Location-Scale Depth

Ivan Mizera and Christine H. Müller

Abstract. The paper introduces a halfspace depth in the location-scale model, along the lines of the general theory given by Mizera on the basis of the idea by Rousseeuw and Hubert, complemented by a new likelihood-based principle for designing criterial functions. The most tractable version of the proposed depth, the Student depth, turns out to be nothing but the bivariate halfspace depth interpreted in the Poincaré plane model of the Lobachevski geometry. This fact implies many fortuitous theoretical and computational properties, in particular equivariance with respect to the Möbius group and favorable time complexities of algorithms. It also opens a way to introduce some other depth notions in the location-scale context, for instance, location-scale simplicial depth. A maximum depth estimator of location and scale—the Student median—is introduced. Possible applications of the proposed concepts are investigated on data examples.

1. Introduction

This paper proposes a notion of depth in the univariate location-scale model and its possible applications. The new depth is introduced in Section 3 as an instance of the general theory of halfspace depth elaborated by Mizera (2002) on the basis of the idea outlined by Rousseeuw and Hubert (1999). This theory is complemented in Section 2 by a likelihood-based principle for designing of criterial functions in various statistical models.

The core of the paper, starting with Section 4, is devoted to the most tractable version of the new concept, the Student depth, and also to the maximum depth location and scale estimator based on it, the Student median. After Section 4, a casual reader may go directly to Section 8 which contains several data-analytic examples. To avoid logical gaps, however, we suggest rather

Key words and phrases. Depth contours; Exploratory data analysis; Location-scale model; Median; Möbius equivariance; Robust estimation.

Ivan Mizera is Associate Professor, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G2G1, Canada. Christine H. Müller is Professor, Institute for Mathematics, Carl von Ossietzky University Oldenburg, Postfach 2503, D-26111, Oldenburg, Germany. The research of Ivan Mizera was supported by the Natural Sciences and Engineering Research Council of Canada, the research of Christine H. Müller by the grant MU 1031/6-1 of the Deutsche Forschungsgemeinschaft during her sabbatical semester in Edmonton. Both authors thank all referees, the associate editor and the editor for constructive remarks that led to removal of several ambiguities and helped to make the presentation more focused.

to read sections in their normal order: Section 5 explores the underlying hyperbolic geometry—it turns out that the Student depth is nothing but the halfspace depth in the Poincaré plane model of the Lobachevski geometry, inheriting all favorable properties of the bivariate location halfspace depth; Section 6 studies equivariance properties with respect to the Möbius group; Section 7 surveys some further statistical (asymptotics, robustness) and computational facts. Conclusions and future directions are briefly summarized in Section 9; the Appendix contains all proofs.

2. Depth via likelihood-based criterial functions

The definition of the depth in general models is motivated by theoretical considerations with a decision-theoretic flavor. Those are more thoroughly explained, together with details on the examples considered below, by Mizera (2002); here we give only an accelerated overview.

Our starting point are **data** composed of **datapoints** z_i (as usual, $i = 1, 2, \ldots, n$). For every datapoint z_i, we consider a **criterial function** F_i; given a **fit** represented by ϑ, the criterial function F_i evaluates the lack of fit of ϑ to the particular datapoint z_i. That is, we consider $\tilde{\vartheta}$ fitting z_i better than ϑ, if $F_i(\tilde{\vartheta}) < F_i(\vartheta)$.

Such criterial functions may be derived from intuitive considerations. For instance, in linear regression with datapoints $z_i = (x_i, y_i)$, a natural choice is $F_i(\vartheta) = (y_i - x_i^T\vartheta)^2$, or $F_i(\vartheta) = |y_i - x_i^T\vartheta|$. Both choices are equivalent, since only the order on ϑ imposed by F_i is essential. In the canonical example of the multivariate location model considered by Tukey (1975), the criterial functions may be either $F_i(z_i) = \|z_i - \vartheta\|$ or their squares.

General halfspace depth can be defined as a measure of data-analytic admissibility—the simplest version of this principle, in the spirit of Rousseeuw and Hubert (1999), defines depth of ϑ as the proportion of the datapoints whose omission causes ϑ to become a *nonfit*, a fit than can be uniformly dominated by another one. We refer again to Mizera (2002) for the more elaborate version of what is called *global depth* therein, as well as for further technical details of its properties and in particular its relationship to the more operational *tangent depth*, the result of a transition from the optimality-based principle to its first-order reformulation. A good analogy is that of maximum likelihood prescription, and the related estimating equation(s) obtained by taking derivatives and equating them to zero. The equations are often equivalent to the original optimization problem, but even if they are not, they generally represent an interesting prescription of their own.

Taking derivatives in the optimization problem explains why the following definition involves gradients $\nabla_\vartheta F_i(\vartheta)$, in ϑ, of the criterial functions. In this paper, we define the (tangent) **depth** of a fit ϑ to be

$$d(\vartheta) = \inf_{u \neq 0} \#\{i \colon u^T \nabla_\vartheta F_i(\vartheta) \geq 0\}, \tag{1}$$

where $\#$ stands for the relative proportion in the index set—its cardinality divided by n. We suppress the dependence on the data in the depth notation. Tukey (1975) and others consider

cardinalities instead of proportions; however, it is a minor difference whether depth assumes values 0, 1, 2, ..., n or 0, $1/n$, $2/n$, ..., 1, and relative proportions allow for the unified treatment of population distributions later. For the same reason, we use in (1) less intuitive "inf" instead of equivalent "min".

In the linear regression example, we may work our way from substituting criterial functions $F_i(\vartheta) = \frac{1}{2}(y_i - x_i^T\vartheta)^2$ into the formula (1) to the standard expressions of the regression depth,

$$\begin{aligned} d(\vartheta) &= \inf_{u\neq 0} \#\{i\colon\ -u^T x_i(y_i - x_i^T\vartheta) \geq 0\} \\ &= \inf_{u\neq 0} \#\{i\colon u^T x_i(y_i - x_i^T\vartheta) \geq 0\} \\ &= \inf_{u\neq 0} \#\{i\colon u^T x_i\ \mathrm{sgn}(y_i - x_i^T\vartheta) \geq 0\} \\ &= \inf_{u\neq 0} \#\{i\colon\ \mathrm{sgn}(u^T x_i)\ \mathrm{sgn}(y_i - x_i^T\vartheta) \geq 0\}, \end{aligned}$$

as defined by Rousseeuw and Hubert (1999). (In the spirit of the equivalence of criterial functions modulo the order they impose, the factor 1/2 involved in F_i is merely a convenience multiplier, to give the gradients a neat form $-x_i(y_i - x_i^T\vartheta)$. The choice $F_i(\vartheta) = |y_i - x_i^T\vartheta|$ yields the same depth.)

Similar calculations show the above-considered criterial functions in the multivariate location model lead to the standard definition of the halfspace depth: the minimal proportion of datapoints lying in any closed halfspace whose boundary contains ϑ, or, equivalently, the minimal proportion of datapoints whose omission leaves ϑ outside the convex hull of the remaining ones. Note that in this special case datapoints z_i and fits ϑ live in the same space; generally, however, formula (1) defines depth of fits, not datapoints.

The theoretical innovation brought by the present paper is the use of likelihood considerations for designing criterial functions. As a motivating example, consider again the linear regression model. In this model with i.i.d. Gaussian disturbances (for simplicity with a fixed known scale set equal to one), the standard expression for the negative log of the likelihood reads

$$-\log L(\vartheta) = \sum_{i=1}^{n} \left(\frac{1}{2}(y_i - x_i^T\vartheta)^2 + \log\sqrt{2\pi} \right).$$

Apart from the constant $\log\sqrt{2\pi}$, which does not depend on ϑ and hence may be omitted, we obtained the sum of functions of ϑ, each of them dependent only on one datapoint. Actually, they are identical with the criterial functions we considered in the linear regression model.

This suggests the following principle: the negative log-likelihood for the i.i.d. model is always a sum of contributions each involving one particular datapoint; hence we may adopt these contributions for criterial functions. The principle not only gives some additional justification for the instances already known, but provides a vehicle to move beyond the limits of intuitive considerations that typically led to those. In this paper, we want to illustrate this thesis on a

novel instance, the univariate location-scale model. For other applications of this principle, see Müller (2003).

3. Location-scale depth

Let us think for a moment that datapoints y_i are realizations of i.i.d. random variables with a density f, determined up to location parameter μ and scale parameter σ. The form of the resulting negative log-likelihood,

$$\sum_{i=1}^{n} \left(-\log f\left(\frac{y_i - \mu}{\sigma}\right) + \log \sigma \right),$$

suggests, according to the just formulated principle, criterial functions

$$F_i(\mu, \sigma) = -\log f\left(\frac{y_i - \mu}{\sigma}\right) + \log \sigma. \tag{2}$$

To avoid technical complications, let us suppose that $f(\tau) > 0$ for all τ, the assumption satisfied by most distributions used in modeling location-scale data. On substituting (2) into (1), we obtain the following expression for the depth:

$$d(\mu, \sigma) = \inf_{u \neq 0} \# \left\{ i \colon (u_1, u_2) \begin{pmatrix} (-\log f)'\left(\frac{y_i - \mu}{\sigma}\right)\left(-\frac{1}{\sigma}\right) \\ (-\log f)'\left(\frac{y_i - \mu}{\sigma}\right)\left(-\frac{y_i - \mu}{\sigma^2}\right) + \frac{1}{\sigma} \end{pmatrix} \geq 0 \right\}, \tag{3}$$

where the expression in the braces is interpreted in the spirit of formula (1) as the inner product in matrix notation.

We assumed $\sigma > 0$ so far; to see what to do with $\sigma = 0$, imagine all datapoints lying in a single point c; for typical instances of f, the formula (3) yields zero depth for all (μ, σ) with $\sigma > 0$ in such a case. The likelihood philosophy suggests that $\mu = c$ and $\sigma = 0$ provide the single best fit for the data then—and when σ approaches 0, then the values of the criterial function tend to 0 too. Therefore, it is natural to assign depth $n/n = 1$ to $(c, 0)$, and 0 to other values of (μ, σ).

Let us introduce functions $\psi(\tau) = (-\log f(\tau))' = -f'(\tau)/f(\tau)$ and $\chi(\tau) = \tau\psi(\tau)$, in analogy with M-estimation in location-scale models as presented by Huber (1981). Starting from (3), we arrive after some algebra to the following definition.

Definition 1. The **location-scale depth** of $(\mu, \sigma) \in \mathbb{R} \times [0, \infty)$, with respect to the datapoints $y_1, y_2, \ldots, y_n$ from $\mathbb{R}$, is

$$\begin{aligned} d(\mu, \sigma) &= \inf_{u \neq 0} \# \left\{ i \colon (u_1, u_2) \begin{pmatrix} \psi(\tau_i) \\ \chi(\tau_i) - 1 \end{pmatrix} \geq 0 \right\}, && \text{for } \sigma > 0, \\ &= \#\{i \colon y_i = \mu\}, && \text{for } \sigma = 0, \end{aligned} \tag{4}$$

where τ_i is a shorthand for $(y_i - \mu)/\sigma$ and ψ and χ depend on a fixed density f as specified above.

In order to elucidate the dependence on f, let us assume that f is strictly unimodal with mode 0, the assumption again satisfied by many distributions used in modeling location-scale data. Together with the requirement that f is everywhere positive, this assumption implies that $\operatorname{sgn}(\psi(\tau)) = \operatorname{sgn}(\tau)$. Let $\zeta(\tau) = \tau/\psi(\tau)$ for $\tau \neq 0$; for $\tau = 0$, set $\zeta(0) = \liminf_{\tau\to 0} \zeta(\tau)$. (In all practical cases ζ has a limit at 0; for what follows, it is important only that $\zeta(0) > 0$.)

THEOREM 1. *If* $\operatorname{sgn}(\psi(\tau)) = \operatorname{sgn}(\tau)$ *and* $\zeta(0) > 0$, *then the location-scale depth is equal to*

$$
\begin{aligned}
d(\mu,\sigma) &= \inf_{u\neq 0} \#\left\{ i\colon (u_1,u_2) \begin{pmatrix} \tau_i \\ \tau_i^2 - \zeta(\tau_i) \end{pmatrix} \geq 0 \right\}, && \text{for } \sigma > 0, \\
&= \#\{i\colon y_i = \mu\}, && \text{for } \sigma = 0,
\end{aligned} \tag{5}
$$

where τ_i *has the same meaning as in Definition 1 and* ζ *relates to* ψ, χ *and* f *as specified above.*

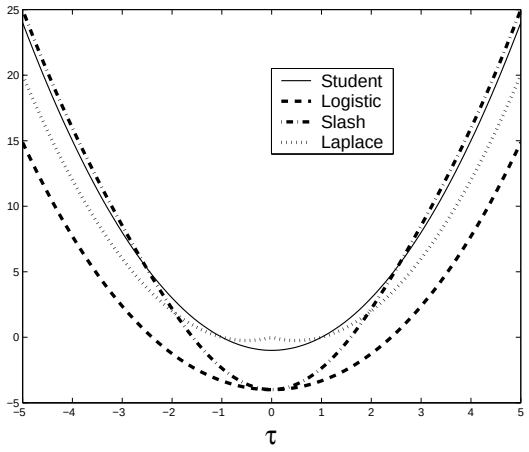
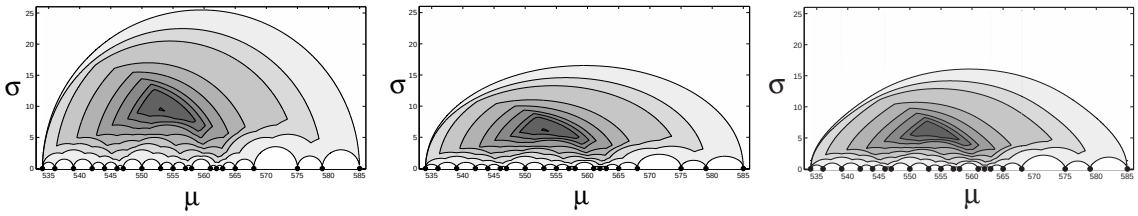


FIGURE 1. Illustration to Theorem 1: the plot of $\tau^2 - \zeta(\tau)$ against τ for various f: t (including Gaussian), logistic, slash, Laplace.

The dependence on f still remains, but Theorem 1 reduced it to a single term $\tau^2 - \zeta(\tau)$, shown in Figure 1 for various f. The simplest form $\zeta(\tau) = 1$ corresponds to f equal to the standard Gaussian density, as well as to all distributions from the t family; other choices are f logistic, slash, and Laplace (double-exponential). The slash version is plotted multiplied by two, which corresponds to a simple reparametrization $\sigma \mapsto \sigma/2$.

Figure 2 shows the contours of location-scale depth for f set equal to t, logistic and slash density, using the same dataset for all three plots. All three panels appear qualitatively similar, although this may not be always the case; see Section 9. The plots of depth contours bear some visual similarity to "Tukey's graphical method of computing Hodges-Lehmann estimate"; see Fig. 13 of Fisher (1983) and the related references therein.

To gain first insights, we may look at simplified models obtained by regarding one of the parameters as a constant. In the first of those models, σ is fixed and μ free. After analogous



Student (and Gaussian) Logistic Slash

FIGURE 2. Contours of the location-scale depth for different f, using the same artificial dataset. The contours were obtained by computing depth for a fine grid of values; note qualitative similarity, but also possible differences.

steps as above, we obtain that the **location likelihood depth** is, for given f,

$$d_\sigma(\mu) = \inf_{u\neq 0} \# \left\{ i\colon u\, \psi\left(\frac{y_i-\mu}{\sigma}\right) \geq 0 \right\}. \tag{6}$$

For any strictly unimodal f, we have $\operatorname{sgn}(\psi(t)) = \operatorname{sgn}(t)$; this converts (6) to the usual definition of univariate location depth

$$\inf_{u\neq 0} \#\{i\colon u\, \operatorname{sgn}(y_i-\mu) \geq 0\} = \min\{\#\{i\colon y_i \leq \mu\}, \#\{i\colon y_i \geq \mu\}\}. \tag{7}$$

Note that $d_\sigma(\mu)$ does not depend on σ.

The second simplified model has μ fixed and σ free. The resulting **scale depth** is

$$\begin{aligned} d_\mu(\sigma) &= \inf_{u\neq 0} \# \left\{ i\colon u\, \left(\chi\left(\frac{y_i-\mu}{\sigma}\right) - 1\right) \geq 0 \right\} \\ &= \min\left\{ \# \left\{ i\colon \chi\left(\frac{y_i-\mu}{\sigma}\right) \leq 1 \right\}, \# \left\{ i\colon \chi\left(\frac{y_i-\mu}{\sigma}\right) \geq 1 \right\} \right\}. \end{aligned} \tag{8}$$

If f is unimodal, then $\chi(t) \geq 0$. If, moreover, $\chi(\tau)$ decreases when $\tau < 0$ and increases when $\tau > 0$ (this often holds—always if ψ is monotone) and is symmetric (which is the case when f is symmetric), then there is k such that

$$d_\mu(\sigma) = \min\{\#\{i\colon |y_i-\mu| \leq k\sigma\}, \#\{i\colon |y_i-\mu| \geq k\sigma\}\}.$$

When f is taken to be the density of standard Gaussian, t or Laplace distribution, then $k = 1$ and σ has depth zero if and only if $[-\sigma, \sigma]$ contains either all datapoints or no datapoint. The depth depends on μ now; the fit with maximal scale depth corresponds to the quantity known as median absolute deviation (MAD) about the fixed location μ.

The extension of all the depth notions to general probabilities and even measures is straightforward: the proportion of the sample points in a given set is replaced by the measure of this

set. The definition for finite samples is embedded into the general scheme via empirical probabilities. For the specific details of the application of this well-known principle, see Mizera (2002) or Rousseeuw and Hubert (1999).

THEOREM 2. *The location-scale, location (likelihood), and scale depths satisfy for all μ and σ,*

$$d(\mu,\sigma) \le d_\sigma(\mu) \quad \textit{and} \quad d(\mu,\sigma) \le d_\mu(\sigma), \tag{9}$$

for any given measure P (including any empirical probability supported by finite-sample data).

4. The Student depth and its applications

Our approach to likelihood-based procedures is rather operational: we do not firmly believe in the postulated model, but rather use it as a guideline to derive a procedure possibly applicable in a wider context. Definition 1 introduces not one, but a family of depths, depending on the choice of the underlying density f. Among these densities, all with similar unimodal shape, we favor those possessing better tractability and computability than others. It cannot be said that we pay no attention to the modeling realism, but our focus is rather on the final result than initial premises; once the procedure is derived, we tend to forget the initial parametric assumptions, and rather investigate its behavior in the broader context. Such an attitude is not new—just recall the approach of Huber (1967) to maximum likelihood estimation, for instance.

It is hardly that unexpected that the most tractable version of location-scale depth is that involving the standard Gaussian density f. In such a case,

$$d(\mu,\sigma) = \inf_{u\neq 0} \# \left\{ i\colon (u_1,u_2) \begin{pmatrix} \tau_i \\ \tau_i^2 - 1 \end{pmatrix} \ge 0 \right\}. \tag{10}$$

It is tempting to think that what we deal here with is just the bivariate location depth with respect to the datapoints lifted on a parabola—but one has to keep in mind that τ_i depend on μ and σ, so when the parameters change, the position of lifted points changes too.

Interestingly, the same depth is obtained when f is taken to be the density of any t distribution with ν degrees of freedom:

$$d(\mu,\sigma) = \inf_{u\neq 0} \# \left\{ i\colon (u_1,u_2) \begin{pmatrix} \tau_i \\ \frac{\nu}{\nu+1}(\tau_i^2 - 1) \end{pmatrix} \ge 0 \right\};$$

the equality to (10) follows after absorbing the constant $\nu/(\nu+1)$ into the u term. This suggests that we may view the standard Gaussian distribution as t with $\nu = \infty$ here.

In fact, formula (10) allows for a unified treatment of all σ, without a need to consider the case $\sigma = 0$ separately. To this end, note that, still formally assuming $\sigma > 0$, we can rewrite (10) as

$$d(\mu,\sigma) = \inf_{u\neq 0} \# \left\{ i\colon \left(\frac{u_1}{\sigma}, \frac{u_2}{\sigma^2}\right) \begin{pmatrix} (y_i - \mu) \\ (y_i - \mu)^2 - \sigma^2 \end{pmatrix} \ge 0 \right\}.$$

DEFINITION 2. The **Student depth** of $(\mu, \sigma) \in \mathbb{R} \times [0, \infty)$, with respect to a (probability) measure P on $\mathbb{R}$ is

$$d(\mu, \sigma, P) = \inf_{(u_1,u_2)^T \neq 0} P\left\{y\colon u_1(y-\mu) + u_2\left((y-\mu)^2 - \sigma^2\right) \geq 0\right\}. \tag{11}$$

The Student depth with respect to the data $y_1, y_2, \ldots, y_n$ is obtained by applying the definition to the empirical probability measure P_n supported by the datapoints.

Although the definition is formulated for general measures this time, we may return back to the more comprehensible sample notation, and also suppress the dependence on P or P_n in the notation and write simply $d(\mu, \sigma)$, if no confusion may arise. All theorems formulated for samples remain valid in the more general setting, with proportions (here denoted by $\#$) replaced by appropriate measures.

What are the possible data-analytic uses of the new concept? The constantly growing literature on the subject records numerous applications of various brands of multivariate location depth and a growing number of the application of the regression depth. Thus, some first observations can be made along the general lines.

One important direction are maximum depth estimators—deepest fits. They can be considered as medians in the underlying models, since in the univariate location case, the deepest fit is the sample median. Starting from the Tukey median in the multivariate location model, it is quite remarkable how the known instances fit the mosaic; for instance, the median character of the deepest regression is quite evident from Rousseeuw and Hubert (1999), Van Aelst, Rousseeuw, Hubert, and Struyf (2002).

Maximum depth estimators have a few handicaps, possessed already by their univariate sample median prototype. There may be problems with the uniqueness of the deepest fit—formally it is more appropriate to define the **maximum depth estimator** as the *set* of all deepest fits. In the univariate case, this ambiguity may be resolved by taking the midpoint of the median interval; analogous strategies in more sophisticated models are more demanding, but not prohibitively.

Depth can also be used in various testing applications, as those of Rousseeuw and Struyf (2002); a nice general perspective in the multivariate location context was given by Chaudhuri and Sengupta (1993). The maximal depth attained in the particular setting often plays a prominent role here, but this would require considerable theoretical development.

What we find more appealing is that the very special feature of the present context—the two-dimensionality of our parametric space—allows for graphical representation of depth contours. For any $\delta \in [0, 1]$, we define, abusing slightly the language, the **depth contour** to be the set of (μ, σ) such that $d(\mu, \sigma) \geq \delta$. Trivially, the contours are nested: the contour corresponding to δ_1 is contained in that corresponding to δ_2 whenever $\delta_1 \geq \delta_2$. In the bivariate location model, the plot of depth contours can be viewed as a generalization of quantile plotting—this line of applications accompanied depth from its very beginning, see Tukey (1975), Donoho and Gasko (1992), or Rousseeuw, Ruts, and Tukey (1999).

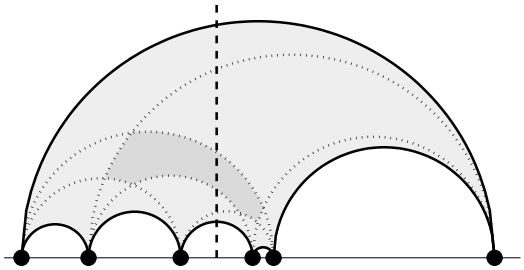


FIGURE 3. Poincaré plane: the horizontal axis corresponds to μ, the vertical to σ, the solid line with datapoints is the line $\sigma = 0$. The shaded areas contain points with the Student depth 1/6 (lighter) and 2/6 (darker) for this artificial datapoints.

The sensitivity of depth contours to the distribution of the data suggests using them as a tool for assessing distributional assumptions. In the style of Chapter 6 of Chambers, Cleveland, Kleiner, and Tukey (1983), we tried to explore how much the plots of the Student depth contours might complement and enhance the use of quantile plots, sharing with them the similar incisive character (compared to histograms and related methods) and the lack of need for elaborate tuning (compared to density estimation). Some examples in this vein are studied in Section 8.

5. The Lobachevski geometry of the Student depth

It turns out that the Student depth is nothing but the bivariate location halfspace depth in the Poincaré plane model of the Lobachevski hyperbolic geometry. To explain this adequately, we have to introduce several notions from non-Euclidean geometry; the reader wishing to get more thorough understanding is advised to consult, for instance, Greenberg (1980).

What we will call the **Poincaré plane** here is the halfplane $\mathbb{PP} = \mathbb{R} \times [0, \infty)$, the parametric space for (μ, σ). A **Poincaré line** ℓ in $\mathbb{PP}$ is an object which is either a halfline $\mu = \text{const}$, $\sigma \geq 0$, or a hemicircumference whose center lies on the line $\sigma = 0$. The complement of ℓ in $\mathbb{PP}$ consists of two connected components; their respective unions with ℓ form two (closed) **Poincaré halfspaces** with boundary ℓ. The (Poincaré) **points** are simply points in $\mathbb{PP}$; we consider them lying in a Poincaré halfspace or on a Poincaré line if they belong to them in the usual set-theoretic sense. (For a connoisseur, our version of the Poincaré plane includes also the ideal points on the line $\sigma = 0$, but not the ∞ endpoint of all vertical Poincaré lines.)

Figure 3 demonstrates how the sample space, the home of datapoints, is embedded into the Poincaré plane, the home of parameters (μ, σ). The Poincaré lines connecting the datapoints delineate the corresponding Poincaré halfspaces; the shaded areas indicate contours with the depth 1/6 and 2/6. The dashed line shows the vertical type of Poincaré line; its μ coordinate is the midpoint of the interval of sample medians.

THEOREM 3. *The Student depth of (μ, σ) is the minimal proportion (infimum of measure P in the general case) of datapoints y_i that lie in any Poincaré halfspace with the point (μ, σ) on*

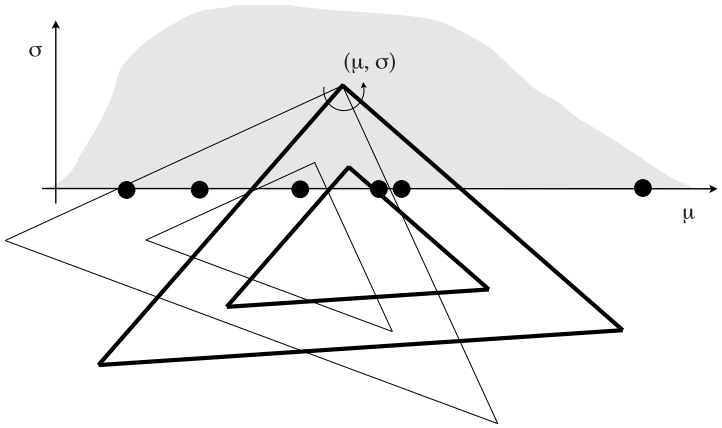


FIGURE 4. The Student depth can be calculated with a right-angle triangular ruler only. The ruler revolves with the vertex placed in the point whose depth is computed (arbitrary values of μ, but only nonnegative values of σ are considered); the number of points inside and outside is recorded; the minimum of all those divided by n gives the value of depth.

its boundary; or, equivalently, the minimal proportion of datapoints y_i that lie in any Poincaré halfspace containing the point (μ, σ).

The second part of Theorem 3 establishes a link to the definition of the halfspace depth originally given by Tukey (1975). It turns out that the Lobachevski geometry happens to be the lucky choice among the non-Euclidean ones: still possessing parallels, albeit possibly in a non-unique fashion. The rewarding outcome is the characterization of contours: a contour on level δ is the intersection of all Poincaré halfspaces whose measure P is greater than $1 - \delta$; in sample cases that means halfspaces containing at least $n - \lceil n\delta \rceil + 1$ datapoints y_i.

Figure 4 shows how the Student depth can be calculated (for smaller data sets, obviously) in the spirit of Tukey (1977): just with the help of a right angle triangular ruler (and perhaps a pencil to record the counts). By the Thales theorem, the set of points on the line $\sigma = 0$ lying in the circle circumscribing $(y_i, 0)$, $(y_j, 0)$ and (μ, σ) is the same as the set of points lying in the right angle with the vertex (μ, σ) and sides passing through $(y_i, 0)$ and $(y_j, 0)$. In view of Theorem 3, one has just to revolve the ruler with its right angle vertex positioned at (μ, σ), starting and ending with the position when one leg is perpendicular to the line $\sigma = 0$ containing datapoints, count the number of points in and outside the angle (including in both cases those lying on the sides), and eventually take the minimum of all counts.

It may be of some interest that the hyperbolic geometry of the Student depth coincides with the Riemannian geometry generated by the Fisher information matrix, the so-called information geometry introduced by Rao (1945) and Jeffreys (1946); see Kass and Voss (1997).

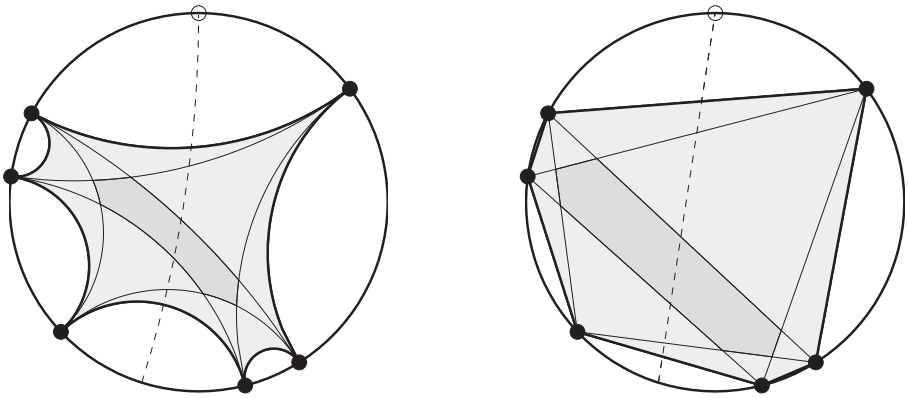


FIGURE 5. The configuration from Figure 3 transformed, as specified by (13) and (14), to the Poincaré (left) and Klein (right) disks.

It is useful to invoke also other models of the Lobachevski geometry. The right panel of Figure 5 shows the **Klein disk** $\mathbb{KD}$ where lines are represented by chords of the boundary circumference; this model allows for better thinking in usual Euclidean-geometric terms, without a need to check every move from the axioms. The transitory model is the **Poincaré disk** $\mathbb{PD}$, shown in the left panel of Figure 5, whose lines are arcs intersecting the boundary circumference in the right angle. After transforming data and parameters into the Klein disk, the Student depth reduces to the standard bivariate halfspace depth, with the added advantage that all datapoints are extremal points of their convex hull.

THEOREM 4. *The Student depth satisfies for any probability measure P:*

(i) *for all (μ,σ), $d(\mu,\sigma) \le (P(\{\mu\})+1)/2$; in particular, if P has continuous cumulative distribution function, then the depth never exceeds $1/2$; also, if no two points in a sample coincide, then the upper bound on the depth is $(n+1)/(2n)$ for n odd, and $1/2 = n/(2n)$ for n even;*

(ii) *all depth contours are connected and closed; they are compact for $\delta > 0$;*

(iii) *if P has connected support and its cumulative distribution function is continuous, then there is a unique (μ,σ) with the maximal depth;*

(iv) *there is (μ,σ) such that $d(\mu,\sigma) \ge 1/3$ (centerpoint theorem).*

A word of caution is appropriate here. The Poincaré lines and halfspaces in the Klein disk coincide with those in the ordinary Euclidean geometry sense; but this coincidence does not extend to notions like length or volume. Congruent segments in hyperbolic geometry may not possess the same Euclidean length; in particular, the hyperbolic distance of any point on the boundary of the Klein disk to any point inside it is infinite. This means that the realization of the potential strategy "transform to Klein—calculate depth—transform back" may be far from obvious, if the depth notion in the middle step is based on the distance or volume—as, for

instance, the Oja simplicial volume depth or Mahalanobis depth, both surveyed by Liu, Parelius, and Singh (1999), or the various L^1 depth versions presented by Zuo and Serfling (2000a), Vardi and Zhang (2000), or Serfling (2002).

The aforementioned strategy is, however, possible for depth notions that are, like the halfspace depth, independent of metric concepts. Such notions include simplicial and majority depth; see Liu et al. (1999). For instance, it is fairly clear what is a triangle in the Lobachevski geometry and when a point lies inside it—and this is all we need for a definition along the lines of Liu (1988, 1990). We define the **location-scale simplicial depth** of (μ, σ) to be the number, divided by $\binom{n}{3}$, of all triangles containing (μ, σ) whose vertices are datapoints. The population version can be defined accordingly. The transformation argument shows that this simplicial depth inherits all favorable properties of the two-dimensional location one, in particular the qualitative properties of contours and the U-statistical structure in the asymptotics, as elucidated by Dümbgen (1992) and Arcones, Chen, and Giné (1994). The location-scale simplicial depth contours, for the same dataset used for Figures 2 and 12, are shown in Figure 6; unlike halfspace depth, the simplicial depth assumes quite a large range of values—we outlined only about every tenth contour.

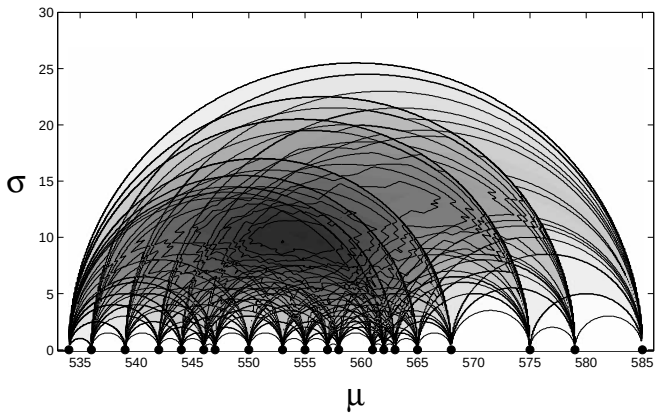


FIGURE 6. The location-scale simplicial depth assumes quite a large range of values—only every tenth contour is shown. (The dataset is the same artificial one as used in Figure 2.)

6. The Möbius equivariance and the Student median

The definition of the location-scale likelihood depth implies that it is location and scale equivariant: if we apply a transformation $g(y) = ay + b$ on the datapoints, then the depth of the transformed parameter $(a\mu + b, a\sigma)$ is the same as that of (μ, σ). This translates to location and scale equivariance of the deepest location and scale, a property shared by many location and scale estimators. However, the Student depth offers more: it is equivariant with respect to the larger Möbius group group containing all rational transformations $g(y) = (ay + b)/(cy + d)$ with

$ad - bc \neq 0$. This fact may be utilized, for instance, when the transition to reciprocal values may occur.

We follow McCullagh (1996), who formulated and showed Möbius equivariance for the Cauchy location-scale maximum likelihood estimators, also in the use of complex numbers as a convenient formalism. Our parametric space, the Poincaré plane, is quite naturally identified with the upper complex halfplane; we further extend it to the whole complex plane adding the reflection along the horizontal axis and identifying complex conjugates. The sample space remains embedded as the real line into the complex plane. McCullagh (1996) calls a statistic T **Möbius equivariant** if it is equivariant with respect to the Möbius group of the transformations: $T(gy) = \bar{g}T(y)$, where $\bar{g}$ is the transformation of the complex plane given by the same formula as g, but interpreted in the complex domain, and equality is up to complex conjugation; gy is understood as g applied coordinate-wise on the collection of datapoints constituting y.

THEOREM 5. *The Student depth satisfies, for any probability measure P and any g in the Möbius group,*

$$d(\mu, \sigma, P) = d(\bar{\mu}, \bar{\sigma}, \bar{P}), \tag{12}$$

where $\bar{P} = P \circ g^{-1}$ denotes the transformation of P under g, and $(\bar{\mu}, \bar{\sigma})$ is, up to complex conjugation, the image of (μ, σ) under $\bar{g}$.

The proof uses the fact that the Möbius group is generated by linear transformations and the reciprocal transformation $1/y$. The equivariance of the Student depth under linear transformations is quite apparent; it is only the equivariance under the reciprocal transformation that needs to be demonstrated. The latter follows by the transformation to the Poincaré or Klein disk, where $1/y$ acts as complex conjugation—symmetry about horizontal coordinate line—and then by the subsequent transformation back. In the complex notation, the isomorphisms between Poincaré models are given by the formulas

$$\mathbb{PP} \to \mathbb{PD}\colon\ z \mapsto \mathrm{i}\,\frac{z - \mathrm{i}}{z + \mathrm{i}} = \frac{1 + \mathrm{i}z}{\mathrm{i} + z}, \qquad \mathbb{PD} \to \mathbb{PP}\colon\ z \mapsto \mathrm{i}\,\frac{\mathrm{i} + z}{\mathrm{i} - z} = \frac{\mathrm{i}z - 1}{\mathrm{i} - z}. \tag{13}$$

The formula for the mapping from $\mathbb{PD}$ to $\mathbb{KD}$ shows that the direction is unchanged and alters only the absolute value by a factor $2/(1 + |z|^2)$; its inverse analogously multiplies the absolute value by $(1 - |z|)/|z|^2$:

$$\mathbb{PD} \to \mathbb{KD}\colon\ z \mapsto \frac{2z}{1 + |z|^2} = \frac{2z}{1 + z\bar{z}}, \qquad \mathbb{KD} \to \mathbb{PD}\colon\ z \mapsto \frac{z(1 - |z|)}{|z|^2} = \frac{1 - \sqrt{z\bar{z}}}{\bar{z}}. \tag{14}$$

We can see in Figure 5 that the Poincaré plane "infinity line" $\sigma = 0$ wraps on the bounding circumference, with zero positioned at its lowermost point; approaching $-\infty$ and ∞ in the Poincaré plane means approaching, from left and right, respectively, the hollow uppermost point in the image, the endpoint of the transformed dashed line from Figure 3.

The Möbius equivariance of the Student depth entails the same equivariance for the maximum Student depth estimator, which we propose to be called the **Student median**. It is defined as

the set of μ and σ having the maximal depth for given data. Its Möbius equivariance implies that it always contains the center of symmetry whenever the distribution is symmetric. Since the (standard) Cauchy distribution is invariant under the reciprocal transformation, its Student median satisfies $\mu = 0$ and $\sigma = 1$ (as confirmed by the Cauchy panel of Figure 7). For random samples from the Cauchy distributions, the Student median estimates the same quantity as the maximum likelihood estimator: the center of symmetry–median for μ, and the median absolute deviation (MAD) about the median μ for σ. For symmetric distributions, the MAD is equal to the semiinterquartile range—the probable error of McCullagh (1996), who gave closed-form formulas for the maximum Cauchy likelihood estimator when $n = 3, 4$; in those special cases, this estimator coincides with the Student median. On the basis of the formula for $n = 4$, McCullagh (1996) suggests that maximum Cauchy likelihood may be an appealing location-scale estimator for very small data sets; this recommendation thus transfers to the Student median as well. For more discussion on estimating location and scale in very small datasets, see Hoaglin, Mosteller, and Tukey (1983) and Rousseeuw and Verboven (2002).

The examples in Section 8 suggest that the location μ of the Student median lies relatively close to the sample median—in particular for data exhibiting symmetry, consistently with theoretical expectations. For asymmetric unimodal distributions, we may observe that the Student median location μ shrinks from the sample median toward the mode. We observed also that the Student median scale σ is usually shrunk down from the MAD. However, we have no exact justification for any of these claims; we can only prove that the maximal Student depth at the sample median is never too low.

THEOREM 6. *If μ is a median of the probability measure P, then* $\max_\sigma d(\mu, \sigma) \geq 1/4$.

7. THEORETICAL AND COMPUTATIONAL PROPERTIES

The notions introduced in the previous sections raise many theoretical questions whose detailed study is beyond the scope of this paper; we just try to survey properties that are either known or do not require substantial technical effort. In accord with our philosophy stated above, we assume only a general probabilistic model for the data (if any): we consider our datapoints to behave as outcomes of independent random variables with the same distribution P. There are many properties that may hold beyond this simplest **i.i.d. sampling model**, but those extensions are not pursued here. Let $\mathbb{P}_n$ denote the corresponding empirical probabilities.

THEOREM 7. *The Student depth satisfies, for any probability measure P: under the i.i.d. sampling model, $d(\mu, \sigma, \mathbb{P}_n) \to d(\mu, \sigma, P)$ uniformly in (μ, σ) almost surely.*

Theorem 7 implies the convergence of depth contours via Theorem 4.1 of Zuo and Serfling (2000b), which extends the results of He and Wang (1997). In particular it holds almost surely that $D^{\delta+\varepsilon} \subset D_n^\delta \subset D^{\delta-\varepsilon}$ for sufficiently large n, uniformly in $\delta \in [0, 1]$ for every $\varepsilon > 0$; here $D^\delta = \{(\mu, \sigma); d(\mu, \sigma, P) \geq \delta\}$ and D_n^δ is defined similarly by replacing P by $\mathbb{P}_n$. Another consequence

is the almost sure convergence of the maximal depth and the maximum depth estimators. The latter holds under some regularity conditions on the depth function $d(\mu, \sigma, P)$, for instance, the condition that the set of maximum depth is a singleton; see Theorem 2 in Mizera and Volauf (2002).

The asymptotic distribution theory of maximum depth estimators is a topic still under intense investigation—see He and Portnoy (1998), Bai and He (1999), Massé (2002, 2004). The standard $\sqrt{n}$ rate of convergence can be established in all known instances—including the Student median (Benoît Laine, personal communication, December 2003). However, the exact expressions for the asymptotic distributions and even asymptotic variances are yet unknown (except when the dimension of the parametric space is one). Simulations, like those performed by He and Portnoy (1998), indicate reasonable efficiencies, at least for low-dimensional parametric spaces.

The results of Mizera (2002) imply, in view of the centerpoint theorem for the Student depth, that the breakdown point of the Student median is not less than $\lceil n/3 \rceil$. This means considerable robustness (although certainly not the highest possible). The influence function in location and regression case was derived by Chen and Tyler (2002) and Van Aelst and Rousseeuw (2000); we do not attempt the application of the similar techniques, albeit we believe it possible.

An important theoretical question, related to the use of the Student depth for investigating distributional properties, is whether every probability measure on a real line is characterized by its Student depth function. Although the positive answer is likely, the problem is in general open. The transformation argument implies that the answer is positive for empirical and atomic distributions, via the results of Struyf and Rousseeuw (1999) and Koshevoy (2002); we believe that the technique of Koshevoy (2001) can be adapted to extend the characterization for all absolutely continuous distributions.

According to Theorem 3, the computation of the contour with the depth $\delta = k/n$ amounts to finding an intersection of all halfspaces containing at least $n-k+1$ points. In the Klein disk, that means finding the intersection of n halfspaces whose boundary contains i-th and $(i+k)$-th point, in the circular order. After constructing the initial polygon, the update for a new halfspace is the "stabbing of a convex polygon" problem, as described in Section 7.9.1 of O'Rourke (1998). Since this needs $O(\log n)$ steps, the computation of the whole contour needs $O(n \log n)$ steps. In addition to the stabbing of the polygon, one has to determine the correct orientation of the intersection, to identify the correct halfspace; however, this does not increase the time complexity.

The time complexity $O(n \log n)$ for one contour translates trivially to that of $O(n^2 \log n)$ for all contours (for graphical purposes this is overly pessimistic, since the number of required contours is usually limited by the graphical resolution of the output device). The reason why complexities $O(n \log n)$ for one and $O(n^2 \log n)$ for all contours are better than $O(n^2 \log n)$ and $O(n^3 \log n)$, reported by Ruts and Rousseeuw (1996) for the bivariate location depth, is that the data in our situation consist entirely of datapoints that are extremal points of their convex hull. This considerably simplifies the algorithms for the bivariate location depth.

Miller et al. (2003) developed an $O(n^2)$ algorithm for simultaneous computing of all location halfspace depth contours. The transformation argument implies that via their algorithm we may compute all Student depth contours in $O(n^2)$ time as well. We do not know yet whether this complexity can be improved in our special situation; note that $O(n \log n)$ is the best possible complexity if a problem requires initial sorting of the data.

Thus, we may conclude that the Student depth enjoys theoretical time complexities of the same or better order than all the cases mentioned above. While the transformation principle provides a theoretical argument, it is better in practical computations to perform all necessary operations directly in the original Poincaré plane, to avoid rounding errors arising in transforming to and from the Klein disk.

It turns out that vertical Poincaré lines are not needed. An intersections of Poincaré halfspaces can be accomplished by taking maxima and minima of the hemicircumference functions $\sigma = ((\mu - y)(\tilde{y} - \mu))^{1/2}$. Let $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ denote the ordered datapoints. For μ satisfying $y_{(k)} \le \mu \le y_{(n-k+1)}$, we define

$$c_k^-(\mu) = \max\left\{\big((\mu - y_{(i)})(y_{(i+k)} - \mu)\big)^{1/2} : i \in M_k(\mu)\right\}, \tag{15}$$

$$c_k^+(\mu) = \min\left\{\big((\mu - y_{(i)})(y_{(n-k+i)} - \mu)\big)^{1/2} : i = 1, 2, \ldots, k\right\}, \tag{16}$$

where $i \in M_k(\mu)$ means that $y_{(i)}$ are the k largest datapoints such that $y_{(i)} < \mu$.

THEOREM 8. *Let $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ be the ordered datapoints. For given δ, the contour of the Student depth is the set of all (μ, σ) such that for $k = \lceil n\delta \rceil$,*

$$y_{(k)} \le \mu \le y_{(n-k+1)} \quad \textit{and} \quad c_k^-(\mu) \le \sigma \le c_k^+(\mu).$$

Analogously, for general P with the cumulative distribution function $F(y) = P((-\infty, y])$, we define for μ satisfying $\delta \le \min\{F(\mu), 1 - F(\mu)\}$,

$$c_{\delta,P}^-(\mu) = \sup\left\{((\mu - q_\beta)(q_{\delta+\beta} - \mu))^{1/2} : \beta \in (F(\mu) - \delta, F(\mu))\right\}, \tag{17}$$

$$c_{\delta,P}^+(\mu) = \inf\left\{((\mu - q_\beta)(q_{1-\delta+\beta} - \mu))^{1/2} : \beta \in (0, \delta)\right\}, \tag{18}$$

where $q_\beta = \min\{y \in \mathbb{R}\colon F(y) \ge \beta\}$ is the β-quantile of P. Note that if P is an empirical distribution, (17)–(18) reduce to (15)–(16).

THEOREM 9. *Let P be a probability measure with the cumulative distribution function F. For given δ, the contour of the Student depth is the set of all (μ, σ) such that*

$$\delta \le \min\{F(\mu), 1 - F(\mu)\} \quad \textit{and} \quad c_{\delta,P}^-(\mu) \le \sigma \le c_{\delta,P}^+(\mu).$$

If the distribution is symmetric about μ_0, then the depth contours are also symmetric about μ_0. If there is a unique deepest point (μ, σ), then μ must be equal to μ_0, and Theorem 9 yields that $\sigma = c_{\delta,P}^-(\mu) = c_{\delta,P}^+(\mu)$, where δ is the depth of (μ, σ). According to Theorem 4(iii), the

deepest point is unique, for instance, if P has connected support and its cumulative distribution function is continuous.

8. Data examples: the Student depth in action

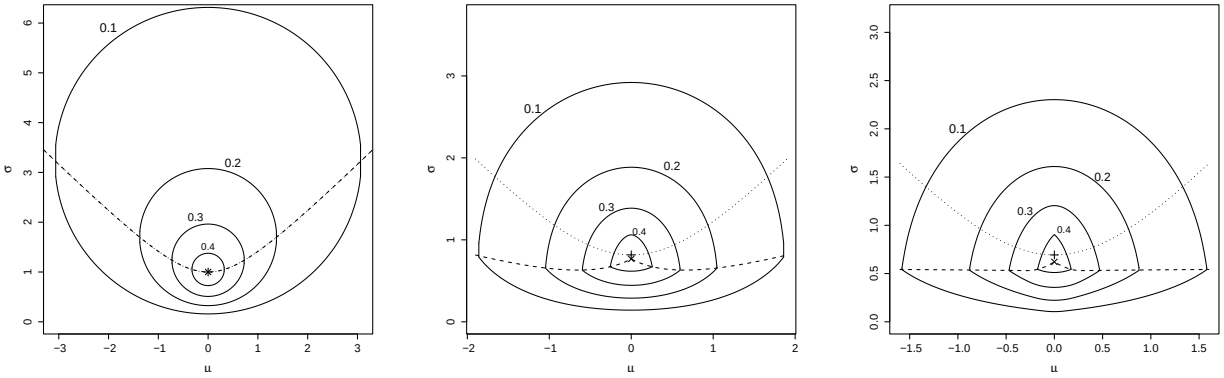
We analyzed several univariate datasets to illustrate the directions formulated at the end of Section 4. The central object of our analyses was the plot of the Student depth contours. For small datasets, we plotted also the original datapoints; this is not practical for larger samples.

According to the visual desiderata formulated by Cleveland (1994), it may be desirable to plot only selected contours. Indeed, according to our limited experience, a smaller number of contours is often better; an extreme possibility is to construct a kind of "Student boxplot" and plot only the contours with the depth of approximately 1/4, 1/2 and 3/4 of the maximal one (and possibly also the first depth contour by a dotted line or so). Another possibility is to render principal contours by thicker lines and several other ones by thinner ones, or indicate the intermediate contours just by shading. The position of the Student median is indicated by "$\times$"; we also report the maximal attained depth. We are still in the process of experimenting what graphical appearance would be ideal for the Student depth plots; to help our readers to form their own preferences, we do not present our plots in a uniform style, but rather with minor alterations.

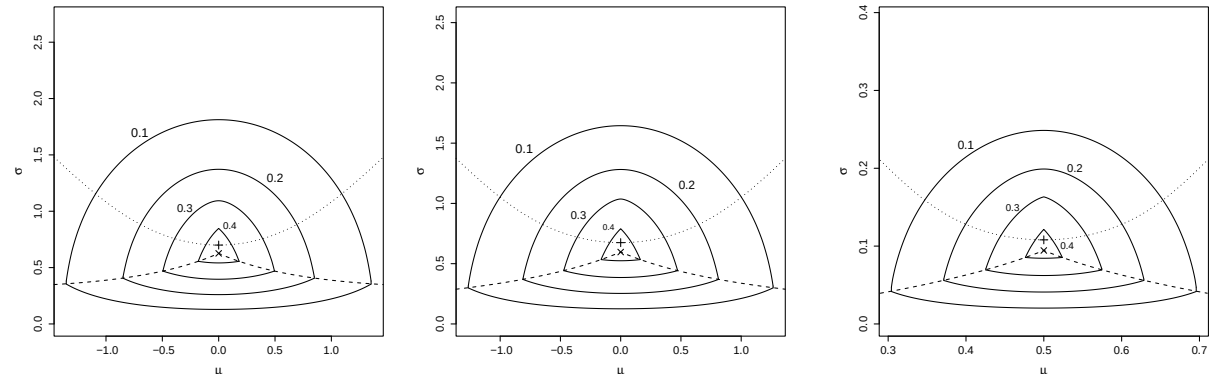
We also plot some additional information. The dashed line indicates the scale σ with the maximal depth among (μ, σ) for a fixed μ. The dotted line plots the median absolute deviation from μ against this μ; the special case when μ is the sample median is marked by "+".

Chambers et al. (1983) analyze the distribution of their datasets by the quantile plots of the original and transformed data, a well-known technique now—the empirical quantiles are plotted against the quantiles from various theoretical distributions, most prominently the Gaussian; this allows for penetrating comparisons of observed and theoretical distributions. Location-scale depth contours plots do not possess a spare dimension, hence our comparison strategy should be different: we compare the observed Student depth contours visually to the theoretical ones plotted for the hypothesized distribution. Due to the location and scale equivariance, we do not have to estimate location and scale parameters—this is an advantage to quantile plots, where, except for the Gaussian case, some value of the scale has to be specified.

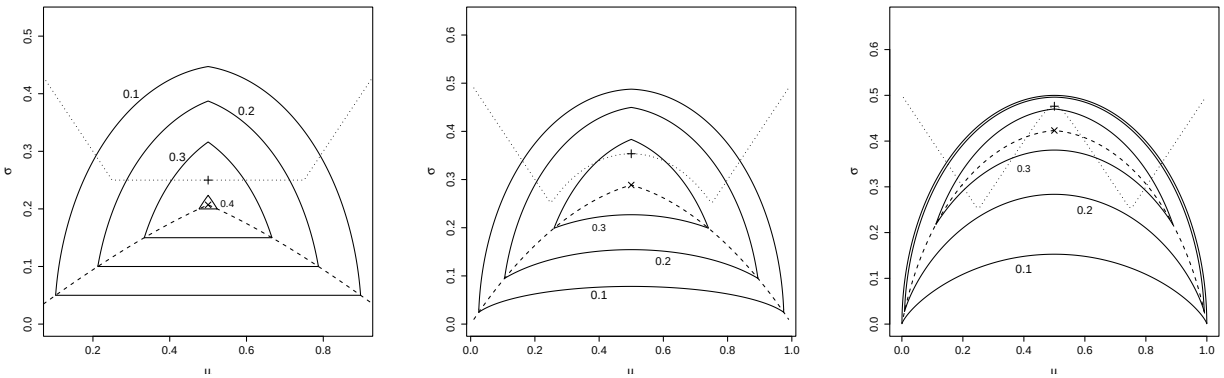
To obtain an initial sampler of theoretical depth contours, we plotted them for selected distributions; the results can be seen in Figures 7 and 8. (In future implementations, we hope to possess an ability to create "model shapes" interactively.) Note that, in particular, the Cauchy distribution is the only one among the displayed distributions with the maximal depth equal to 1/2 (the maximal depth 1/2 can be easily proved for the Cauchy distribution rigorously, via Theorem 9). The pictures indicate that different distributions create different characteristic shapes. The first question is how far those will reveal themselves in the sampled data.



Cauchy, t(1) [*0.500*] t(2) [*0.477*] Laplace [*0.463*]

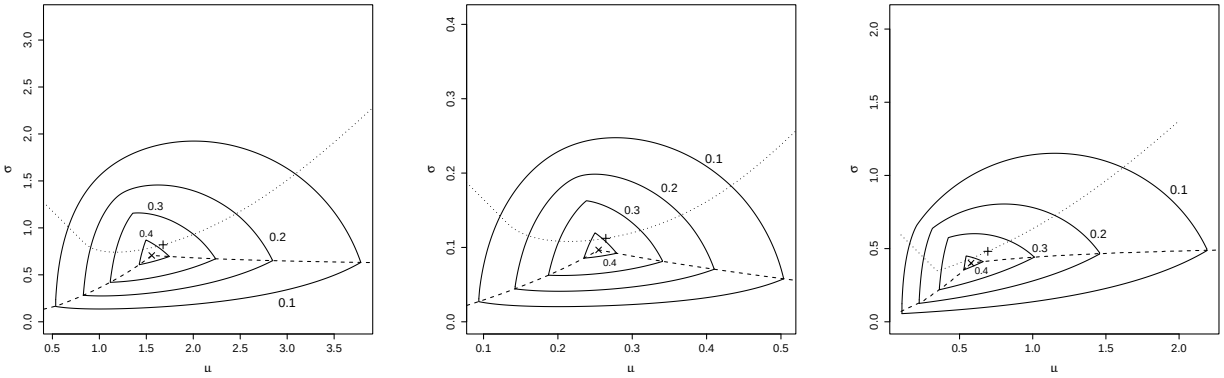


t(10) [*0.455*] Normal [*0.449*] Beta(5,5) [*0.443*]

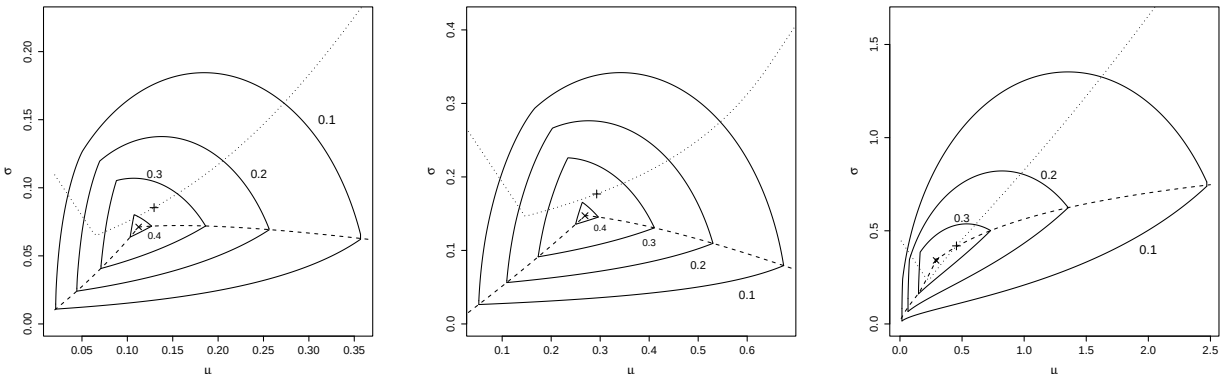


Uniform [*0.414*] Beta(0.5,0.5) [*0.392*] Beta(0.2,0.2) [*0.363*]

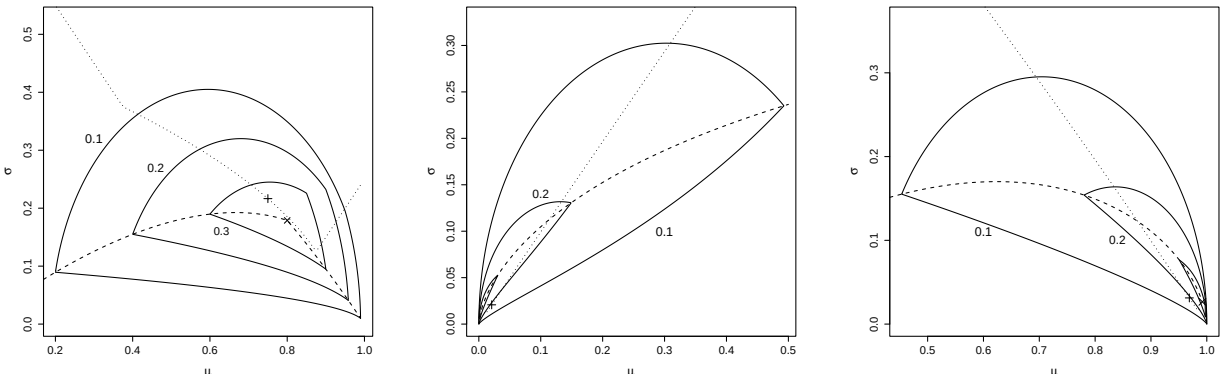
FIGURE 7. The Student depth for selected symmetric distributions. The x-axes correspond to μ and the y-axes to σ; the x:y aspect of all plots is 1:1, except for the three panels in the last line, where it is 2:3. The dashed line indicates the scale σ with the maximal depth for fixed μ; the dotted line plots the median absolute deviation from μ, the special case of the sample median marked by +. Bracketed numbers in italics give maximal attained depth.



Gamma(2,1) [*0.438*] Beta(2,5) [*0.437*] Exponential [*0.425*]



Beta(1,5) [*0.424*] Beta(1,2) [*0.422*] Chi-square(1) [*0.405*]



Beta(1,0.05) [*0.399*] Gamma(0.2,1) [*0.373*] Beta(1,0.2) [*0.372*]

FIGURE 8. The Student depth for selected asymmetric distributions. The x-axes correspond to μ and the y-axes to σ; the x:y aspect of all plots is 2:3, except for the three panels in the first line, where it is 1:1. The dashed line indicates the scale σ with the maximal depth for fixed μ; the dotted line plots the median absolute deviation from μ, the special case of the sample median marked by +. Bracketed numbers in italics give maximal attained depth.

EXAMPLE 1. SIMULATED GAUSSIAN/CAUCHY MIXTURES. To assess this, we start by an artificial example. Two Gaussian/Cauchy mixtures were simulated: the first contained 35 datapoints generated from the Gaussian and 10 from the Cauchy distribution; the second 10 Gaussian and 35 Cauchy datapoints. Thus, the mixing proportions are the same, only the first sample is majority Gaussian and the second majority Cauchy. The contour plots for both samples are displayed in the left and right sides of Figure 9, respectively.

Normal quantile-quantile plots are added in the middle of the first row, to be compared with the contour plots in terms of visual impression. The latter is, in our opinion, quite similar for both samples; they exhibit outliers and/or heavier tails. In a real situation, we would not have any a priori information about the scale, so the different slope of the line is not decisive. The only clue is perhaps that the outliers of the majority Cauchy sample are considerably wilder and perhaps slightly more transparent. The latter appears also to be more linear in the middle of the sample.

The first row exhibits also the Student depth contours of the sampling distribution mixtures. It is surprising how a majority of approximately 2/3 dominates the shapes of the contours. The contour shapes of the leftmost resemble the Gaussian ones from Figure 7; those in the rightmost the Cauchy ones.

The second row of Figure 9 shows the Student depth contours for the simulated data. In the raw form they are not very informative; it is much better to look at the deeper contours under some magnification. (The ability of convenient rescaling, as well as of controlling the aspect of axes, may be an important requirement for the potential routine use of these plots.) The results are shown in the third row of Figure 9. The samples exhibit considerable reproduction of the theoretical pattern of the first line for a relatively small sample size, in particular the majority distribution pattern (Gaussian and Cauchy, respectively) is revealed.

Of course, the question is whether all the plots are not merely an artifacts of the simulation. This is hard to dispute: due to space constraints, we cannot show a large number of simulations, and a quantitative measure of overall similarity is not available. Nevertheless, the last row of Figure 9 presents, as a kind of compromise, the (inner) Student depth contours of three additional simulations from each mixture. (We observed much better agreement for the doubled sample sizes 70 and 20.)

A possible objection to any methodology like this one is that it requires a considerable training. We do not deny this, we only remark that the amount of the required investment may be in the eye of beholder. In particular, our teaching experience reminds us often that even seemingly obvious procedures like quantile plots are not perceived necessarily as such by beginners.

EXAMPLE 2. SEEDED RAINFALL DATA. In this example, we create a sequel to the story begun by Chambers et al. (1983). They end by the conclusion that rather than the gamma distribution for the datapoints y_i, proposed by the earlier authors cited therein, the better fit is achieved

by the Gaussian distribution fitting $y_i^{0.12}$. This conclusion is drawn from the straightness of the corresponding normal quantile-quantile plot, shown in the left panel of Figure 10.

Chambers, Cleveland, Kleiner, and Tukey did not routinely use the standard method of fitting the line to the quantile-quantile plot through the quartile points (they have only one or two pictures with it in their book), since probably they would see then that the plot indicates somewhat heavier tails than Gaussian. Apparently, the power transformation symmetrizes, but not necessarily normalizes the data.

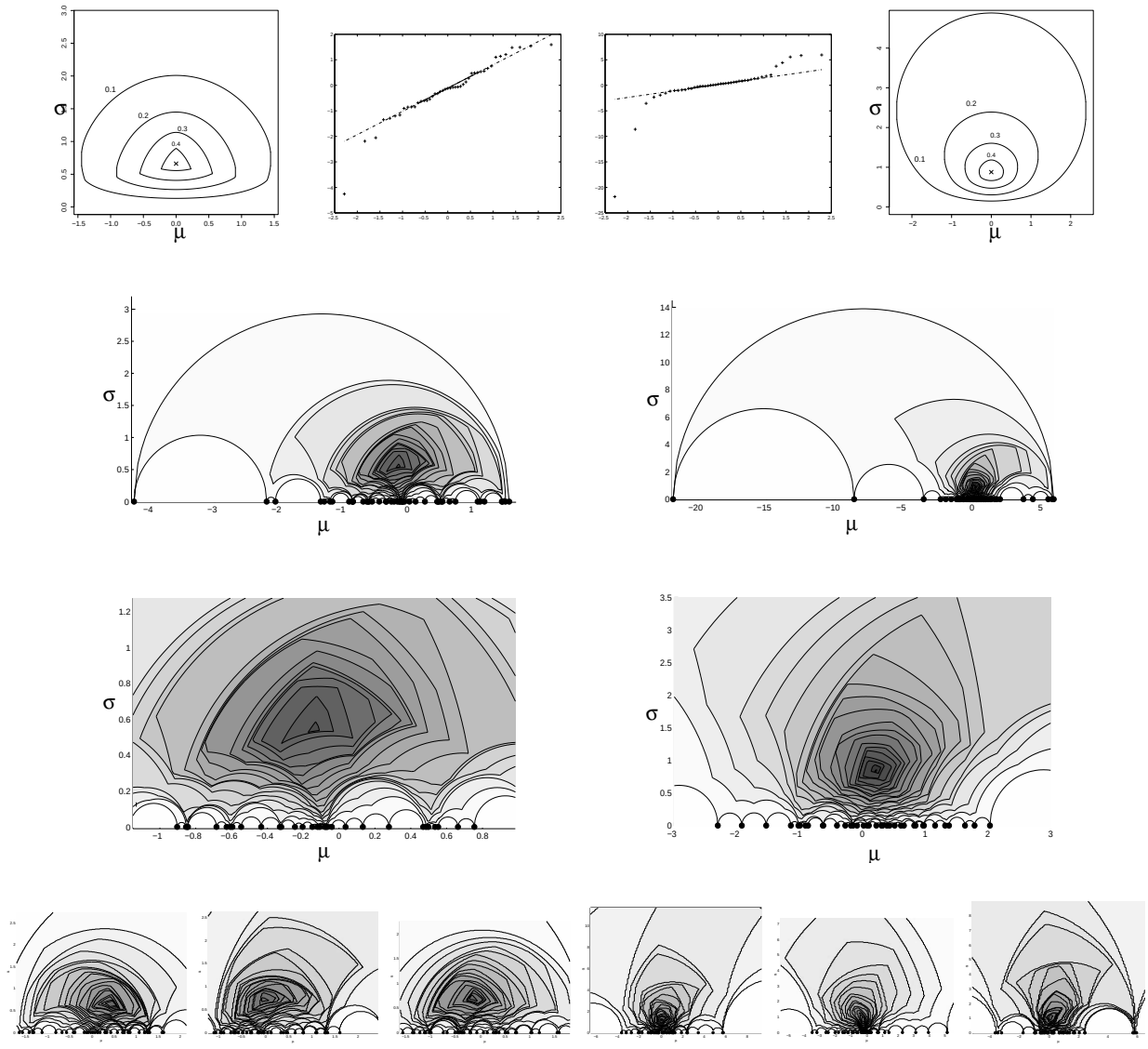


FIGURE 9. 35:10 and 10:35 Gaussian/Cauchy mixtures. Patterns or artifacts?

Alternatives coming to mind are the t distributions, but also several others. We could try a couple of quantile plots; instead, a faster way is to compare the Student contour plot, in the middle panel of Figure 10, with Figure 7. This comparison turns our attention to the Laplace distribution—from all available models this appears as most acceptable, because the inner contours are somewhat more stretched downwards. Actually, after inspecting several quantile plots, the one for the Laplace distribution fits best; see the right panel of Figure 10. In fact, this plot is almost the same as for the t distribution with $\nu = 3$; using only quantile plots, we would not be able to distinguish between this t and the Laplace distribution in this example.

We have to remark that even the Laplace fit may be felt not yet completely satisfactory; thus our story, ending at this point, may have another sequel elsewhere. And after all, the sample size is indeed small. We tried also the Student depth contours for the original untransformed data; our conclusions in this case support those of Chambers et al. (1983) regarding the gamma fits.

Example 3. Intervals between earthquakes. In our last example, we will analyze the data whose distribution is beyond any doubt asymmetric. Recall that our philosophy is consistent with applying a method derived from symmetric likelihoods to asymmetric distributions—once the method was derived, we assess its validity in a nonparametric broader context, without referring to the original working assumptions.

The datapoints are the periods between earthquakes, recorded as number of days between successive serious earthquakes worldwide. For more details and the original source of the dataset, see Hand, Daly, Lunn, McConway, and Ostrowski (1994), who comment on their Dataset 255 that "if earthquakes occur at random, an exponential model for these data should provide a reasonable fit." To assess this graphically, we may try the kernel density situation, as given in the leftmost panel of Figure 11, but this is slightly inappropriate in this situation—we know that the data are positive—and thus would require further adjustments. The next possibility is the quantile-quantile plot, as shown in the rightmost panel of Figure 11, with the rate λ estimated by

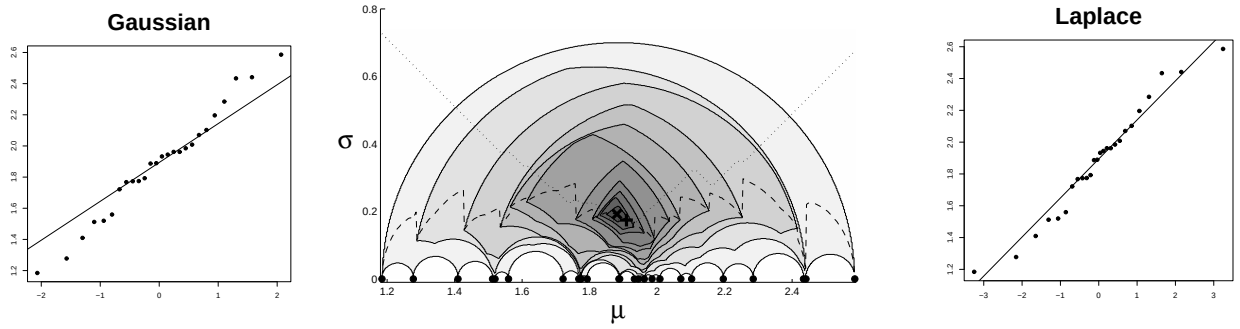


Figure 10. Seeded rainfall data. Peaked or heavy-tailed?

1/mean and on the log scale as recommended by Chambers et al. (1983) to avoid the cluttering of points near the origin. The plot supports the hypothesis of exponentiality.

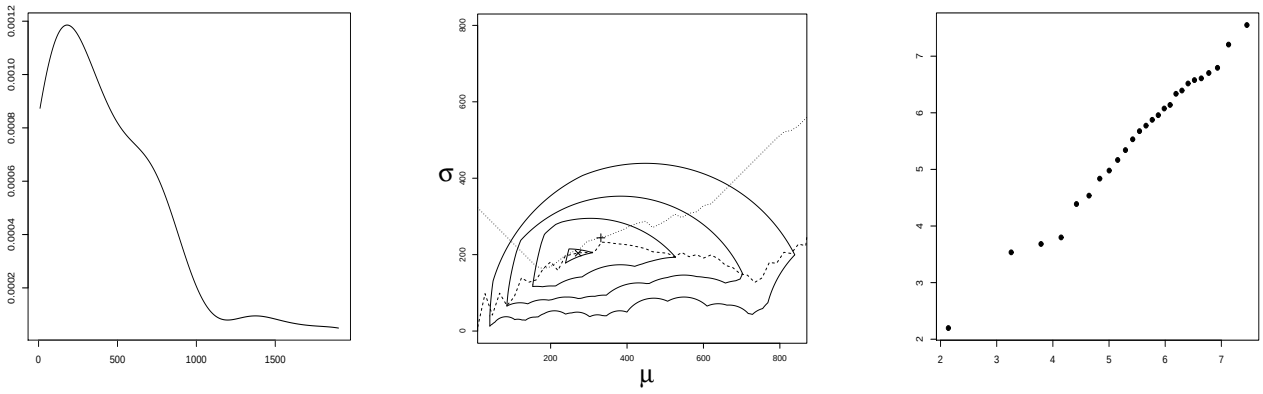


FIGURE 11. Intervals between earthquakes. Are they exponentially distributed? The left panel shows the kernel density estimate, the right panel the quantile-quantile plot on the logarithmic scale. The middle panel shows the contours of the Student depth; note the shift of the Student median location from the median to the mode.

The plot of the Student depth contours in the middle panel supports it too, except for the dashed curve of maximal depth σ is descending rather than ascending; the right-side contours are somewhat closer each to other than those in the model plot in Figure 8. However, these observations may be mere artifacts; confidence bands, or some other exact means, would be needed to conclude whether the dashed curve really descents; it is quite likely that any such band obtained, say, by resampling, would be wider that the amount of descent. Nevertheless, the plot is intended primarily as an exploratory tool here; and in this capacity it is quite informative, albeit its analysis requires some training. In any case, the sample size 62 does not allow for definitive conclusions.

Summarizing our limited experience, we can say that the Student depth plots easily reveal asymmetry, including that present in the core of the data, rather than just in the tails; but they are capable of detecting heavy-tailed behavior too. To get more out of plots, it is better to look rather on deeper contours—which may need some magnification of the central part of the plot. The maximal Student depth contour marks the location of the Student median and thus also gives an idea about the location and scale of data. The contours exhibit different characteristic shapes for different distributions, and therefore they may suggest something about the distribution of the data.

9. Conclusion and open problems

The location-scale depth is not only a non-trivial, novel instance for the general theory of Mizera (2002), but its most tractable version, the Student depth, enjoys remarkable theoretical and computational properties. The maximum depth estimator based on it, the Student median, constitutes a location-scale estimator of median type. Plots of the Student depth contours have some potential to become a graphical tool of exploratory data analysis—although much more experience has to be gathered yet. Some information about the distribution of the data is also provided by the maximal depth. The whole methodology is considerably robust.

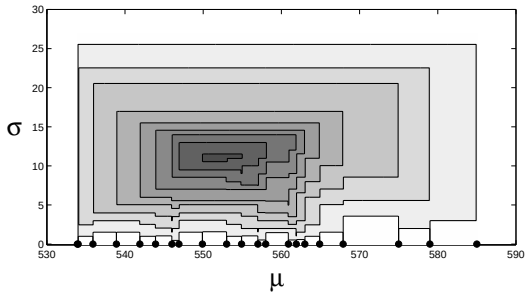


Figure 12. The Laplace version of the location-scale depth has uncertain properties—stemming from the fact that instead of circles we obtain rectangles.

Our paper leaves several directions open for the future research. Definition 1 leads to various versions of location-scale depth, depending on the choice of the normalized density f. It is possible, as suggested by Figure 2, that they are similar in some sense—perhaps some of them are equivalent up to a reparametrization. However, we do not possess any formal insights in this direction; the situation may be not that simple, as indicated by Figure 12 showing the location-scale depth for the Laplace f.

A rather minor direction, in our opinion, concerns exploring the approach that defined simplicial depth, in Section 5, to define other notions of location-scale depth. We already indicated that such a task may be formidable if attempted in a conceptually clean way. Of course, there is always a tempting possibility to simply ignore the hyperbolic geometry in the Klein disk and consider the ordinary Euclidean geometry instead; we believe that this would lead to unpredictable consequences. Such a move, however, may provide a good local approximation—for instance, as a computational shortcut for finding a center of gravity of the deepest contour.

The more promising directions for the future research include deeper theoretical investigation of the Student median, a straightforward but technically somewhat demanding extension of the Student depth to the multivariate location-scale model, and likelihood-based principles for designing criterial functions, resulting halfspace depths and their properties in various models of data analysis.

References

Arcones, M. A., Chen, Z., and Giné, E. (1994), "Estimators related to U-processes with Applications to Multivariate Medians: Asymptotic Normality," *Annals of Statistics*, 22, 1460–1477.

Bai, Z.-D. and He, X. (1999), "Asymptotic Distributions of the Maximal Depth Estimators for Regression and Multivariate Location," *Annals of Statistics*, 27, 1616–1637.

Chambers, J., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Boston: Duxbury Press.

Chaudhuri, P. and Sengupta, D. (1993), "Sign Test in Multidimension: Iinference Based on the Geometry of the Data Cloud," *Journal of the American Statistical Association*, 88, 1363–1370.

Chen, Z. and Tyler, D. E. (2002), "The Influence Function and Maximum Bias of Tukey's Median," *Annals of Statistics*, 30, 1737–1759..

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Summit, N.J.: Hobart Press.

Donoho, D. L. and Gasko, M. (1992), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness," *Annals of Statistics*, 20, 1803–1827.

Dümbgen, L. (1992), "Limit Theorems for the Simplicial Depth," *Statistics and Probability Letters*, 14, 119–128.

Fisher, N. I. (1983), "Graphical Methods in Nonparametric Statistics: A Review and Annotated Bibliography," *International Statistical Review*, 51, 25–58.

Greenberg, M. J. (1980), *Euclidean and Non-Euclidean Geometries: Development and History (2nd ed.)*, San Francisco: W. H. Freeman and Company.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994), *A Handbook of Small Data Sets*, London: Chapman and Hall.

He, X. and Portnoy, S. (1998), "Asymptotics of the Deepest Line," in *Applied Statistical Science III: Papers in Honor of A. K. Md. E. Saleh*, eds. Ahmed, S. E., Ahsanullah, M., and Sinha, B. K., Commack, N.Y.: Nova Science Publications, pp. 71–81.

He, X. and Wang, G. (1997), "Convergence of Depth Contours for Multivariate Datasets," *Annals of Statistics*, 25, 495–504.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983), "Introduction to More Refined Estimators," in *Understanding Robust and Exploratory Data Analysis*, eds. Hoaglin, D. C., Mosteller, F., and Tukey, J. W., New York: John Wiley and Sons.

Huber, P. J. (1967), "The Behaviour of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, eds. Neyman, J. and Le Cam, L., Berkeley: University of California Press, pp. 221–233.

— (1981), *Robust Statistics*, New York: John Wiley and Sons.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London, Ser. A*, 196, 453–461.

Kass, R. E. and Voss, P. W. (1997), *Geometrical Foundations of Asymptotic Inference*, New York: John Wiley and Sons.

Koshevoy, G. A. (2001), "Projections of Lift Zonoids, the Oja Depth and Tukey Depth," unpublished manuscript.

— (2002), "The Tukey Depth Characterizes the Atomic Measure," *Journal of Multivariate Analysis*, 83, 360–364.

Liu, R. Y. (1988), "On a Notion of Simplicial Depth," *Proceedings of the National Academy of Sciences of the USA*, 85, 1732–1734.

— (1990), "On a Notion of Data Depth Based on Random Simplices," *Annals of Statistics*, 18, 405–414.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference," *Annals of Statistics*, 27, 783–840.

Massé, J.-C. (2002), "Asymptotics for the Tukey Median," *Journal of Multivariate Analysis*, 81, 286–300.

— (2004), "Asymptotics for the Tukey Depth Process: Application to a Multivariate Trimmed Mean," *Bernoulli*, in press.

Massé, J.-C. and Theodorescu, R. (1994), "Halfplane Trimming for Bivariate Distributions," *Journal of Multivariate Analysis*, 48, 188–202.

McCullagh, P. (1996), "Möbius Transformation and Cauchy Parameter Estimation," *Annals of Statistics*, 24, 787–808.

Miller, K., Ramaswami, S., Rousseeuw, P., Sellarès, J. A., Souvaine, D., Streinu, I., and Struyf, A. (2003), "Efficient Computation of Location Depth Contours by Methods of Computational Geometry," *Statistics and Computing*, in press.

Mizera, I. (2002), "On Depth and Deep Points: A Calculus," *Annals of Statistics*, 30, 1681–1736.

Mizera, I. and Volauf, M. (2002), "Continuity of Halfspace Depth Contours and Maximum Depth Estimators: Diagnostics of Depth-Related Methods," *Journal of Multivariate Analysis*, 83, 365–388.

Müller, C. H. (2003), "Depth Estimators and Tests Based on the Likelihood Principle with Application to Regression," unpublished manuscript.

O'Rourke, J. (1998), *Computational Geometry in C*, Cambridge: Cambridge University Press.

Rao, C. R. (1945), "Information and Accuracy Attainable in the Estimation of Statistical Parameters," *Bulletin of Calcutta Mathematical Society*, 37, 81–89.

Rousseeuw, P. J. and Hubert, M. (1999), "Regression Depth" (with discussion), *Journal of the American Statistical Association*, 94, 388–402.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999), "The Bagplot: A Bivariate Boxplot," *American Statistician*, 53, 382–387.

Rousseeuw, P. J. and Struyf, A. (2002), "A Depth Test for Symmetry," in *Goodness-Of-Fit Tests and Model Validity*, eds. Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., and Mesbah, M., Birkhäuser, pp. 401–412.

Rousseeuw, P. J. and Verboven, S. (2002), "Robust Estimation in Very Small Samples," *Computational Statistics and Data Analysis*, 40, 741–758.

Ruts, I. and Rousseeuw, P. J. (1996), "Computing Depth Contours of Bivariate Point Clouds," *Computational Statistics and Data Analysis*, 23, 152–168.

Serfling, R. (2002), "A Depth Function and a Scale Curve Based on Spatial Quantiles," in *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, ed. Dodge, Y., Basel: Birkhäuser Verlag, pp. 25–38.

Struyf, A. and Rousseeuw, P. J. (1999), "Halfspace Depth and Regression Depth Characterize the Empirical Distribution," *Journal of Multivariate Analysis*, 69, 135–153.

Tukey, J. W. (1975), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, Quebec: Canad. Math. Congress, pp. 523–531.

— (1977), *Exploratory Data Analysis*, Reading, Massachusets: Addison-Wesley.

Van Aelst, S. and Rousseeuw, P. J. (2000), "Robustness Properties of Deepest Regression," *Journal of Multivariate Analysis*, 73, 82–106.

Van Aelst, S., Rousseeuw, P. J., Hubert, M., and Struyf, A. (2002), "The Deepest Regression Method," *Journal of Multivariate Analysis*, 81, 138–166.

Vardi, Y. and Zhang, C.-H. (2000), "The Multivariate L_1-Median and Associated Data Depth," *Proceedings of the National Academy of Sciences of the USA*, 97, 1423–1426.

Zuo, Y. and Serfling, R. (2000a), "General Notions of Statistical Depth Function," *Annals of Statist*, 28, 461–482.

— (2000b), "Structural Properties and Convergence Results for Contours of Sample Statistical Depth Functions," *Annals of Statistics*, 28, 483–499.

Appendix

Proof of Theorem 1. As there is nothing to prove when $\sigma = 0$, we may assume that $\sigma > 0$. Then the theorem is proved by showing that

$$u_1\psi(\tau_i) + u_2(\tau_i\psi(\tau_i) - 1) \geq 0$$

is equivalent to

$$u_1\tau_i + u_2(\tau_i^2 - \zeta(\tau_i)) \geq 0,$$

which is done either when $\tau_i \neq 0$ by algebraic manipulations, using the assumption $\text{sgn}(\psi(\tau)) = \text{sgn}(\tau)$, or when $\tau_i = 0$ by observing that this case means $-u_2 \geq 0$ which is equivalent to $-u_2\,\zeta(\tau_i) \geq 0$ whenever $\zeta(0) > 0$. □

Proof of Theorem 2. The inequalities follow from the observation that (4) in Definition 1 minimizes over sets that contain all those appearing in (6), for $u_2 = 0$, and all those appearing in (8), for $u_1 = 0$. □

PROOF OF THEOREM 3. The starting point is (11) from Definition 2, rewritten as follows:

$$d(\mu,\sigma) = \inf_{u\neq 0} \#\left\{i\colon u_1\tau_i + u_2(\tau_i^2-1) \geq 0\right\} = \inf_{u\neq 0} \#\left\{i\colon u_1\tau_i + u_2\left(\tfrac{1}{2}(\tau_i^2-1)\right) \geq 0\right\}.$$

It is enough to take the inf in the last expression just over $\|u\| = 1$, that is, over $u = (\cos 2\alpha, \sin 2\alpha)$ with $\alpha \in [0,\pi)$:

$$d(\mu,\sigma) = \inf_{\alpha} \#\{i\colon y_i \in H_\alpha\}, \tag{19}$$

where H_α is the set of all y such that

$$(\cos 2\alpha)\,\frac{y-\mu}{\sigma} + (\sin 2\alpha)\left(\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\right) \geq 0. \tag{20}$$

On solving the quadratic inequality (20) for y, we obtain that

$$H_\alpha = \begin{cases} \left(-\infty, \mu - \sigma\dfrac{\cos\alpha}{\sin\alpha}\right] \cup \left[\mu + \sigma\dfrac{\sin\alpha}{\cos\alpha}, \infty\right), & \text{for } \alpha \in [0, \tfrac{1}{2}\pi), \\ \left(\mu + \sigma\dfrac{\sin\alpha}{\cos\alpha}, \mu - \sigma\dfrac{\cos\alpha}{\sin\alpha}\right) & \text{for } \alpha \in [\tfrac{1}{2}\pi, \pi). \end{cases} \tag{21}$$

In both cases, the boundary is the intersection of the line $\sigma = 0$ with the circumference centered at $(\mu - \sigma(\cos 2\alpha)/(\sin 2\alpha), 0)$, with radius $\sigma/|\sin 2\alpha|$; a straightforward verification shows that also point (μ,σ) lies on this circumference. This concludes the proof of the first part, in view of (19).

For the second part, we have just to show that given any Poincaré halfspace containing (μ,σ), there is another Poincaré halfspace contained in the first one and such that (μ,σ) lies on its boundary. Once this holds, the infimum of the cardinality (or measure) of the points contained in a halfspace taken over all halfspaces with (μ,σ) on their boundary is not smaller than that taken over all halfspaces containing (μ,σ). And the converse inequality is trivial, since any halfspace with (μ,σ) on its boundary contains (μ,σ).

As already mentioned in the main text, the desired property follows from the behavior of parallels in the Lobachevski geometry. Given a halfspace $\tilde{H}$ and a point (μ,σ), either this point lies on the Poincaré line $\tilde{\ell}$ forming the boundary of $\tilde{H}$, and then the property holds trivially, or the point (μ,σ) does not lie on $\tilde{\ell}$ and then there exists a Poincaré line ℓ through this point not intersecting $\tilde{\ell}$. Consequently, there is a Poincaré halfspace $H \subset \tilde{H}$ with boundary ℓ containing (μ,σ). □

PROOF OF THEOREM 4. All the properties follow by transformation to the Klein disk, then by applying some property of the halfspace depth, and then by applying the inverse transformation, if necessary. For (i), see Proposition 5.10 of Mizera (2002). Part (ii) follows from the convexity of the depth contours in the Klein disk; since the isomorphism to Poincaré plane is continuous, their connectedness is preserved. The assumptions of (iii) assert that the distribution function is strictly increasing and P assigns a positive probability to any nonempty open interval; consequently, the transform of P assigns a nonzero probability to any strip with nonempty

interior in the Klein disk; Proposition 7 of Mizera and Volauf (2002) then implies (iii); see also Proposition 3.5 of Massé and Theodorescu (1994). The centerpoint theorem (iv) for the Student depth could be proved also via results of Mizera (2002), but here it follows by the transformation argument more directly, as the corollary of the standard centerpoint theorem in the bivariate location model. □

Proof of Theorem 5. The Möbius group is generated by the linear (affine) transformations and the reciprocal transformation. The equivariance under linear transformations is obvious from the definition; hence it remains only to prove the theorem for $gy = 1/y$.

We do this by straightforward verification. Given data y, we transform them into the Poincaré disk. If the datapoint is $z = x + 0\mathrm{i}$, then simple algebra using left part of the formula (13) shows that its transformation,

$$\frac{2x}{x^2+1} + \mathrm{i}\frac{x^2-1}{x^2+1},$$

is the complex conjugate of the transformation of $1/x$. That is, in the Poincaré disk, the datapoints corresponding to $1/y$ are those flipped about the real line. It follows that parameters from the inside of the Poincaré disk retain their depth when flipped in the same way: in other words, the depth of $a + \mathrm{i}b$ in the Poincaré disk is with respect to the original data y the same as the depth of $a - \mathrm{i}b$ under $1/y$.

Now we have to calculate what does this mean in the original Poincaré plane. If a parameter $\mu + \mathrm{i}\sigma$ transforms to $a + \mathrm{i}b$ (beware: the formula makes sense only for the original Poincaré, that is, upper halfplane, so we have to start with $\sigma > 0$ at this point), then under the inverse transformation, expressed by the right part of the formula (13), it transforms back to itself; while $a - \mathrm{i}b$ transforms to $(\mu + \mathrm{i}\sigma)/(\mu^2 + \sigma^2)$. A simple verification shows that $1/(\mu + \mathrm{i}\sigma) = (\mu - \mathrm{i}\sigma)/(\mu^2 + \sigma^2)$, the same parametric value up to complex conjugation. □

Proof of Theorem 6. The theorem follows from Theorem 9, via the elementary inequalities $\beta_1 \le \beta_2$ and $1 - \delta + \beta_1 \ge \beta_2 + \delta$, holding whenever $\delta \le 1/4$, $0 \le \beta_1 \le \delta$, and $1/2 - \delta \le \beta_2 \le 1/2$. Consequently, $\mu - q_{\beta_1} \ge \mu - q_{\beta_2}$ and $q_{1-\delta+\beta_1} - \mu \ge q_{\delta+\beta_2} - \mu$; this results in $c^+_{\delta,P}(\mu) \ge c^-_{\delta,P}(\mu)$ if $\mu = q_{1/2}$, in view of (17)–(18). □

Proof of Theorem 7. According to Theorem 3, the Student depth is computed as the infimum of measures of certain intervals in the real line. Under the i.i.d. sampling model, the almost sure uniform convergence of the measures of those intervals follows from the Glivenko-Cantelli theorem; the almost sure uniform convergence of the infima follows. □

For notational simplicity, we assume in the proofs of Theorems 8 and 9 that $y_1 \le y_2 \le \dots y_n$. It is convenient to represent the Student depth for $\sigma > 0$ in the vein of (20) as

$$
\begin{aligned}
n\, d(\mu,\sigma) &= \inf_{\alpha\in[-\pi,\pi]} \operatorname{card}\left\{ i\colon\ \sin(\alpha)\left(\frac{y_i-\mu}{\sigma}\right) + \cos(\alpha)\left(\left(\frac{y_i-\mu}{\sigma}\right)^2 - 1\right) \le 0 \right\} \\
&= \inf_{\alpha\in[-\pi,\pi]} \sum_{y_i<\mu} 1\{\sin(\alpha)+\cos(\alpha)a_i(\mu,\sigma)\ge 0\} + \sum_{y_i=\mu} 1\{-\cos(\alpha)\le 0\} \\
&\quad + \sum_{y_i>\mu} 1\{\sin(\alpha)+\cos(\alpha)a_i(\mu,\sigma)\le 0\} \\
&= \begin{cases}
\sum_{y_i<\mu} 1\{-\tan(\alpha)\ge a_i(\mu,\sigma)\} + \sum_{y_i>\mu} 1\{-\tan(\alpha)\le a_i(\mu,\sigma)\}, \\
\qquad \text{if } \cos(\alpha)<0 \\
\sum_{y_i<\mu} 1\{-\tan(\alpha)\le a_i(\mu,\sigma)\} + \operatorname{card}\{i:\ y_i=\mu\} + \sum_{y_i>\mu} 1\{-\tan(\alpha)\ge a_i(\mu,\sigma)\}, \\
\qquad \text{if } \cos(\alpha)>0 \\
\sum_{y_i<\mu} 1\{\sin(\alpha)\ge 0\} + \operatorname{card}\{i:\ y_i=\mu\} + \sum_{y_i>\mu} 1\{\sin(\alpha)\le 0\}, \\
\qquad \text{if } \cos(\alpha)=0
\end{cases}
\end{aligned}
$$

where

$$
a_i(\mu,\sigma) = a_{\mu,\sigma}(y_i) = \frac{y_i-\mu}{\sigma} - \frac{\sigma}{y_i-\mu} \tag{22}
$$

and $1\{\dots\}$ abbreviates the indicator function $1_{\{\dots\}}(\alpha)$. The following lemma follows from routine algebraic calculations.

LEMMA 1.
(a) If $y < \tilde{y}$ and $(\mu-y)(\tilde{y}-\mu) < 0$, then $a_{\mu,\sigma}(y) \lessgtr a_{\mu,\sigma}(\tilde{y}) \iff (\mu-y)(\tilde{y}-\mu) \lessgtr \sigma^2$.
(b) If $y < \tilde{y}$ and $(\mu-y)(\tilde{y}-\mu) > 0$, then $a_{\mu,\sigma}(y) \lessgtr a_{\mu,\sigma}(\tilde{y}) \iff (\mu-y)(\tilde{y}-\mu) \lessgtr \sigma^2$.

LEMMA 2. *Let l be a nonnegative integer, $a_1 \le a_2 \le \dots \le a_m$, $a_{m+l+1} \le a_{m+l+2} \le \dots \le a_n$, $k \le \min\{m, n-m-l\}$ and let*

$$
\begin{aligned}
d^*(\alpha) &= \sum_{n=1}^{m} 1\{\sin(\alpha) \ge -\cos(\alpha)\, a_i\} + l\ 1\{-\cos(\alpha)\le 0\} \\
&\quad + \sum_{n=m+1+l}^{n} 1\{\sin(\alpha) \le -\cos(\alpha)\, a_i\}.
\end{aligned}
$$

(a) Then $\min_{\alpha\in[-\pi,\pi]} d^*(\alpha) \le \min\{m, n-m-l\}$.
(b) If $a_i \ge a_{k+i}$ for all $i = m-k+l+1,\dots,m$ and $a_i \le a_{n-k+i}$ for all $i = 1,\dots,k$, then $\min_{\alpha\in[-\pi,\pi]} d^*(\alpha) \ge k$.
(c) If $a_i < a_{k+i}$ for some i, then $\min_{\alpha\in[-\pi,\pi]} d^*(\alpha) < k$.
(d) If $a_i > a_{n-k+i}$ for some i, then $\min_{\alpha\in[-\pi,\pi]} d^*(\alpha) < k$.

PROOF OF LEMMA 2. First, note that

$$d^*(\alpha) = \begin{cases} \sum_{i=1}^{m} 1\{-\tan(\alpha) \geq a_i\} + \sum_{i=m+l+1}^{n} 1\{-\tan(\alpha) \leq a_i\}, & \text{if } \cos(\alpha) < 0, \\ \sum_{i=1}^{m} 1\{-\tan(\alpha) \leq a_i\} + \sum_{i=m+l+1}^{n} 1\{-\tan(\alpha) \geq a_i\} + l, & \text{if } \cos(\alpha) > 0, \\ \sum_{i=1}^{m} 1\{\sin(\alpha) \geq 0\} + \sum_{i=m+l+1}^{n} 1\{\sin(\alpha) \leq 0\} + l, & \text{if } \cos(\alpha) = 0. \end{cases}$$

(a) For α with $-\tan(\alpha) < \min\{a_1, a_{m+l+1}\}$, we have $d^*(\alpha) = n - m - l$ for $\cos(\alpha) < 0$ and $d^*(\alpha) = m + l$ for $\cos(\alpha) > 0$. If $-\tan(\alpha) > \max\{a_m, a_n\}$ then we have $d^*(\alpha) = m$ for $\cos(\alpha) < 0$ and $d^*(\alpha) = n - (m + l) + l = n - m$ for $\cos(\alpha) > 0$.

(b) For α with $\cos(\alpha) = 0$ we have $d^*(\alpha) \geq \min\{m + l, n - m\} \geq k$. Now regard any α with $\cos(\alpha) \neq 0$.

If $a_1 \leq a_i \leq -\tan(\alpha) \leq a_{i+1} \leq a_m$ and $\cos(\alpha) < 0$, then there are two possibilities. One possibility is that $n - k + i + 1 \leq n$ so that $-\tan(\alpha) \leq a_{i+1} \leq a_{n-k+i+1}$ which implies $d^*(\alpha) \geq i + n - (n - k + i) = k$. The other possibility is that $n - k + i + 1 > n$ so that $i \geq k$ which implies $d^*(\alpha) \geq i \geq k$.

If $a_1 \leq a_i \leq -\tan(\alpha) \leq a_{i+1} \leq a_m$ and $\cos(\alpha) > 0$, then there are also two possibilities. One possibility is that $i \geq m - k + 1 + l$ so that $a_{k+i} \leq a_i \leq -\tan(\alpha)$ which implies $d^*(\alpha) \geq m - i + (k + i) - (m + l) + l = k$. The other possibility is that $i < m - k + 1 + l$ which implies $d^*(\alpha) \geq m - i + l \geq k$.

If $a_{m+l+1} \leq a_{m+i} \leq -\tan(\alpha) \leq a_{m+i+1} \leq a_n$ and $\cos(\alpha) < 0$, then again there are two possibilities. One possibility is that $m + i \geq n - k + 1$ so that $m + i = n - k + \tilde{i}$ and thus $a_{m+i-n+k} = a_{\tilde{i}} \leq a_{n-k+\tilde{i}} = a_{m+i} \leq -\tan(\alpha)$. This implies $d^*(\alpha) \geq m + i - n + k + n - (m + i) = k$. The other possibility is that $m + i < n - k + 1$ so that $i \leq n - m - k$ which implies $d^*(\alpha) \geq n - (m + i) \geq k$.

If $a_{m+l+1} \leq a_{m+i} \leq -\tan(\alpha) \leq a_{m+i+1} \leq a_n$ and $\cos(\alpha) > 0$, then one possibility is that $m + i + 1 \leq m + k$ so that $m + i + 1 = k + \tilde{i}$ and thus $-\tan(\alpha) \leq a_{m+i+1} = a_{k+\tilde{i}} \leq a_{\tilde{i}} = a_{m+i+1-k}$. This implies $d^*(\alpha) \geq m - (m + i - k) + (m + i) - (m + l) + l = k$. The other possibility is $m + i + 1 > m + k$ so that $i \geq k$ which implies $d^*(\alpha) \geq (m + i) - (m + l) + l \geq k$.

If $-\tan(\alpha) \notin (\min\{a_1, a_{m+l+1}\}, \max\{a_m, a_n\})$, then $d^*(\alpha) \geq k$ follows from a).

(c) If $a_i < a_{k+i}$ then there exists α with $a_{k+i} > -\tan(\alpha) > a_i$ and $\cos(\alpha) > 0$ so that $d^*(\alpha) \leq m - i + (k + i - 1) - (m + l) + l = k - 1$.

(d) If $a_i > a_{n-k+i}$ then there exists α with $a_i > -\tan(\alpha) > a_{n-k+i}$ and $\cos(\alpha) < 0$ so that $d^*(\alpha) \leq i - 1 + n - (n - k + i) = k - 1$. □

PROOF OF THEOREM 8. Set $m = \text{card}\{i\colon y_i < \mu\}$ and $l = \text{card}\{i\colon y_i = \mu\}$. If $0 = \sigma \in [c_k^-(\mu), c_k^+(\mu)]$ then $(\mu - y_i)(y_{k+i} - \mu) = 0$ for $i = m - k + 1, \ldots, m$ which implies

$$n\, d(\mu, \sigma) = \inf_{(u_1, u_2) \neq 0} \text{card}\left\{i :\ u_1(y_i - \mu) + u_2((y_i - \mu)^2 - \sigma^2) \geq 0\right\} \geq l \geq k.$$

If $0 = \sigma \notin [c_k^-(\mu), c_k^+(\mu)]$ then $l < k$ so that with $u_1 = 0$ and $u_2 = -1$

$$n\, d(\mu, \sigma) \leq \text{card}\left\{i :\ -(y_i - \mu)^2 \geq 0\right\} = l < k.$$

Now let $\sigma > 0$. Set $a_i = a_i(\mu, \sigma)$. Lemma 1(a) asserts that $a_i < a_j$ if $y_i < y_j < \mu$ or $\mu < y_i < y_j$, since in both cases $(\mu - y_i)(y_j - \mu) < 0 < \sigma^2$ is always satisfied. Hence the assumptions of Lemma 2 are satisfied and therefore $n\, d(\mu, \sigma) = \min_{\alpha \in [-\pi, \pi]} d^*(\alpha)$, where $d^*(\alpha)$ is the function considered in Lemma 2.

If $\sigma \in [c_k^-(\mu), c_k^+(\mu)]$, then

$$\sigma^2 \geq (\mu - y_i)(y_{k+i} - \mu) \quad \text{for all } i = m - k + 1, \ldots, m,$$
$$\sigma^2 \leq (\mu - y_i)(y_{n-k+i} - \mu) \quad \text{for all } i = 1, \ldots, k.$$

This is equivalent to $a_{k+i} \leq a_i$ for all $k = m - k + 1 + l, \ldots, m$ and $a_i \leq a_{n-k+i}$ for all $i = 1, \ldots, k$ according to Lemma 1(b). Hence Lemma 2(b) yields $n\, d(\mu, \sigma) \geq k$.

Now let $\sigma \notin [c_k^-(\mu), c_k^+(\mu)]$. Then there exists k with

$$(\mu - y_i)(y_{k+i} - \mu) > \sigma^2 \quad \text{or} \quad \sigma^2 > (\mu - y_i)(y_{n-k+i} - \mu).$$

According to Lemma 1(b), this is equivalent to

$$a_{k+i} > a_i \quad \text{or} \quad a_i > a_{n-k+i},$$

so that Lemma 2(c) and 2(d) gives $n\, d(\mu, \sigma) < k$. □

Proof of Theorem 9. We prove the theorem only for distributions whose density has connected support. However, a combination of this of its proof and that of Theorem 8 for the sample case yields the theorem for any general distribution P. In what follows, b stands for $F(\mu) = P((-\infty, \mu])$.

If $0 = \sigma \in [c_{\delta,P}^-(\mu), c_{\delta,P}^+(\mu)]$, then $\delta = 0 \leq d(\mu, \sigma, P)$ since a continuous distribution means $c_{\delta,P}^-(\mu) > 0$ for $\delta > 0$. If $0 = \sigma \notin [c_{\delta,P}^-(\mu), c_{\delta,P}^+(\mu)]$ then $\delta > 0$ and with $u_1 = 0$ and $u_2 = -1$ we obtain

$$d(\mu, \sigma, P) \leq P\left(\left\{y\colon\ -(y - \mu)^2 \geq 0\right\}\right) = 0 < \delta.$$

Let $\sigma > 0$ and define

$$d^*(\alpha) = \begin{cases} P\left(\{y \in (-\infty, \mu)\colon\ -\tan(\alpha) \geq a_{\mu,\sigma}(y)\}\right) \\ \qquad + P\left(\{y \in (\mu, \infty)\colon\ -\tan(\alpha) \leq a_{\mu,\sigma}(y)\}\right), \text{ if } \cos(\alpha) < 0, \\ P\left(\{y \in (-\infty, \mu)\colon\ -\tan(\alpha) \leq a_{\mu,\sigma}(y)\}\right) \\ \qquad + P\left(\{y \in (\mu, \infty)\colon\ -\tan(\alpha) \geq a_{\mu,\sigma}(y)\}\right), \text{ if } \cos(\alpha) > 0, \\ P\left(\{y \in (-\infty, \mu)\colon\ \sin(\alpha) \geq 0\}\right) \\ \qquad + P\left(\{y \in (\mu, \infty)\colon\ \sin(\alpha) \leq 0\}\right), \text{ if } \cos(\alpha) = 0, \end{cases}$$

where $a_{\mu,\sigma}(y)$ was defined in (22). Then we have $d(\mu, \sigma, P) = \min_{\alpha \in [-\pi, \pi]} d^*(\alpha)$.

Let $\sigma \in [c_{\delta,P}^-(\mu), c_{\delta,P}^+(\mu)]$. First, note that for α with $\cos(\alpha) = 0$ we have $d^*(\alpha) \geq \min\{b, 1-b\} \geq \delta$. Now consider α with $\cos(\alpha) \neq 0$. The condition on σ provides

$$(\mu - q_\beta)(q_{1-\delta+\beta} - \mu) \geq \sigma^2 \text{ for } \beta \in (0, \delta),$$
$$(\mu - q_\beta)(q_{\delta+\beta} - \mu) \leq \sigma^2 \text{ for } \beta \in (b - \delta, b).$$

This is equivalent to

$$(\mu - q_{\beta-1+b+\delta})(q_{b+\beta} - \mu) \geq \sigma^2 \text{ for } \beta \in (1-b-\delta, 1-b)\,,$$
$$(\mu - q_{\beta+b-\delta})(q_{b+\beta} - \mu) \leq \sigma^2 \text{ for } \beta \in (0, \delta)\,.$$

Using Lemma 1(b) we obtain

$$a_{\mu,\sigma}(q_{\beta-1+b+\delta}) \leq a_{\mu,\sigma}(q_{b+\beta}) \text{ for } \beta \in (1-b-\delta, 1-b)\,, \tag{23}$$
$$a_{\mu,\sigma}(q_{\beta+b-\delta}) \geq a_{\mu,\sigma}(q_{b+\beta}) \text{ for } \beta \in (0, \delta)\,. \tag{24}$$

For $y \in (-\infty, \mu)$ we have $(\mu - y)(q_{\beta-1+b+\delta} - \mu) < 0 < \sigma^2$ so that Lemma 1(a) implies

$$a_{\mu,\sigma}(q_{\beta-1+b+\delta}) \lessgtr a_{\mu,\sigma}(y) \Longleftrightarrow q_{\beta-1+b+\delta} \lessgtr y.$$

The same holds for $q_{\beta+b-\delta}$ and an analogous result holds for $y \in (\mu, \infty)$ and $q_{b+\beta}$. Now let α any value with $-\tan(\alpha) = a_{\mu,\sigma}(q_{b+\beta})$ for some $\beta \in (0, 1-b)$. If $\cos(\alpha) < 0$ and $\beta > 1-b-\delta$ we obtain with (23)

$$\begin{aligned} d^*(\alpha) &= P\left(\{y \in (-\infty, \mu) \colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &\qquad + P\left(\{y \in (\mu, \infty) \colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &\geq P\left(\{y \in (-\infty, \mu) \colon a_{\mu,\sigma}(q_{\beta-1+b+\delta}) \geq a_{\mu,\sigma}(y)\}\right) \qquad (25) \\ &\qquad + P\left(\{y \in (\mu, \infty) \colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &= P\left(\{y \in (-\infty, \mu) \colon q_{\beta-1+b+\delta} \geq y\}\right) + P\left(\{y \in (\mu, \infty) \colon q_{b+\beta} \leq y\}\right) \\ &= (\beta - 1 + b + \delta) + 1 - (b + \beta) = \delta. \end{aligned}$$

If $\cos(\alpha) < 0$ and $\beta \leq 1-b-\delta$ we obtain

$$\begin{aligned} d^*(\alpha) &\geq 0 + P\left(\{y \in (\mu, \infty) \colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &= 1 - (b + \beta) = 1 - b - \beta \geq \delta. \end{aligned}$$

Analogously, if $\cos(\alpha) > 0$ and $\beta < \delta$ we obtain with (24)

$$\begin{aligned} d^*(\alpha) &= P\left(\{y \in (-\infty, \mu) \colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &\qquad + P\left(\{y \in (\mu, \infty) \colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &\geq P\left(\{y \in (-\infty, \mu) \colon a_{\mu,\sigma}(q_{\beta+b-\delta}) \leq a_{\mu,\sigma}(y)\}\right) \qquad (26) \\ &\qquad + P\left(\{y \in (\mu, \infty) \colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &= P\left(\{y \in (-\infty, \mu) \colon q_{\beta+b-\delta} \leq y\}\right) + P\left(\{y \in (\mu, \infty) \colon q_{b+\beta} \geq y\}\right) \\ &= b - (\beta + b - \delta) + b + \beta - b = \delta. \end{aligned}$$

If $\cos(\alpha) < 0$ and $\beta \geq \delta$ we obtain similarly

$$d^*(\alpha) \geq 0 + P\left(\{y \in (\mu, \infty) \colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) = b + \beta - b \geq \delta.$$

Note that $\lim_{\beta\downarrow 0} a_{\mu,\sigma}(q_{b+\beta}) = -\infty$. If the support of P has no upper bound, then $\lim_{\beta\uparrow 1-b} a_{\mu,\sigma}(q_{b+\beta}) = \infty$ so that for every α there exists $\beta \in (0, 1-b)$ with $-\tan(\alpha) = a_{\mu,\sigma}(q_{b+\beta})$ since the support is connected. Hence in this case, we can conclude $d(\mu,\sigma,P) = \min_{\alpha\in[-\pi,\pi]} d^*(\alpha) \geq \delta$. If the support of P has an upper bound, then for every α with $-\tan(\alpha) \geq a_{\mu,\sigma}(q_1)$ we have with (23)

$$\begin{aligned} d^*(\alpha) &= P\left(\{y \in (-\infty,\mu)\colon\ -\tan(\alpha) \geq a_{\mu,\sigma}(y)\}\right) \\ &\geq P\left(\{y \in (-\infty,\mu)\colon a_{\mu,\sigma}(q_{b+1-b}) \geq a_{\mu,\sigma}(y)\}\right) \\ &\geq P\left(\{y \in (-\infty,\mu)\colon a_{\mu,\sigma}(q_{1-b-1+b+\delta}) \geq a_{\mu,\sigma}(y)\}\right) = \delta \end{aligned}$$

for $\cos(\alpha) < 0$ and

$$d^*(\alpha) \geq P\left(\{y \in (\mu,\infty)\colon\ -\tan(\alpha) \geq a_{\mu,\sigma}(y)\}\right) \geq 1-b$$

for $\cos(\alpha) > 0$. Hence, also for bounded support, we have $d(\mu,\sigma,P) = \min_{\alpha\in[-\pi,\pi]} d^*(\alpha) \geq \delta$.

If $\sigma \notin [c^-_{\delta,P}(\mu), c^+_{\delta,P}(\mu)]$ then there exists $\beta \in (1-b-\delta, 1-b)$ with $(\mu - q_{\beta-1+b+\delta})(q_{b+\beta}-\mu) < \sigma^2$ or $\beta \in (0,\delta)$ with $(\mu - q_{\beta+b-\delta})(q_{b+\beta}-\mu) > \sigma^2$. According to Lemma 1(b) this means

$$a_{\mu,\sigma}(q_{\beta-1+b+\delta}) > a_{\mu,\sigma}(q_{b+\beta}) \text{ or } a_{\mu,\sigma}(q_{\beta+b-\delta}) < a_{\mu,\sigma}(q_{b+\beta}).$$

Since $a'_{\mu,\sigma}(y) = \frac{1}{\sigma} + \frac{\sigma}{(y-\mu)^2} > 0$ the function $a_{\mu,\sigma}$ is strictly increasing in y. Hence for $a_{\mu,\sigma}(q_{\beta-1+b+\delta}) > a_{\mu,\sigma}(q_{b+\beta})$, we have using α with $-\tan(\alpha) = a_{\mu,\sigma}(q_{b+\beta})$ and $\cos(\alpha) < 0$

$$\begin{aligned} d^*(\alpha) &= P\left(\{y \in (-\infty,\mu)\colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &\quad + P\left(\{y \in (\mu,\infty)\colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &< P\left(\{y \in (-\infty,\mu)\colon a_{\mu,\sigma}(q_{\beta-1+b+\delta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &\quad + P\left(\{y \in (\mu,\infty)\colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &= P\left(\{y \in (-\infty,\mu)\colon q_{\beta-1+b+\delta} \geq y\}\right) + P\left(\{y \in (\mu,\infty)\colon q_{b+\beta} \leq y\}\right) \\ &= (\beta - 1 + b + \delta) + 1 - (b+\beta) = \delta. \end{aligned} \tag{27}$$

Thereby the strict inequality holds since the support of the distribution is connected. Analogously for $a_{\mu,\sigma}(q_{\beta+b-\delta}) < a_{\mu,\sigma}(q_{b+\beta})$ we have using α with $-\tan(\alpha) = a_{\mu,\sigma}(q_{b+\beta})$ and $\cos(\alpha) > 0$

$$\begin{aligned} d^*(\alpha) &= P\left(\{y \in (-\infty,\mu)\colon a_{\mu,\sigma}(q_{b+\beta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &\quad + P\left(\{y \in (\mu,\infty)\colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &< P\left(\{y \in (-\infty,\mu)\colon a_{\mu,\sigma}(q_{\beta+b-\delta}) \leq a_{\mu,\sigma}(y)\}\right) \\ &\quad + P\left(\{y \in (\mu,\infty)\colon a_{\mu,\sigma}(q_{b+\beta}) \geq a_{\mu,\sigma}(y)\}\right) \\ &= b - (\beta + b - \delta) + (b+\beta) - b = \delta. \quad \square \end{aligned}$$