```
[1] 0.07293381
```

THE SHORTH PLOT

TECHNICAL REPORT

GÜNTHER SAWITZKI

Abstract. We introduce the shorth plot for exploratory diagnostics of distributions. The shorth plot is a graphical representation of the length of the shorth, the shortest interval covering a certain fraction of the distribution. Localising the shorth, i.e. requiring it to contain specific data points, makes it usable for diagnostics.

The shorth can be defined as a functional which has an immediate empirical version. The empirical length of the shorth converges to the theoretical value with rate $n^{-\frac{1}{2}}$.

Contents

Date: June 1992, published as technical report [Saw92].
For quotation, please use the general reference [Saw94].
Revised: August 2007
Typeset, with minor revisions: October 7, 2007.
Key words and phrases. shorth, distribution diagnostics, data analysis, probability mass concentration.
Private Version

.

1. Distribution Diagnostics

Exploratory diagnostics is one of the basic tasks in data analysis. Graphical displays are essential tools. If the task can be narrowed down, specialised displays may be available. For example, if there is a model distribution F to be compared with, from a mathematical point of view the empirical distribution F_n is a key instrument, and its graphical representations such as PP-plots

$$x \mapsto (F(x), F_n(x))$$

or QQ-plots

$$\alpha \mapsto \left(F^{-1}(\alpha), F_n^{-1}(\alpha)\right)$$

are tools of first choice. If we consider the overall scale and location, box & whisker plot is a valuable tool. This tool loses its sharpness for large data sets, as the rules for identifying "out" and "far out" points seem to reflect data sizes which were typical at the time box & whisker plots were introduced, but this is a detail which might be fixed. The main limitation is that box & whisker plots give a global view and ignore any local structure. In particular, they are not an appropriate tool if it comes to analyse the modality of a distribution. More specialised tools are needed in this case, such as the silhouette and the excess density plot, both tools introduced in [MS91].

Here we look for general purpose tools for the analysis of a distribution. While we have some instruments for specific tasks, the situation is not satisfactory if it comes to general purpose tools. PP-plots and QQ-plots need considerable training to be used as diagnostic tools, as they do not highlight qualitative features.

Focussing on the density, in contrast to the distribution function, leads to density estimators and their visual representations, such as histograms and kernel density plots. These however introduce another complexity, such as the choice of cut points or bandwidth choice. The qualitative features revealed or suggested by density estimation based methods may critically depend on bandwidth choice. Moreover, estimating density is a more specific task than understanding the shape of a density. Density estimation based methods are prone to pay for the initial smoothing steps in terms of slow convergence or large fluctuation, or disputable choices of smoothing.

We will use the length of the shorth to analyse the qualitative shape of a distribution. The shorth is the shortest interval containing half of a distribution. The length of the shorth is a functional which is easy to estimate, with convergence of rate $n^{-\frac{1}{2}}$, and gives a graphical representation which is easy to interpret.

We will start with the classical definition of a shorth. To overcome the handicaps of other methods, we have to extend the classical definition to supply localisation, and to allow for multi scale analysis.

2. The Length of the Shorth

The shorth is the shortest interval containing half of a distribution. More general, the α-shorth is the the shortest interval containing an α fraction of the distribution. The shorth was originally introduced in the Princeton robustness study as a candidate for a robust location estimator, using the mean of a shorth as an estimator for a mode [ABH$^+$72].

As a a location estimator, it performed poorly. The mean of a shorth as an estimator of location has an asymptotic rate of only $n^{-1/3}$, with non-trivial limiting distribution. [ABH$^+$72, p. 50] [SW86, p. 767]. Moreover, the shorth interval is not well defined, since there may be several competing intervals.

However, as Grübel [Grü88] has pointed out, the length of the shorth has a convergence rate of $n^{-\frac{1}{2}}$ with a Gaussian limit. The critical conditions are that the shorth interval is sufficiently pronounced (see [Grü88, section 3.3]). Essentially this means that the shorth interval must not be in a flat part of the distribution. While the shorth position is not a good candidate as a location estimator, the length of the shorth qualifies as a reasonable candidate for scale estimation.

The length of the shorth is a functional which can be localised, thus providing a tool for local diagnostics. We define:

Definition 1. The shorth length at point x for coverage level α is

$$S_\alpha(x) = \min\{|I| : I = [a, b],\ x \in I,\ P(I) \geq \alpha\}.$$

We get the length of the shorth as originally defined by taking $\inf_x S_{0.5}(x)$.

The definition has a functional form which can be applied to theoretical as well as empirical distributions. The definition in terms of a theoretical probability $P(\,\cdot\,)$ has an immediate empirical counterpart, the empirical length of the shorth

$$S_{n,\alpha}(x) = \min\{|I| : I = [a, b],\ x \in I,\ P_n(I) \geq \alpha\}$$

where $P_n(\,\cdot\,)$ is the empirical distribution.

The get a picture of the optimisation problem behind the shorth length, we consider the bivariate function

$$a, b \longmapsto I = [a, b] \longmapsto \big(|I|, P(I) \big) \qquad \text{where } a \le b.$$

This is defined in the half space $a \le b$ above the diagonal. The level curves of $|I|$ are deterministic parallels to the diagonal. The level curves of $P(I)$ depend on the distribution. The shorth at level α minimises $|I|$ in the area above the level curve at level α, i.e. $P(I) \ge \alpha$. Going to the empirical version replaces the level curves of of $P(I)$ by those of of $P_n(I)$. The theoretical curves for the Gaussian distribution and for a Gaussian sample are shown in figure 1. Localising the shorth at a point x restricts optimisation to the top left quadrant anchored at $a = b = x$.

paragraph added

The increasingly flat level curves for $P([a, b]) = const$ in figure 1 illustrate why the location of the shorth does not have satisfactory stattistical properties while the length of the shorth has good asymptotic behaviour.

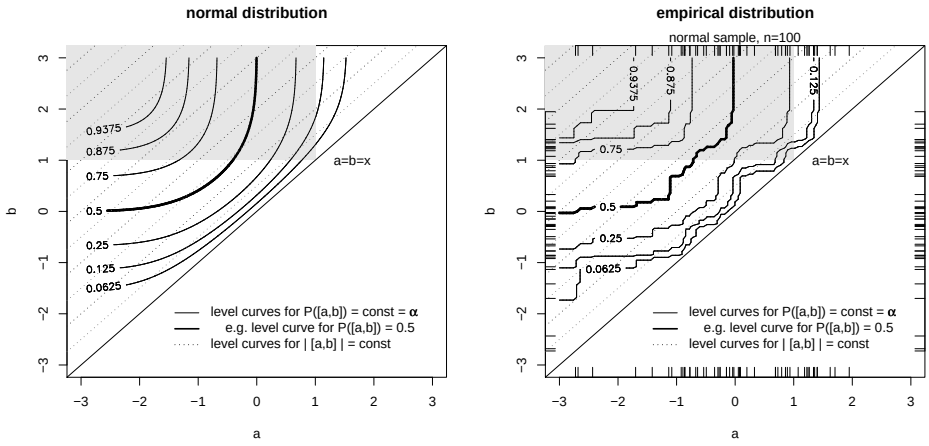


FIGURE 1. The shorth length as an optimisation problem: minimize $|[a, b]|$, under the restriction $P([a, b]) \ge \alpha$. Localising at x restricts the optimisation to the quadrant top left of x (shaded grey).

The inverse optimisation problem is to find the maximal coverage which can be achieved by a given length:

$$\delta \longmapsto \sup_y \{P(I) : |I| \ge \delta\}$$

together with its localised and/or empirical version. The relation between the shorth length and its inverse is extensively used in Grübel's analysis of the asymptotics of the (unlocalised) shorth. A modified version of Grübel's proof carries over to the localised shorth, provided

there are no flat parts in the distribution, giving a $n^{-\frac{1}{2}}$ asymptotics of the empirical shorth length to the theoretical shorth length.

ToDo: extend discussion of concentration function

2.1. **Elementary Properties.** Here and in the following, we assume a sample $X_1, \ldots, X_n$ from some distribution P with distribution function F. Let P_n be the empirical distribution and F_n the empirical distribution function. The k-th order statistics is denoted by $X_{(k)}$.

ToCheck: ok?

Remark 1. *(invariance)* For all α

$$x \mapsto S_\alpha(x)$$

is invariant under shift transformations and equivariant under scale transformations, that is for $x' = ax + b$

$$S'_\alpha(x') = aS_\alpha(x).$$

Remark 2. *(monotonicity)* For all x,

$$\alpha \mapsto S_\alpha(x)$$

is monotonously non decreasing in α.

If F is continuous with density f, additional properties are guaranteed:

Remark 3. *(minimising intervals)* If F is continuous, then for any $\alpha \in [0, 1]$ and any $x \in \mathbb{R}$, there is a (possibly infinite) interval I such that $x \in I$, $P(I) = \alpha$ with $|I| = S_\alpha(x)$.
If $\alpha < 1$, the interval is finite and contained in $[x - S_\alpha(x), x + S_\alpha(x)]$.

ToDo: add proof concl. 5

Remark 4. *(continuity)* For a continuous distribution function F,

$$(x, \alpha) \mapsto S_\alpha(x)$$

is continuous.

Remark 5. *(strict monotonicity)* For a continuous distribution function F, for each $x \in \mathbb{R}$,

$$(x, \alpha) \mapsto S_\alpha(x)$$

is strictly increasing in α on $(0, 1)$.

In the limit, $\lim_{\alpha \to 0} S_\alpha(x) = 0$.

In particular, for the empirical version, $S_{n,\alpha}(x) = 0$ for $\alpha \leq \frac{1}{n}$.

2.2. **Computing the Empirical Shorth Length.** To use empirical distribution functions, the discontinuous case is of interest.

Remark 6. *(interpolation)* If $x_0 < x < x_1$ and $P(\ (x_0, x_1)\ ) = 0$, then

$$S_\alpha(x) = \big(S_\alpha(x_0) + \Delta_0\big) \wedge \big(S_\alpha(x_1) + \Delta_1\big),$$

where $x = x_0 + \Delta_0 = x_1 - \Delta_1$.

Remark 7. *(algorithm)* For $\alpha, 0 \le \alpha \le 1$, let

$$\Delta_\alpha = \min\left\{k : \frac{k+1}{n} \ge \alpha\right\}.$$

Then

$$S_{n,\alpha}(X_i) = \min\{X_{(j+\Delta_\alpha)} - X_{(j)} :\ 1 \le j \le i \le j + \Delta_\alpha \le n\}.$$

Using a stepwise algorithm, a further reduction of complexity is possible: let

$$\mathscr{C}_i := \{I : X_{(i)} \in I, P(I) \ge \alpha\}$$

be the set of candidate intervals at $X_{(i)}$ for level α. Then

$$S_{n,\alpha}(X_i) = \min\{|I| : I \in \mathscr{C}_i\}.$$

Unless boundary corrections apply, we have

$$\mathscr{C}_i = \mathscr{C}_{i-1} \setminus \{[X_{(i-1-\Delta_\alpha)}, X_{(i-1)}]\} \cup \{[X_{(i)}, X_{(i+\Delta_\alpha)})]\}.$$

This gives an algorithm with linear complexity in n.

An additional reduction is possible using the monotonicity in α (remark 5), but this may not be worth the effort.

3. The Shorth Plot

Definition 2. The shorth plot is the graph of

$$x \mapsto S_\alpha(x)$$

for a selection of coverages α.

The empirical shorth plot is

$$x \mapsto S_{n,\alpha(x)}.$$

See figure 2.

Mass concentration now can be represented by the graph of $x \mapsto S_\alpha(x)$. A small length of the shorth signals a large mass concentration. To make the interpretation easier, we prefer to invert the orientation of the axis so that it is aligned with density axis. This will be used in the subsequent figures.

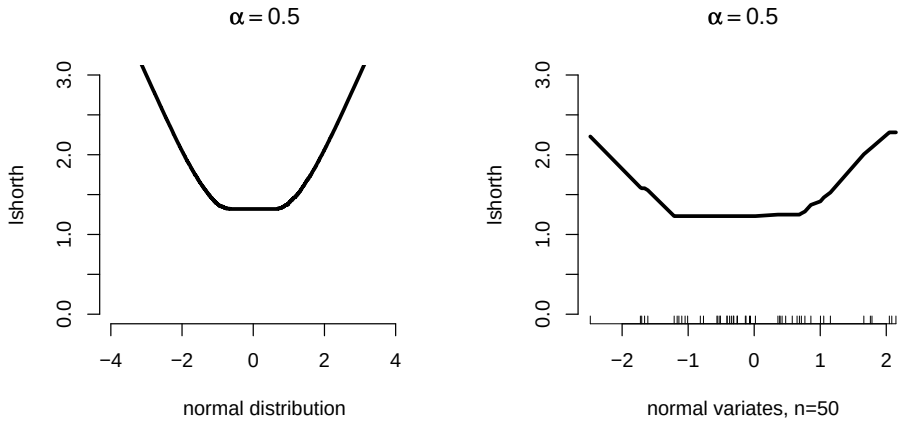


FIGURE 2. Theoretical shorth length and shorth length for a sample of 50 normal variates for $\alpha = 0.5$.

3.1. **Display Details.** Several choices can to be made for the visual representation. The common conception seems to view a distribution represented by its density. From a mathematical point of view, plotting $x \mapsto 1/S_\alpha(x)$ would be first choice, since this is approximately proportional to the local average density. This however is an infinitesimal approximation. It tends to overemphasise peaks (see figure 3), and becomes useless for point masses. Using just a downward orientation for the y-axis avoids the need to special case point masses while keeping the qualitative impression.

To make comparison between different data sets easier, we can use the classical shorth length $\min_x S_{0.5}(x)$ as a scale estimator and remove the scale dependency by taking the quotient. This quotient

$$x \mapsto \frac{S_\alpha(x)}{\min_{x'} S_{0.5}(x')}$$

is called the standardised short length . The only difference is the scale labeling. In these notes, we do not use standardized shorth length, but use the original scales.

Instead of the exact interpolation as in remark 6, we use a linear interpolation. The loss of information is negligible.

Figure 4 gives examples of the shorth plot for the uniform, normal and log-normal distribution with varying sample sizes.

ToDo: replace last line by theoretical plots

Varying the coverage level α as in figure 5 gives an impression of the mass concentration. Small coverage levels (the top curves in figure 5) give information about the local behaviour, in particular near modes.

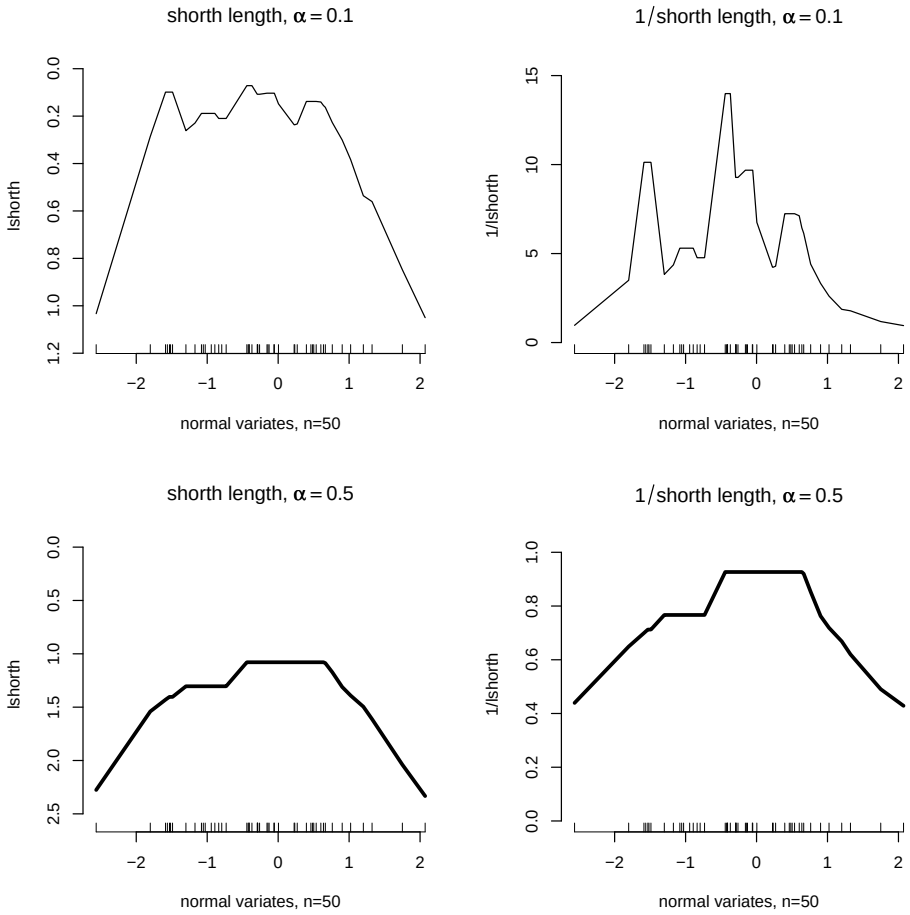


FIGURE 3. Shorth length and 1/Shorth length for a sample of 50 normal variates. The axis for the shorth is pointing downward. Note: different scales are used.

High coverage levels give information about skewness the overall distribution shape. A dyadic scale with steps chosen based on the sample size, e.g., $0.125, 0.25, 0.5, 0.75, 0.875$, is a recommended choice. The monotonicity (remark 2) allows the multiple scales to be displayed simultaneously without overlaps, thus giving a multi resolution image of the distribution.

ToDo: replace 10000 by theor

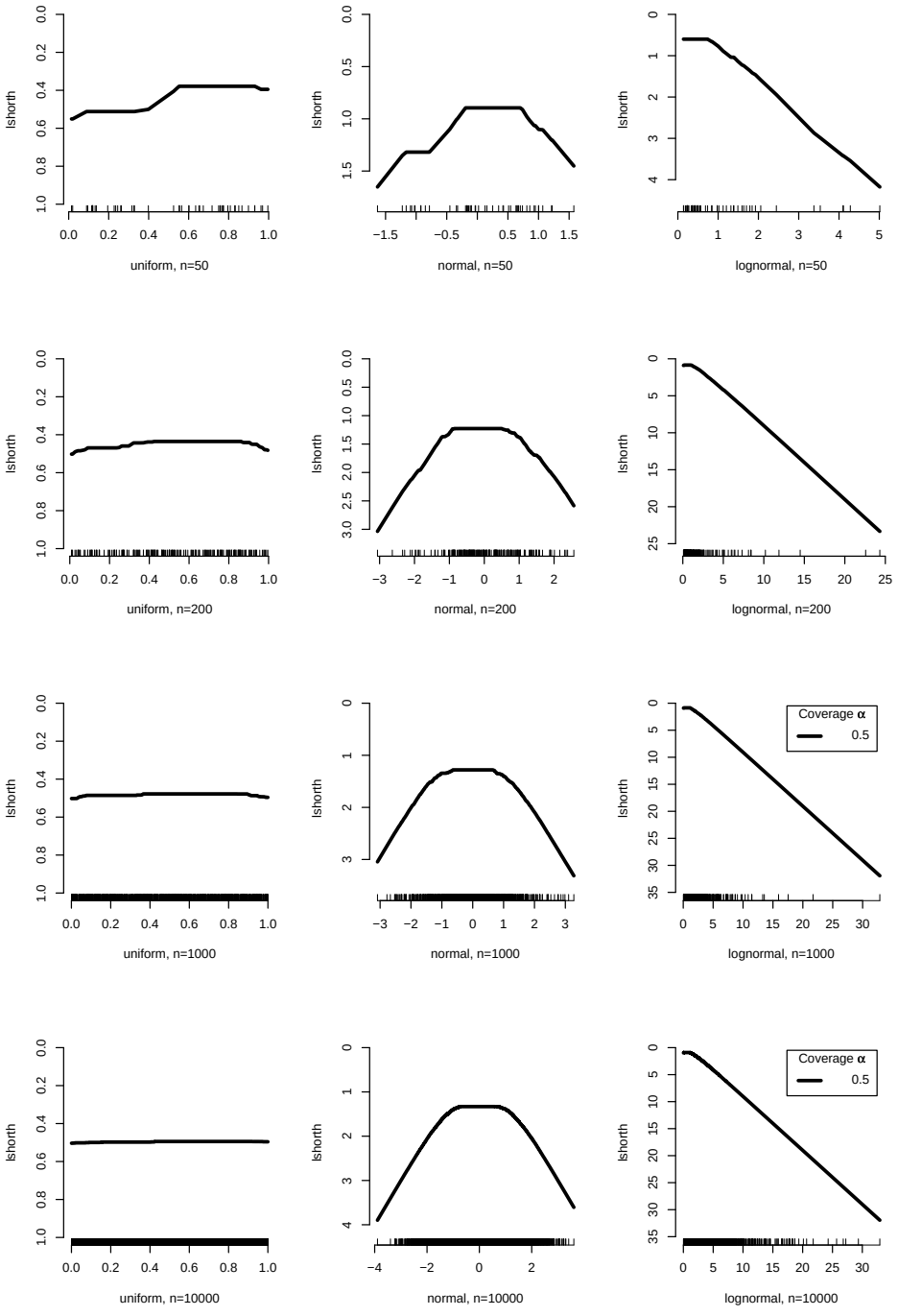


FIGURE 4. $x \mapsto S_{n,0.5}(x)$ for a uniform, a normal, and a log-normal distribution with varying sample sizes. Note: Different scales are used for the shorth length.

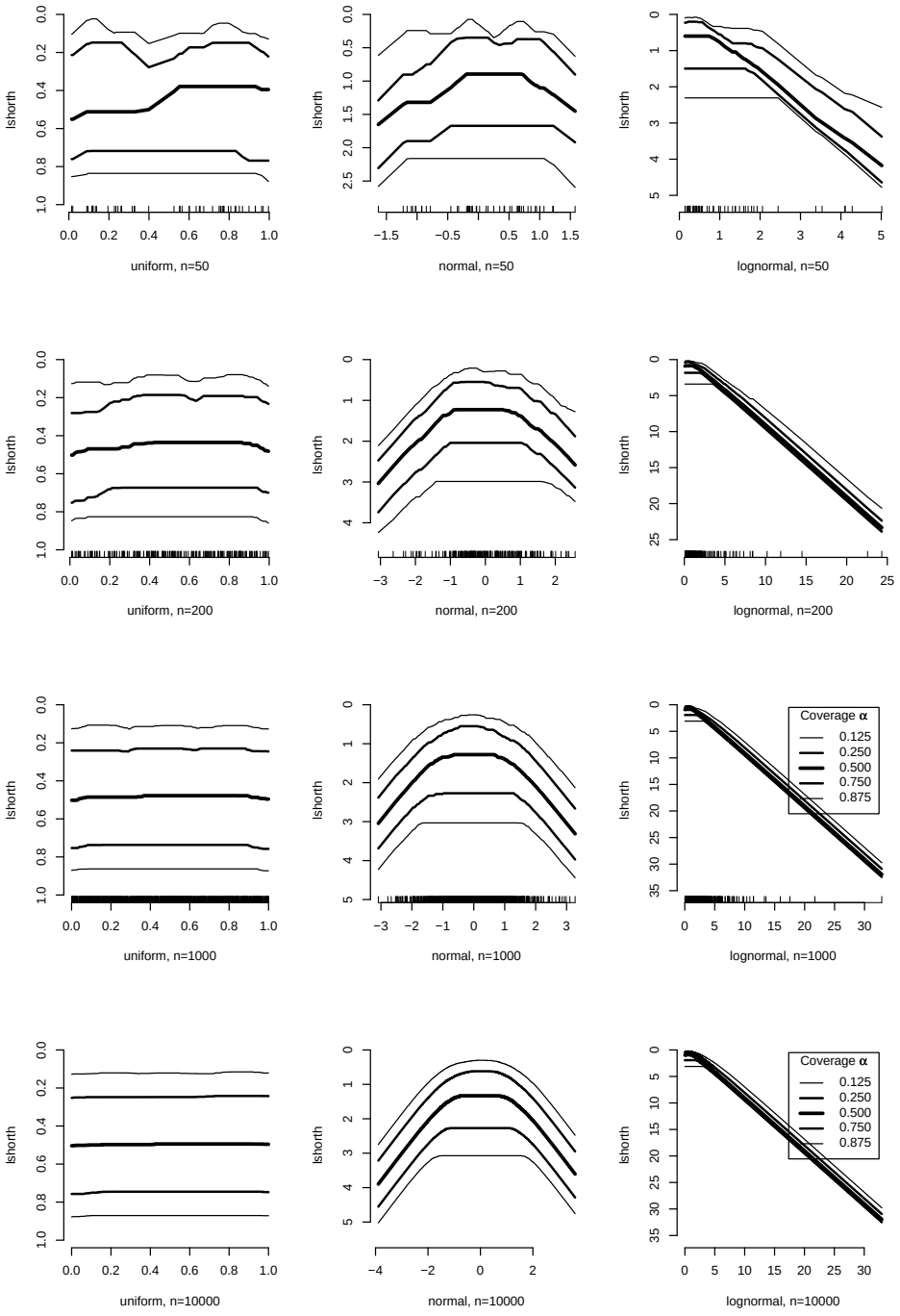


FIGURE 5. $x \mapsto S_{n,\alpha}(x)$ for a uniform, a normal and a log-normal distribution with varying sample sizes.

4. Examples

4.1. **Old Faithful Geyser.** As a first example, we use the eruption durations of the Old Faithful geyser. The data is just one component of a bivariate time series data set. Looking at a one dimensional marginal distribution ignores the process structure. However these data have been used repeatedly to illustrate smoothing algorithms (figure 6), and we reuse it to illustrate our approach (see figure 7). This is a good natured data set showing two distinct nodes with sizeable observation counts, and some overall skewness.

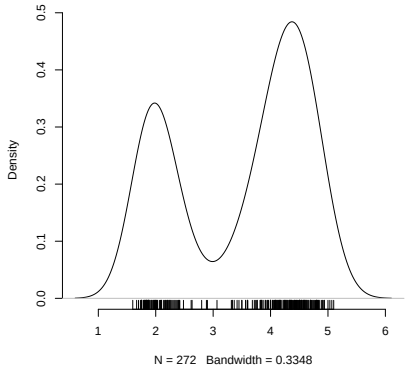


Figure 6. (left) Eruption durations of the Old Faithful geyser: density estimation (R defaults)

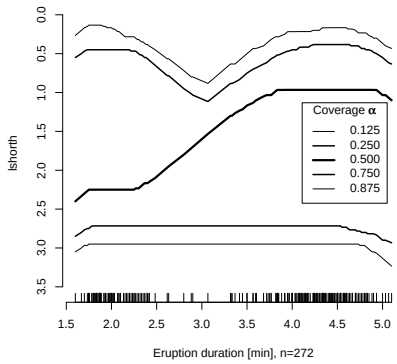


Figure 7. (right) Eruption durations of the Old Faithful geyser: shorth plot

The high coverage levels contours of the shorth plot ($\alpha > 50\%$) just show the overall range of the data. The 50% level indicates a pronounced skewness. The top levels ($25\%, 12.5\%$) reveal that we have two modes, with a comparable coverage range. This is not obvious from the density plot, since the density plot mixes information about local heights attributable to modes with information about the mixture proportions. Density estimators with varying bandwidth or histograms with varying parameters could reveal these details (see figure 8). The multi scale property of the shorth plot allows to combine the aspects in one picture.

figure added to illustrate this

4.2. **Melbourne Temperature Data.** R. Hyndman pointed out the bifurcation to bimodality in the Melbourne temperature data set [HBG96]. We use an extended version of the data set[1] and analyze the day by

[1]Melbourne temperature data 1955-2007, provided by the Bureau of Meteorology, Victorian Climate Services Centre, Melbourne.

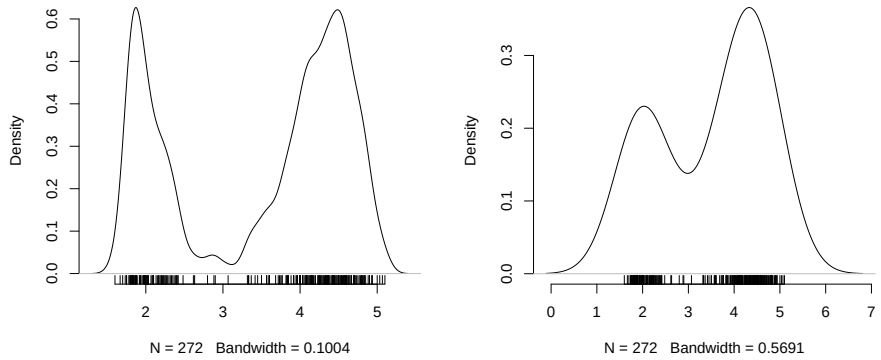


FIGURE 8. Eruption durations of the Old Faithful geyser: density estimation with varying bandwidth. The density plot mixes information about local heights attributable to modes with information about the mixture proportions.

day difference in temperature at 15h (the daily report reference time) conditioned on today's emperature and pressure at the reference time. The shorth plot view is in figure 9. The full picture reveals a cusp-type bifurcation. Figure 10 shows the shorth plot for the temperature difference at 15h (the daily report reference time) to next day's temperature, conditioned on today's temperature and pressure. It reveals the modality split as well as the skewness which is pressure dependent.

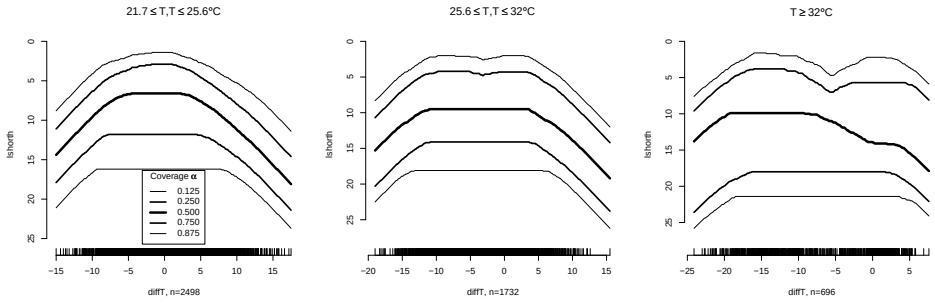


FIGURE 9. Shorth plot at coverage levels $\alpha = 0.125, 0.25, 0.5, 0.75, 0.875$ for Melbourne day by day temperature difference at 15:00h conditioned at today's temperature. A bifurcation to bimodality occurs at high temperatures.

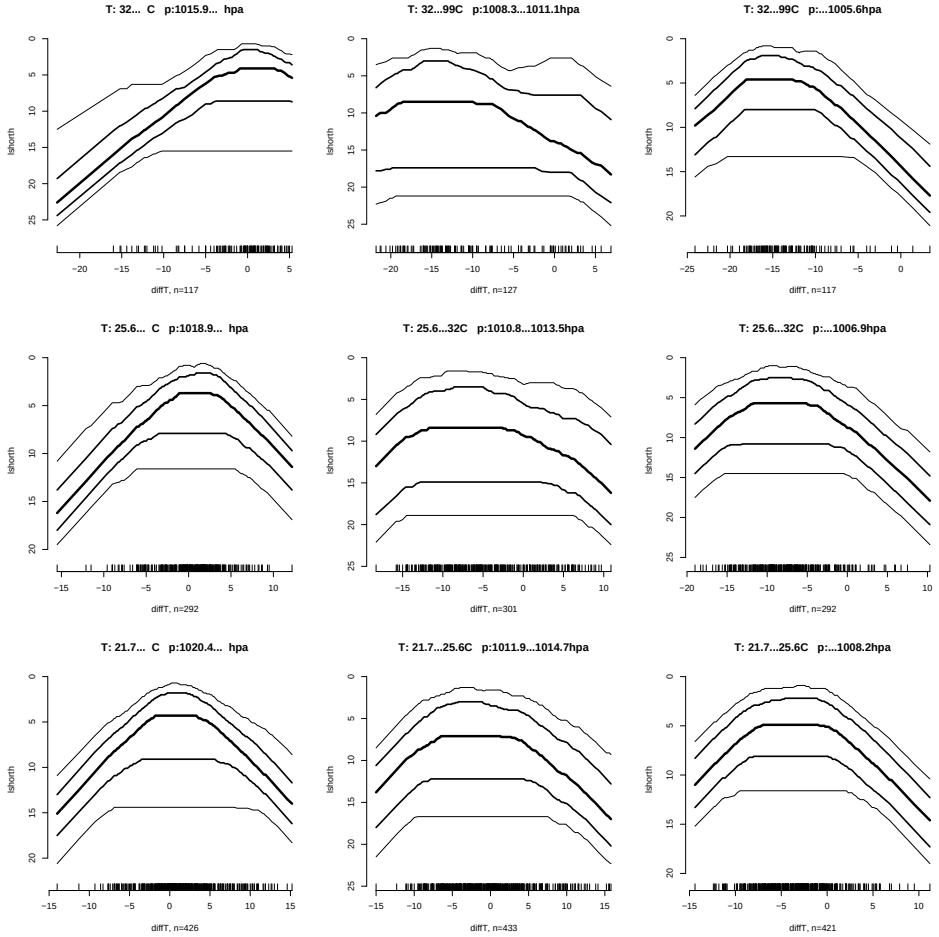
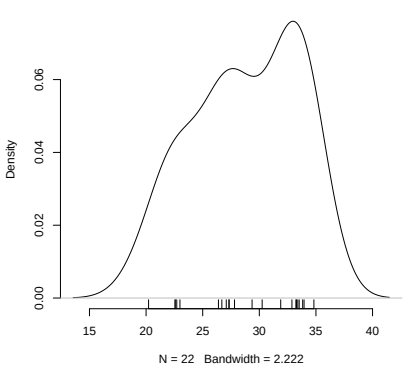


FIGURE 10. Shorth plot at coverage levels at coverage levels $\alpha = 0.125, 0.25, 0.5, 0.75, 0.875$ far Melbourne day by day temperature difference at 15:00h conditioned at today's temperature and pressure. Plot matrix of shorth plots for varying temperature ranges (vertical) and varying pressures (horizontal).

4.3. **Chondrite Data.** The chondrite data set was used in to illustrate a strategy to hunt for modes in [GG80]. This data set is pressing the limits for the shorth plot since it has a very low sample size with presumably three modes (see figure 12).

Using the methods of [GG80] would allow to reveal a third mode, but it is subject to discussion whether this is over-using the data dependency. The shorth, as a general purpose method, gives figure 12. Of course it would be possible to isolate a third mode by including a smaller coverage level.



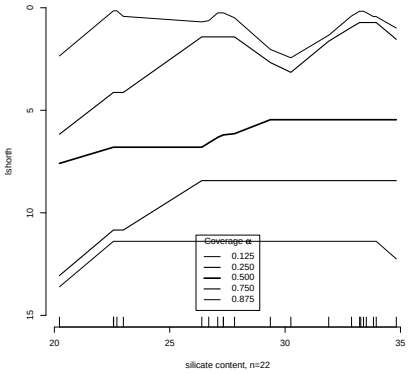


FIGURE 11. (Left) Silicate in chondrite: density estimation

FIGURE 12. (Right) Silicate in chondrite: shorth plot. Note: different scales are used.

For comparison, we add the silhouette plot suggested in [MS91] as figure 13. The silhouette plot, specialised at detecting modes, clearly outperforms the shorth plot for this extremely small data set. But although the short plot is a general purpose plot, it hints at a third mode at all levels. If it goes to level 12.5% it can trace the third mode, and clearly identifies it for lower coverage levels.

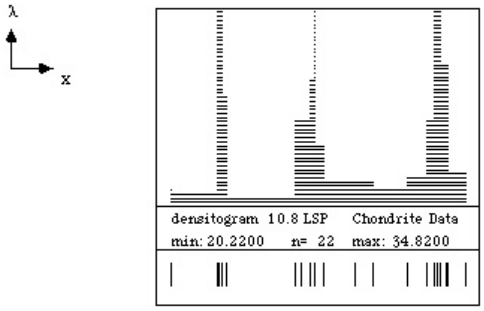


FIGURE 13. Silicate in chondrite: silhouette plot

4.4. **Hartigan's Hat.** This example is a mixture in proportions 3 : 2 : 3 of a uniform on $\left(0, \frac{1}{4}\right)$, on $\left(\frac{1}{4}, \frac{3}{4}\right)$, and on $\left(\frac{3}{4}, 1\right)$ used in [HH85] to illustrate the dip test of unimodality. See figure 14. For a general analysis, it is a challenge because it combines bimodality with flat parts in the distribution, and a relatively low density in the "dip". Only 25% of the distribution fall in the "dip", and thus it must be hidden in higher coverages for the shorth plot.

In this situation, kernel density estimation performs poorly, since it is heavily degraded by boundary effects which cannot cope with the flat parts of the distribution. The shorth plot hints at the flat parts on the outside, but has difficulties identifying the flat middle part.

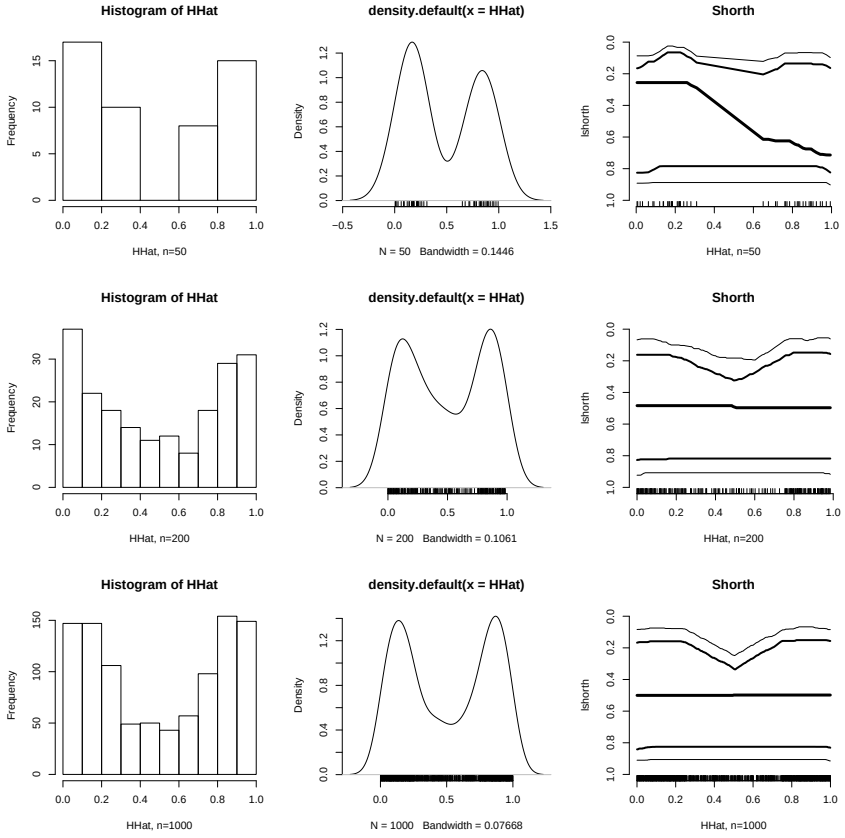


FIGURE 14. Hartigan's Hat

5. Extensions

The shorth length is a well defined concept in one dimension. The generalisation to higher dimensions is the volume of a container covering a proportion of the data. In higher dimensions however there is no distinct class of containers to be considered. So additional choices have to be taken, such as using spheres or ellipsoids of minimal volume.

As with QQ-plots and PP-plots, the generalisation to the two sample case is immediate.

An open question is whether the shorth length approach can be carried over to a regression context.

6. Related Approaches

The shorth plot is related to kernel density estimation with variable bandwidth. It can be seen as a k-nearest neighbour approach. But in contrast to density estimation, it focusses on a concentration functional. Density is an infinitesimal concept. Mass concentration however is a local concept, but not an infinitesimal concept. As a consequence, density has no empirical counterpart, whereas mass concentration has. This makes the shorth length easier to handle for data analytical purposes.

The relation to mass concentration is shared with the silhouette and the excess density plot ([MS91]) or Hyndman's highest density regions ([Hyn96]). The view however is complementary. Silhouette and the excess density focus on concentration, but the shorth on local spread. Silhouette and the excess density target at detecting modality and are model based for a global model (e.g., unimodal vs. bimodal). The shorth however has a local perspective, and is model independent.

P. A. and J. W. Tukey suggested a "balloon plot" in [TT81] (in [Bar81], reprinted in [Tuk88]). This is most closely related to the shorth plot. The main difference is that the balloons are centred at data points. The shorth plot does not use this centring, thus avoiding unnecessary random fluctuation.

The SIzer approach by Chaudhuri and Marron [CM99] is related in spirit to the shorth plot. It tries to present a multiscale representation for smoothing while controlling the artifacts of smoothing by going to a probability scale. The shorth plot avoids the initial smoothing step and is strictly data based.

7. Summary

The shorth plot is a means to investigate mass concentration. It is easy to compute, avoids the bandwidth selection problems, and allows scanning for local as well as for global features of the distribution. The good rate of convergence of the shorth estimator makes it useful already at moderate sample sizes.

References

[ABH+72] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey, *Robust estimation of location: Survey and advances*, Princeton Univ. Press, Princeton, N.J., 1972.

[Bar81] Vic H. Barnett (ed.), *Looking at multivariate data*, Wiley, Chichester [u.a.], 1981 (eng).

[CM99] Probal Chaudhuri and James S. Marron, *Sizer for exploration of structures in curves*, Journal of the American Statistical Association **94** (1999), no. 447, 807–823.

[GG80] I. J. Good and R. A. Gaskins, *Density estimation and bump-hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data (with discussion)*, Journal of the American Statistical Association **75** (1980), 42–73.

[Grü88] R. Grübel, *The length of the shorth*, Annals of Statistics **16.2** (1988), 619–628.

[HBG96] Rob J. Hyndman, David M. Bashtannyk, and Gary K. Grunwald, *Estimating and visualizing conditional densities*, J. Comput. Graph. Statist. **5** (1996), no. 4, 315–336.

[HH85] J. A. Hartigan and P. M. Hartigan, *The dip test of unimodality*, Annals of Statistics **13** (1985), 70–84.

[Hyn96] Rob. J. Hyndman, *Computing and graphing highest density regions*, The American Statistician **50** (1996), no. 2, 120–126.

[MS91] D.W. Müller and Günther Sawitzki, *Excess mass estimates and tests for multimodality*, Journal of the American Statistical Association **86** (1991), 738–746.

[Saw92] Günther Sawitzki, *The shorth plot*, Tech. report, StatLab Heidelberg, 1992.

[Saw94] ———, *Diagnostic plots for one-dimensional data*, Computational Statistics. Papers Collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg. (Peter Dirschedl and Rüdiger Ostermann, eds.), Physica-Verlag, Heidelberg, 1994, pp. pp. 237–258.

[SW86] Galen R. Shorack and Jon A. Wellner, *Empirical processes with applications to statistics*, Wiley, New York, 1986 (eng).

[TT81] P.A. Tukey and J. W. Tukey, *Data-driven view selection: Agglomeration and sharpening*, Looking at multivariate data (Vic H. Barnett, ed.), 1981.

[Tuk88] John W. Tukey, *The collected works of john w. tukey - graphic: 1965–1985*, vol. V, Chapman & Hall, 1988.

```
$URL: svn+ssh://gsawitzki@svn.r-forge.r-project.org/svnroot/lshorth/Rnw/TheShorthPlot.Rnw.tex $
$Revision: 42 $
$Id: TheShorthPlot.Rnw.tex 42 2007-10-07 16:51:02Z gsawitzki $
```

Günther Sawitzki
StatLab Heidelberg
Im Neuenheimer Feld 294
D 69120 Heidelberg

E-mail address: gs@statlab.uni-heidelberg.de

URL: http://lshorth.r-forge.r-project.org/