Measuring the stability of histogram appearance when the anchor position is changed

JEFFREY S. SIMONOFF

Department of Statistics and Operations Research, New York University, New York, NY 10012, USA. e–mail: jsimonoff@stern.nyu.edu

FREDERIC UDINA

Departament d'Economia, Universitat Pompeu Fabra, 08008 Barcelona, Spain. e–mail: udina@upf.es

Abstract: Although the histogram is the most widely used density estimator, it is well–known that the appearance of a constructed histogram for a given bin width can change markedly for different choices of anchor position. In this paper we construct a stability index G that assesses the potential changes in the appearance of histograms for a given data set and bin width as the anchor position changes. If a particular bin width choice leads to an unstable appearance, the arbitrary choice of any one anchor position is dangerous, and a different bin width should be considered. The index is based on the statistical roughness of the histogram estimate. We show via Monte Carlo simulation that densities with more structure are more likely to lead to histograms with unstable appearance. In addition, ignoring the precision to which the data values are provided when choosing the bin width leads to instability. We provide several real data examples to illustrate the properties of G. Applications to other binned density estimators are also discussed.

Keywords: Bin width, Frequency polygon, Gini index, Linear binning, Lorenz curve, Monte Carlo simulation

1. Introduction

It is well–established statistical practice that data analysts should look at their data graphically when analyzing it. For univariate data, such graphical display leads to density estimation. Although researchers have proposed many different density estimators through the years (see Scott, 1992, Wand and Jones, 1995, and Simonoff, 1996, for general discussion), the simplest estimator, the histogram, remains the most widely used. This is due in large part, no doubt, to the fact that virtually all statistical packages provide histograms as a standard method of data examination. The histogram also has the advantages of ease and simplicity of construction, simplicity of interpretation (including for the statistically unsophisticated), and lack of requirement of advanced graphics tools.

Let $\{x_1, \ldots, x_N\}$ be a set of data values. A fixed bin width histogram is defined by dividing the region of interest into a set of K equisized bins, each with bin width h, determined by the bin edges $\{b_1, \ldots, b_{K+1}\}$ (where $b_{j+1} - b_j = h$ for all j). The histogram estimate of the underlying density $f(x)$ within a given bin is

$$\hat{f}(x) = \frac{n_j}{Nh}, \qquad x \in (b_j, b_{j+1}],$$

where n_j is the number of observations falling in the j^{th} bin $(b_j, b_{j+1}]$. When needed, we will define $n_0 = n_{K+1} \equiv 0$ as the counts of the adjacent empty bins.

The bin edges $\mathbf{b}$ of the histogram (or, more precisely, the bin width h and the *anchor position* b_1) completely determine its appearance (any value that fixes the position of the bin edges for a given h could be defined as the anchor position, but b_1 seems the most natural choice). The bin width h acts as a smoothing parameter, as it controls the degree of smoothness of the estimate, with larger values of h resulting in histograms with a smoother appearance. All density estimators include some form of smoothing parameter, and a good deal of research has focused on choosing it for different estimators, often based on an assessment of accuracy using the integrated squared error of the estimator,

$$ISE = \int_{-\infty}^{\infty} [\hat{f}(x) - f(x)]^2 dx,$$

and its expected value, mean integrated squared error ($MISE$).

Unfortunately, such analysis does not address questions regarding the appearance of the histogram, for two reasons. First, the ISE measure is not particularly effective at quantifying how well a density estimate approximates the appearance of a true density, despite its natural role as a measure of accuracy; see Kooperberg and Stone (1991), Marron (1996) and Marron and Tsybakov (1995), among others, for a discussion of this point. Second, asymptotic analysis shows that anchor position of a histogram has a lower order asymptotic effect on $MISE$ compared with the bin width, and can therefore be ignored. Despite this, from a practical point of view, shifting the bin edges by changing the anchor position can have an effect on the appearance of the resultant histogram for finite samples. Many authors have focused on this as one of the biggest drawbacks of using the histogram (see, e.g., Fisher, 1989; Härdle, 1991, Section 1.4; Härdle and Scott, 1992; Izenman, 1991; Samiuddin, Jones and El–Sayyad, 1993; Scott, 1992, Section 4.3; Silverman, 1986, Section 2.2; Wand and Jones, 1995, Section 1.2).

Figure 1 illustrates the problem. This figure is based on values for the number of visas that were issued by the U.S. Immigration and Naturalization Service in 1991 for the purpose of adoption by U.S. residents for 39 countries or regions of origin (Chatterjee, Handcock and Simonoff, 1995, p. 13). The data are very long–tailed, and have been logged (base 10).

The three histograms in Figure 1 all have bin width $h = .276$, with anchor positions .48, .5 and .59, respectively. While all three plots agree on the existence of a major mode around 1.8, they disagree on the height of that mode, on the location and symmetry of a minor mode at low values, and on the existence and location of possible modes at high values (Chatterjee *et al.*, p. 16, present a histogram with $h = .2$ that appears to be a combination of the third histogram at the low end and the first at the high end).

There has been little systematic examination of anchor position effects in the literature. Simonoff (1995) found that the average ISE (what can be termed quantitative accuracy) of the histogram estimate is insensitive to anchor position, unless a discontinuity (or near discontinuity) of the density is crossed by a bin (that is, if the discontinuity occurs inside a bin rather than at a bin edge; see also Scott, 1992, pp. 65–66). On the other hand, the appearance of histograms (as quantified by the number of observed modes, a measure of qualitative accuracy) can be very sensitive to anchor position. Scott (1992, p. 111), in the context of the frequency polygon (the close cousin of the histogram where the estimated density values are the linear interpolants of the heights at the bin centers), noted that the anchor position can be thought of as a nuisance parameter, and suggested choosing it for a given bin width to make the resultant estimate as smooth as possible.

We agree that the anchor position is a nuisance parameter. Rather than pick a particular (arbitrary) value for that parameter, however, we propose to construct a measure to assess how sensitive the appearance of the histogram is to *any* possible choice. That is, the measure is a function of the data and h, not any particular anchor choice.

If the appearance of the histogram doesn't change very much for a given h as the anchor position changes, the analyst is free to choose the anchor however he or she wishes, without worrying about the effect of that choice. We will term such a bin width a *stable* bin width. However, if the appearance is sensitive to anchor position, the impressions from a histogram using any particular choice cannot be trusted, since a different choice could lead to very different impressions (we will term such a bin width an *unstable* bin width).

It is important to note that the stability measure is **not** a measure of the accuracy of the histogram as an estimate of the true density f, but rather of the consistency of the representation of that density as anchor position is changed. That is, a stable bin width is not necessarily one that gives an accurate impression of the true density, but rather one where the impressions don't change very much with anchor position. Thus, a data analyst would use the index as a secondary tool, after first choosing the bin width to provide an accurate impression of the true density (based on ISE, or some other measure). If the chosen bin width is stable, an anchor position is chosen, and the estimate is constructed.

If the bin width is unstable, however, any choice of anchor position is dangerous. Instead, a different, more stable, bin width should be chosen. If the new bin width is close to the original one, it is likely that the histogram is as accurate as one based on the original choice, and little is lost; but if there are no stable choices near the original choice, it is likely that no histogram will be satisfactory, and a different density estimator should be used.

Section 2 describes the construction of the stability index. It is based on a measure of the statistical roughness of the estimate, since that is related to general impressions of its shape. The index itself is a variant of the Gini index, based on the area under a constructed Lorenz curve. The properties of the index are investigated using Monte Carlo simulations in Section 3, where it is shown that the shape of the true density, and the precision to which the data are given, both have a strong effect on the stability of histograms. Section 4 gives several real data examples that illustrate the use of the index. The paper concludes with a discussion of extension to other binned density estimators.

2. Motivation and definition of the stability index

In this section we define the stability index and discuss its properties. Given the data set and a fixed bin width h, the goal is to define a measure G of the similarity of the histograms that result from all possible anchor choices. The problem can be considered in two parts:

(1) how to measure the (dis)similarity between histograms, and

(2) how to combine so many possible histograms into a single measure.

Two histograms are considered "similar" if they have roughly the same shape. When data analysts refer to the shape of the density estimate, they are typically thinking of modes, bumps and dips in the estimate — that is, f'. Our measure of the similarity is a global quantity that is sensitive to changes in the shape of the density, $R(f') = \int f'(x)^2 dx$ (in this usual notation, the R stands for statistical roughness). $R(f')$ is related to shape, but also occupies a central position in the theory of histograms, as it determines the optimal accuracy of a histogram estimate (in terms of asymptotic $MISE$ [$AMISE$]) and the optimal bin width. If f has squared–integrable and absolutely continuous derivative, the optimal bin width for histograms is given by

$$h^\star = \left[\frac{6}{R(f')}\right]^{1/3} N^{-1/3} \tag{2.1}$$

(see Scott, 1992, Chapter 3, for the details). One drawback of using the global measure $R(f')$ is that different–looking histograms can have similar values of $R(f')$, but its use avoids the need for handling shifts and translations of the estimates or of the bumps (see Marron and Tsybakov, 1995).

Given a histogram, the natural estimator of f' is the step function obtained by differencing the bin counts; that is,

$$\widehat{f'(x)} = \frac{n_{j+1} - n_j}{Nh^2}, \qquad x \in (b_{j+1} - h/2, b_{j+1} + h/2).$$

This way, the estimate for $R(f')$ is

$$S = \frac{1}{N^2 h^3} \sum_{j=0}^{K} (n_{j+1} - n_j)^2.$$

Scott and Terrell (1987) used this estimate of $R(f')$ when they developed the biased cross–validation bin width choice, based on (2.1). Asymptotically, for N large and $h \to 0$,

$$E(S) = R(f') + \frac{2}{Nh^3} + O(h)$$

(Scott, 1992, equation 3.48) and

$$V(S) = \frac{12R(f)}{N^2h^5} + O(N^{-2}h^{-4})$$

(Scott and Terrell, 1987, equation 3.20). These asymptotics are not useful in our setting, because we want to deal with small and medium sample sizes, in which case the slow convergence rates of these asymptotic formulae give poor performance of the approximations. Also note that we don't consider S as a random variable, changing with the random sample, but rather are interested in changes in S due to shifting the anchor position. Finally, the asymptotics are based on ignoring the anchor position as being a higher order effect, when anchor position is the key issue here. So, we will use the quantity $R(f')$ simply as a number that reflects changes in the shape of the histogram as the anchor is shifted.

Define the shifted histogram H_t to have bin edges $\{b_1 - t, \ldots, b_{K+2} - t\}$, $t \in [0, h)$. S_t will denote the corresponding $R(f')$ estimate. S_t as a function of t is a step function and, given the discreteness of the problem, a good way to compute the variability of S_t as t changes is to take T evenly spaced values of t and consider the corresponding S_t values (T is the number of anchor positions examined, and would be large).

For simplicity we will denote here $S_i = S_{ih/T}$, $i = 1, \ldots, T$. To decide if this set of T numbers is highly variable (suggesting instability) we use a variant of the Gini index (see Marshall and Olkin, 1979). Define $q_0 = 0$ and, for $i = 1, \ldots, T$

$$q_i = \sum_{j=1}^{i} S_{(j)} \Big/ \sum_{j=1}^{T} S_j$$

where $S_{(j)}$ is the j^{th} order statistic. Take the pairs $(i/T, q_i)$, $i = 0, \ldots, T$ to draw the so–called Lorenz curve in the unit square and define the stability index G as twice the area below this curve. Note that if the numbers S_i are very similar, this curve will be close to the diagonal of the unit square. Then

$$G = \sum_{i=1}^{T} \frac{q_i + q_{i-1}}{T} = \frac{1}{T}\left[2\sum_{i=1}^{T} q_i - 1\right] = \frac{1}{T}\left[\frac{2}{\sum_{i=1}^{T} S_i}\sum_{i=1}^{T}(T-i+1)S_{(i)} - 1\right].$$

Transforming the last expression by using $S_{[i]}$ to denote the inverse order statistic, so that $S_{[1]} \geq \ldots \geq S_{[T]}$, we can write

$$G = \frac{1}{T}\left[\frac{2\sum iS_{[i]}}{\sum S_i} - 1\right],$$

which is simpler and useful for computation. The G index also can be written as

$$G = 1 - \frac{1}{2T\sum S_i}\sum_i\sum_j |S_i - S_j| = \frac{1}{T\sum S_i}\sum_i\sum_j \min(S_i, S_j).$$

The form of G immediately implies several properties of it:

(1) $G \in (0, 1]$, higher values showing more similarity among the shifted histograms. So, G has an absolute scale that doesn't depend on N or h. Values greater than .85 are usually interpreted as representing stable bin widths, as we will explain in sections 3 and 4.

(2) As we have defined it, G is determined only by the data set and the bin width. Dependency on T can be shown to be negligible if T is large enough. We used $T = 100$ in our computations.

(3) To compute G, computation of the bin counts of the T different histograms can be done with a single loop through the data, resulting in $O(N)$ computation time. See the appendix for more technical details.

(4) For a given data set, G as a function of h is a step function. The examples and simulations in the next sections show that the length of the steps and the jumps between steps are usually small. At the typical graphing scale, G usually can be drawn as a continuous curve.

(5) Looking at the pattern of G as a function of h can give some guidance about how to choose the bin width within a reasonable range. At least, it can be used to reject strongly some value of h that gives very unstable histogram appearance.

3. Monte Carlo investigation of the properties of G

In this section we investigate the relationship between G and various properties of the data. In this way, it is possible to see what kinds of properties (and data analytic choices) are associated with lower or higher values of G, and hence less or more stability of histogram appearance.

Table 1 gives a description of the underlying distributions being examined here. All are Gaussian mixtures (except for the discrete distribution), some of which were drawn from Marron and Wand (1992). For each distribution, sample sizes $N = 20$, 100 and 500 were treated. Two hundred bin widths were examined, equally spaced in the range $[.1h_{OS}, h_{OS}]$, where $h_{OS} = 1.5$ (for $N = 20$), 1 (for $N = 100$) and .6 (for $N = 500$), respectively, are the oversmoothed choices of bin width based on using roughly $(2N)^{1/3}$ bins (the value determined by Terrell and Scott, 1985, to be the minimum number of bins that should be used for a given sample size). For each bin width, $T = 100$ anchor positions corresponding to $x_{(1)} - .01ih$, $i = 1, \ldots, 100$, were used to calculate G. This was repeated 400 times for each (distribution, sample size, bin width) triple. Pseudo–random uniform deviates were generated using the algorithm of Wichmann and Hill (1982), which were then transformed to be Gaussian using the Box–Muller transformation.

Figure 2 gives results for the Gaussian density. In this plot (and in Figures 3–8) the three curves correspond to the average values of G for given bin widths, connected by lines, for $N = 20$ (solid line), $N = 100$ (dotted line) and $N = 500$ (dashed line). In order to make the curves for different sample sizes comparable, the horizontal axis is scaled to be the bin width as a proportion of the oversmoothed choice (since a data analyst shouldn't take a bin width larger than that value). The vertical line represents the approximate position of the $AMISE$–optimal bin width, based on (2.1) (the position is approximate because the values of h_{OS} used do not correspond to exactly $(2N)^{1/3}$ bins).

Since the Gaussian density is relatively featureless (being unimodal, symmetric and not kurtotic), Figure 2 gives a good sense of what G looks like for data that are likely to lead to histograms with stable appearance. As was noted in Section 2, the jaggedness in the curves is partly due to a lack of continuity as a function of h, but the appearance is still reasonably smooth, as the discontinuities cover small vertical distances. The plot suggests general patterns for well–behaved data:

(a) Increasing sample size is associated with increasing stability. This makes sense, since as $N \to \infty$, the effect of the anchor position becomes progressively less important.

(b) The minimum average G is about .8; for $N = 100$ and 500, it is over .85. This reinforces the impression that .8 – .85 are high values (suggesting stability).

(c) The bin width that is most unstable occurs at about .3 – .4 times the oversmoothed choice. This is considerably smaller than the choice that minimizes $AMISE$, so for this distribution undersmoothing leads to more instability.

Figure 3 refers to a density with more structure, being strongly kurtotic (with a narrow mode). In this figure curves for $N = 1000$ (dotted and dashed line) are also given. Figure 3(a) uses the proportion of oversmoothed bin width as the horizontal axis, and is directly comparable to Figure 2. It is similar to that figure, except that the minimizing values of the index curves are at different points on the horizontal axis. These minimizing values correspond to the same bin width in an absolute sense, as Figure 3(b) shows, corresponding to $h \approx .4$. In this half of the figure, a line representing the $AMISE$–optimal bin width for each sample size is given, since it becomes progressively smaller with increasing sample size. Recall that the narrow mode in this density comes from a normal density with standard deviation .1; thus, $h = .4$ is roughly the width of the mode. That is, the troublesome bin width roughly equals the range of the structure of interest. More important, the problems corresponding to $h \approx .4$ don't disappear with increasing sample size, as they reflect a fundamental characteristic of the distribution itself (although $h = .4$ becomes too large to be a candidate bin width for large enough sample size).

The minimizer of $AMISE$ for this density is smaller than .4 for any reasonable sample size. That is, oversmoothing, which has been suggested as a conservative approach to smoothing parameter selection (Terrell, 1990), can be a bad idea, in terms of histograms and the stability of the histogram. This bin width corresponds to using about 15 bins, a reasonable choice for a data analyst to make. The index is still not too small in an absolute sense, however, suggesting that this density does not lead to overly unstable histogram appearance.

Figure 4 refers to the "shoulders" density, one with one major mode and two bumps on either side of the mode. This has more structure than the Gaussian, but the index curves are very similar to those in Figure 2. The $R(f')$ shape measure is not very sensitive to changes in shape corresponding to small bumps (that is, it is not very sensitive to small bumps appearing and disappearing, being much more sensitive to higher and narrower modes appearing and disappearing). If the consistency of appearance of small bumps was of particular importance, the shape measure could be changed to account for this, by weighting:

$$R_w(f') = \int_{-\infty}^{\infty} [f'(x)]^2 w(x) dx,$$

where $w(\cdot)$ is a weight function that could be a function of the underlying density, giving more weight in lower density regions. We do not investigate this possibility further here.

Figure 5 refers to a trimodal density, which of course has more structure. This is reflected in a different pattern of the index versus bin width. The index falls below .8 for bin width equal to about 1.05 for $N = 20$ (Figure 5(b)), and is even worse for $N = 100$ (this corresponds to about 6 bins). Thus, larger samples can be more, rather than less, problematic when there is detailed structure in the data. The reason for this is that such structure often will not be apparent for small samples, for any bin width or anchor position choice. The figure also shows that this bin width is consistent with oversmoothing for all sample sizes studied.

Figure 6 refers to the claw density, a density with a great deal of very fine structure. For small samples, this structure cannot be discerned, and the stability index curve is similar to that for the Gaussian density. As the sample size increases, instability increases, focusing ultimately at $h = .25$ (Figure 6(b)), or about 24 bins. Since the $AMISE$–optimal bin width is $h = .953N^{-1/3}$, $h = .25$ corresponds to the $AMISE$ choice for $N \approx 55$. Thus, a bin width chosen based on $AMISE$ considerations can lead to greater instability in histogram appearance for this density. A periodicity in the index related to bin width is also apparent in this plot. Apparently fine structure leads to this periodicity for larger bin widths, for reasons that are not clear.

All the densities so far were symmetric, and smoothly approached zero at both the upper and lower extremes. Figure 7 refers to a strongly skewed density, with a sharp drop at the low end. The index shows that small bin widths lead to relatively stable histogram appearance, but as the bin width increases, the instability increases. Further, for these larger bin widths, the effect does not diminish with increasing sample size. This is probably due to the sharp drop at the low end of the density (close to a discontinuity). A wider bin is more likely to cross the (near) discontinuity, resulting in the combination of both a high probability and near–zero probability region in the same bin (with resultant instability in the estimate). A similar pattern occurs concerning estimation accuracy, in that the histogram becomes increasingly inefficient with respect to $MISE$ with increasing sample size when a bin crosses a discontinuity in the density (Scott, 1992, pp. 65–66; Simonoff, 1995).

The final distribution examined is a discrete one. Naturally, no one should use a histogram for discrete data, but this distribution allows examination of questions related to the precision of the data. By "precision" we mean the accuracy to which the values are reported; for example, data that are rounded to the nearest integer have precision at the level of integers. The discrete distribution used here has a precision of $\frac{2}{3}$ (this was done so that it would cover the same range as the continuous densities). Figure 8 shows that this sort of discreteness can have a big effect on the stability of the histogram. The natural bin width to choose for these data is $h = \frac{2}{3}$, since that is the gap between the distribution values; this turns the histogram into a (probability) frequency distribution. For that choice, $G = 1$, as all histograms are identical. However, the instability can increase dramatically (and quickly) as the bin width moves away from the natural value. The stability index dips dramatically at $h = .625$ and $h = .725$, only 6% lower and 9% higher, respectively, than the best choice. Thus, the presence of repeats in the data can have a large effect on histogram stability for certain bin widths, and care must be taken with that choice. It is dangerous to choose a bin width that is not consistent with the precision of the data (a noninteger for data rounded to the nearest integer, for example).

4. Application to real data

In this section, we apply the stability index to several real data sets. The results of the previous section show that values of G below .8 or so indicate potential instability, but it would be useful to be able to attach an indicator of the strength of evidence for or against instability provided by an observed value of G for a given data set. What is necessary is some way to evaluate whether an observed G is unusually small, given the general distribution of the data.

We propose using Monte Carlo to construct such a measure of evidence for a given data set and bin width choice, as follows:

(1) Construct a histogram using some anchor position, such as $x_{(1)} - h/2$ (the algorithm is quite insensitive to this choice). This yields observed bin counts n_j. Define the frequency estimate of the probability of falling in a given histogram bin to be $\overline{p}_j = n_j/N$.

(2) Create a simulated histogram by generating counts n_j^* based on a Multinomial($N, \overline{\mathbf{p}}$) distribution. Calculate S. Repeat this T times, and calculate G (notation is as in Section 2).

(3) Repeat step (2) B times, getting a "null" distribution for G. An observed G can then be compared to this distribution to see if it is surprisingly small (or large, for that matter).

Note that in this construction, the anchor is never actually moved. This means that any observed variation in G comes from the inherent variability in the histogram, rather than from moving the anchor. An unusually small G therefore is due to the anchor problem, not the properties of histogram estimation itself. Although the construction is similar to what would be done to determine a Monte Carlo tail probability, we will term the analogous number the "evidence level," rather than significance level, with small evidence levels indicating unstable bin widths.

The first data set gives the durations of 222 eruptions of the "Old Faithful" geyser in Yellowstone National Park in August 1978 and August 1979 (Weisberg, 1982). Many authors have examined these data (or a subset of them), with a bimodal distribution of the eruption durations being found (Scott, 1992, p. 18, gives a histogram for a subset of these data with $h = .5$). Figure 9 gives a plot of G versus the bin width for 200 values in (.16, .5) ($h = .5$ being the oversmoothed choice). The gray rectangles in this plot (and those of Figures 10–12) define values of h where the Monte Carlo evidence level of the index was less than .01 (based on 400 Monte Carlo replications for each bin width). The most important message from the plot is that $G > .84$, showing great stability in histogram appearance for all choices of h.

Given this, the index plot has many spikes in it, due to sharp changes in the index for close bin widths. This is because the data are given only to the nearest tenth; as was seen in the last section, it is potentially dangerous to choose the bin width to a greater precision than the data have. The sharp dips in the index correspond to where the evidence level is small (i.e., histograms like this typically have G values higher than the observed value), but since the values are still large in an absolute sense, even those bin widths are probably reasonably stable.

The second data set gives the concentrations of PCBs in 37 U.S. bays and estuaries in 1985 (Chatterjee *et al.*, 1995, p. 164). The data are very long–tailed, with most bays having low concentrations (less than 50 parts per billion) and a few being much higher (to as much as 750 parts per billion). This kind of data set is susceptible to unstable histogram appearance, as Figure 10 shows. Virtually all bin widths less than 70 have index value less than .8, and virtually all have Monte Carlo evidence level less than .01. Chatterjee *et al.*, p. 165, give a histogram with $h = 50$ ($G = .80$), but then note the long tails and suggest working in the logged scale.

The third data set is the logged adoption visa data set discussed in Section 1. Figure 11 gives a stability index plot for these data. Most bin width choices are stable (including, for example, $h = .562$, the choice using (2.1) assuming a Gaussian distribution for f, and estimating the population standard deviation using the scale estimate of Janssen *et al.*, 1995), but the choice used in Section 1 ($h = .276$) has $G = .66$, and is quite unstable.

The fourth data set gives the ages in years of the 105 players in the National Basketball Association who played the guard position during the 1992–1993 season (Chatterjee *et al.*, 1995, p. 201). These data are discrete, having been rounded to the nearest year. The index plot given in Figure 12 shows that the natural choice $h = 1$ has $G = 1$, but close values of h have much smaller values of G (the spikiness of the plot is a direct consequence of the discrete nature of the data). Figure 13 gives three histograms for these data with $h = 1.38$ ($G = .68$) with anchors 21, 21.5 and just below 22, respectively. The three plots give very different impressions of the number of modes in the data, and the relative heights of those modes. None look very much like the natural histogram ($h = 1$) in Figure 14, which shows that the modal values of age are 24, 27 and 30, with the first two modes having slightly higher probability than the third.

The final data set examined is the well–known Buffalo snowfall data (Parzen, 1979). Scott (1992, p. 110) used these data, with $h = 13.5$, to illustrate the sensitivity of histogram appearance to anchor position, but the stability index does not support this ($G = .85$). The reason is that these data are similar to the "shoulders" density summarized in Figure 4; when underlying structure comes from small secondary modes compared with a large primary mode, the changes in appearance of the histogram in Scott's figures correspond to relatively small changes in $R(f')$.

5. Extension to other binned density estimators

Any density estimator based on binning can present anchor dependency problems. Some binned estimators, however, are devised specifically to suppress this problem. Scott (1985) introduced the averaged shifted histogram (ASH) as a method to suppress the noise effect of anchor shifting. Härdle and Scott (1992) generalized this method to use with a general kernel function and named it WARPing (weighted averaging of rounded points). These methods can be seen as approximations to the kernel density estimator, and are clearly better than any histogram–based estimator. In this section we examine other simple density estimators that present bin edge problems.

The simplest improvement to the histogram is the frequency polygon. The linear interpolant of histogram heights at the bin centers, the frequency polygon has the form

$$\hat{f}_{fp}(x) = (Nh)^{-1}\left[\frac{n_i + n_{i+1}}{2} + \left(\frac{n_{i+1} - n_i}{h}\right)(x - b_{i+1})\right], \quad x \in [b_{i+1} - h/2, b_{i+1} + h/2], \quad i = 0, \ldots, K. \tag{5.1}$$

The frequency polygon is superior to the histogram in terms of $MISE$, achieving the rate $AMISE = O(N^{-4/5})$ (taking $h = O(N^{-1/5})$, rather than the optimal $O(N^{-1/3})$ rate for histograms), and this improved accuracy carries over to small samples (Simonoff and Hurvich, 1993), but its appearance (in terms of modes, bumps and dips) is identical to that of the histogram, and it therefore has the identical anchor stability properties for a given h.

An alternative to the frequency polygon that is also piecewise linear, but can achieve 11.5% smaller optimal $AMISE$, was introduced by Jones *et al.* (1995), and has the form

$$\hat{f}_{L1}(x) = (2Nh)^{-1}\left[\frac{n_{i+1} + 2n_i + n_{i-1}}{2} + \left(\frac{n_{i+1} - n_{i-1}}{h}\right)(x - b_{i+1} + h/2)\right], \quad x \in [b_i, b_{i+1}], \quad i = 1, \ldots, K.$$

This estimate is the linear interpolant of the averages of two adjacent bin heights at right bin edges, and will therefore be called the average frequency polygon here.

A more complicated estimator, the linearly binned frequency polygon, replaces the cell counts n_i in (5.1) with linear bin counts

$$\ell_i = \sum_{j=1}^{n}(1 - h^{-1}|x_j - b_i - h/2|)_+$$

where + subscript denotes positive part (Jones and Lotwick, 1983; Jones, 1989). This can be seen as each data point splitting its unit mass between the two nearest bin centers, in inverse proportion to the distances to them. The estimator is a discretized kernel estimator with triangular kernel function (Jones, 1989), and can achieve 5.8% lower optimal $AMISE$ than $\hat{f}_{L1}$.

Figure 15 illustrates the sensitivity to anchor position of these three types of frequency polygon for two of the data sets discussed in Section 4 (the geyser eruption and logged adoption visa data sets, respectively). The plots give the stability index values for the three frequency polygons, estimating f' from each estimated density in the natural way (the curve for the ordinary frequency polygon is identical to that for the histogram). Both alternative frequency polygons are less susceptible to unstable bin widths, which supports the informal impressions in Jones *et al.* (1995, Section 3). The average frequency polygon, which is no more difficult to calculate and interpret than the ordinary frequency polygon, leads to noticeably more stable bin widths, and deserves further study and use. Monte Carlo examination of the frequency polygons confirms this pattern.

In these comparisons we used the stability index as defined in section 2. It can be argued, however, that for frequency polygons $R(f'')$ is a better measure of shape than $R(f')$; in particular, the $AMISE$ and the optimal bandwidth depend on $R(f'')$ for such estimators, and for smooth enough curves $R(f'')$ is the most natural measure of curvature. Although values of $R(f')$ and $R(f'')$ are obviously different, and the resulting stability indices are also different, both simulation evidence and application to real data sets shows that bandwidths with high or low index values are roughly the same, whichever functional is used to reflect the shape changes. Thus, using $R(f')$ in Figure 15 allows direct comparison with the earlier results, without changing the results appreciably.

LISP–STAT code to calculate G, and dynamic graphics demonstrating the methods discussed here, are available via anonymous `ftp` at the address `ftp.upf.es`, in the directory `pub/stat/anchor-position`. The material can also be accessed using a World Wide Web (WWW) browser at the address

$$\mathtt{ftp://ftp.upf.es:/pub/stat/anchor-position}.$$

Acknowledgments

We would like to thank Steve Marron for helpful discussion of this material (and for suggesting that one of us contact the other about this topic). The research of the second author was partially supported by a Spanish DGICYT grant PB92.1037.

Appendix

Fast computation of the stability index

The index G can be computed with a minimum of looping through the data. A naive approach would be to form T histograms to obtain the values S_t, which would raise the amount of computation to $O(NT)$. Constructing a single binning on a finer grid can give all the information needed, which reduces the order of the computation time to the maximum of N and $T(K+1)$, where K is the number of bins. In this appendix, it is convenient to label the bins from 0 to $K-1$, rather than from 1 to K.

Given data $\{x_i\}_{i=1\ldots N}$, a number T of anchor positions, and bin width h, let $K = 1 + \left\lceil \frac{x_{(N)} - x_{(1)}}{h} \right\rceil$ be the minimum integer such that all data points are in $[x_{(1)}, x_{(1)} + (K-1)h]$ (all histograms will thus have K bins). We will consider first the bin edges

$$d_0 = x_{(1)} - h, \quad d_j = d_0 + jh/T \quad j = 0, \ldots, (K+1)T$$

and compute the bin counts

$$m_j = \#\{x_i | x_i \in (d_j, d_{j+1}]\} \quad j = 0, \ldots, (K+1)T - 1$$

(this is the only loop over the data). A further loop over these bin counts gives the quantities

$$M_j = \sum_{k=0}^{T-1} m_{j+k}, \quad j = 0, \ldots, KT,$$

computed using the recursive relation $M_{-1} = 0, M_j = M_{j-1} - m_{j-1} + m_{j+T-1}$. In the same loop, the quantities

$$D_j = M_j - M_{j-T} \quad j = T, \ldots, KT,$$

can be computed and stored.

Now consider the histogram H_t with bin edges $\{d_t, d_{t+T}, \ldots, d_{t+KT}\}$ $(t = 0, \ldots, T-1)$. For $j = 0, \ldots, K-1$, the j^{th} bin will have a count of

$$n_j^t = \sum_{k=0}^{T-1} m_{t+jT+k} = M_{t+jT}.$$

S_t is then computed as

$$S_t = [n_0^t]^2 + \sum_{j=0}^{K-2} (n_{j+1}^t - n_j^t)^2 + [n_{K-1}^t]^2$$
$$= M_t^2 + \sum_{j=0}^{K-2} D_{t+(j+1)T}^2 + M_{t+(K-1)T}^2.$$

All these sums can be computed in the same loop as when the M_j and D_j are computed, so only one loop $i = 1, \ldots, N$ and one loop $j = 0, \ldots, T(K+1)$ are needed.

References

Chatterjee, S., Handcock, M.S. and Simonoff, J.S., *A casebook for a first course in statistics and data analysis* (John Wiley, New York, 1995).

Fisher, N.I., Smoothing a sample of circular data, *J. Structural Geology*, 11 (1989) 775–778.

Härdle, W., *Smoothing techniques with implementation in S* (Springer–Verlag, New York, 1991).

Härdle, W. and Scott, D.W., Smoothing by weighted average of rounded points, *Comput. Statist.*, 7 (1992) 97–128.

Izenman, A.J., Recent developments in nonparametric density estimation, *J. Amer. Statist. Assoc.*, 86 (1991) 205–224.

Janssen, P., Marron, J.S., Veraverbeke, N. and Sarle, W., Scale measures for bandwidth selection, *J. Nonparam. Statist.*, 5 (1995) to appear.

Jones, M.C., Discretized and interpolated kernel density estimates, *J. Amer. Statist. Assoc.*, 84 (1989) 733–741.

Jones, M.C. and Lotwick, H.W., On the errors involved in computing the empirical characteristic function, *J. Statist. Comput. Simul.*, 17 (1983) 133–149.

Jones, M.C., Samiuddin, M., Al–Harbey, A.H. and Maatouk, T.A.H., Piecewise linear smoothed histograms, unpublished manuscript (1995).

Marshall, A.W. and Olkin, I., *Inequalities: theory of majorization and its applications* (Academic Press, New York, 1979).

Marron, J.S., Assessing bandwidth selectors with visual error criteria, *J. Amer. Statist. Assoc.*, 91 (1996) to appear.

Marron, J.S. and Tsybakov, A.B., Visual error criteria for qualitative smoothing, *J. Amer. Statist. Assoc.*, 90 (1995) 499–507.

Marron, J.S. and Wand, M.P., Exact mean integrated squared error, *Ann. Statist.*, 20 (1992) 712–736.

Parzen, E., Nonparametric statistical data modeling, *J. Amer. Statist. Assoc.*, 74 (1979) 105–131.

Samiuddin, M., Jones, M.C. and El–Sayyad, G.M., On bin–based density estimation, *J. Statist. Comput. Simul.*, 47 (1993) 241–252.

Scott, D.W., Average shifted histograms: effective nonparametric density estimators in several dimensions, *Ann. Statist.*, 13 (1985) 1024–1040.

Scott, D. W., *Multivariate density estimation* (John Wiley, New York, 1992).

Scott, D. W. and Terrell, G. R., Biased and unbiased cross–validation in density estimation, *J. Amer. Statist. Assoc.*, 82 (1987) 1131–1146.

Silverman, B., *Density estimation for statistics and data analysis* (Chapman and Hall, London, 1986).

Simonoff, J.S., The anchor position of histograms and frequency polygons: quantitative and qualitative smoothing, *Commun. Statist. — Simul. Comput.*, 24 (1995) 691–710.

Simonoff, J.S., *Smoothing methods in statistics* (Springer–Verlag, New York, 1996).

Simonoff, J.S. and Hurvich, C.M., A study of the effectiveness of simple density estimation methods, *Comput. Statist.*, 8 (1993) 259–278.

Terrell, G.R., The maximal smoothing principle in density estimation, *J. Amer. Statist. Assoc.*, 85 (1990) 470–477.

Terrell, G.R. and Scott, D.W., Oversmoothed nonparametric density estimates, *J. Amer. Statist. Assoc.*, 80 (1985) 209–214.

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

Weisberg, S., *Applied linear regression, 2nd. ed.* (John Wiley, New York, 1985).

Wichmann, B.A. and Hill, I.D., AS183: An efficient and portable pseudo–random number generator, *Appl. Statist.*, 31 (1982) 188– 192.

Table 1. Distributions used in the Monte Carlo simulations.

Distribution	*Form*
Gaussian	$N(0,1)$
Kurtotic unimodal	$\frac{2}{3}N(0,1)+\frac{1}{3}N(0,\left(\frac{1}{10}\right)^2)$
"Shoulders"	$\frac{4}{5}N(0,1)+\frac{1}{10}N(-1.8,\left(\frac{2}{5}\right)^2)+\frac{1}{10}N(1.8,\left(\frac{2}{5}\right)^2)$
Trimodal	$\frac{1}{3}N(0,1)+\frac{1}{3}N(-2,\left(\frac{1}{3}\right)^2)+\frac{1}{3}N(2,\left(\frac{1}{3}\right)^2)$
Claw	$\frac{1}{2}N(0,1)+\sum_{i=0}^{4}\frac{1}{10}N(i/2-1,\left(\frac{1}{10}\right)^2)$
Strongly skewed	$\sum_{i=0}^{7}\frac{1}{8}N(3\{\left(\frac{2}{3}\right)^i-1\},\left(\frac{2}{3}\right)^{2i})$
Discrete	$P(\pm 3)=.033$; $P(\pm 2\frac{1}{3})=.067$; $P(\pm 1\frac{2}{3})=.1$; $P(\pm 1)=.133$; $P(\pm\frac{1}{3})=.167$

Figure 1. Histograms of logged adoption visa data. All histograms have the same bin width $h = .276$, with different anchor positions.

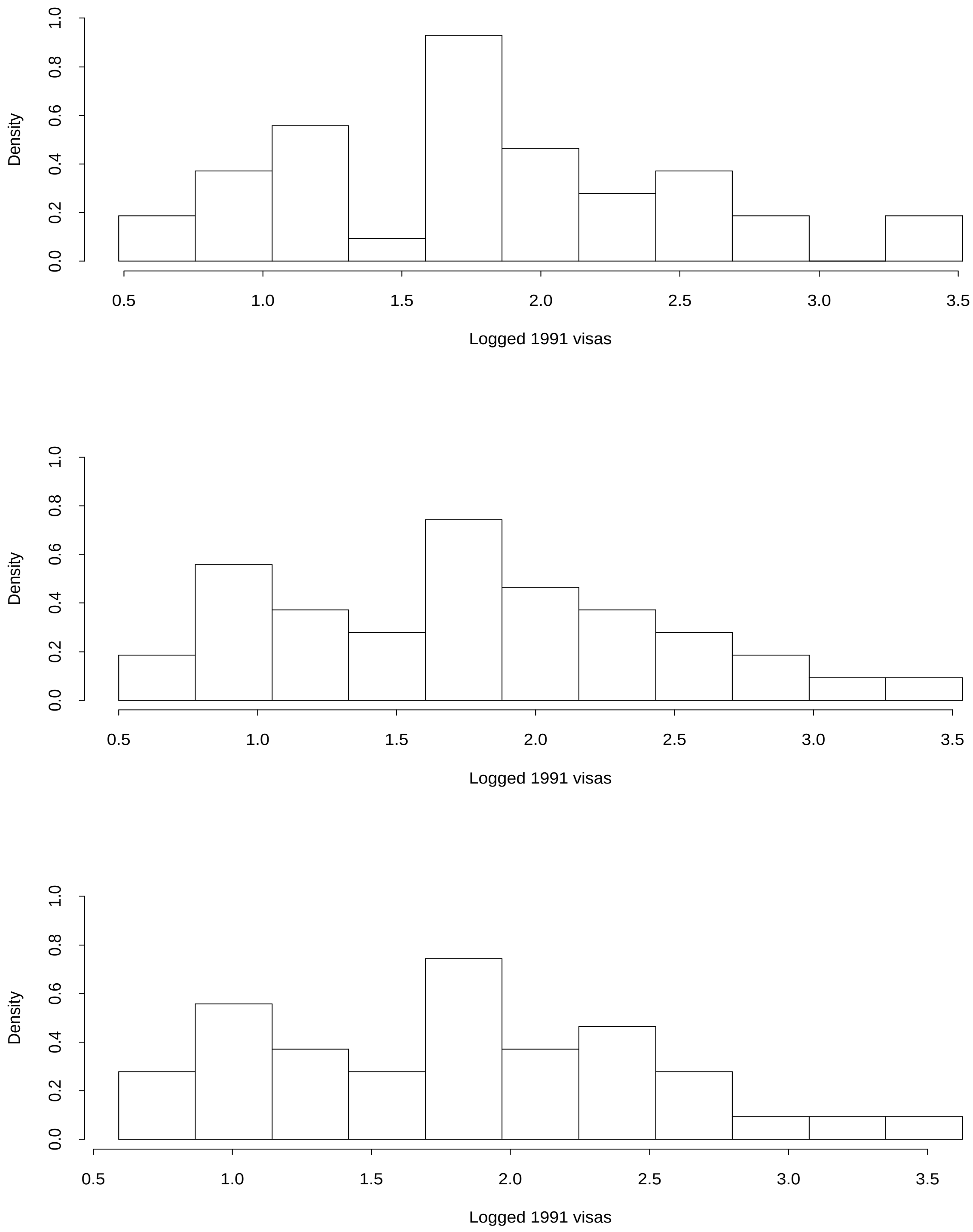


Figure 2. Stability index plot for Gaussian density. $N = 20$ (solid line), $N = 100$ (dotted line), $N = 500$ (dashed line).

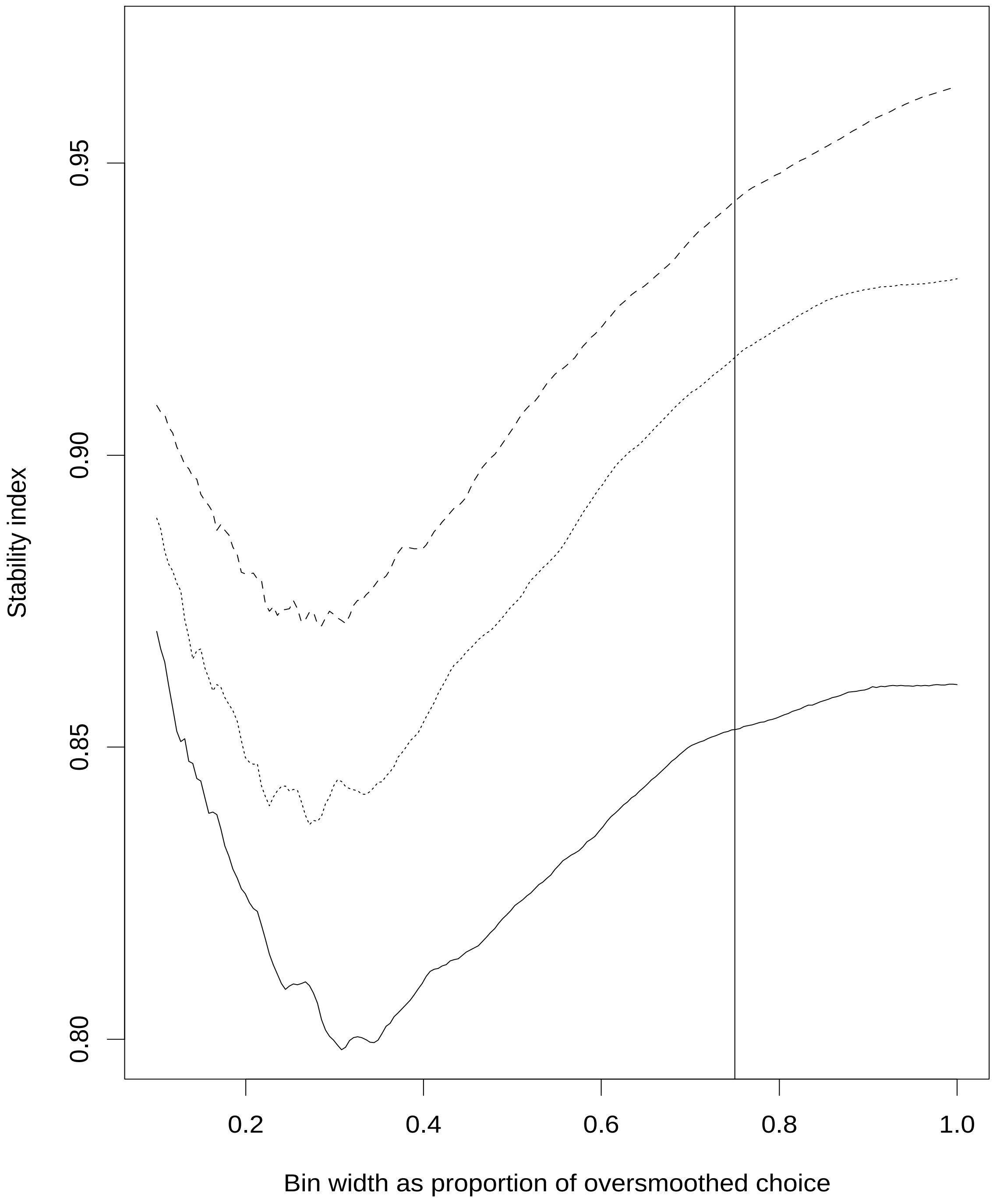
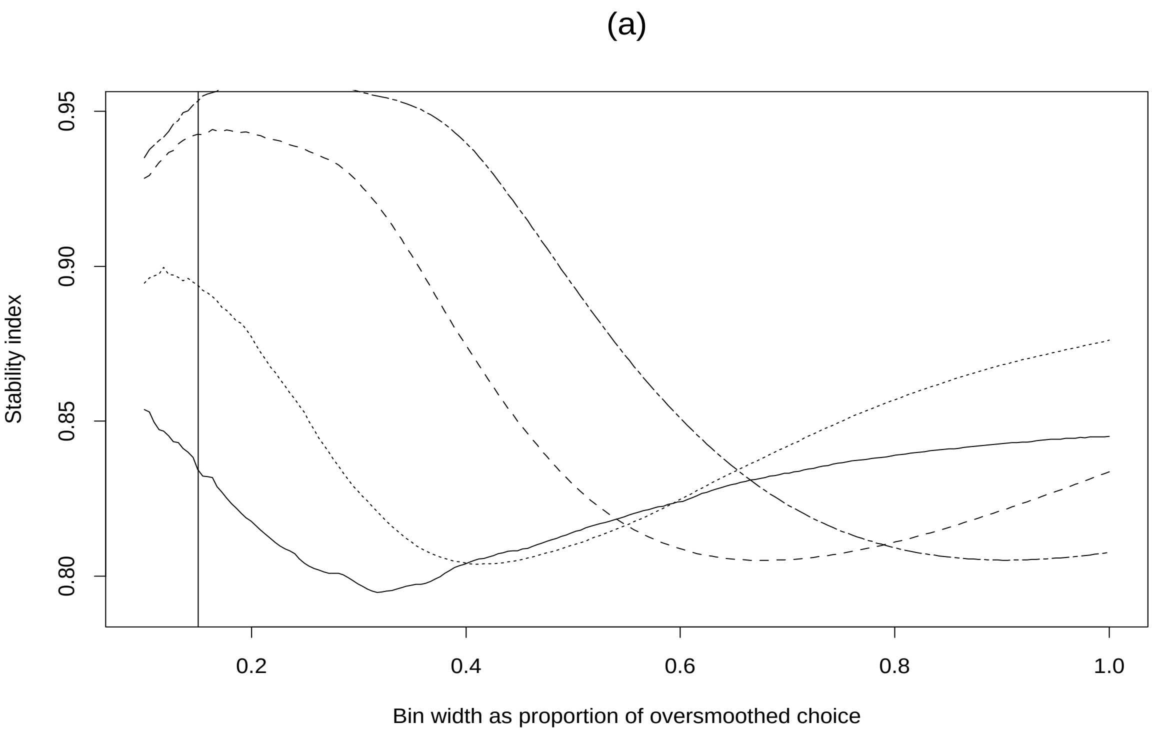


Figure 3. Stability index plots for kurtotic unimodal density. $N = 20$ (solid line), $N = 100$ (dotted line), $N = 500$ (dashed line), $N = 1000$ (dotted and dashed line).



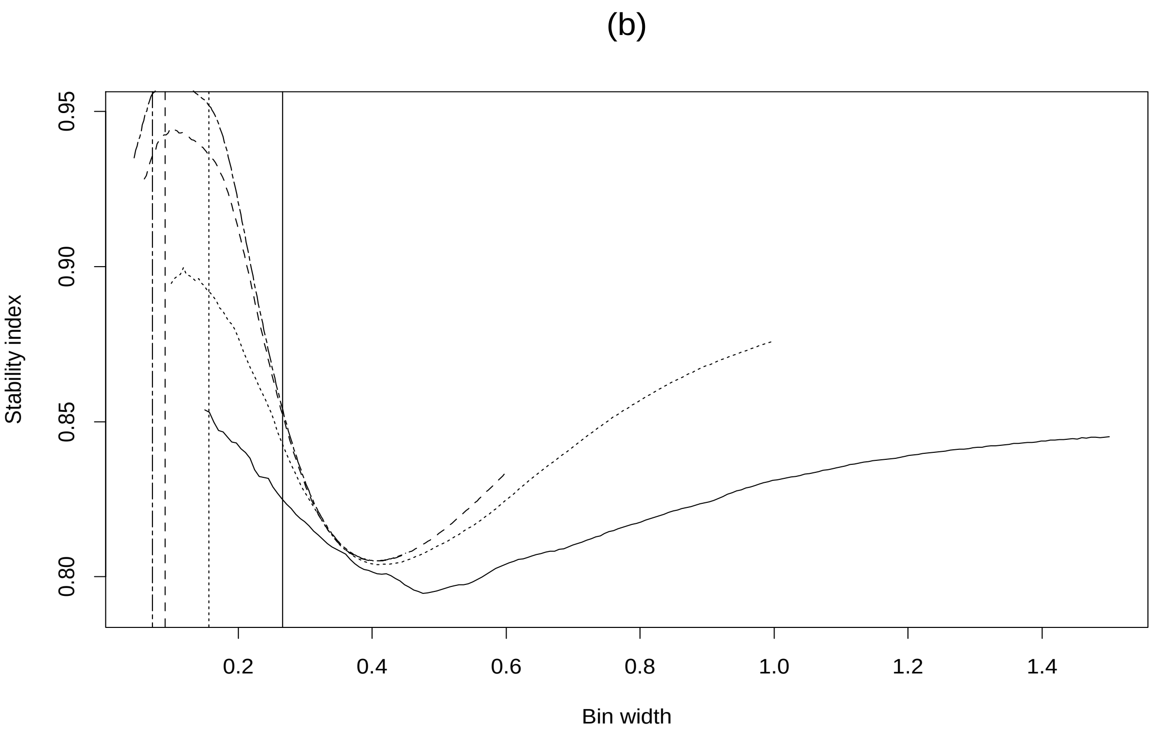


Figure 4. Stability index plot for "shoulders" density.

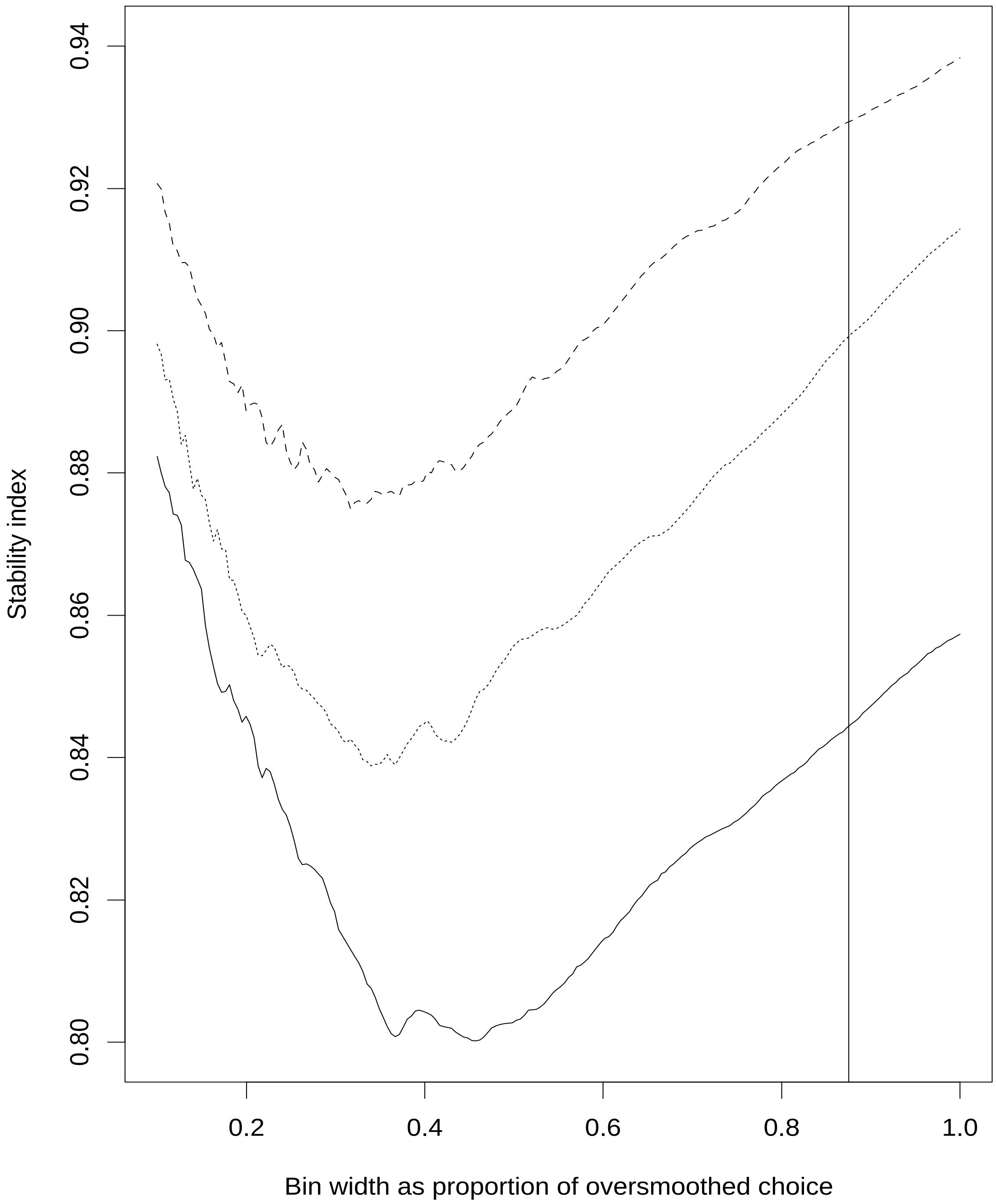


Figure 5. Stability index plots for trimodal density.

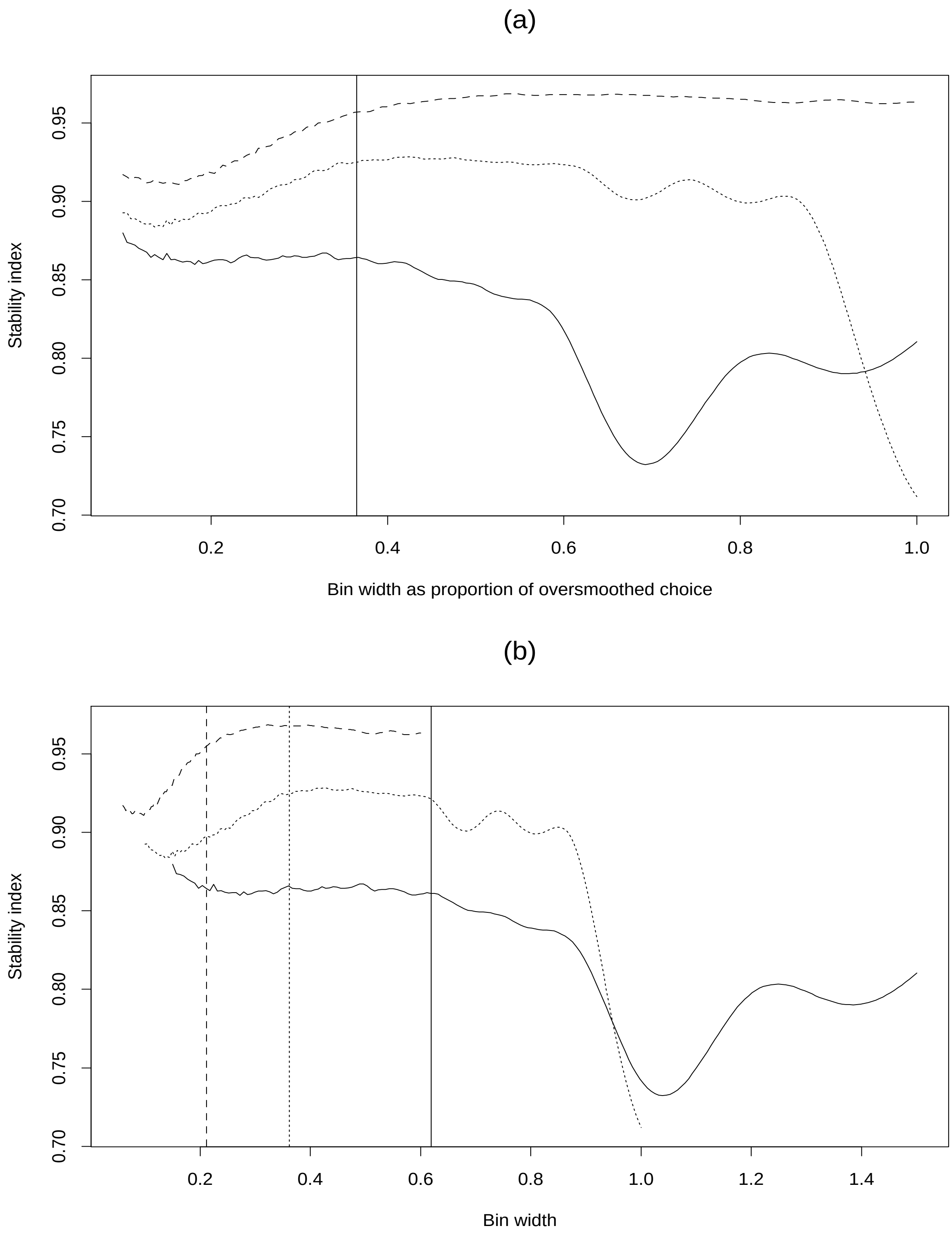


Figure 6. Stability index plots for claw density.

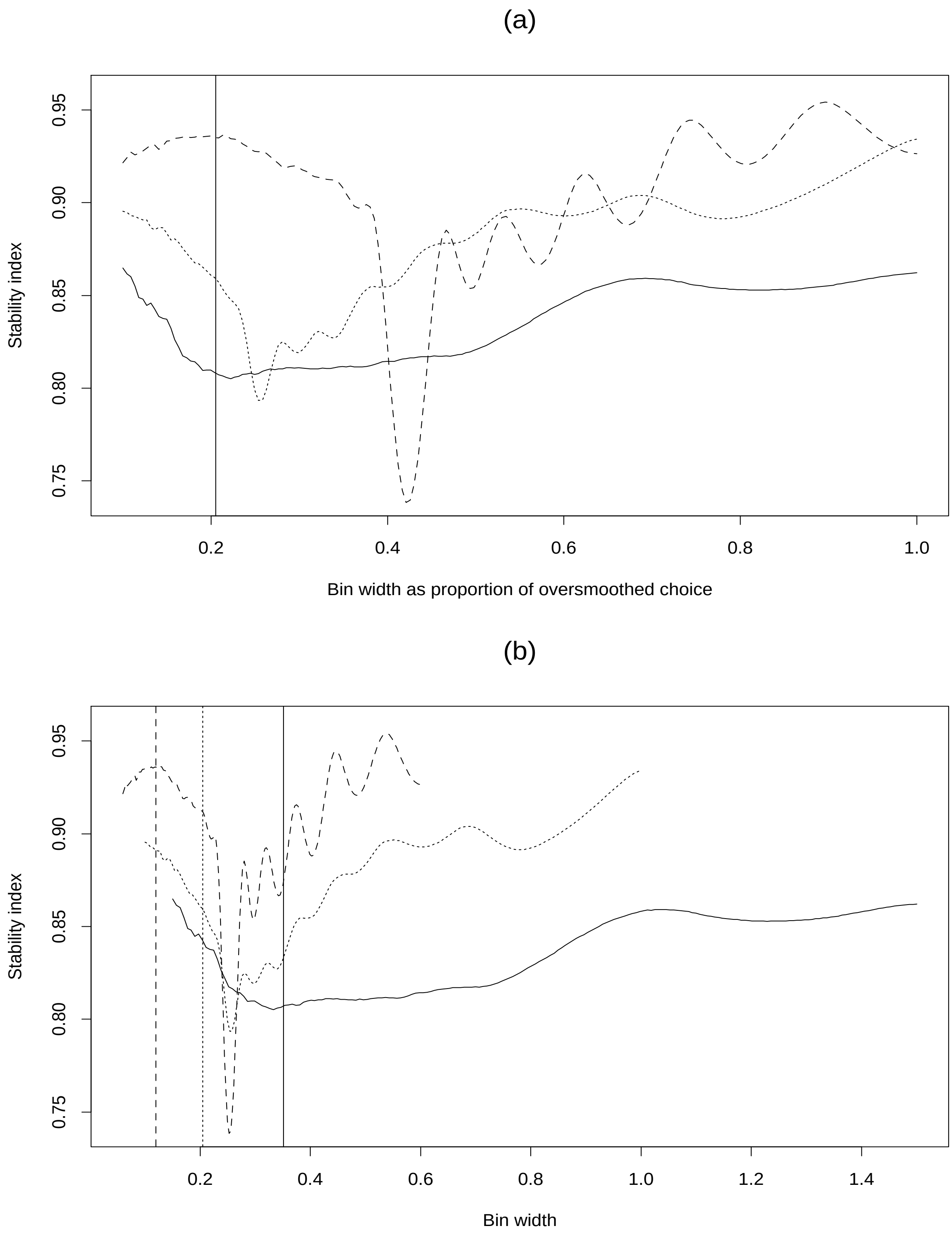


Figure 7. Stability index plot for strongly skewed density.

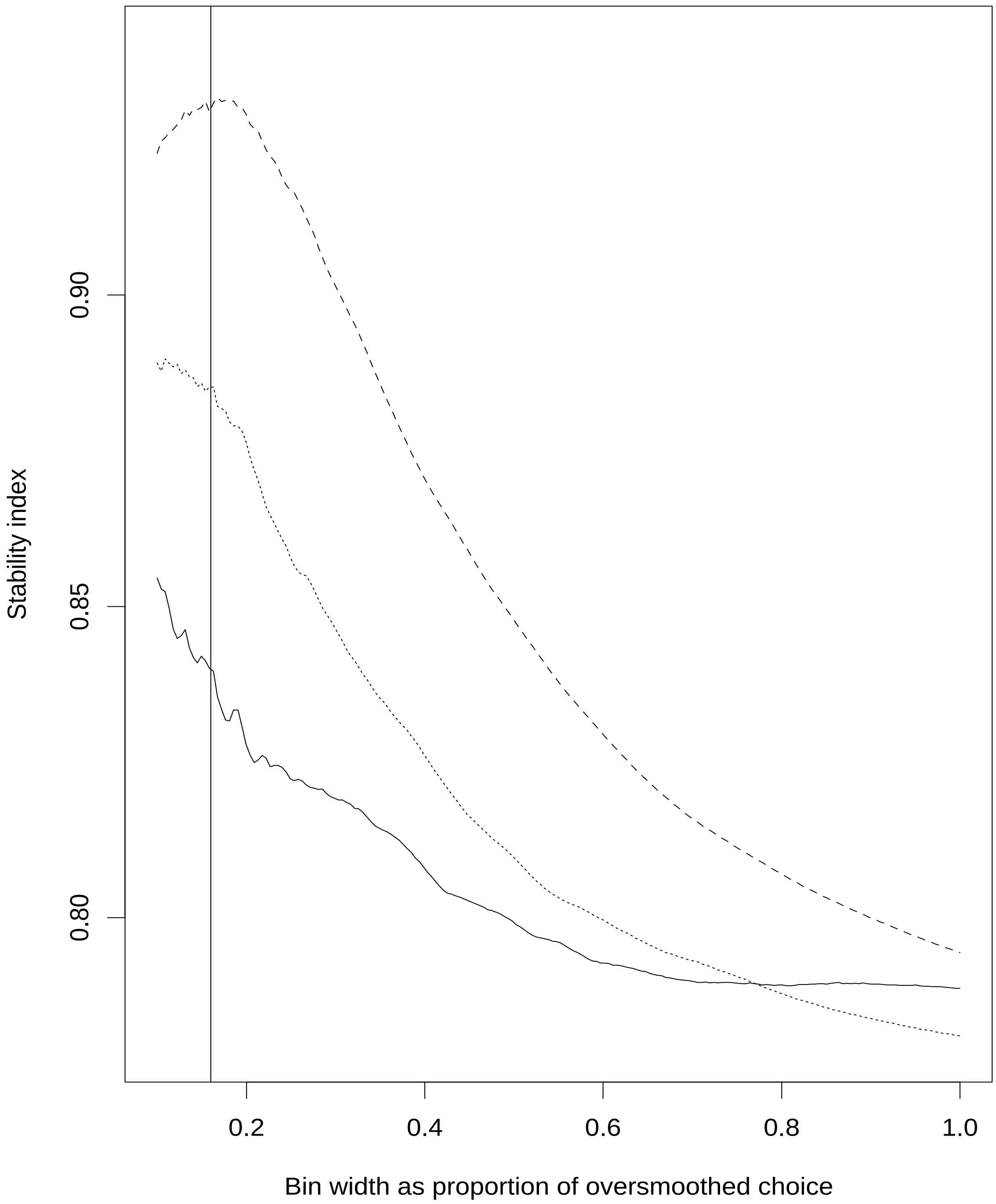
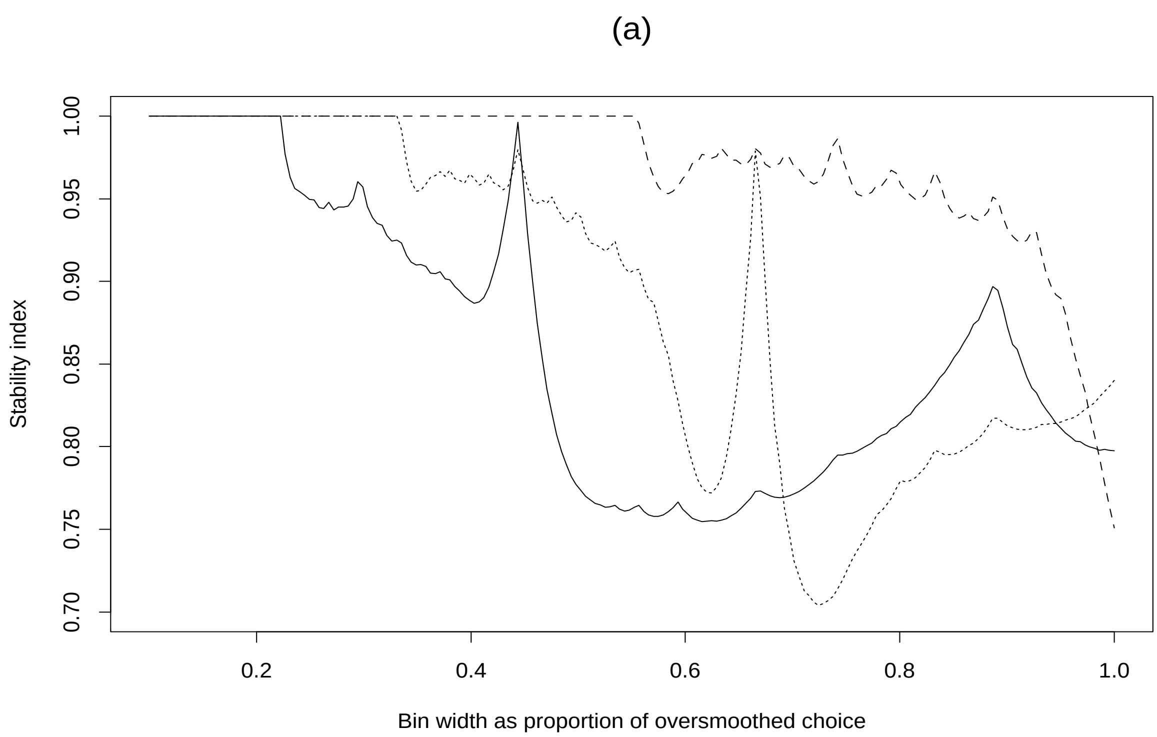


Figure 8. Stability index plots for discrete distribution.



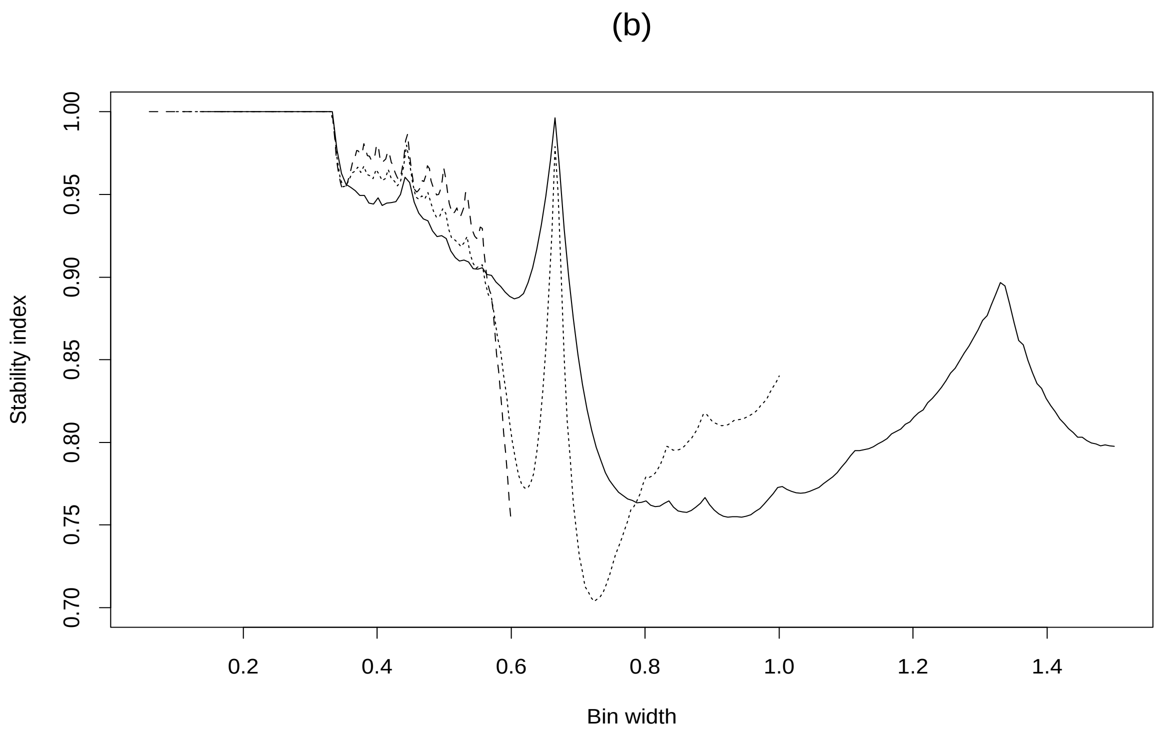


Figure 9. Stability index plot for "Old Faithful" geyser data. Shaded areas correspond to bin widths with low simulated evidence level.

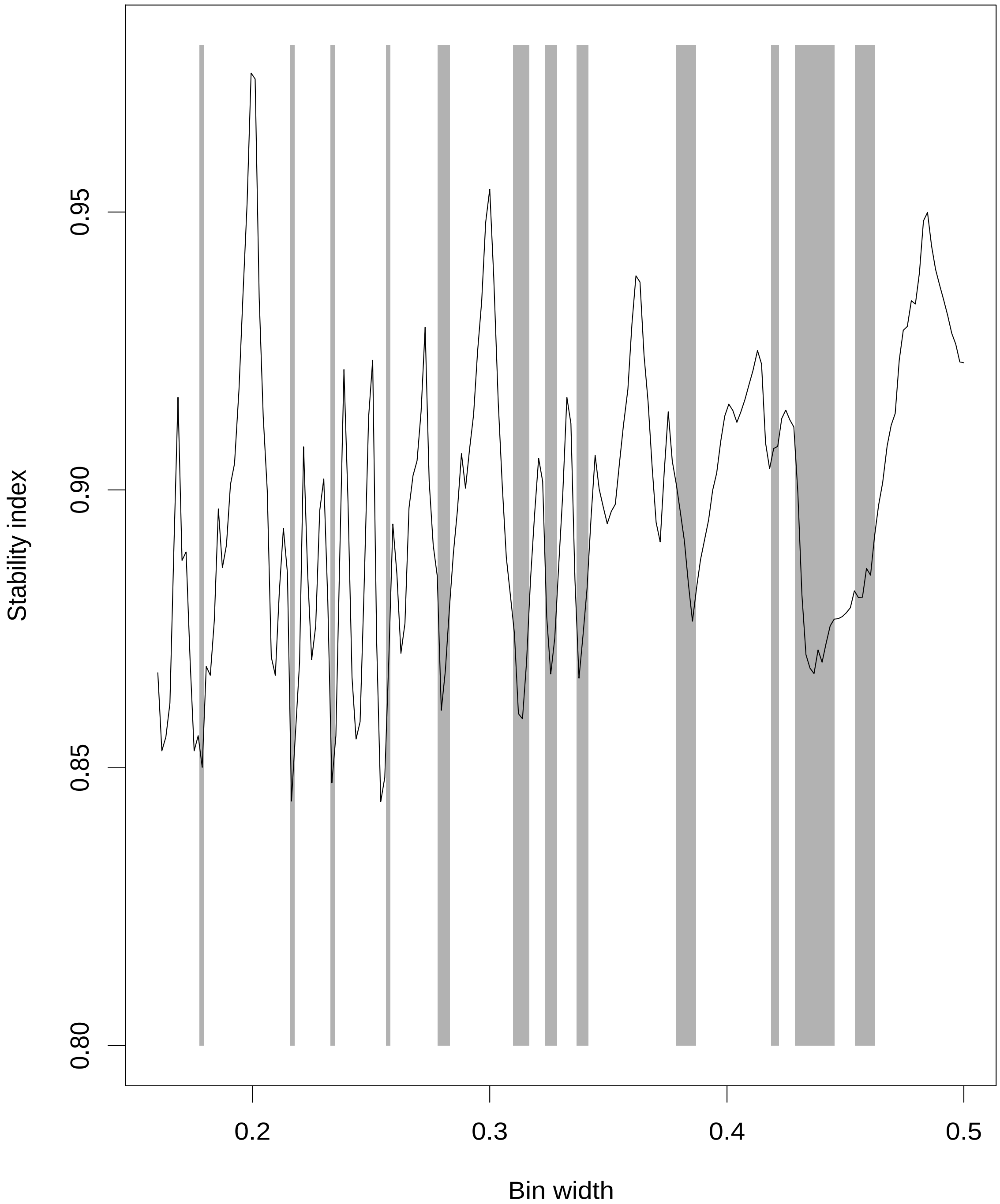


Figure 10. Stability index plot for PCB data.

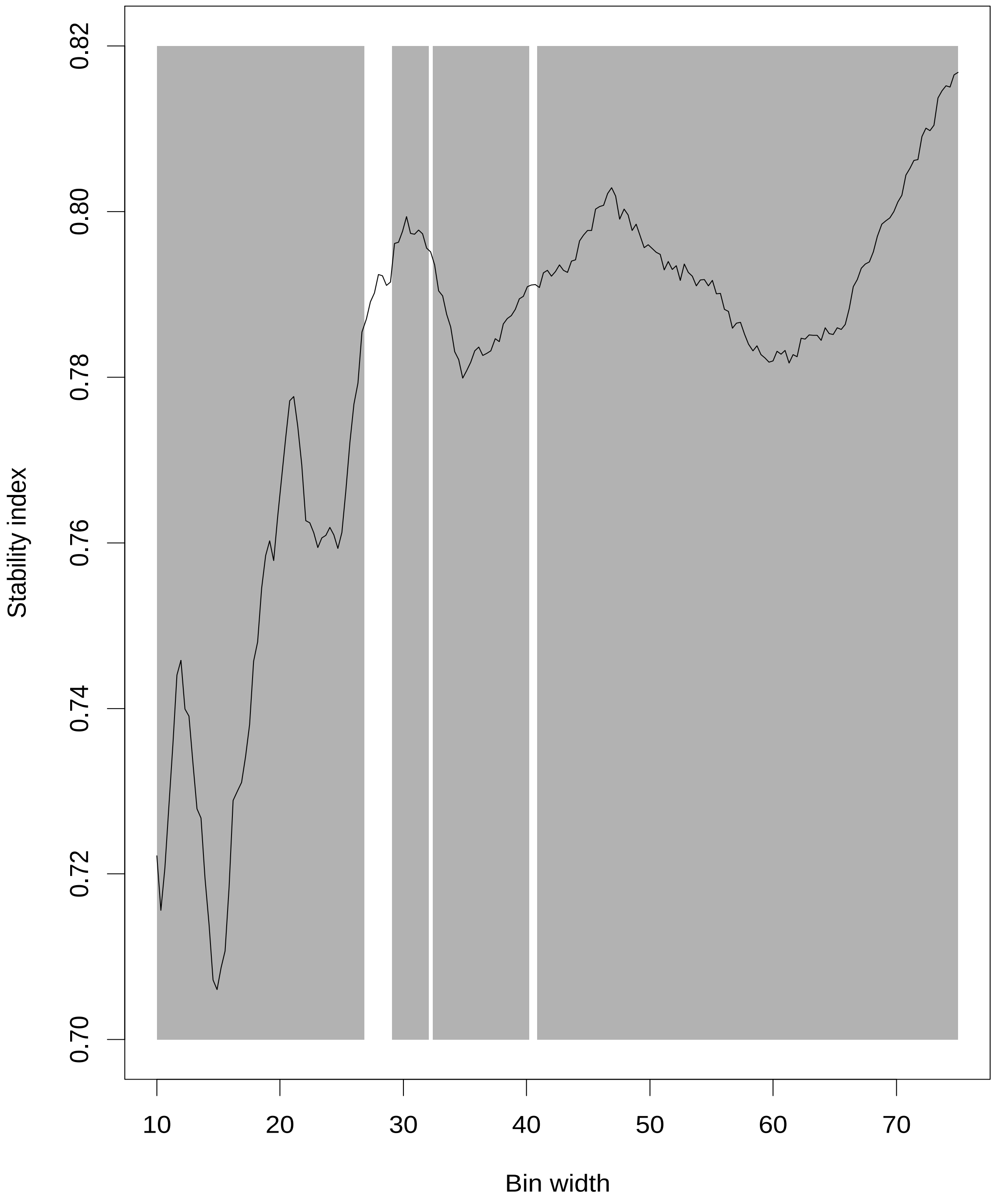


Figure 11. Stability index plot for logged adoption visa data.

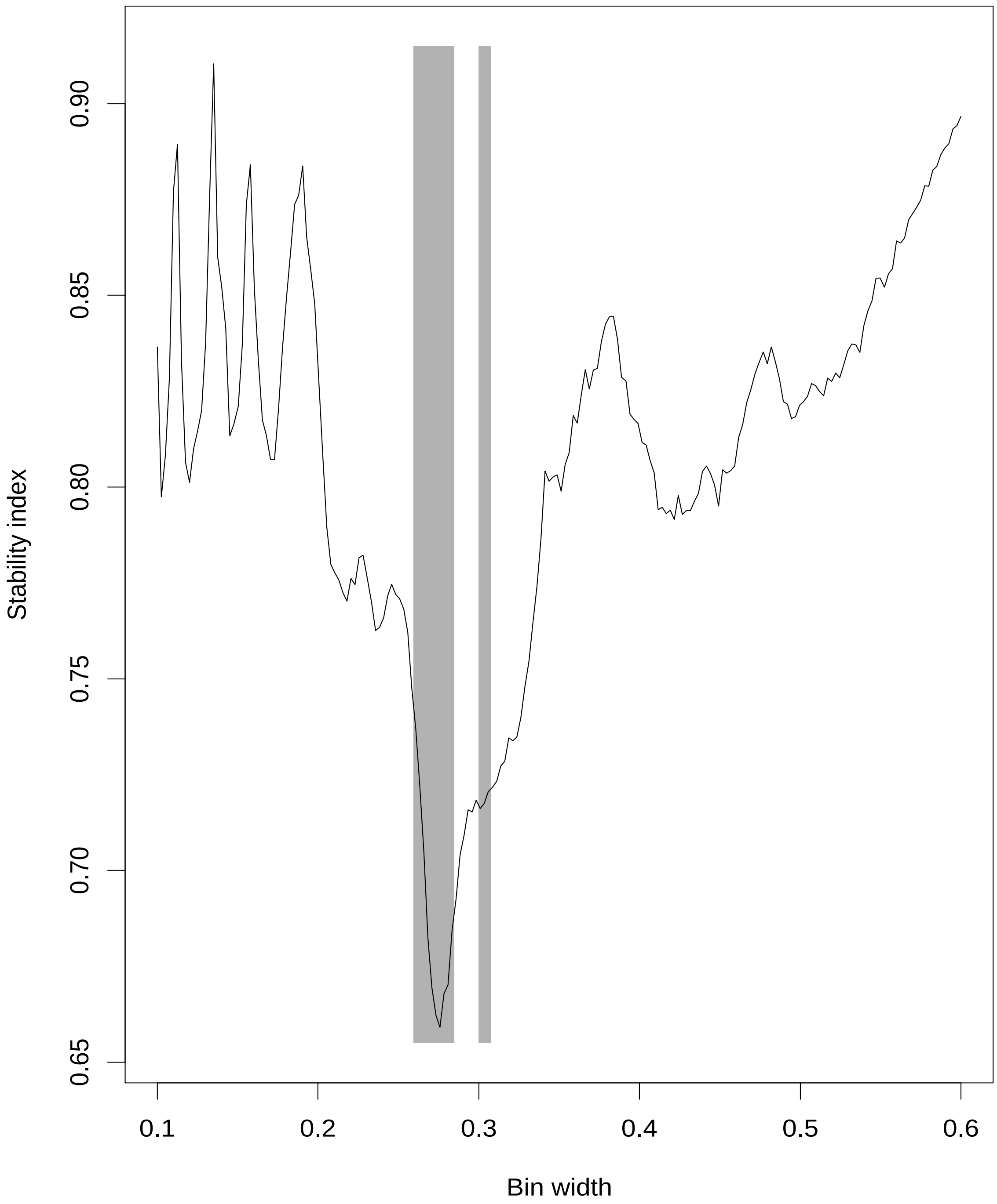


Figure 12. Stability index plot for NBA age data.

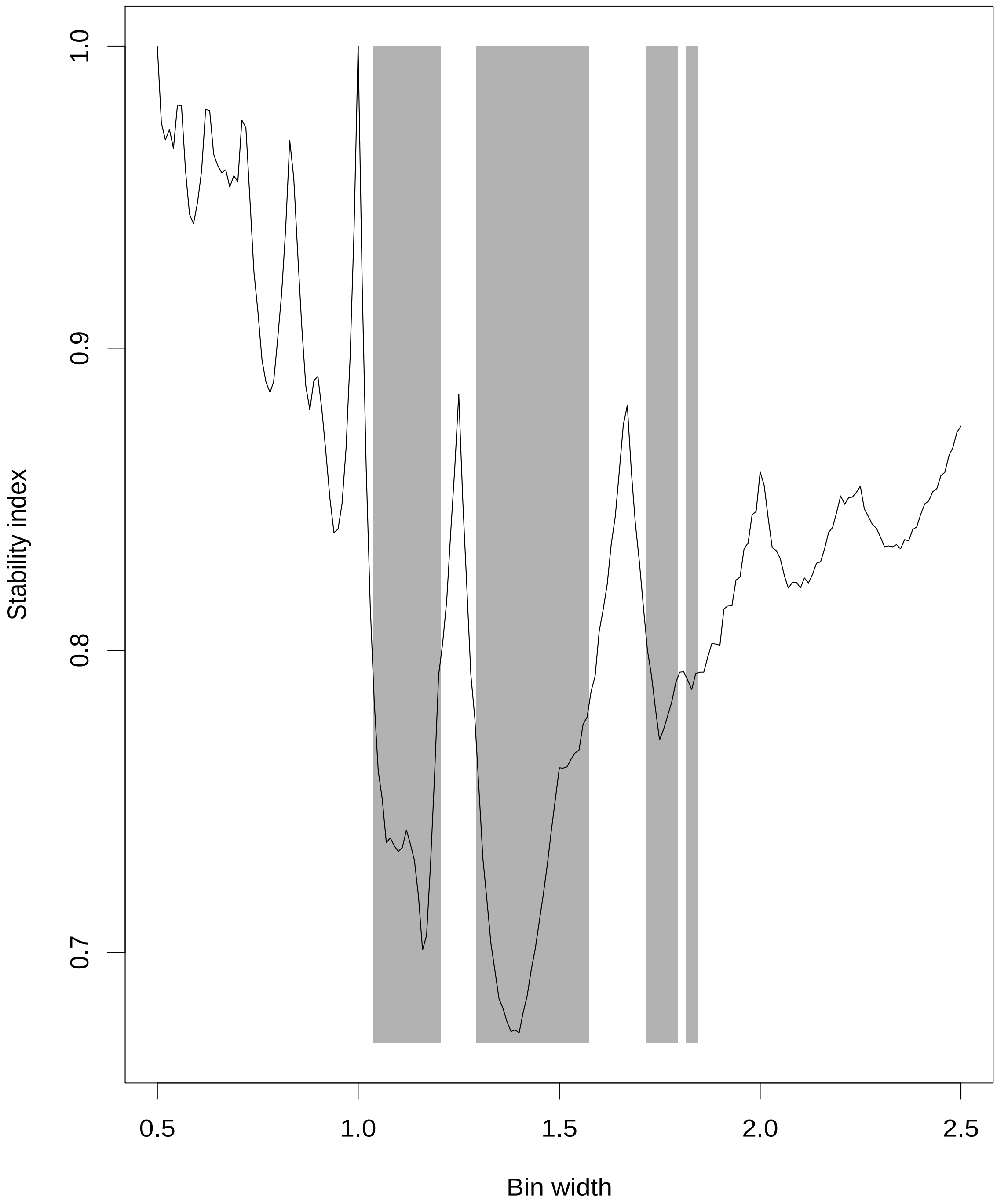


Figure 13. Histograms of NBA age data. All histograms have bin width $h = 1.38$, with different anchor positions.

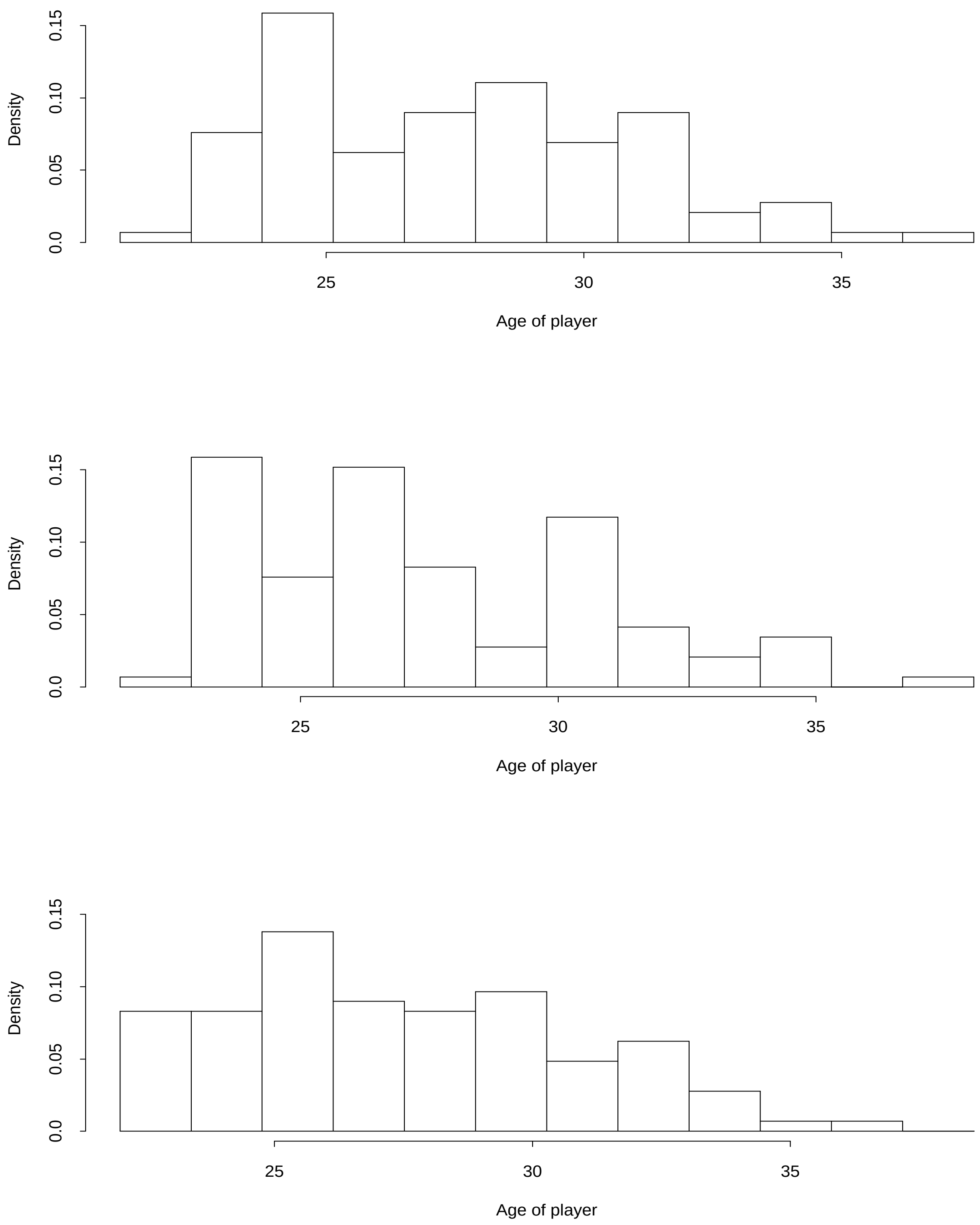


Figure 14. "Natural" histogram of NBA age data with $h = 1$.

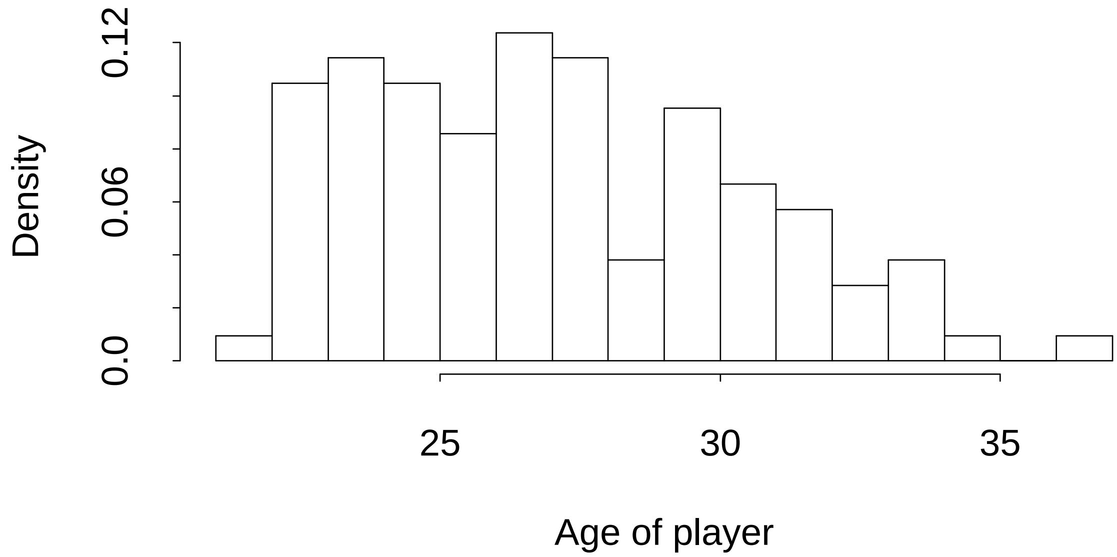


Figure 15. Stability index plots for frequency polygon estimators for (a) geyser eruption data, and (b) logged adoption visa data. Frequency polygon (dotted line), average frequency polygon (dashed line), linearly binned frequency polygon (dotted and dashed line).