

THE SHORTH PLOT

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

1. Distribution Diagnostics	1
2. The Length of the Shorth	2
2.1. Elementary Properties	4
2.2. Computing the Shorth Length	5
3. The Shorth Plot	6
3.1. Display Details	6
4. Examples	11
4.1. Old Faithful Geyser	11
4.2. Melbourne Temperature Data	11
4.3. Chondrite Data	11

Date: October 1992, published as technical note [Saw92].

For quotation, please use the general reference [Saw94].

Revised: August 2007

Typeset: September 2, 2007.

Key words and phrases. shorth, distribution diagnostics, data analysis.

Private Version

4.4. Hartigan's Hat	13
5. Extensions	14
6. Related Approaches	14
7. Summary	14
References	15
Index	16

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 representation such as *PP*-plots

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

or *QQ*-plots

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

are tools of first choice. If we consider the overall scale and location, box & whisker plots are 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 being 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 these initial 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 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 $P()$ 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 $P_n()$ is the empirical distribution.

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

$$a, b \mapsto I = [a, b] \mapsto |I|, P(I) \quad a \leq b$$

This is defined in the half space 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. Going to the empirical version replaces the level curves of $P(I)$ by those 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 $a = b = x$.

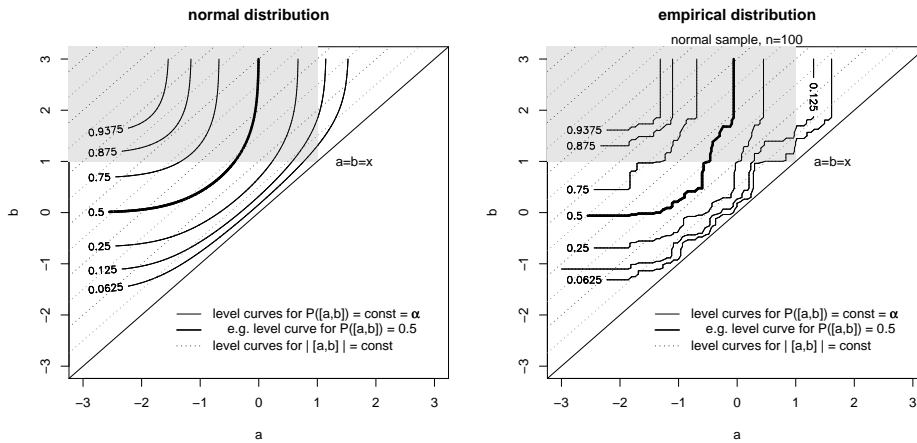


FIGURE 1. The shorth length as an optimisation problem: minimize $|[a, b]|$, under the restriction $P([a, b]) \geq \alpha$. Localising at x restricts the optimisation to the quadrant top left of x .

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

$$\delta \mapsto \sup_y \{P(I) : |I| \geq \delta\}$$

together with its localised and/or empirical version. The relation between the shorth 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.

2.1. Elementary Properties. Here and in the following, we assume a sample X_1, \dots, 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 $X_{(k)}$.

ToDo: 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 $\alpha < 1$

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 $(0, 1)$.

To use empirical distribution functions, the discontinuous case is of interest.

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

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

ToDo: add contour curves for theoretical and empirical

2.2. Computing the Shorth Length.

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

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

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

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

$$\Delta_\alpha = \min\{k : \frac{k+1}{n} \geq \alpha\}.$$

Then

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

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

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

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

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

Unless boundary corrections apply,

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

giving an algorithm with linear complexity in n .

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.

ToDo: show theoretical and empirical shorth plot for Gaussian

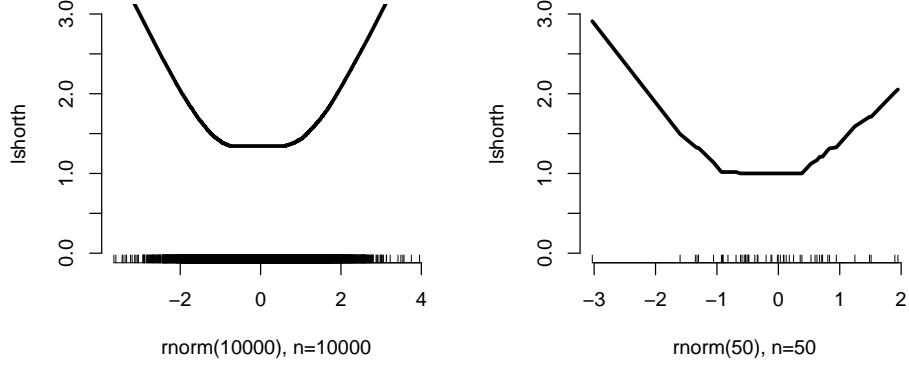


FIGURE 2. Shorth length and $1/\text{Shorth}$ length for a sample of 50 normal variates. Note: different scales are used.

3.1. Display Details. Several choices can 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 (See figure 3). This however is an infinitesimal approximation. It tends to overemphasise peaks, 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 called the standardised shorth length and we use

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

for the (standardised) shorth plot (See figure 4).

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

Figure 5 gives examples of the shorth plot for the uniform, normal and log-normal distribution with varying sample sizes.

Varying the coverage level α as in figure 6 gives an impression of the mass concentration. Small levels give information about the local behaviour, in particular near modes. High levels give information about skewness the overall distribution shape. A dyadic scale, e.g.,

ToDo: show 50% shorth for normal and normal sample
ToDo: reduce to neg

ToDo: replace last line by theoretical plots

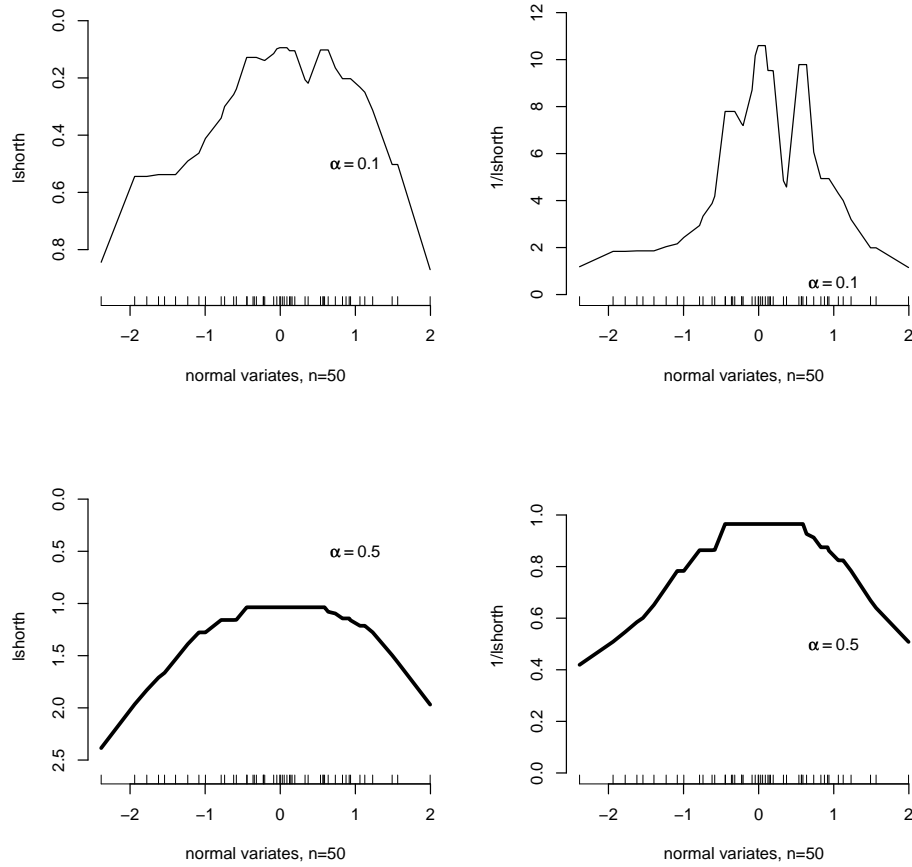


FIGURE 3. Shorth length and $1/\text{Shorth}$ length for a sample of 50 normal variates. The axis for the shorth is pointing downward. Note: different scales are used.

0.125, 0.25, 0.5, 0.75, 0.875 with steps chosen based on the sample size 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
1000 by theor

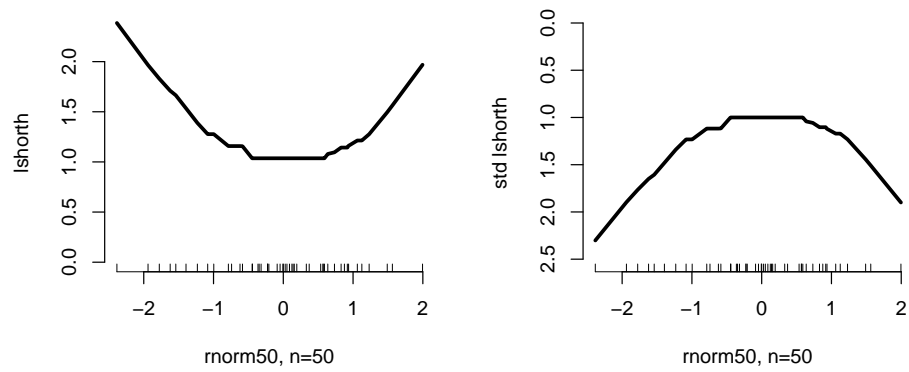


FIGURE 4. Shorth length and standardised shorth length for the same sample of 50 normal variates as above.

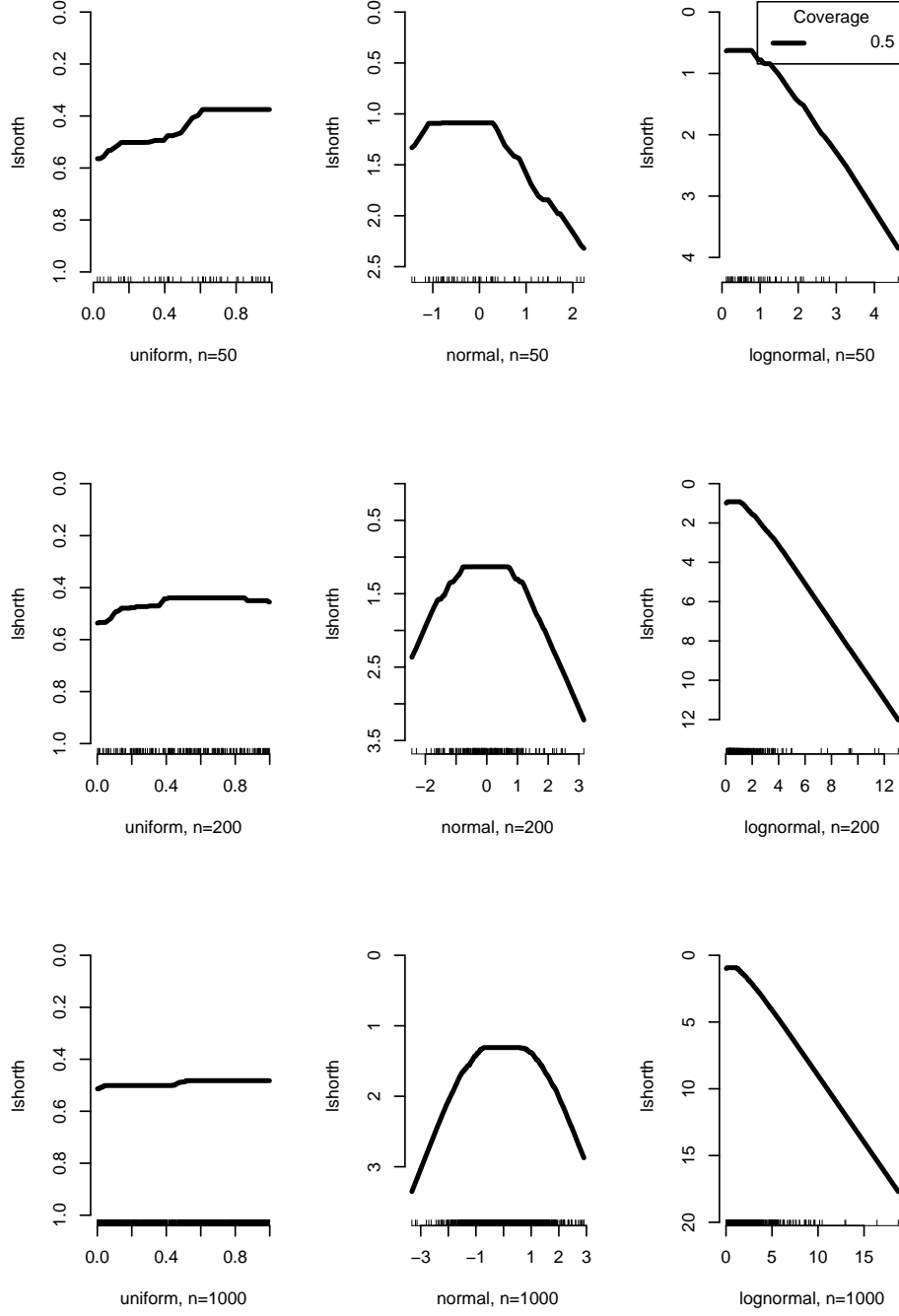


FIGURE 5. $x \mapsto -S_{n,0.5}(x)$ for a uniform, a normal, and a log-normal distribution with sample sizes 50, 100, 200. Note: Different scales are used for the shorth length.

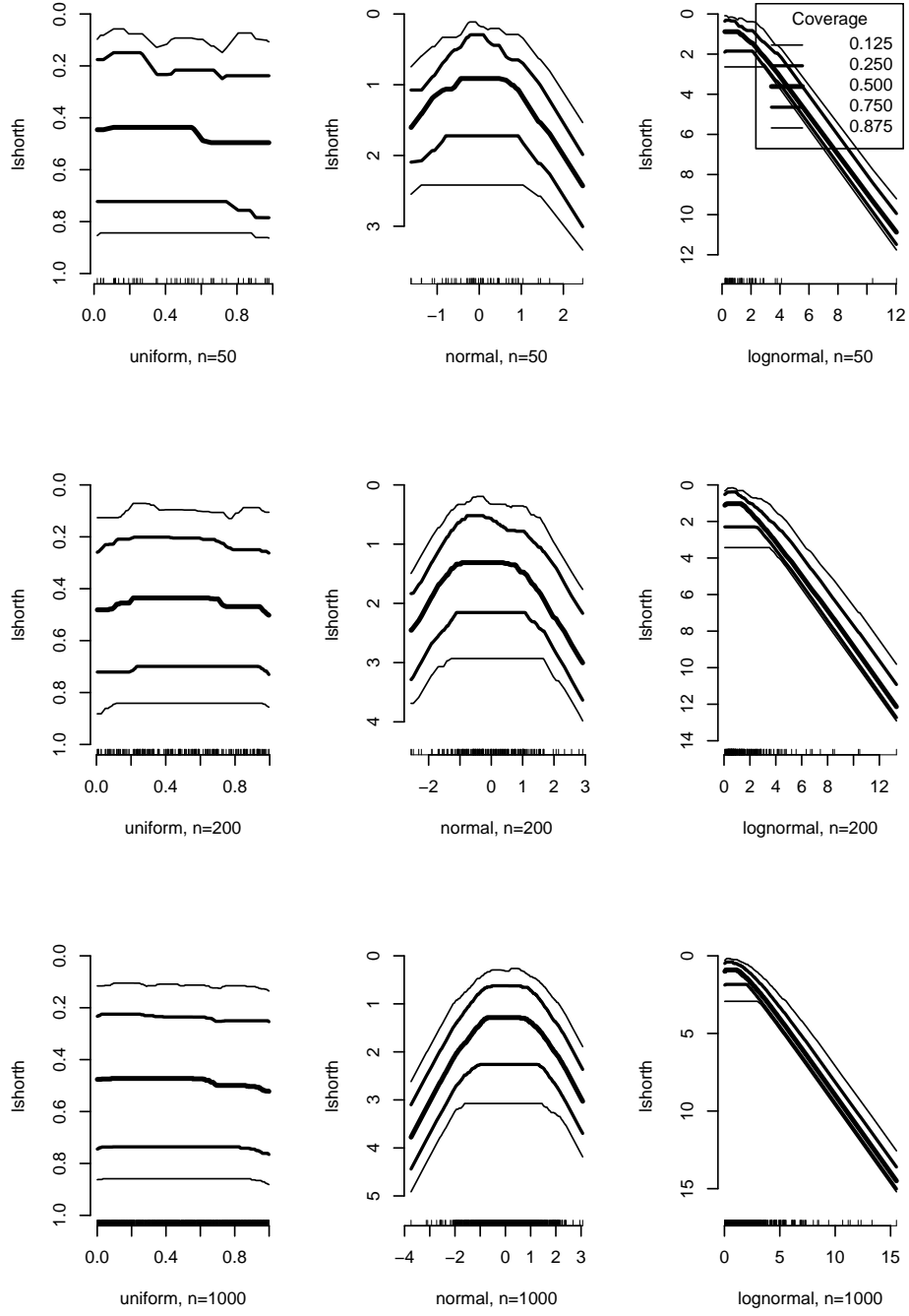


FIGURE 6. $x \mapsto -S_{n,\alpha}(x)$ for a uniform, a normal and a log-normal distribution with varying sample sizes.

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 process data set. Looking at a one dimensional marginal distribution ignores the process structure. However these data have been used repeatedly to illustrate smoothing algorithms, and we reuse it to illustrate our approach. See figure 8.

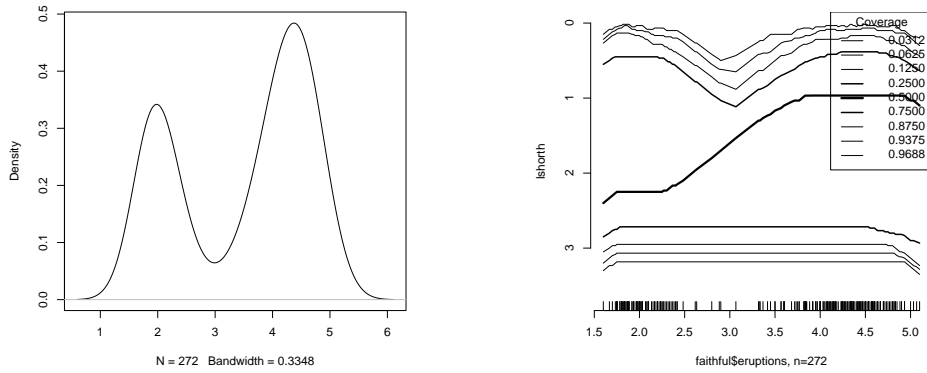


FIGURE 7. Eruption durations of the Old Faithful geyser: (left) density estimation (R defaults)

FIGURE 8. (right) 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 to a pronounced skewness. The top levels reveal that we have two modes, with a comparable coverage range. This is not obvious from the density plot, since it 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. The multi scale property of the shorth plot allows to combine the pictures.

4.2. Melbourne Temperature Data.

4.3. Chondrite Data. The second example is the chondrite data set, used in to illustrate a strategy to hunt for modes in [GG80]. See figure 10. Using the methods of [GG80] would allow to reveal a third mode, but of course it is subject to discussion whether this is over-using the data dependency. The shorth, as a general purpose method, gives figure

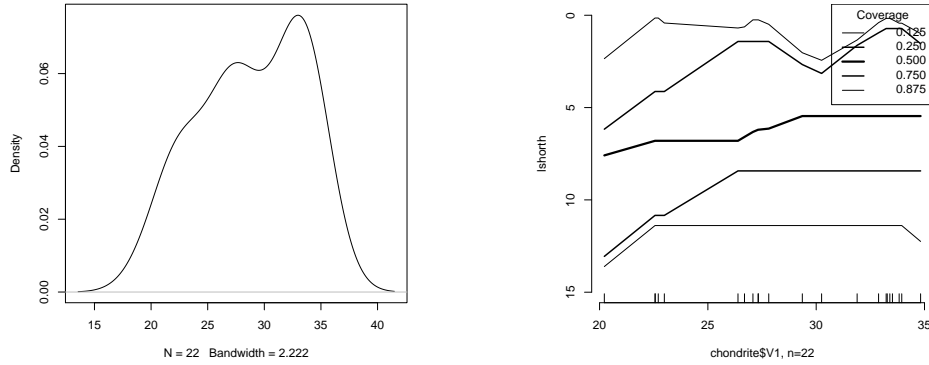


FIGURE 9. (Left) Silicate in chondrite: density estimation

FIGURE 10. (Right) Silicate in chondrite: shorth plot

10. Of course it would be possible to isolate a third mode by including a smaller coverage level.

For comparison, we add the silhouette plot suggested in [MS91] as figure 11. The silhouette plot, specialised at detecting modes, clearly outperforms the shorth plot for this extremely small data set. But although the shorth plot is a general purpose plot, it hints a third mode at all levels. if it goes to level 12.5% it can trace it, and clearly identifies it for lower levels.

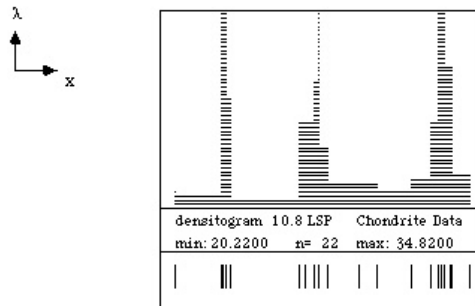


FIGURE 11. Silicate in chondrite: silhouette plot

4.4. Hartigan’s Hat. The third example is a mixture in proportions $3 : 2 : 3$ of a uniform on $(0, \frac{1}{4})$, on $(\frac{1}{4}, \frac{3}{4})$, and on $(\frac{3}{4}, 1)$ used in [HH85] to illustrate the dip test of unimodality. See figure 12.

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. Note that in this situation, only 25% of the distribution fall in the “dip”, and thus it must be hidden in higher coverages.

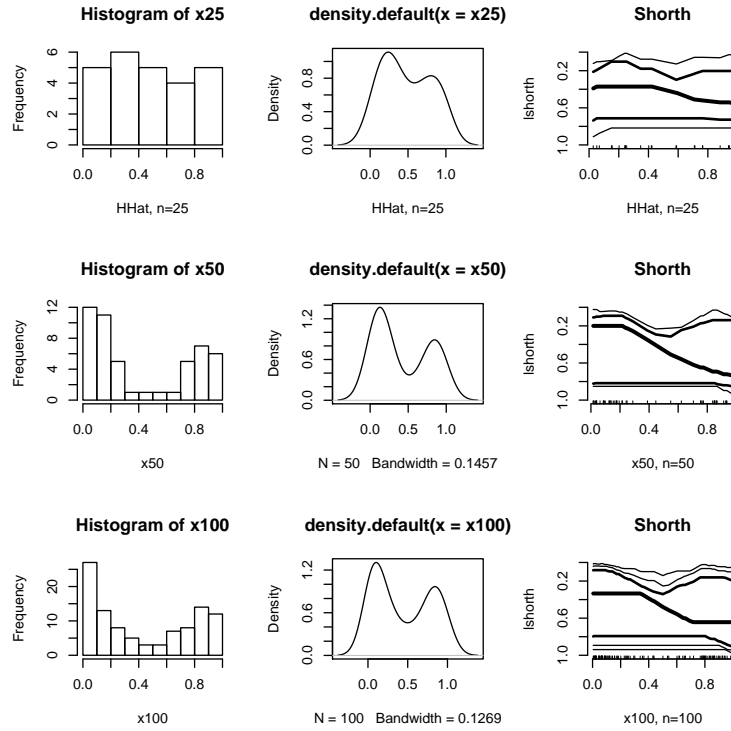


FIGURE 12. Hartigan’s Hat

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]). 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.

ToDo: discuss
sizer
ToDo: discuss
hyndman highest
density regions

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 for moderate sample size.

REFERENCES

- [ABH⁺72] ANDREWS, D. F. ; BICKEL, P. J. ; HAMPEL, F. R. ; HUBER, P. J. ; ROGERS, W. H. ; TUKEY, J. W.: *Robust Estimation of Location: Survey and Advances*. Princeton, N.J. : Princeton Univ. Press, 1972
- [Bar81] BARNETT, Vic (Hrsg.): *Looking at multivariate data*. Chichester [u.a.] : Wiley, 1981. – XVI, 374 S. S. – ISBN 0-471-28039-9
- [GG80] GOOD, I. J. ; GASKINS, R. A.: Density estimation and bump-hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data (with discussion). In: *Journal of the American Statistical Association* 75 (1980), S. 42–73
- [Grü88] GRÜBEL, R.: The Length of the Shorth. In: *Annals of Statistics* 16.2 (1988), S. 619–628
- [HH85] HARTIGAN, J. A. ; HARTIGAN, P. M.: The dip test of unimodality. In: *Annals of Statistics* 13 (1985), S. 70–84
- [MS91] MÜLLER, D.W. ; SAWITZKI, G.: Excess Mass Estimates and Tests for Multimodality. In: *Journal of the American Statistical Association* 86 (1991), S. 738–746
- [Saw92] SAWITZKI, Günther: The Shorth Plot / StatLab Heidelberg. 1992. – Forschungsbericht
- [Saw94] SAWITZKI, Günther: Diagnostic Plots for One-Dimensional Data. In: PETER DIRSCHIEDL, Rüdiger O. (Hrsg.): *Computational Statistics. Papers collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reissensburg*. Physica-Verlag/Springer, Heidelberg, 1994, S. pp. 237–258
- [SW86] SHORACK, Galen R. ; WELLNER, Jon A.: *Empirical processes with applications to statistics*. New York : Wiley, 1986. – XXXVII, 938 S. S. – ISBN 0-471-86725-X
- [TT81] TUKEY, P.A. ; TUKEY, J. W.: Data-Driven View Selection: Agglomeration and Sharpening. In: H.BARNETT, Vic (Hrsg.): *Looking at multivariate data*, 1981
- [Tuk88] TUKEY, John W. ; CLEVELAND, William S. (Hrsg.): *The Collected Works of John W. Tukey - Graphic: 1965–1985*. Bd. V. Chapman & Hall, 1988

INDEX

ToCheck

- 2: ok?, 5

ToDo

- 2: add contour curves for
theoretical and empirical, 6
- 2: add proof concl. 5, 5
- 3: reduce to neg, 7
- 3: replace 1000 by theor, 8
- 3: replace last line by theoretical
plots, 7
- 3: show 50% shorth for normal
and normal sample, 7
- 3: show theoretical and empirical
shorth plot for Gaussian, 6
- 6: discuss hyndman highest
density regions, 15
- 6: discuss silhouette, 15
- 6: discuss sizer, 15

```
$Source: /u/math/j40/cvsroot/lectures/src/shorth/Rnw/lshorth.Rnw.tex,v $  
$Revision: 1.6 $  
$Date: 2007/08/31 15:24:37 $  
$Name: $  
$Author: j40 $
```

```
$Log: lshorth.Rnw.tex,v $  
Revision 1.6  2007/08/31 15:24:37  j40  
basic plot added for definition
```

```
Revision 1.4  2007/08/30 15:03:54  j40  
sync
```

```
Revision 1.3  2007/08/30 12:27:09  j40  
Tilburg sync
```

```
Revision 1.2  2007/08/24 22:21:21  j40  
sync
```

```
Revision 1.1  2007/08/01 19:13:42  j40  
shorth tex added
```

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/`