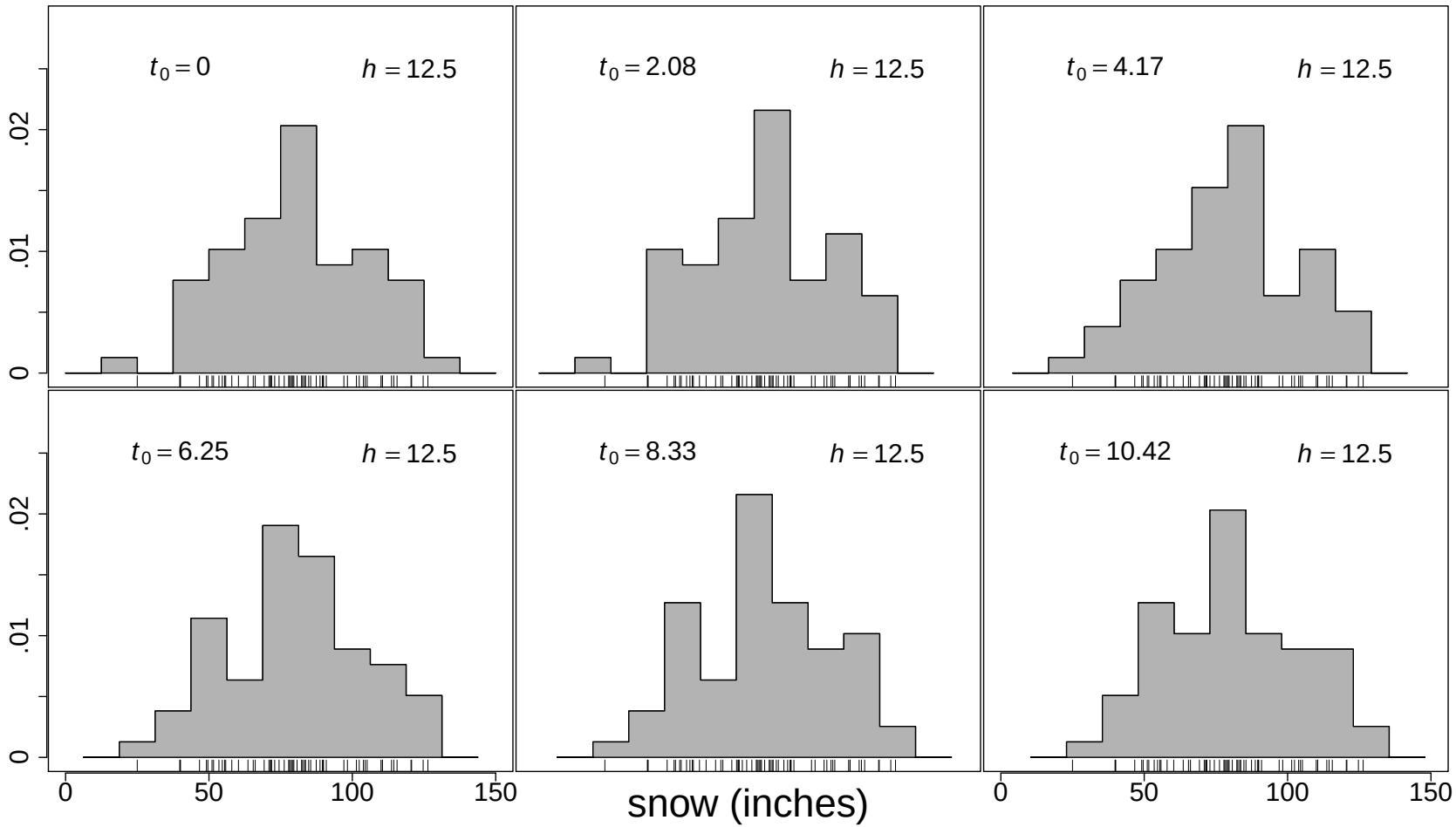

Figure: Six shifted histograms of the Buffalo snowfall data. All have a bin width of 12.5 inches, but different bin origins $t_0 = kh/6,\ k = 0, 1, \ldots, 5$.

Other Modifications of Histograms

- Wand advocates linear binning, where a data point does add 1 to just one bin, but splits it among the adjacent bins. This results is a modest and fixed improvement in *AMISE*.
- Minnotte investigated adjusting the raw histogram counts so that the usual FP preserves the area of the original histogram:

$$\tilde{\nu}_k = \sum_{j=-\infty}^{\infty} c_j \, \nu_{k+j} \,, \text{ where } \; c_j = 2^{\frac{1}{2}} \, (2^{\frac{3}{2}} - 3)^{|j|} = 2^{\frac{1}{2}} (-.1716)^{|j|}$$

- Minnotte followed this up by using higher order splines with these adjusted counts and found the convergence rates improved accordingly.
- In 1-D, linear binning improved *AMISE* by 11% while area-matching achieved 4.4%.
- Scott, Sagae, and Papkov investigated fitting splines that matched the sample moments in all the bins. Again, higher order splines converged at higher rates.

Polynomial histogram

- One of the attractions of a histogram is that you can compute the bin counts on the fly (assuming you can construct a good mesh *a priori*)
- Scott, Sagae, and Papkov investigated the value of compute the sample moments in each bin at the same time.
- For example, if the **sample mean** in a bin is **greater than the midpoint**, then intuitively the density estimate should be **increasing** in that bin.
- Using just the bin count and bin mean, we define the *linear polynomial histogram* (LPH)

$$\hat{f}(x) = \hat{f}_{LPH}(x) = a + bx \quad \text{for } x \in B_0 = (-h/2, h/2)$$

Polynomial histogram (continued)

- The area of the LPH is ah, so the conditional density over B_0 is

$$\hat{f}_{B_0}(x) = \hat{f}(x|x \in B_0) = \frac{a + bx}{ah} \qquad x \in B_0$$

- The LPH should match the area and bin mean, giving us the two constraints

$$\int_{B_0} \hat{f}(x)dx = \frac{\nu_0}{n} \qquad \text{and}$$

$$\int_{B_0} x\,\hat{f}_{B_0}(x)dx = \bar{x}_0 \qquad \text{where} \quad \bar{x}_0 = \frac{1}{\nu_0}\sum_{x_i \in B_0} x_i$$

- The solution to these equations is

$$a = \frac{\nu_0}{nh} \qquad \text{and} \qquad b = \frac{12\,\nu_0\,\bar{x}_0}{nh^3}$$

Linear Polynomial Histogram Theory

- *Suppose f'' is absolutely continuous and $R(f''') < \infty$. Then for the linear polynomial histogram estimator, $\hat{f}_{LPH}(x)$,*

$$\begin{aligned} \text{AMISE}(h) &= \frac{2}{nh} + \frac{1}{720} h^4 R(f''); \text{ hence,} \\ h^* &= 360^{1/5} R(f'')^{-1/5} n^{-1/5} \\ \text{AMISE}^* &= (625/2304)^{1/5} R(f'')^{1/5} n^{-4/5} \end{aligned}$$

- Shares the asymptotics of a FP.
- Now the leading coefficient of $AMISE^*$ is 0.565, which is 45.9% greater than for the frequency polygon.
- However, the optimal bin width is over twice as wide (2.056).
- Advantage over the frequency polygon is that these asymptotics are all appropriate **within the histogram bins**, not across adjacent bins. Thus the boundary problems do not exist if known.

Polynomial Histogram Examples

- We compare the bin-by-bin LPH, a continuous version, and a bin-by-bin quadratic polynomial histogram (QPH) for 3 sample sizes. The QPH incorporates the bin sample variances.

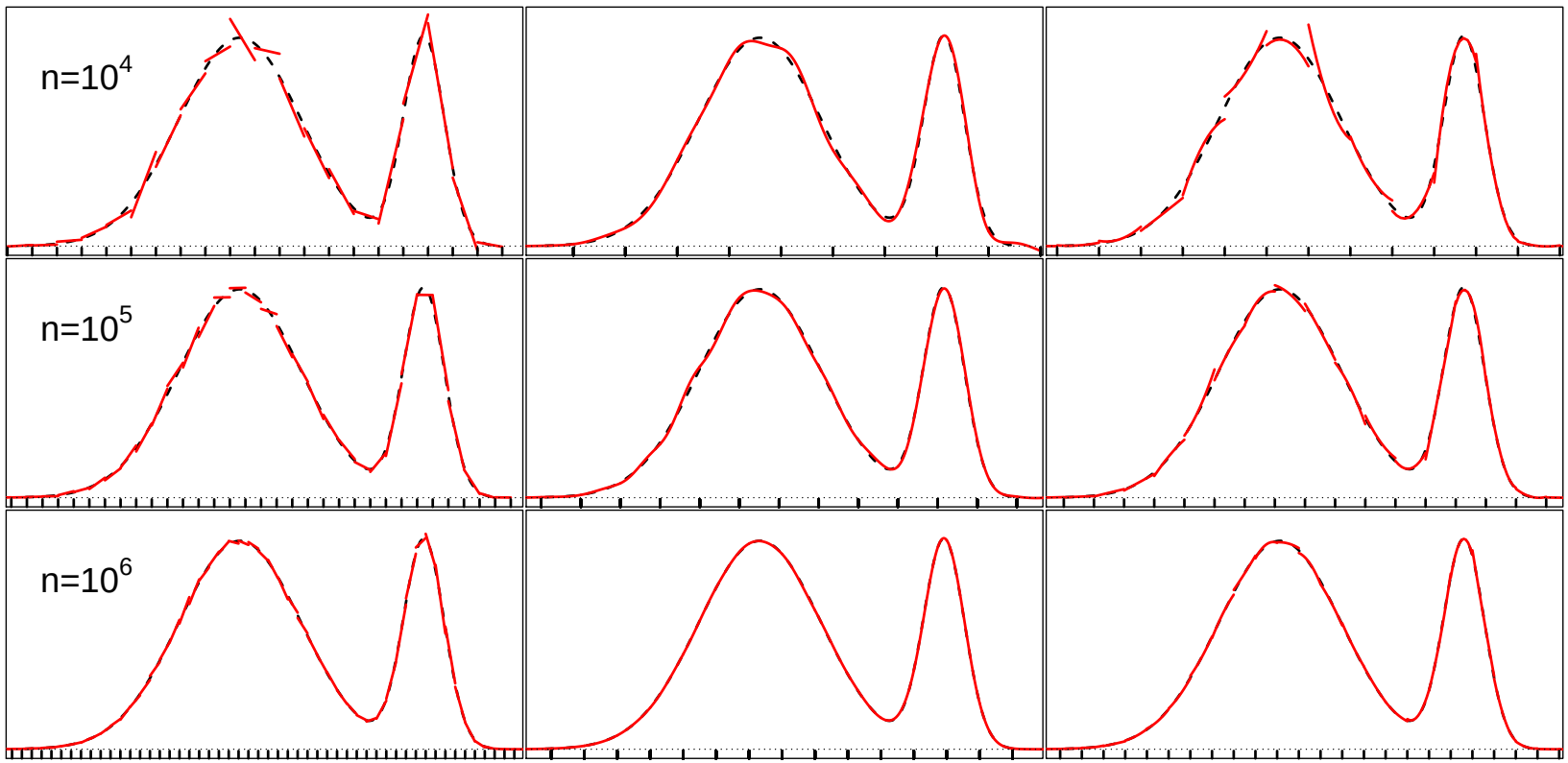

Figure: For 3 sample sizes from a mixture density, examples of the piecewise LPH, the continuous LPH, and the piecewise QPH estimates.

How Much Information Is There In a Few Bins?

- My son and I re-visited some data published in *USA Today* on 10/13/2006.
- The data came from a phone survey of $n = 1{,}207$ subjects.
- Appeared on the front page in the daily graphical feature *USA TODAY Snaphots*®.
- Men were asked how long they were romantically involved before marriage. (Of any interest to any of you?)
- Only four bin counts: 181, 147, 651, and 228.
- Possible parametric models:
  - Normal?
  - Uniform?
  - Negative exponential (with the interesting consequence that being engaged for two years does not change the likelihood of marriage...memoryless process)

How Much Information Is There In a Few Bins?

- We reproduce the original graphic, and two alternatives

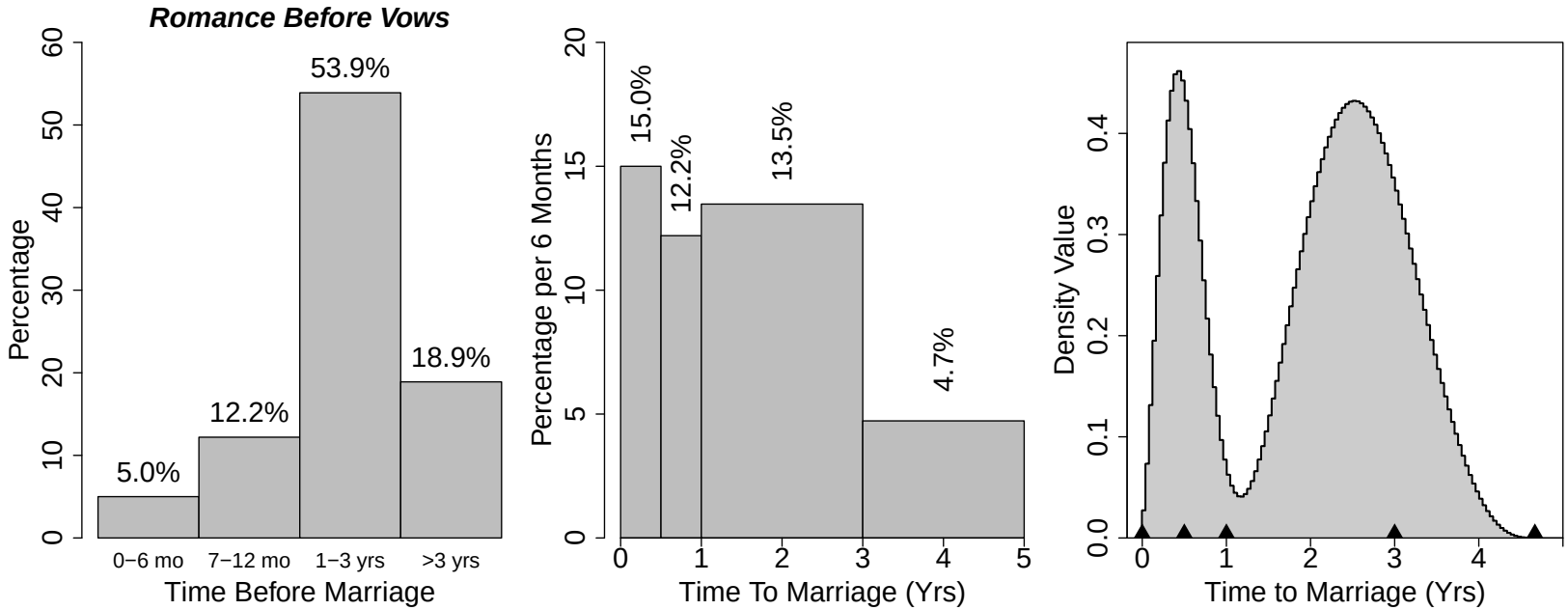

Figure: (Left) Barchart of raw improperly normalized binned marriage survey data. (Middle) Example of a barchart that is properly normalized. (Right) A penalized histogram of the data matching the 4 bin proportions.

- Check totals; width of 4th bin?

Wrapping Up

- Frequency polygons are but one possible modification of the basic histogram.
- The statistical efficiency of the FP is much improved over the histogram, getting ever closer to the magic $O(n^{-1})$ parametric rate of convergence
- The computational efficient of the FP and essentially the same as for the histogram
- The FP suffers from possible boundary effects
- The wider histogram bins used in the construction of the FP can make the choice of the bin origin t_0 more relevant for graphical purposes.
- We examine this question of the choice of t_0 in the next chapter.

Chapter V: Averaged Shifted Histograms

- Scott proposed a simple device for eliminating the bin edge problem of the frequency polygon while retaining many of the computational advantages of a density estimate based on bin counts.
- Rather than choosing among alternative choices of t_0, simply average the shifted frequency polygons.
- Note the average of piecewise linear functions is also piecewise linear (ASFP).
- Almost equivalent is to average shifted histograms and then treat that piecewise constant function as a histogram to linearly interpolated with a FP. Easier to analyze.

Construction of Averaged Shifted Histogram (ASH)

- Consider a collection of m histograms, $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m$, each with bin width h, but with bin origins

$$t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \ldots, \frac{(m-1)h}{m}$$

- The (*naive* or unweighted) averaged shifted histogram is defined as

$$\hat{f}(\cdot) = \hat{f}_{\mathrm{ASH}}(\cdot) = \frac{1}{m} \sum_{i=1}^{m} \hat{f}_i(\cdot)$$

- The ASH will look like a histogram but with bin width h/m

ASH with Buffalo Snowfall Data (How Many Modes?)

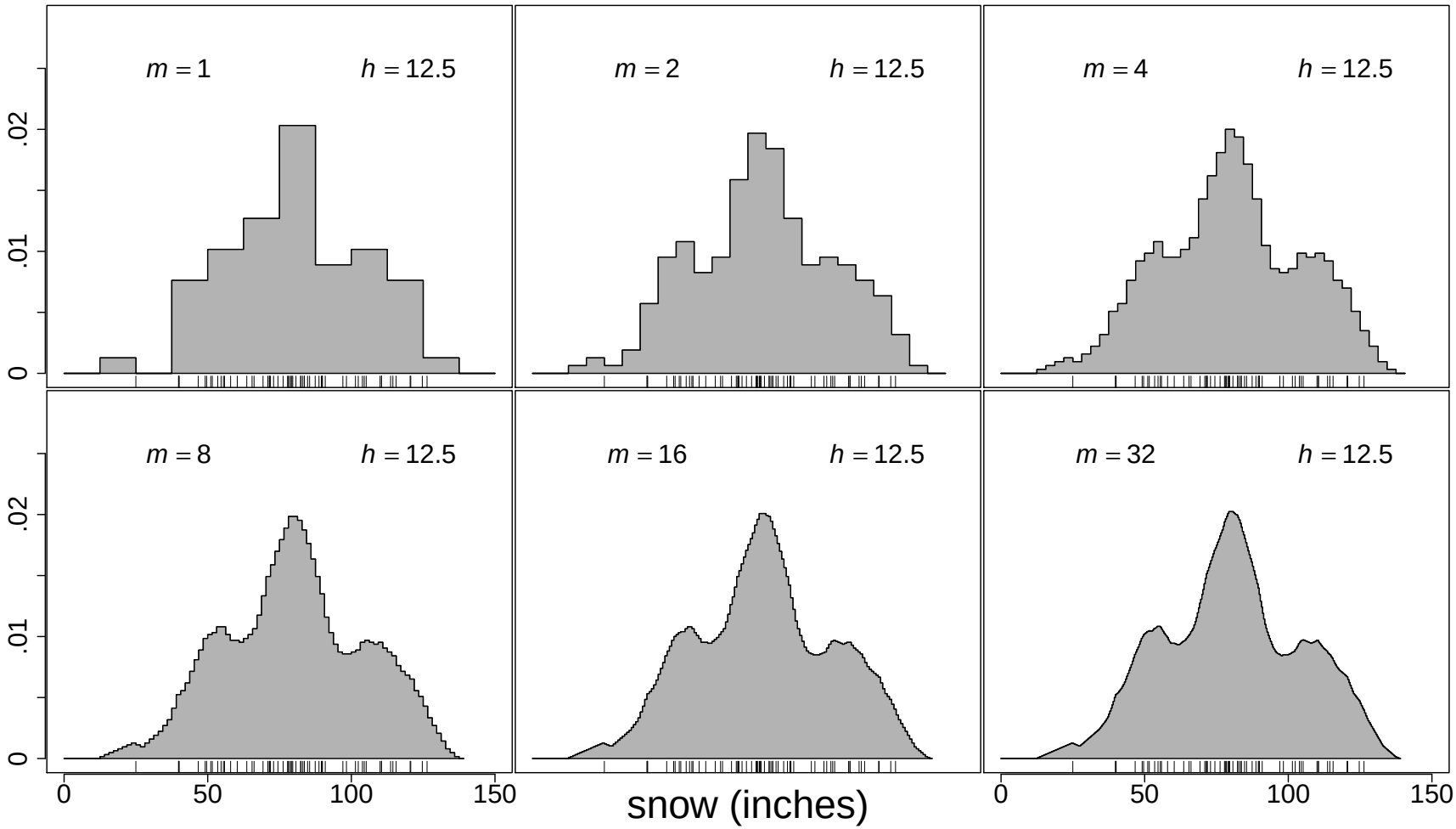

Figure: Naive averaged shifted histograms of the Buffalo snowfall data with bin width $h = 12.5$ inches.

Multivariate Averaged Shifted Histograms

- Multivariate ASHs are constructed by averaging shifted multivariate histograms, each with bins of dimension $h_1 \times h_2 \times \cdots \times h_d$.
- If every possible multivariate histogram is constructed by coordinate shifts that are multiples of $\delta_i \equiv h_i/m_i, i = 1, \ldots, d$, then the multivariate ASH is the average of $m_1 \times m_2 \times \cdots \times m_d$ shifted histograms.
- In the bivariate case, we have

$$\hat{f}_{ASH}(\cdot,\cdot) = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \hat{f}_{ij}(\cdot,\cdot)$$

ASH Examples of the Blood Fat - Heart Disease Data

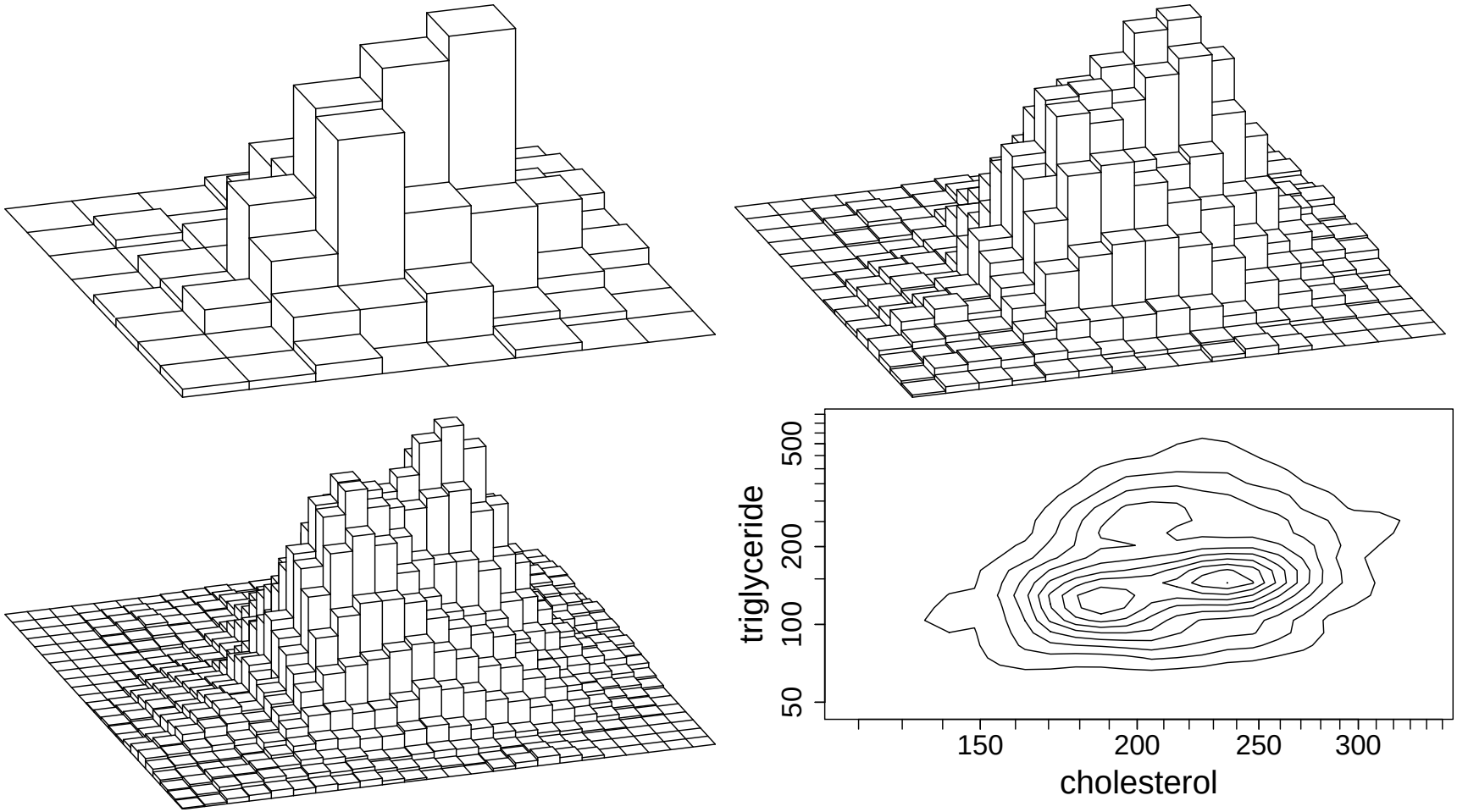

Figure: Bivariate averaged shifted histograms of the lipid dataset for 320 diseased males; $m = 1$, 2, 3, and 3 again for the contour plot

Details of the Univariate ASH

- Let $\delta \equiv h/m$ denote the "apparent" bin width of the ASH
- The univariate ASH is piecewise constant over the intervals $[k\delta, (k+1)\delta)$ where $\delta \equiv h/m$
- It will be convenient to refer to these narrower intervals as the bins B_k and let

$$v_k = \text{bin count in bin } B_k, \quad \text{where } B_k \equiv [k\delta, (k+1)\delta)$$

- To obtain a bin count for the original histogram, we add m adjacent bin counts at the finer resolution.

Details of the Univariate ASH (continued)

- Consider the ASH estimate in bin B_0, which is the average of the m shifted histograms

$$\frac{v_{1-m} + \cdots + v_0}{nh}, \frac{v_{2-m} + \cdots + v_0 + v_1}{nh}, \ldots, \frac{v_0 + \cdots + v_{m-1}}{nh}$$

- Hence, a general expression for the naive ASH is

$$\begin{aligned}\hat{f}(x; m) &= \frac{1}{m} \sum_{i=1-m}^{m-1} \frac{(m - |i|) v_{k+i}}{nh} \\ &= \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) v_{k+i} \quad \text{for } x \in B_k\end{aligned}$$

Details of the Univariate ASH (continued)

- The weights on the bin counts take on the shape of an isosceles triangle with base $(-1, 1)$.
- Other shapes (that are smoother) may be preferred
- The general ASH uses arbitrary weights, $w_m(i)$:

$$\text{General ASH}: \quad \hat{f}(x; m) = \frac{1}{nh} \sum_{|i|<m} w_m(i)\, v_{k+i} \quad \text{for } x \in B_k$$

- A convenient way to define general weights is to sample a bf kernel, which is any probability density defined on $(-1, 1)$:

$$w_m(i) = m \times \frac{K(i/m)}{\sum_{j=1-m}^{m-1} K(j/m)} \qquad i = 1 - m, \ldots, m - 1$$

$$\text{e.g.} \quad K(t) = \frac{15}{16}(1 - t^2)_+^2 = \frac{15}{16}(1 - t^2)^2 I_{[-1,1]}(t)$$

Pseudo-Code for the ASH

BIN1(x, n, a, b, $nbin$) **Algorithm**: (* Bin univariate data *)

$\delta = (b - a)/nbin$
for $k = 1, nbin$ $\{v_k = 0\}$
for $i = 1, n\{$
$k = (x_i - a)/\delta + 1$ (* integer part *)
if $(k \in [1, nbin])$ $v_k = v_k + 1\}$
return $(\{v_k\})$

Pseudo-Code for the ASH

ASH1 $(m, v, nbin, a, b, n, w_m)$ **Algorithm**: (* Univariate ASH *)

$\delta = (b - a)/nbin$
$h = m\delta$
for $k = 1, nbin\{f_k = 0\}$
for $k = 1, nbin\{$
    if$(v_k = 0)$ next k
    for $i = \max(1, k - m + 1), \min(nbin, k + m - 1)\{$
        $f_i = f_i + v_k w_m(i - k)\}\}$
for $k = 1, nbin\{f_k = f_k/(nh); t_k = a + (k - 0.5)\delta\}$
return $(\mathbf{x} = \{t_k\}, \mathbf{y} = \{f_k\})$ (* Bin centers and ASH heights *)

These functions and more are available on my web site and in the R ash library.

ASH Examples with Biweight Kernel Weights

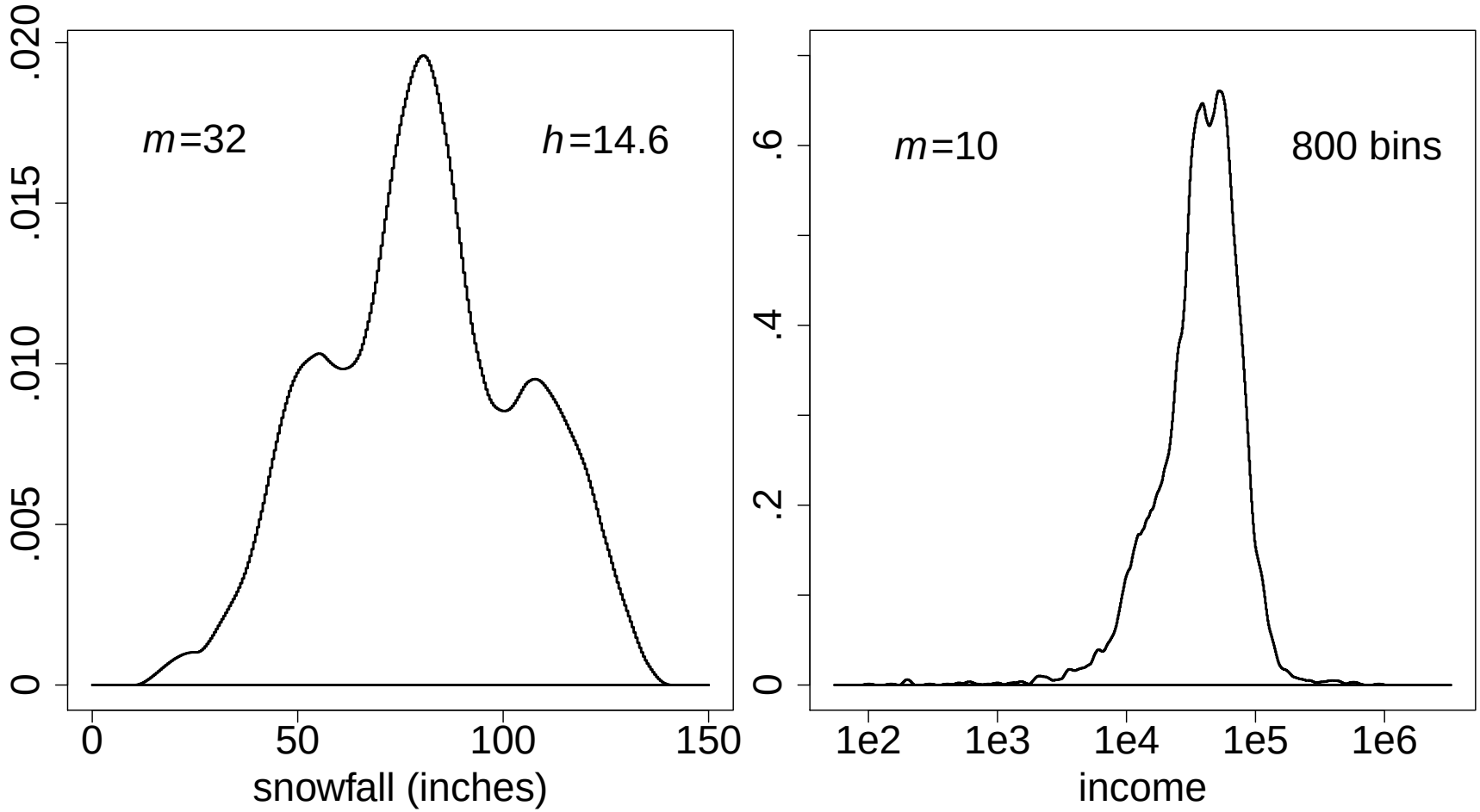

Figure: Examples of ASH with biweight kernel applied to the Buffalo snowfall and German household income data sets.

Some Brief Asymptotics for the ASH

- For the naive ASH (isosceles triangle kernel), we have

$$AMISE = \frac{2}{3nh}\left(1 + \frac{1}{2m^2}\right) + \frac{h^2}{12m^2}R(f') + \frac{h^4}{144}\left(1 - \frac{2}{m^2} + \frac{3}{5m^4}\right)R(f'').$$

- Plugging in $m = 1$ gives the histogram result precisely.
- Plugging in $m = \infty$ completely remove the histogram-like bias based on $R(f')$
- However, any $m > 5$ or 10 essentially does the same.

$$h^*_{m=\infty} = \left[24/(nR(f''))\right]^{1/5} \quad \left(= 2.576\,\sigma\, n^{-1/5} \text{ for } N(\mu, \sigma^2)\right)$$

Equivalent sample sizes with ASH

- Equivalent Sample Sizes Required for AMISE $\approx 1/400$ for $N(0,1)$ Data:

Estimator	$N(\bar{x}, s^2)$	ASH	FP-ASH	FP	Histogram
Sample Size	100	436	436	546	2,297

- The AMISE for the FP-ASH (naive version) has no histogram-like bias:

$$AMISE(h) = \frac{2}{3nh} + \frac{h^4}{144}\left(1 + \frac{1}{m^2} + \frac{9}{20m^4}\right) R(f'')$$

- Inserting $m = 1$ gives the ordinary frequency polygon formula.
- The AMISE of the linear blend of the naive ASH equals

$$\frac{2^d}{3^d n\ h_1 \cdots h_d} + \frac{1}{720}\sum_{i=1}^{d} \delta_i^4 R(f_{ii}) + \frac{1}{144}\int_{\Re^d}\left[\sum_{i=1}^{d} h_i^2\left(1 + \frac{1}{2m_i^2}\right) f_{ii}\right]^2$$

The Limiting ASH as a Kernel Estimator

- The parameter m in the ASH is a nuisance parameter, but much less so than the bin origin.
- We study the limiting behavior of the ASH as $m \to \infty$
- Suppose h and n fixed and m increasing
- Isolate the effect of a single data point $x_j \in B_{k+i}$ on the ASH estimate $\hat{f}(x)$, at a fixed point $x \in B_i$

$$1 - \frac{|i|}{m} = 1 - \frac{|i| \cdot \delta}{m \cdot \delta} = 1 - \frac{|x - x_j|}{h} + O\left(\frac{\delta}{h}\right), \quad \text{if } |x - x_j| < h$$

- This may be summarized by the new density formula

$$\lim_{m \to \infty} \hat{f}(x; m) = \frac{1}{nh} \sum_{j=1}^{n} \left(1 - \frac{|x - x_j|}{h}\right) I_{[-1,1]}\left(\frac{x - x_j}{h}\right),$$

Limiting ASH as a Kernel Estimator

- Defining a *kernel function* $K(\cdot)$ to be an isosceles triangle density,

$$K(t) = (1 - |t|) I_{[-1,1]}(t)\,,$$

- the limiting ASH may be written as

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

- The multivariate kernel corresponding to the multivariate naive ASH is

$$\hat{f}(\mathbf{x}) = \frac{1}{nh_1 h_2 \cdots h_d} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right) \right\},$$

the so-called **product kernel estimator**.

Example with the Silica Data ($n = 22$)

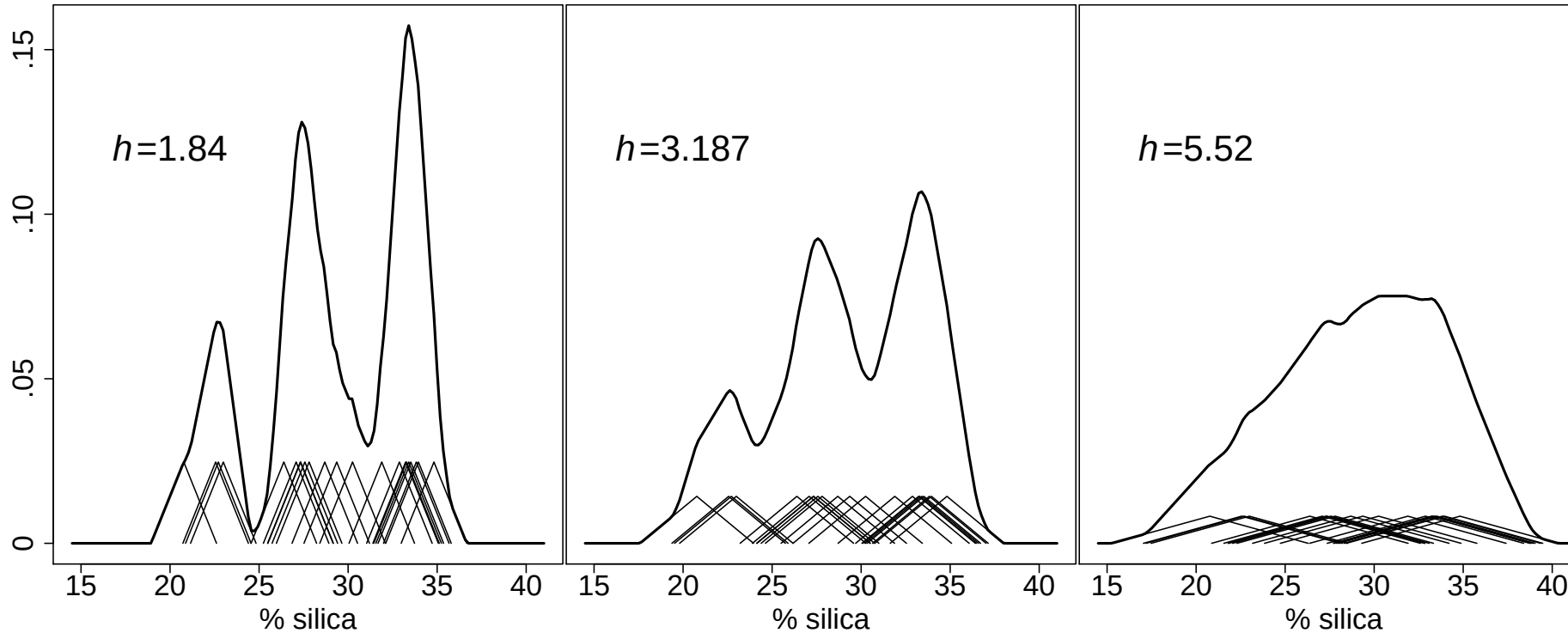

Figure: Triangle kernel estimates of the silica dataset showing the individual kernels.

Wrapping Up

- The product kernel estimator does not result in a kernel density estimator where the dimensions factor as well (i.e. independent). Why? Because of the summation.
- Kernel methods go back to Fix and Hodges (1956), Rosenblatt (1956), Parzen (1962), then an explosion of authors and research, including yours truly with the histogram in 1979 and the discrete maximum penalized likelihood estimator in 1980. Papers on the FP, ASH, oversmoothing appeared shortly thereafter.
- The first edition of Multivariate Density Estimation appeared in 1992. The second edition in 2015.

Wrapping Up (continued)

- The ASH is a form of discrete convolution of weights and bin counts.
- Chamayou independently described the ASH in 1980.
- Härdle and Scott generalized the ASH into Weighted Average of Shifted Points, or **WARP'ing**.
- The ASH has been used by Wegman to smooth parallel coordinate plots.
- Debbie Swayne used the 1-d ASH in xgobi to show the density of points computed in the grand tour. John Salch extended this idea to projections in 2 and 3 dimensions and demonstrated the visualization on an SGI workstation