GEMS:

GEOMETRIC METHODS IN STATISTICS

Stanislav Nagy

podzimní setkání KPMS 2024

Charles University, Prague
Department of Probability and Mathematical Statistics

ERC CZ grant LL2407

GeMS: Geometric Multivariate and Non-Euclidean Statistics

- October 2024 – October 2029,
- funded by MŠMT ČR,
- S. Nagy (PI), I. Mizera, D. Pokorný (MÚ UK), J. Kynčl (KAM UK),
- 2 post-docs and 3 PhD students.

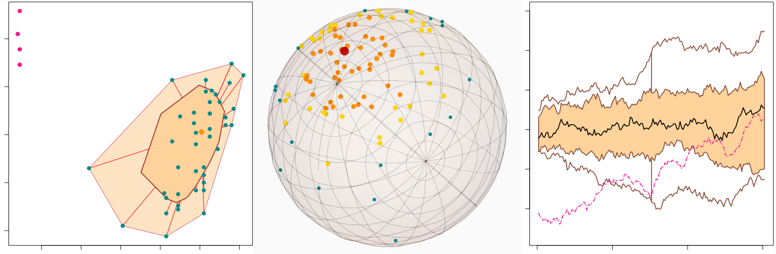
More info at

https://gems.karlin.mff.cuni.cz

Multivariate nonparametric statistics

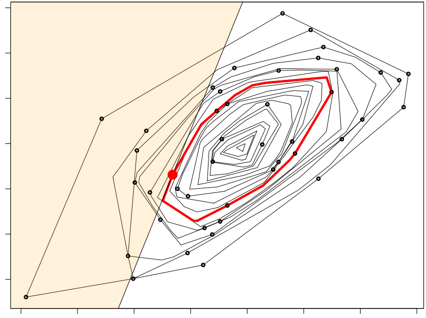
Nonparametric statistics:

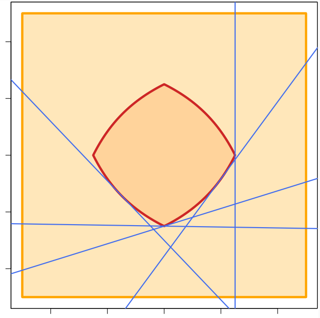
- Inference without assumptions.
- On the real line using the ordering — median, quantiles, ranks...
- What are ranks or quantiles for multivariate (non-Euclidean) data?



Statistical depth

Statistical depth function: Ordering data in multivariate spaces.



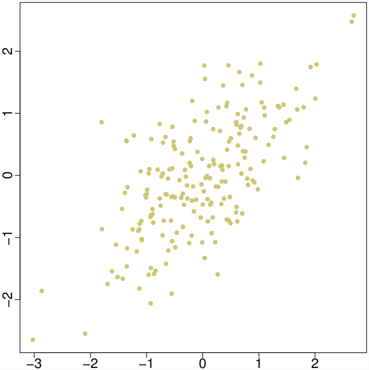


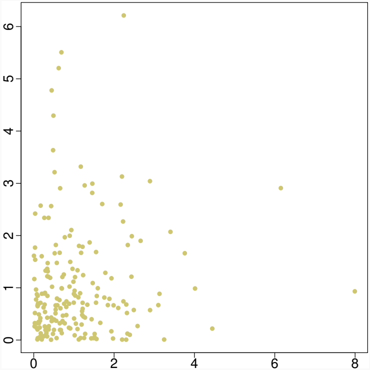
Introduced in 1975 (Tukey); studied intensively since the 1990s.

Statistical depth function

For $\mathcal{P}\left(\mathbb{R}^d\right)$ Borel probability measures on $\mathbb{R}^d$, a depth is

$$D\colon \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1]\colon (x,P) \mapsto D(x,P).$$

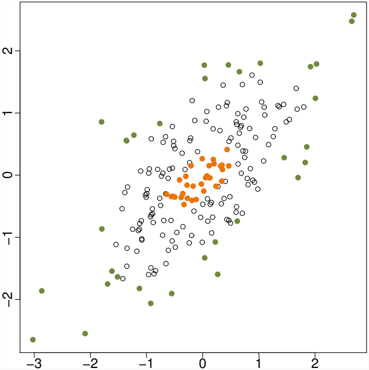


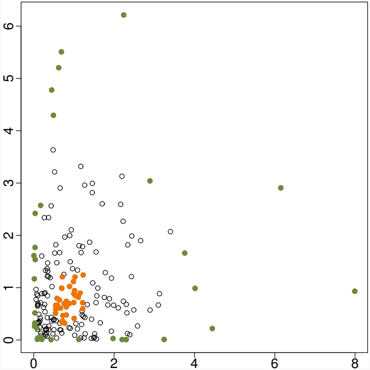


Statistical depth function

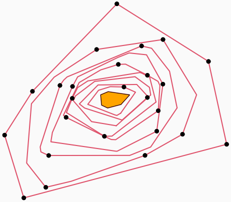
For $\mathcal{P}(\mathbb{R}^d)$ Borel probability measures on $\mathbb{R}^d$, a depth is

$$D\colon \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1]\colon (x,P) \mapsto D(x,P).$$





Principal goal: Distribution-characterizing depths



Find a depth that:

- C1 characterizes probability distributions uniquely,
- C2 is highly (e.g., affine) equivariant,
- C3 induces robust medians, and
- C4 is fast to compute.

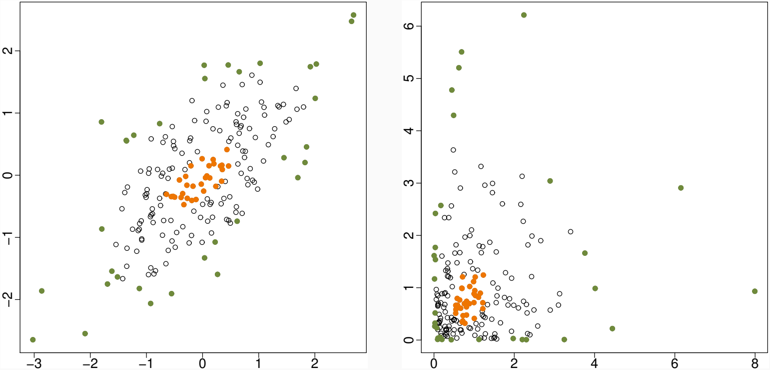
Impact: A universal framework of multivariate nonparametrics.

➥ Data exploration/statistical estimation and testing/visualisation free of parametric assumptions for complex datasets.

Halfspace depth

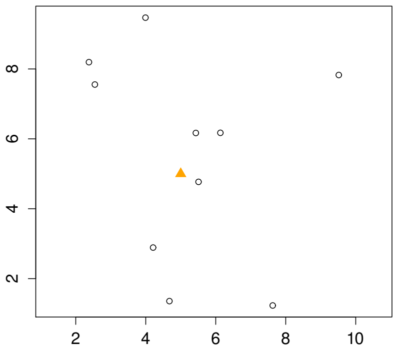
Halfspace depth (Tukey, 1975) of a point $x \in \mathbb{R}^d$ w.r.t. $P \in \mathcal{P}\left(\mathbb{R}^d\right)$

$$D(x; P) = \inf_{H \in \mathcal{H}(x)} P\left(H\right).$$



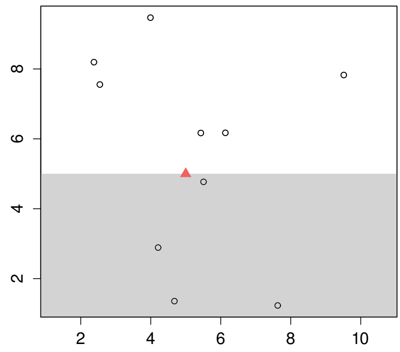
Halfspace depth

$$D\left(x; P_n\right) = \min \frac{\text{\# of observations in a halfspace that contains } x}{n}$$

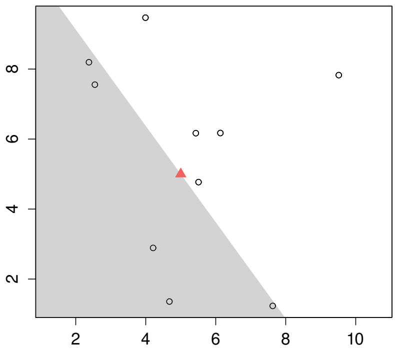


Halfspace depth

$$D(x; P_n) = \min \frac{\text{\# of observations in a halfspace that contains } x}{n}$$

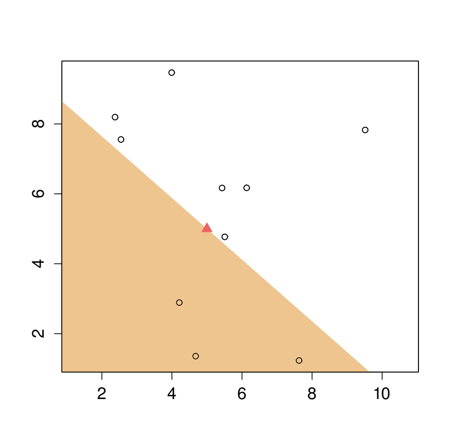


$$D(x; P_n) = \min \frac{\text{\# of observations in a halfspace that contains } x}{n}$$



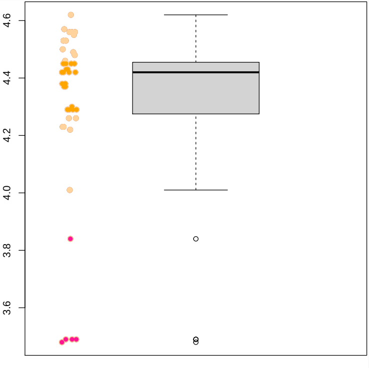
Halfspace depth

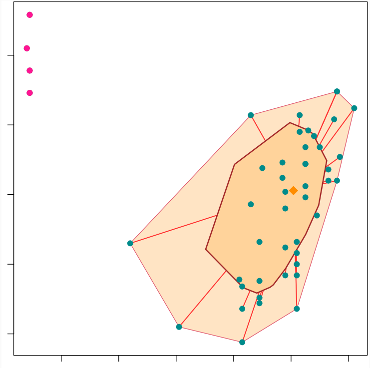
$$D(x; P_n) = \min \frac{\text{\# of observations in a halfspace that contains } x}{n}$$



Application: Bagplot

Bagplot: A multivariate boxplot (Rousseeuw et al., 1999)

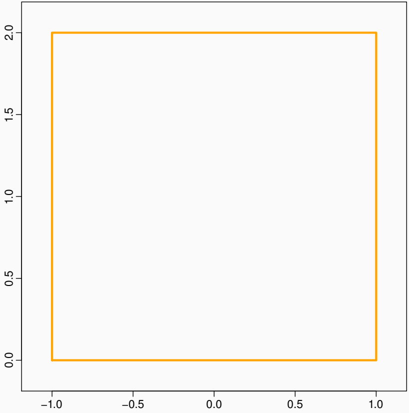




Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

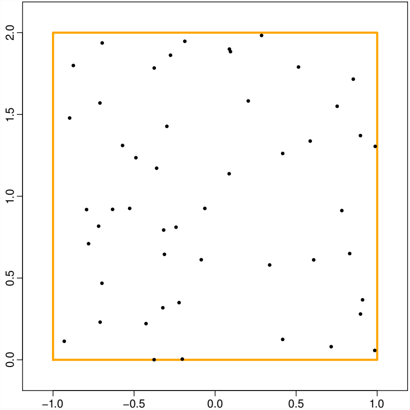
$$P_\delta = \left\{x \in \mathbb{R}^d : D(x; P) \geq \delta\right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

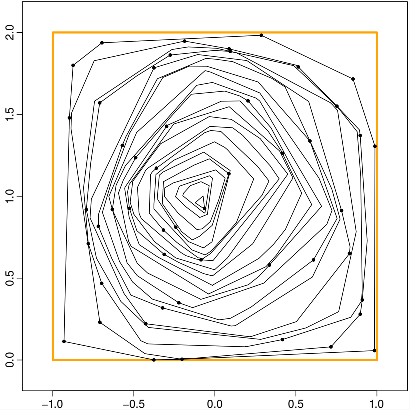
$$P_\delta = \left\{ x \in \mathbb{R}^d : D(x; P) \geq \delta \right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

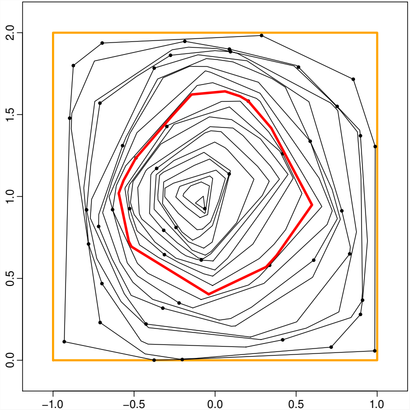
$$P_\delta = \left\{ x \in \mathbb{R}^d : D(x; P) \geq \delta \right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

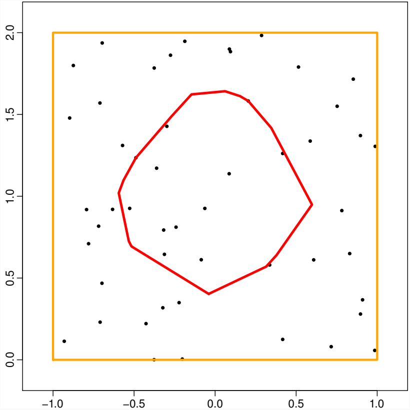
$$P_\delta = \left\{ x \in \mathbb{R}^d : D(x; P) \geq \delta \right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

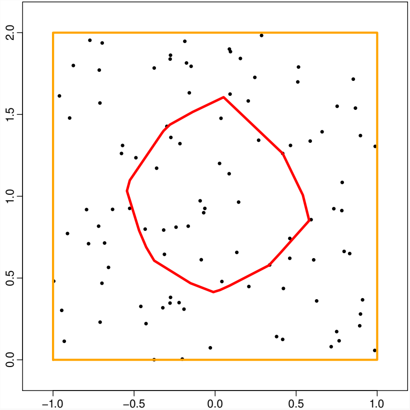
$$P_\delta = \left\{ x \in \mathbb{R}^d : D(x; P) \geq \delta \right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

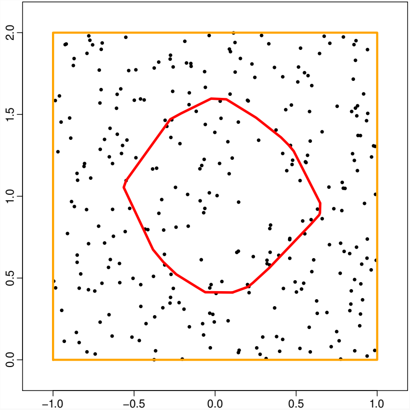
$$P_\delta = \left\{ x \in \mathbb{R}^d : D(x; P) \geq \delta \right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

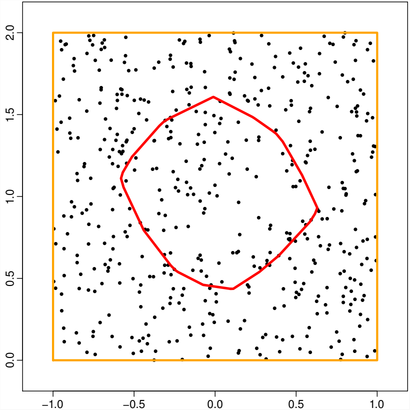
$$P_\delta = \left\{x \in \mathbb{R}^d : D(x; P) \geq \delta\right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0,1]$

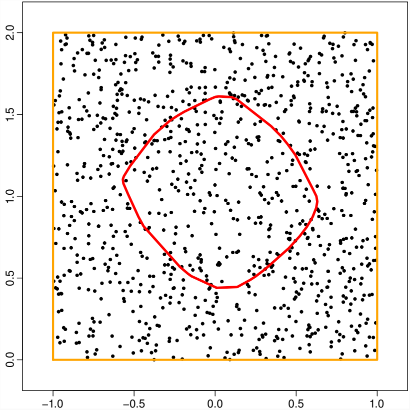
$$P_\delta = \left\{x \in \mathbb{R}^d : D(x; P) \geq \delta\right\} \text{ is convex}$$



Depth: Level sets

$D(\cdot; P)$ is always quasi-concave, i.e. for each $\delta \in [0, 1]$

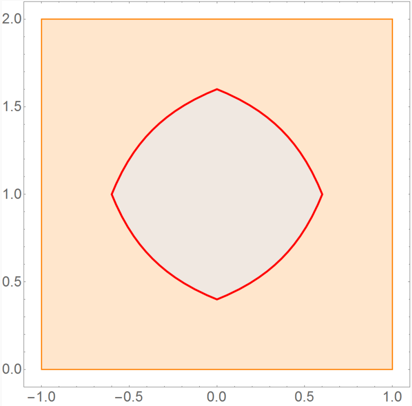
$$P_\delta = \left\{x \in \mathbb{R}^d : D(x; P) \geq \delta\right\} \text{ is convex}$$



Depth: Level sets

We can write (Rousseeuw and Struyf, 1999; Zuo and Serfling, 2000)

$$P_\delta = \left\{x \in \mathbb{R}^d : D(x; P) \geq \delta\right\} = \bigcap \{H \in \mathcal{H} : P(H) > 1 - \delta\}.$$

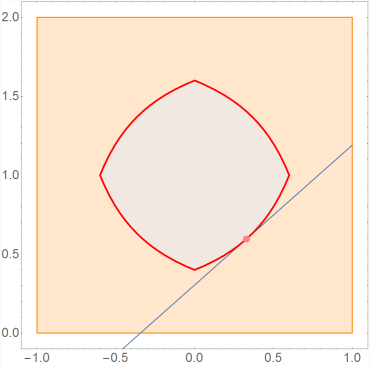


Depth: Asymptotic normality

Let $P_n \in \mathcal{P}\left(\mathbb{R}^d\right)$ be the empirical measure of n i.i.d. variables from P.

$\sqrt{n}\,(D(x;P_n) - D(x;P))$ is asymptotically normal

$$\iff \quad \text{the contour of } D(\cdot;P) \text{ is smooth at } x$$

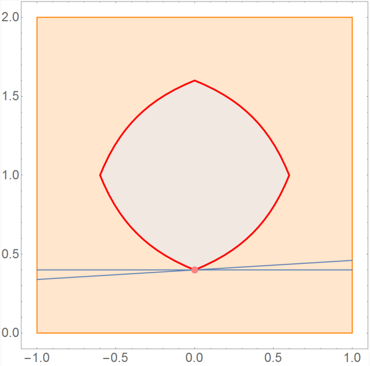


Depth: Asymptotic normality

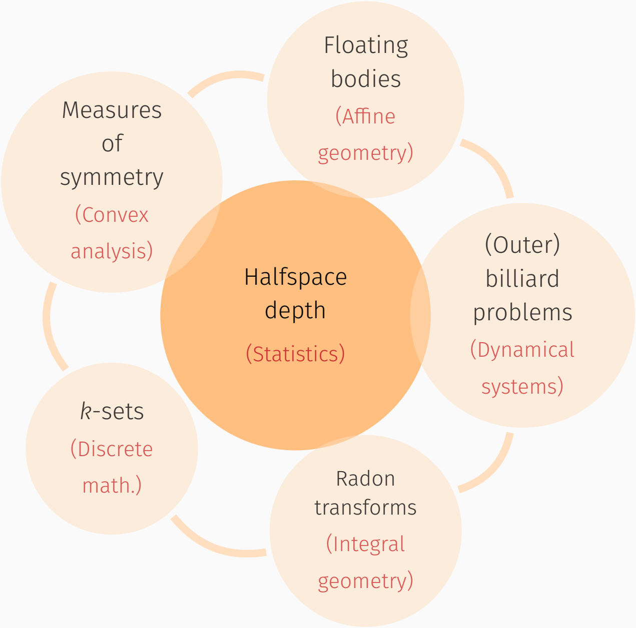
Let $P_n \in \mathcal{P}\left(\mathbb{R}^d\right)$ be the empirical measure of n i.i.d. variables from P.

$\sqrt{n}\left(D(x; P_n) - D(x; P)\right)$ is asymptotically normal

$\iff$ the contour of $D(\cdot; P)$ is smooth at x



Nonparametrics outside statistics: Halfspace depth



Open problems

Some open problems on the halfspace depth

Smoothness: When are halfspace depth contours smooth?

Reconstruction: How to reconstruct P from its depth?

Computation: How fast can the depth, or its median, be computed?

Non-standard data: Can depths be used also in non-$\mathbb{R}^d$ spaces?

Testing/Inference: Depth-based testing procedures?

GeMS PROJECT: SOME OPEN PROBLEMS

STANISLAV NAGY

Basic definitions and notation. Let $\mathcal{P}\left(\mathbb{R}^d\right)$ be the set of all probability measures on $\mathbb{R}^d$. For $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ and $x \in \mathbb{R}^d$, the halfspace (Tukey) depth [7] of x with respect to (w.r.t.) P is defined as

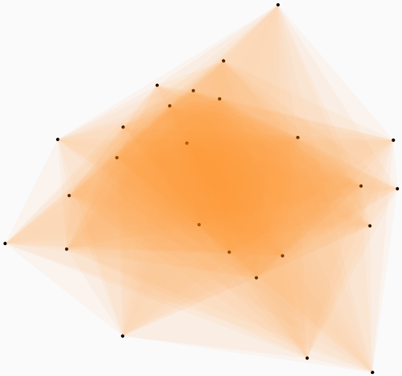
$$hD(x; P) = \inf_{\mathfrak{H} \in \mathcal{H}(x)} P(\mathfrak{H}),$$

https://gems.karlin.mff.cuni.cz/pdf/Depth_problems_update.pdf

Simplicial depth

Simplicial depth (Liu, 1988) of $x \in \mathbb{R}^d$ w.r.t. $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ is

$$SD(x; P) = \mathrm{P}\left(x \in \triangle(X_1, \ldots, X_{d+1})\right).$$



How to compute simplicial depth?

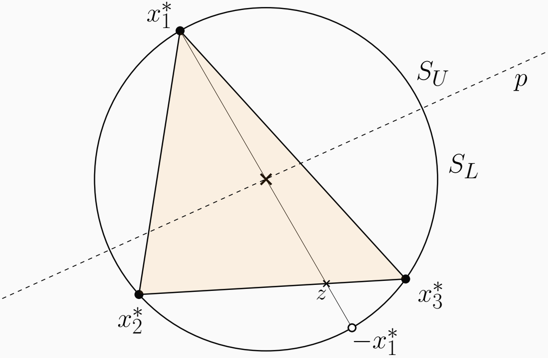
Simplicial depth (Liu, 1988) of $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$ w.r.t. $P \in \mathcal{P}(\mathbb{R}^2)$ is

$$SD(x;P) = \mathrm{P}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \triangle\left(\begin{pmatrix} X_{1,1} \\ X_{1,2} \end{pmatrix}, \begin{pmatrix} X_{2,1} \\ X_{2,2} \end{pmatrix}, \begin{pmatrix} X_{3,1} \\ X_{3,2} \end{pmatrix}\right)\right)$$
$$= \iint \iint \iint \mathbb{I}\left[x \in \triangle\right] \mathrm{d}\, P(x_{1,1}, x_{1,2})\, \mathrm{d}\, P(x_{2,1}, x_{2,2})\, \mathrm{d}\, P(x_{3,1}, x_{3,2}).$$

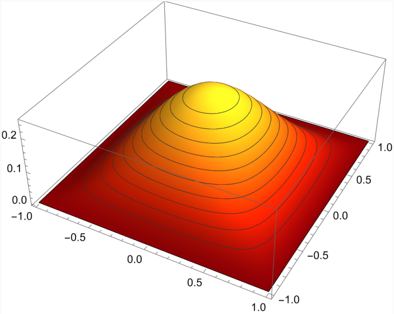
- $d \times (d+1)$ integrals in $\mathbb{R}^d$.
- Impossible to calculate already for Gaussian distributions in $\mathbb{R}^2$.
- How does the population $SD(\cdot; P)$ even look like?

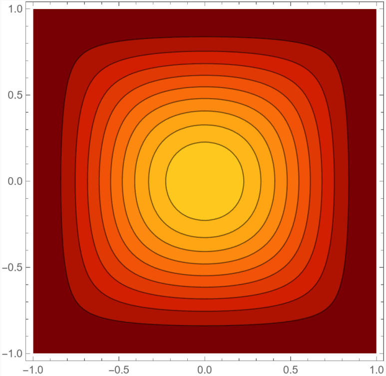
We want to compute the simplicial depth in $\mathbb{R}^2$ exactly:

$$SD(x; P) = \mathrm{P}\left(x \in \triangle(X_1, X_2, X_3)\right).$$

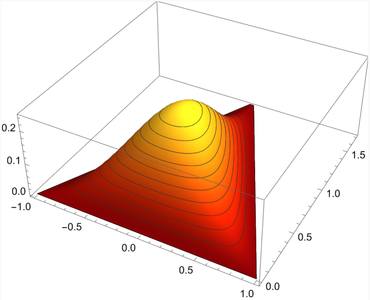


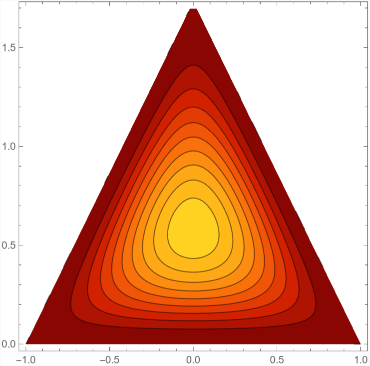
Simplicial depth of a square



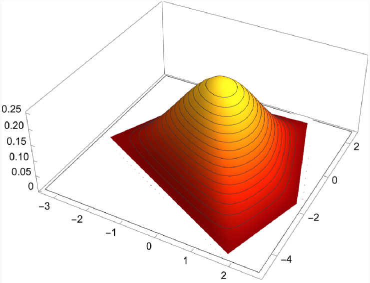


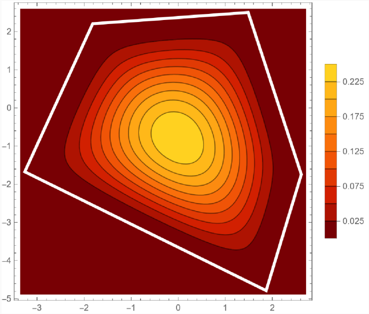
Simplicial depth of a triangle





Simplicial depth of polygons





(Undergraduate) Student collaboration

Journal of Multivariate Analysis
Volume 205, January 2025, 105375

Explicit bivariate simplicial depth

Erik Mendroš, Stanislav Nagy

Statistics & Probability Letters
Volume 209, June 2024, 110094

Conditions for equality in Anderson's theorem

Filip Bočinec, Stanislav Nagy

More student collaboration

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 33, NO. 2, 699–713
https://doi.org/10.1080/10618600.2023.2257781

Taylor & Francis
Taylor & Francis Group

Check for updates

On Exact Computation of Tukey Depth Central Regions

Vít Fojtík[a], Petra Laketa[a], Pavlo Mozharovskyi[b], and Stanislav Nagy[a]

[a]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic; [b]LTCI, Telecom Paris, Institut Polytechnique de Paris, Paris, France

Stat
The ISI's Journal for the Rapid Dissemination of Statistics Research

ORIGINAL ARTICLE | Full Access

A note on the convergence of lift zonoids of measures

František Hendrych, Stanislav Nagy

First published: 09 January 2022 | https://doi.org/10.1002/sta4.453

Foreign collaboration and travel

(Several) long-term collaborators of GeMS:

- **ULB, Belgium:** Germain van Bever;
- **Télécom Paris, France:** Pavlo Mozharovskyi;
- **Aalto, Finland:** Pauliina Ilmonen;
- **Uni. of Cologne, Germany:** Rainer Dyckerhoff;
- **Uni. Toulouse, France:** Frédéric Ferraty;
- **Uni. of Bristol, UK:** George Wynne;
- **Uni. of Turku, Finland:** Joni Virta;
- **Northeastern Uni., Boston, US:** Sara López-Pintado;
- **Uni. of Lazio, Italy:** Giovanni Porzio;
- **Uni. Málaga, Spain:** Antonio Elías.

Junior positions in GeMS

Research team positions:

- **Postdoc 1:** Gauthier Louis Thurin (Bordeaux)

  Gilberto Chavez-Martínez (Montreal)
- **Postdoc 2:** Gaëtan Louvet (Brussels)

  Manuel Hernández-Banadik (Montevideo)
- **PhD students 1 and 2:** Filip Bočinec, Erik Mendroš
- **PhD student 3:** ? (starting from October 2025 or later)

Interested? Ask.

GeMS: Geometric Methods in Statistics

`GeMS.karlin.mff.cuni.cz`

Mathematical Forum, K1, December 4 (Wednesday), 15:40