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. Localizing 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: October 1992, published as technical note [Saw92].
For quotation, please use the general reference [Saw94].
Revised: August 2007
Typeset: August 25, 2007.
Key words and phrases. shorth, distribution diagnostics, data analysis.

.

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, specialized 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

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

are tools of first choice. If we focus on the overall scale and location and scale , box & whisker plots are a valuable tool. This tool looses 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 analyze the modality of a distribution. More specialized tools are needed in this case, such as the silhouette and the 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 satifactory 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 analyze 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 localization, 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 Princton 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 perfomed 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 Gausssian 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 localized, thus providing a tool for local diagnostics. We define:

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

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

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 therotical as well as empirical distributions. The definition in terms of a theoretical probabiltiy $P()$ has an immediate empirical counterpart, the

empirical length of the shorth

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

where $P_n()$ is the empirical distribution.

A modified version of Grübel's proof carries over to the localized 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 $X_1, \ldots, X_n$ 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 $X_{(k)}$.

To get an impression of the shorth process, it might be helpful to read these remarks in three processes: first, for the "vanilla mode", that is a continuous, strictly increasing distribution function. Using the usual compactification helps to unify the cases of finite or infintite intervals. In a second pass, the continuity details may be added. In a final pass, the empirical distribution funtion can be considered as a special case.

Remark 1. *(invariance)* For all α

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

is invariant under shift transformations and equivariant under scale transformations.

Remark 2. *(continuity)* For a contiuous distribution function F,

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

is continuous.

Remark 3. *(monotonicity)* For all x,

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

is monotonously non decreasing in α.

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

Remark 4. *(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 5. *(algorithm)* For $\alpha, 0 \leq \alpha \leq 1$, let

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

Then

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

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

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

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,

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

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

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

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.

3.1. **Display Details.** Several choices have 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. Using just $x \mapsto -S_\alpha(x)$ avoids the need to special case point masses while keeping the qualitative impression.

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

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

for the (standardized) shorth plot.

Instead of the exact interpolation as in remark 4, we use a linear interpolation. The loss of information is negligible.

Figure 1 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 2 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. $0.125, 0.25, 0.5, 0.27, 0.875$ with steps chosen based on the sample size is a recommended choice. The monotonicity (remark 3) allows the multiple scales to be displayed simultaneously without overlaps, thus giving a multi resolution image of the distribution.

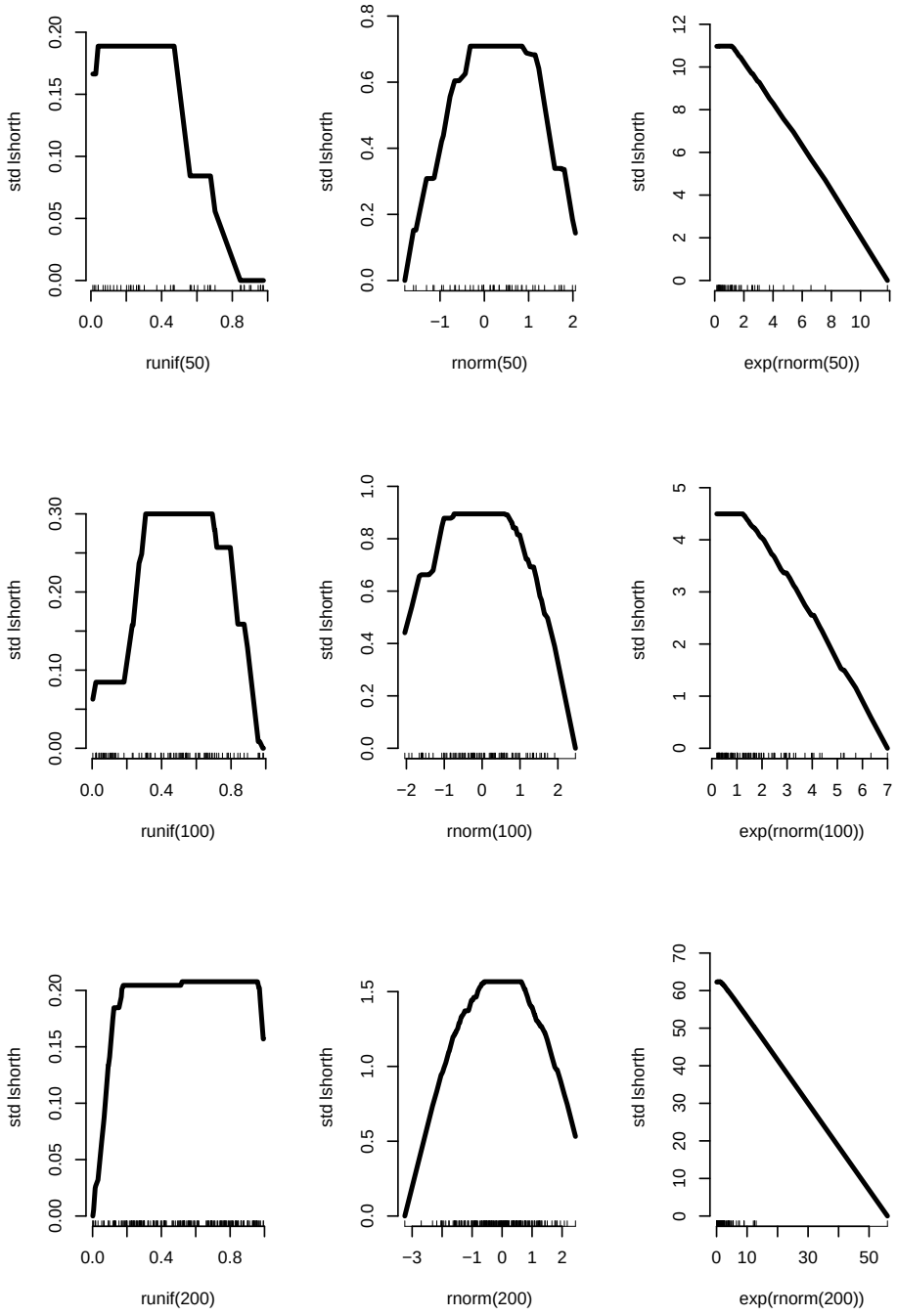

FIGURE 1. $x \mapsto -\frac{S_\alpha(x)}{\min_{x\prime} S_{0.5}(x\prime)}$ for a uniform, a normal, and a log-normal distribution with sample sizes $50, 100, 200$. Note: Different scales are used for the standardized shorth length.

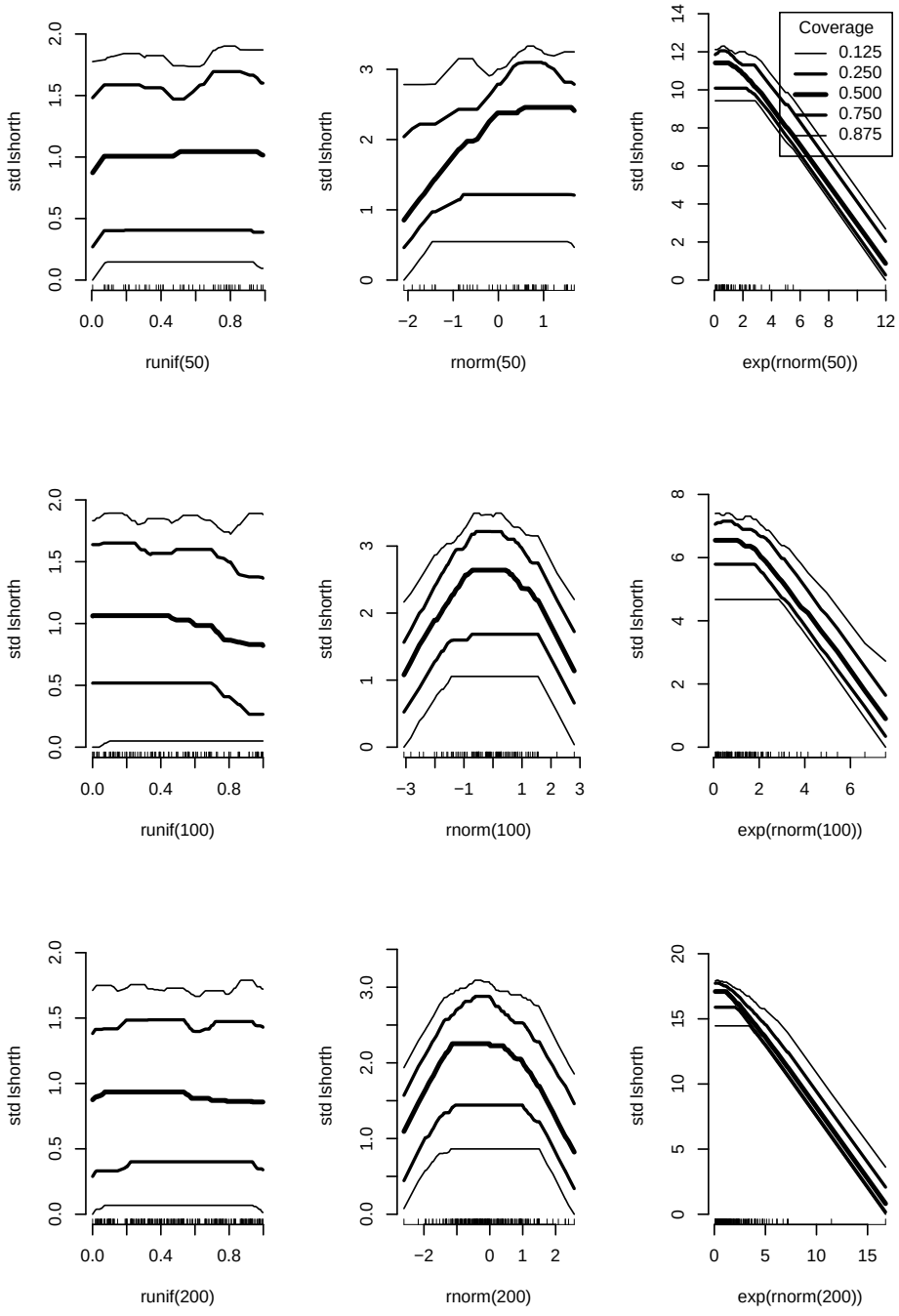

FIGURE 2. $x \mapsto -\frac{S_\alpha(x)}{\min_{x\prime} S_{0.5}(x\prime)}$ for a uniform, a normal and a log-normal distribution with sample sizes $50, 100, 200$.

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 4.

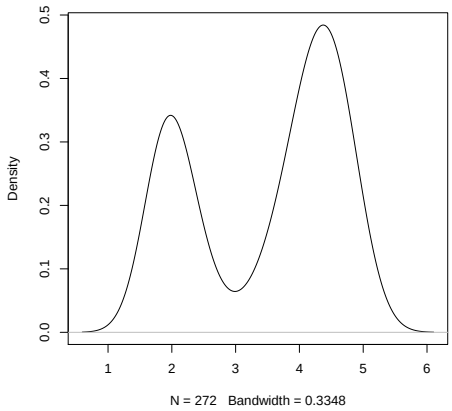

Figure 3. Eruption durations of the Old Faithful geyser: density estimation (R defaults)

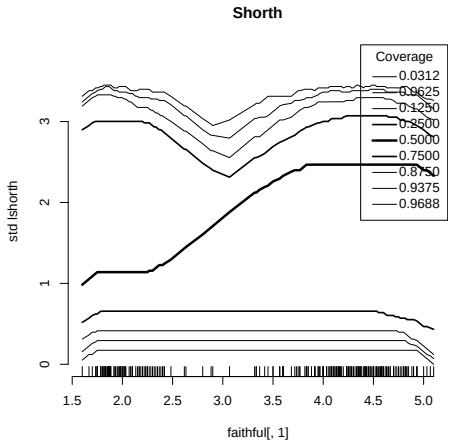

Figure 4. Eruption durations of the Old Faithful geyser

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. **Chondrite Data.** The second example is the chondrite data set, used in to illustrate a strategy to hunt for modes in [GG80]. See figure 6. Using the methods of [GG80] would allow to reveal a third mode, but

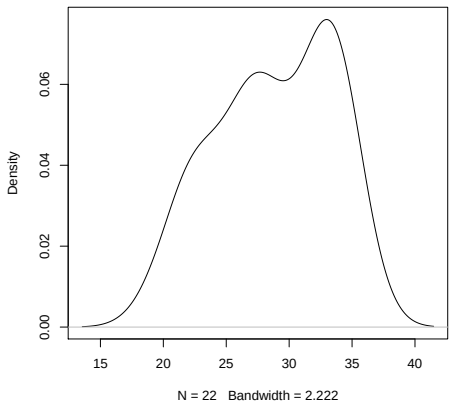

FIGURE 5. Silicate in chondrite: density estimation

of course it is subject to discussion whether this is over-using the data dependency. The shorth, as a general purpose method, gives figure 6. 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 7 The silhouette plot, specialized 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 a third mode at all levels. if it goes to level 12.5% it cna trace it, and clearly identifies it for lower levels.

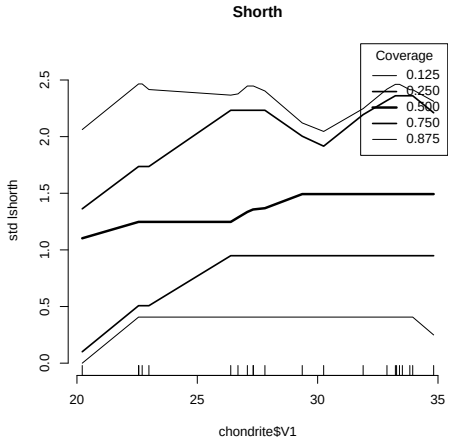

FIGURE 6. Silicate in chondrite

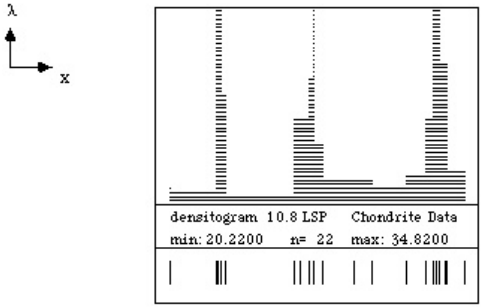

FIGURE 7. Silicate in chondrite: silhouette plot

4.3. **Hartigan's Hat.** The third 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 8.

In this situation, kernel density estimation performs poorely, 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.

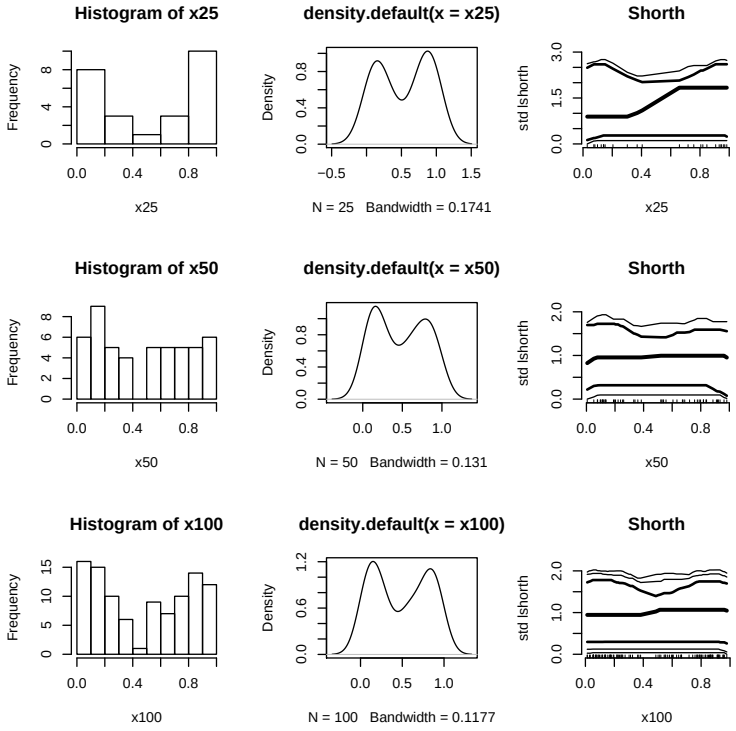

FIGURE 8. Hartigan's Hat

5. Extensions

The shorth length is a well defined concept in one dimension. The generalization to higher dimensions is the volume of a container covering a proportion of the date. 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.

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 bandwith. 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.

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

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 Dirschedl, Rüdiger O. (Hrsg.): *Computational Statistics. Papers collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg.* 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: Aglomeration and Sharpening. In: H.Barnett, Vic (Hrsg.): *Looking at multivariate data*, 1981

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/