Mach Learn (2015) 100:677–699
DOI 10.1007/s10994-015-5524-x

# Half-space mass: a maximally robust and efficient data depth method

**Bo Chen**[1] · **Kai Ming Ting**[2] · **Takashi Washio**[3] · **Gholamreza Haffari**[1]

Received: 29 March 2015 / Accepted: 2 July 2015 / Published online: 5 August 2015


**Abstract** Data depth is a statistical method which models data distribution in terms of center-outward ranking rather than density or linear ranking. While there are a lot of academic interests, its applications are hampered by the lack of a method which is both robust and efficient. This paper introduces *Half-Space Mass* which is a significantly improved version of half-space data depth. *Half-Space Mass* is the only data depth method which is both robust and efficient, as far as we know. We also reveal four theoretical properties of *Half-Space Mass*: (i) its resultant mass distribution is concave regardless of the underlying density distribution, (ii) its maximum point is unique which can be considered as median, (iii) the median is maximally robust, and (iv) its estimation extends to a higher dimensional space in which the convex hull of the dataset occupies zero volume. We demonstrate the power of *Half-Space Mass* through its applications in two tasks. In anomaly detection, being a maximally robust location estimator leads directly to a robust anomaly detector that yields a better detection accuracy than half-space depth; and it runs orders of magnitude faster than $L_2$ depth, an existing maximally robust location estimator. In clustering, the *Half-Space Mass* version of K-means overcomes three weaknesses of K-means.

Editors: João Gama, Indrė Žliobaitė, Alípio M. Jorge, and Concha Bielza.

✉ Bo Chen
bo.chen@monash.edu

Kai Ming Ting
kaiming.ting@federation.edu.au

Takashi Washio
washio@ar.sanken.osaka-u.ac.jp

Gholamreza Haffari
gholamreza.haffari@monash.edu

1 Faculty of Information Technology, Monash University, Clayton, VIC 3168, Australia

2 School of Engineering and Information Technology, Federation University Australia, Churchill, VIC 3842, Australia

3 The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibarakishi, Osaka 5670047, Japan

**Keywords** Half-space mass · Mass estimation · Data depth · Robustness

## 1 Introduction

> "Most important for the selection of a depth statistic in applications are the questions of computability and - depending on the data situation - robustness." - Karl Mosler (2013)

Data depth (Liu et al. 1999) is a statistical method which models data distribution in terms of center-outward ranking rather than density or linear ranking. In 1975, Tukey (1975) proposed a way to define multivariate median in a data cloud, known as half-space depth or Tukey depth. Since then it has been extensively studied. Donoho and Gasko (1992) have revealed the breakdown point of Tukey median; Zuo and Serfling (2000) have compared it to various competitors and Dutta et al. (2011) have investigated the properties of half-space depth. Meanwhile, the concept of data depth has been adopted for multivariate statistical analysis since it provides a nonparametric approach that does not rely on the assumption of normality (Liu et al. 1999).

Despite its popularity, the following characteristics of half-space depth have hampered its applications. As demonstrated by a simple example in Fig. 1, the "deepest point", or half-space median, is not guaranteed to be unique. A set of discrete data points has a layered depth distribution, which is not concave. Moreover, half-space depth is not a maximally robust depth method, i.e., its distribution is easily distrubed by outliers. While a maximally robust method exists, e.g., $L_2$ depth (Mosler 2013), it is computationally expensive. No current data depth method is both computationally efficient and robust, as far as we know.

We introduce half-space mass, a significantly improved version of half-space depth, which is both efficient and maximally robust. We reveal four theoretical properties of half-space mass:

(i) It is concave in a user defined region that covers the source density distribution or the data cloud. An example is shown in Fig. 1.
(ii) It has a unique maximum point, which can be regarded as a multi-dimensional median.
(iii) Its median, which has a breakdown point equal to $\frac{1}{2}$, is maximally robust.
(iv) It extends the information carried in a dataset to a higher dimensional space in which such dataset has a zero-volume convex hull.

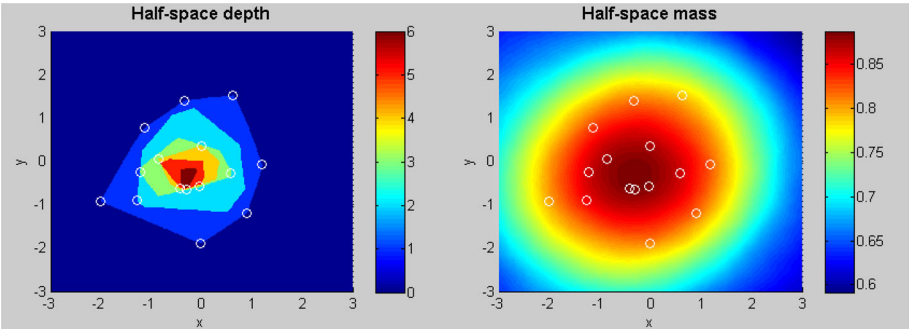


**Fig. 1** Distributions half-space depth and half-space mass of a simple dataset. *White circle* markers denote the data points while the color indicates the depth/mass value at each location of the space

The key contributions of this paper are the formal definition of half-space mass and the uncovering of its theoretical properties backed up with their proofs. To demonstrate its applicability to real life problems, half-space mass is applied to two tasks: anomaly detection and clustering. We provide a comparison with two existing data depth methods: half-space depth (Tukey 1975) and $L_2$ depth (Mosler 2013). Based on half-space mass, we create a clustering algorithm reminiscent of the K-means algorithm (Jain 2010).

Our empirical evaluations show that half-space mass has the following advantages compared to its contenders:

- Its maximal robustness leads directly to better performance in anomaly detection than half-space depth.
- Compared to the existing maximally robust $L_2$ depth, it runs orders of magnitude faster.
- Compared to the distance-based K-means clustering method, the half-space mass-based version overcomes three weaknesses of K-means (Tan et al. 2014) to find clusters of varying densities and sizes, as well as in the presence of noise.

The rest of the paper is organized as follows. Section 2 introduces the formal definitions of half-space mass as well as the proposed implementation. Sections 3 and 4 provide its theoretical properties and proofs, respectively. Section 5 discusses the relationship between half-space mass and other data depth methods. Section 6 describes applications of half-space mass in anomaly detection and clustering. Section 7 reports the empirical evaluations. Section 8 discusses its relation to mass estimation and Sect. 9 concludes the paper.

## 2 Half-space mass

### 2.1 Definitions

The proposed half-space mass is formally defined in this section. The key notations are provided in Table 1.

**Table 1** Notations

| | |
|---|---|
| $\Re^d$ | A $d$-dimensional real space |
| $\ell$ | A direction in $\Re^d$ |
| $x$ | A one-dimensional point in $\Re$ |
| $\mathbf{x}$ | A point in $\Re^d$ |
| $D$ | A dataset, where $\|D\| = n$ |
| $\mathbf{X}$ | A point in $D$ |
| $\mathcal{D}$ | A subset of $D$, where $\|\mathcal{D}\| = \psi$ |
| $t$ | Number of half-spaces sampled for estimation |
| $R$ | A convex region covering a source density $F$ or a dataset $D$ |
| $\lambda$ | A parameter that determines the size of $R$ |
| $P_F(\cdot)$ | A probability mass function of a probability density distribution $F$ |
| $P_D(\cdot)$ | An empirical probability mass function of a dataset $D$ |
| $HM(\cdot\|F)$ | Half-space mass function given $F$ |
| $HM(\cdot\|D)$ | Half-space mass function given $D$ |

Let $F(\mathbf{x})$ be a probability density on $\mathbf{x} \in \Re^d$, $d \geq 1$; $R \subset \Re^d$ be a convex and closed region covering the domain of $F$; and $H$ be a closed half-space formed by separating $\Re^d$ with a hyperplane that intersects $R$. Note that the probability mass of $H$ computed with respect to $F$ is $0 \leq P_F(H) = P_F(H \cap R) \leq 1$.

**Definition 1** Half-space mass ($HM$) of a point $\mathbf{x} \in \Re^d$ with respect to $F$ is defined as:

$$\begin{aligned} HM(\mathbf{x}|F) &= E_{\mathcal{H}(\mathbf{x})}[P_F(H)] \\ &= \lim_{\mathbb{H}(\mathbf{x}) \to \mathcal{H}(\mathbf{x})} \frac{1}{|\mathbb{H}(\mathbf{x})|} \sum_{H \in \mathbb{H}(\mathbf{x})} P_F(H) \end{aligned}$$

where $\mathcal{H}(\mathbf{x}) := \{H : \mathbf{x} \in H\}$ is a set of all closed half-spaces $H$ which contains the query point $\mathbf{x}$ and $\mathbb{H}(\mathbf{x}) \subset \mathcal{H}(\mathbf{x})$.

The definition of half-space mass can be conceptualized as the expectation of the probability mass of a randomly selected half-space $H$, which is defined for $R$ and contains the query point $\mathbf{x}$, given that every half-space is equally likely. This definition happens to have certain similarity to that of half-space depth (Tukey 1975). While half-space depth takes the minimum of probability mass of a random half-space containing query point $\mathbf{x}$ as the depth value (see its definition in Table 2 in Sect. 5), half-space mass takes the expectation of it. This key difference enables half-space mass to have more desirable properties, which will be discussed in Sects. 3 and 4.

Practically an i.i.d. sample $D$ is usually given instead of the source density distribution $F$. The sample version of $HM(\mathbf{x}|F)$ is obtained by replacing $F$ with $D$ as follows.

**Definition 2** Half-space mass ($HM$) of a point $\mathbf{x} \in \Re^d$ with respect to a given dataset $D$ is defined as:

$$\begin{aligned} HM(\mathbf{x}|D) &= E_{\mathcal{H}(\mathbf{x})}[P_D(H)] \\ &= \lim_{\mathbb{H}(\mathbf{x}) \to \mathcal{H}(\mathbf{x})} \frac{1}{|\mathbb{H}(\mathbf{x})|} \sum_{H \in \mathbb{H}(\mathbf{x})} P_D(H) \end{aligned}$$

where $P_D(H)$ is the empirical probability measure of $H$ with respect to $D$, i.e., the proportion of data points in $D$ that lie in $H$. Note that $0 \leq P_D(H) \leq 1$.

$HM(\mathbf{x}|D)$ can be estimated by sampling $t$ half-spaces from $\mathcal{H}(\mathbf{x})$ for each query point $\mathbf{x}$. By selecting $\mathbb{H}(\mathbf{x}) \subset \mathcal{H}(\mathbf{x})$ with size $|\mathbb{H}(\mathbf{x})| = t$, this estimator is defined as:

$$\begin{aligned} \widehat{HM}(\mathbf{x}|D) &= \frac{1}{|\mathbb{H}(\mathbf{x})|} \sum_{H \in \mathbb{H}(\mathbf{x})} P_D(H) \\ &= \frac{1}{t} \sum_{i=1}^{t} P_D(H_i) \end{aligned} \tag{1}$$

where $H_i$ are elements of $\mathbb{H}(\mathbf{x})$.

We also propose a computation-friendly version to estimate $HM(\mathbf{x}|D)$. Instead of using the whole dataset $D$ to calculate $P_D(H_i)$ in (1), a small subsample $\mathcal{D}_i \subset D$ with size $|\mathcal{D}_i| = \psi \ll |D|$ is randomly selected from $D$ without replacement for $i = 1, \ldots, t$. Let $R_i$ be a convex region covering $\mathcal{D}_i$, $H_i(\mathbf{x})$ be a randomly selected half-space containing $\mathbf{x}$ and intersecting $R_i$, for $i = 1, \ldots, t$.

**Definition 3** A computation-friendly estimator for $HM(\mathbf{x}|D)$ is defined as:

$$\widetilde{HM}(\mathbf{x}|D) = \frac{1}{t}\sum_{i=1}^{t} P_{\mathcal{D}_i}(H_i(\mathbf{x})) = \frac{1}{t\psi}\sum_{i=1}^{t}\sum_{j=1}^{\psi} I(\mathbf{X}_j \in H_i(\mathbf{x}))$$

where $I(\cdot)$ is an indicator function and $\mathbf{X}_j$ is a point in $\mathcal{D}_i$.

## 2.2 Implementation

In general, half-space mass is a concave function in $R$, as will be shown in Sects. 3 and 4; therefore it provides distinct center-outward ordering in the region $R$, while concavity outside of $R$ is not guaranteed.

When concavity needs to be guaranteed in a region larger than the convex hull of $D$, a larger $R$ would be desirable. To this end, we propose a projection-based algorithm to estimate $HM(\mathbf{x}|D)$ in which the region $R$ or $R_i$ is determined by a size parameter $\lambda$. It is the ratio of diameters between $R$ and the convex hull of $D$ along every direction. The value of $\lambda$ should be more than or equal to 1. When $\lambda = 1$, $R$ or $R_i$ is the convex hull of $D$ or $\mathcal{D}_i$. The bigger $\lambda$ is, the larger $R$ or $R_i$ expands from the convex hull of $D$ or $\mathcal{D}_i$.

Algorithm 1 is the training procedure of $\widetilde{HM}(\cdot|D)$. The half-space is implemented as follows: a random subsample $\mathcal{D}_i$ is projected onto a random direction $\ell$ in $\Re^d$, $t$ times. For each projection, a split point $s$ is randomly selected between a range adjusted by $\lambda$; and then the number of points that fall in either sides of $s$ are recorded.

Algorithm 2 is the testing procedure when $\widetilde{HM}(\mathbf{x})$ is ready. Given a query point $\mathbf{x}$, it is projected onto each of the $t$ directions, and the number of training points that fall on the same side as $\mathbf{x}$ are averaged and output as estimated value of the half-space mass for $\mathbf{x}$.

## 2.3 Parameter setting

Here we provide a general guide for setting the parameters. The parameter $t$ affects the accuracy of the estimation. The larger $t$ is, the more accurate the estimation is. In high

**Algorithm 1**: Training algorithm of $\widetilde{HM}(\cdot|D)$.

**input** : $D$ - Training dataset; $t$ - number of half-spaces; $\psi$ - subsample size; $\lambda$ - size parameter of $R$
**output**: $\widetilde{HM}(\cdot)$ with $\{\ell_i, s_i, m_i^l, m_i^r\}$, for $i = 1, \ldots, t$
**for** $i = 1, \ldots, t$ **do**
  Generate a random direction $\ell_i$ in $\Re^d$, the data space of $D$.
  Generate a subsample $\mathcal{D}_i$ by randomly selecting $\psi$ points from $D$ without replacement.
  Project $\mathcal{D}_i$ onto $\ell_i$, denoted by $\mathcal{D}_i^{\ell_i}$.
  $max_i \leftarrow \max(\mathcal{D}_i^{\ell_i})$, $min_i \leftarrow \min(\mathcal{D}_i^{\ell_i})$, $mid_i \leftarrow \frac{max_i + min_i}{2}$.
  Randomly select $s_i$ in $(mid_i - \frac{\lambda}{2}(max_i - min_i), mid_i + \frac{\lambda}{2}(max_i - min_i))$.
  $m_i^l \leftarrow \frac{|\{x \in \mathcal{D}_i^{\ell_i} \mid x < s_i\}|}{\psi}$
  $m_i^r \leftarrow \frac{|\{x \in \mathcal{D}_i^{\ell_i} \mid x \geq s_i\}|}{\psi}$
**end**

**Algorithm 2**: Testing algorithm of $\widetilde{HM}(\mathbf{x})$.

**input** : $\mathbf{x}$ - Query point
**output**: The estimated value $\widetilde{HM}(\mathbf{x})$ for $\mathbf{x}$
$HM = 0$
**for** $i = 1, \ldots, t$ **do**
  Project $\mathbf{x}$ onto $\ell_i$, denoted by $\mathbf{x}^{\ell_i}$
  **if** $\mathbf{x}^{\ell_i} < s_i$ **then**
    $HM \leftarrow HM + m_i^l$
  **else**
    $HM \leftarrow HM + m_i^r$
  **end**
**end**
**return** $HM/t$

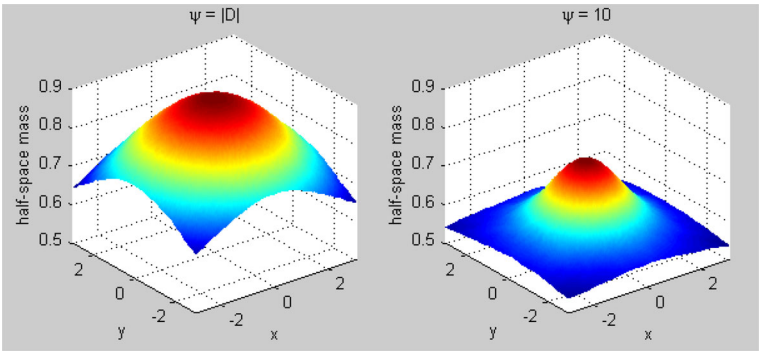


**Fig. 2** A comparison of distributions of half-space mass using $\psi = |D|$ and $\psi = 10$, on a dataset $D$ of 10,000 points generated from a bivariate Gaussian. Both distributions are generated using $t = 5000$ and $\lambda = 1$

dimensional datasets or datasets which are elongated significantly in some direction than others, $t$ shall be set to a large value, in order to gather sufficient information from all directions.

When the computation-friendly version $\widetilde{HM}(\mathbf{x}|D)$ is used, it is worth pointing out that $R_i$ could be significantly smaller than $R$, especially when subsample size $\psi \ll |D|$. Thus a small $\psi$ would produce a more concentrated distribution than that produced with a large $\psi$, as shown in Fig. 2. This is the case where $\lambda > 1$ could be used for some applications. Another effect of a small $\psi$ value when $\lambda = 1$ is that, it limits the range of $\widetilde{HM}(\mathbf{x}|D)$ values. Note that by Definition 3 when $\lambda = 1$, $\frac{1}{\psi} \leq P_{\mathcal{D}_i}(H_i(\mathbf{x})) \leq \frac{\psi-1}{\psi}$, thus $\frac{1}{\psi} \leq \widetilde{HM}(\mathbf{x}|D) \leq \frac{\psi-1}{\psi}$.

For the rest of this paper, we use Algorithms 1 and 2 to estimate half-space mass. The parameter $\lambda$ is set to 1 by default unless mentioned otherwise.

## 3 Properties of half-space mass

We list four theoretical properties of half-space mass in this section, which are concavity in region $R$, unique median, the median having breakdown point equal to $\frac{1}{2}$, and extension across dimension. Proofs of the lemma and theorems stated in this section can be found in Sect. 4.

### 3.1 Concavity

**Lemma 1** *$HM(x|F)$ under Definition 1 is a concave function for any finite $F$ in any finite $R$ in a univariate real space $\Re$.*

Using this lemma, we can obtain the following theorem on the concavity of the multi-dimensional half-space mass distribution.

**Theorem 1** *$HM(\mathbf{x}|F)$ under Definition 1 is a concave function for any finite $F$ in any finite, convex and closed $R \subset \Re^d$.*

Similarly, $HM(\mathbf{x}|D)$ is also concave in the convex region $R$ covering $D$.

### 3.2 Unique median

Based on Theorem 1, a unique location in $R$ which has the maximum half-space mass value is guaranteed, as stated in the following theorem:

**Theorem 2** *The "center" of a given density $F$ based on half-space mass $\mathbf{x}^* := \arg\max_{\mathbf{x}} HM(\mathbf{x}|F)$ is a unique location in $R$, given that $F$ covers an area more than a straight line in $\Re^d$.*

### 3.3 Breakdown point

For a given dataset $D$ of size $n$ and a location estimator $T$, the breakdown point $\epsilon(T, D)$ is defined in the following way as in Donoho and Gasko (1992), which is the minimum proportion of strategically chosen contaminating points required to render the estimated location arbitrarily far away from the original estimation:

$$\epsilon(T, D) = \min\left(\frac{m}{n+m} : \sup_{Q^{(m)}} ||T(D \cup Q^{(m)}) - T(D)||_2 = \infty\right) \tag{2}$$

where $Q^{(m)}$ is a set of contaminating data points of size $m$.

We define a location estimator based on half-space mass as follows: $T(D) := \arg\max_{\mathbf{x}} HM(\mathbf{x}|D)$. It is a maximally robust estimator with properties given in the following theorem:

**Theorem 3** *The breakdown point of $T$, $\epsilon(T, D) > \frac{n-1}{2n-1} \to \frac{1}{2}$ as $n \to \infty$.*

### 3.4 Extension across dimension

Dutta et al. (2011) reveal that, for a size $n$ dataset in a $d > n$ dimensional space, since the $d$-dimensional volume of the convex hull of such dataset is going to be zero, half-space depth will behave anomalously having 0 measures almost everywhere in $\Re^d$. In such cases, half-space depth does not carry any useful statistical information.

On the other hand, the definition of half-space mass enables it not only to rank locations outside the convex hull of the training dataset in the lower dimensional space where this convex hull has positive volume, but also to extend the ranking of locations to a higher dimensional space where the convex hull has zero volume.

As demonstrated in Fig. 3, the training data points are located on a straight line, thus the volume of the convex hull of them in $\Re^2$ is zero. This renders half-space depth to have zero

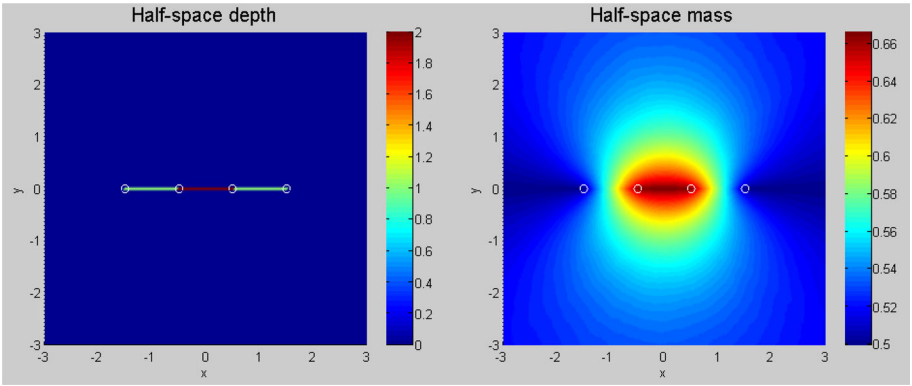


**Fig. 3** Distributions of half-space depth and half-space mass in $\Re^2$ with 4 training data points on a one-dimensional line shown in *white circle markers*. The color indicates the depth/mass values

measures almost everywhere unless the query point lies in the line segment. On the other hand, half-space mass is able to rank almost every location in $\Re^2$ based on their closeness to the center of the dataset. This ability of half-space mass to extend information carried in a dataset to a higher dimensional space could be very useful to high dimensional problems, especially when the sample size is limited.

## 4 Proofs

This section provides the proofs for the lemma and theorems given in the last section. The proofs for Lemma 1, Theorems 1, 2 and 3 are presented in the following four subsections.

### 4.1 Proof of Lemma 1

Given $R = [r_l, r_u]$, $\mathcal{H}(x)$ is a set of all half-spaces containing $x$ formed by splitting $\Re$ at any point $s \in R$. Then, $HM(x|F)$ is represented as follows.

$$\begin{aligned}
HM(x|F) &= \lim_{\mathbb{H}(x)\to\mathcal{H}(x)} \frac{1}{|\mathbb{H}(x)|} \sum_{H\in\mathbb{H}(x)} P_F(H) \\
&= \lim_{\mathbb{H}(x)\to\mathcal{H}(x)} \frac{1}{|\mathbb{H}(x)|} \sum_{H\in\mathbb{H}(x)} \left( I(s < x) \int_s^{r_u} F(y)dy + I(s \geq x) \int_{r_l}^{s} F(y)dy \right) \\
&= \lim_{\Delta s\to 0} \frac{1}{r_u - r_l} \Delta s \left( \sum_{i=1}^{m_x} \int_{s_i}^{r_u} F(y)dy + \sum_{i=m_x+1}^{m} \int_{r_l}^{s_i} F(y)dy \right) \\
&= \frac{1}{r_u - r_l} \left( \int_{r_l}^{x} \int_{s}^{r_u} F(y)dyds + \int_{x}^{r_u} \int_{r_l}^{s} F(y)dyds \right)
\end{aligned}$$

where $\Delta s = (r_u - r_l)/|\mathbb{H}(x)|$; $m$ and $m_x$ are $|\mathbb{H}(x)|$ and the number of $H \in \mathbb{H}(x)$ whose splitting point $s$ is $<x$, respectively. Since $HM(x|F)$ is a double integrated function of the finite $F(x)$, it is twice differentiable.

$$\begin{aligned}
\frac{dHM(x|F)}{dx} &= \lim_{\Delta x\to 0}\frac{HM(x+\Delta x|F)-HM(x|F)}{\Delta x}\\
&= \lim_{\Delta x\to 0}\frac{1}{r_u-r_l}\frac{1}{\Delta x}\left(\int_{r_l}^{x+\Delta x}\int_{s}^{r_u}F(y)dyds+\int_{x+\Delta x}^{r_u}\int_{r_l}^{s}F(y)dyds\right.\\
&\quad\left.-\int_{r_l}^{x}\int_{s}^{r_u}F(y)dyds-\int_{x}^{r_u}\int_{r_l}^{s}F(y)dyds\right)\\
&= \lim_{\Delta x\to 0}\frac{1}{r_u-r_l}\frac{1}{\Delta x}\int_{x}^{x+\Delta x}\left(\int_{s}^{r_u}F(y)dy-\int_{r_l}^{s}F(y)dy\right)ds\\
&= \lim_{\Delta x\to 0}\frac{1}{r_u-r_l}\frac{1}{\Delta x}\int_{x}^{x+\Delta x}\left(C_R-2\int_{r_l}^{s}F(y)dy\right)ds\\
&= \frac{1}{r_u-r_l}\left(C_R-2\int_{r_l}^{x}F(y)dy\right)\\
\Rightarrow \frac{d^2HM(x|F)}{dx^2} &= -\frac{2}{r_u-r_l}F(x)\le 0 \qquad (3)
\end{aligned}$$

where $C_R=\int_{r_l}^{s}F(y)dy+\int_{s}^{r_u}F(y)dy$. Since the double differential of $HM(x|F)$ is non-positive, $HM(x|F)$ is concave.

### 4.2 Proof of Theorem 1

Let $\mathcal{H}_\ell(\mathbf{x})\subset\mathcal{H}(\mathbf{x})$ be a set of all half-spaces in $\mathcal{H}(\mathbf{x})$ whose splitting hyperplanes are perpendicular to direction $\ell$ in $\Re^d$. Let $\mathcal{L}$ be a set of all directions $\ell\in\Re^d$. Define

$$HM(\mathbf{x}|F,\ell):=\lim_{\mathbb{H}_\ell(\mathbf{x})\to\mathcal{H}_\ell(\mathbf{x})}\frac{1}{|\mathbb{H}_\ell(\mathbf{x})|}\sum_{H\in\mathbb{H}_\ell(\mathbf{x})}P_F(H)$$

where $\mathbb{H}_\ell(\mathbf{x})$ is a subset of $\mathcal{H}_\ell(\mathbf{x})$. From Definition 1, $HM(\mathbf{x}|F)$ can be decomposed as

$$\begin{aligned}
HM(\mathbf{x}|F) &= E_{\mathcal{L}}[HM(\mathbf{x}|F,\ell)]\\
&= \lim_{\mathbb{L}\to\mathcal{L}}\sum_{\ell\in\mathbb{L}}HM(\mathbf{x}|F,\ell)P_\ell
\end{aligned}$$

where $P_\ell:=P(H\in\mathbb{H}(\mathbf{x})\text{ s.t. }H\in\mathbb{H}_\ell(\mathbf{x}))$ is the probability of a random half-space $H$ from $\mathbb{H}(\mathbf{x})$ belonging to the set $\mathbb{H}_\ell(\mathbf{x})$ and $\mathbb{L}\subset\mathcal{L}$ is the set of all directions $\ell$ corresponding to $\mathbb{H}(\mathbf{x})$.

$HM(\mathbf{x}|F,\ell)$ is equivalent to the univariate mass distribution on $\ell$ where $F$ is projected onto $\ell$. Accordingly, from Lemma 1, for all $\mathbf{x}\in R$, it is concave in the direction of $\ell$ and constant in the direction vertical to $\ell$. Thus, $HM(\mathbf{x}|F,\ell)$ is concave in $R$. Since the summation of multiple concave functions are also concave, $HM(\mathbf{x}|F)$ is concave in $R$.

### 4.3 Proof of Theorem 2

Here we prove Theorem 2 by contradiction.

Suppose there exists more than one location in $R$ that has the maximum half-space mass value, say $\mathbf{x}_1$ and $\mathbf{x}_2$. Let $\mathbf{x}^\ell$ denote the projection of $\mathbf{x}$ on a line along direction $\ell$ in $\Re^d$, $F^\ell$ denote the projection of density $F$ on $\ell$. Let $L=\{\mathbf{x}_1+c(\mathbf{x}_2-\mathbf{x}_1)|c\in(0,1)\}$ denote the

segment that connects $\mathbf{x}_1$ and $\mathbf{x}_2$, and $L^\ell = \{\mathbf{x}_1^\ell + c(\mathbf{x}_2^\ell - \mathbf{x}_1^\ell) | c \in (0,1)\}$ denote the projection of $L$. The concavity and the upper bound by the maximum value lead to the following:

$$HM\Big(c\mathbf{x}_1 + (1-c)\mathbf{x}_2|F\Big) = cHM(\mathbf{x}_1|F) + (1-c)HM(\mathbf{x}_2|F), \forall c \in (0,1) \tag{4}$$

The one-dimensional half-space mass of $F$ projected on $\ell$ is also concave in the projection of $R$, thus

$$\begin{aligned} &HM\Big(c\mathbf{x}_1^\ell + (1-c)\mathbf{x}_2^\ell|F^\ell\Big) \\ &\geq cHM(\mathbf{x}_1^\ell|F^\ell) + (1-c)HM(\mathbf{x}_2^\ell|F^\ell),\ \forall \ell, \forall c \in (0,1) \end{aligned} \tag{5}$$

Since $HM(\mathbf{x}|F) = E_{\mathcal{L}}[HM(\mathbf{x}^\ell|F^\ell)], \forall \mathbf{x}$, combining (4) and (5) we have

$$\begin{aligned} &HM\Big(c\mathbf{x}_1^\ell + (1-c)\mathbf{x}_2^\ell|F^\ell\Big) \\ &= cHM(\mathbf{x}_1^\ell|F^\ell) + (1-c)HM(\mathbf{x}_2^\ell|F^\ell),\ \forall \ell, \forall c \in (0,1) \end{aligned} \tag{6}$$

Equation (6) shows that $HM(\mathbf{x}^\ell|F^\ell)$ is linear for all $\mathbf{x}^\ell \in L^\ell$; thus whenever $HM(\mathbf{x}^\ell|F^\ell)$ is twice differentiable, by (3) we have

$$\begin{aligned} (6) &\Rightarrow \frac{d^2 HM(\mathbf{x}^\ell|F^\ell)}{d(\mathbf{x}^\ell)^2} = -\frac{2}{r_u - r_l} F^\ell(\mathbf{x}^\ell) = 0,\ \forall \ell, \forall \mathbf{x}^\ell \in L^\ell \\ &\Rightarrow F^\ell(\mathbf{x}^\ell) = 0, \forall \ell,\ \forall \mathbf{x}^\ell \in L^\ell \end{aligned} \tag{7}$$

where $r_u - r_l$ is the length of the projection of $R$ on $\ell$.

But since $F$ covers an area more than a straight line, there will always exist an $\ell$ and $\mathbf{x}$ such that $\mathbf{x}^\ell \in L^\ell$ and $F^\ell(\mathbf{x}^\ell) > 0$, which will contradict with (7). Therefore, there is one unique location that has the maximum half-space mass value in $R$.

### 4.4 Proof of Theorem 3

Suppose for a size $n$ dataset $D$, a contaminating set $Q$ of size $n-1$ is strategically chosen. Let $U$ denote the convex hull of $D$, and $U^\ell$ denote its projection segment on a line along direction $\ell$, assuming $U$ has a finite volume in $\Re^d$.

For any $\ell$, the median point of the projection of $D \cup Q$ on $\ell$ will lie within $U^\ell$. Because if it lies outside of $U^\ell$, then at least $n$ out of $2n-1$ points are on one side of the median which contradicts the definition of median. Since Ting et al. (2013) have shown that the univariate mass is maximised at its median, the maximum value of $HM(\mathbf{x}^\ell|D^\ell \cup Q^\ell)$ occurs in the segment $U^\ell$ for all $\ell$.

For a given query point $\mathbf{x}$, let $\mathcal{L}_{\mathbf{x}}^- = \{\ell : \mathbf{x}^\ell \notin U^\ell\}$ denote the set of directions in $\Re^d$ on which the projection of $\mathbf{x}$ lies outside of the projection of the convex hull of $D$, and $\mathcal{L}_{\mathbf{x}}^+ = \{\ell : \mathbf{x}^\ell \in U^\ell\}$ denote the rest of the directions.

For any $\ell \in \mathcal{L}_{\mathbf{x}}^-$, the one-dimensional mass $HM(\mathbf{x}^\ell|D^\ell \cup Q^\ell)$ increases while $\mathbf{x}^\ell$ moves a small enough distance towards $U^\ell$, since it is a concave function with the maximum value occurs somewhere in the segment $U^\ell$.

Let $\mathcal{H}_{\mathcal{L}_{\mathbf{x}}^-}(\mathbf{x}) \subset \mathcal{H}(\mathbf{x})$ be a set of all half-spaces in $\mathcal{H}(\mathbf{x})$ whose splitting hyperplanes are perpendicular to directions $\ell \in \mathcal{L}_{\mathbf{x}}^-$ in $\Re^d$, and $\mathcal{H}_{\mathcal{L}_{\mathbf{x}}^+}(\mathbf{x})$ be defined in the same way. By Definition 1, $HM(\mathbf{x}|D \cup Q)$ can be decomposed into the sum of two parts as follows:

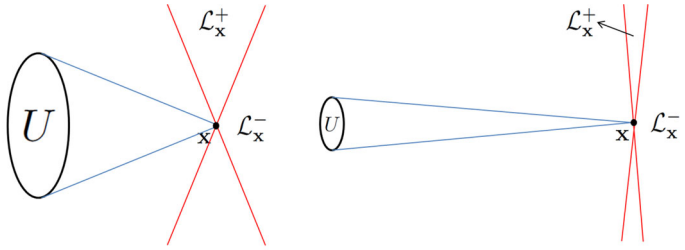


**Fig. 4** Demonstration of $\mathcal{L}^-_{\mathbf{x}}$ and $\mathcal{L}^+_{\mathbf{x}}$ in $\Re^2$. As the distance between $\mathbf{x}$ and $U$ increases to infinity, the solid angle of $U$ over $\mathbf{x}$ goes to 0, thus $\mathcal{L}^+_{\mathbf{x}}$ shrinks to a single direction

$$\begin{aligned} HM(\mathbf{x}|D \cup Q) &= E_{\mathcal{L}}\left[HM(\mathbf{x}^{\ell}|D^{\ell} \cup Q^{\ell})\right] \\ &= P_{\mathcal{L}^-_{\mathbf{x}}} E_{\mathcal{L}^-_{\mathbf{x}}}\left[HM(\mathbf{x}^{\ell}|D^{\ell} \cup Q^{\ell})\right] + P_{\mathcal{L}^+_{\mathbf{x}}} E_{\mathcal{L}^+_{\mathbf{x}}}[HM(\mathbf{x}^{\ell}|D^{\ell} \cup Q^{\ell})\Big] \end{aligned}$$

where $P_{\mathcal{L}^-_{\mathbf{x}}} := P(H \in \mathcal{H}(\mathbf{x}) \text{ s.t. } H \in \mathcal{H}_{\mathcal{L}^-_{\mathbf{x}}}(\mathbf{x}))$ is the probability of a random half-space $H$ from $\mathcal{H}(\mathbf{x})$ belonging to $\mathcal{H}_{\mathcal{L}^-_{\mathbf{x}}}(\mathbf{x})$; and $P_{\mathcal{L}^+_{\mathbf{x}}}$ is defined similarly.

Note that as the distance between $\mathbf{x}$ and $U$ goes to infinity, for a random direction $\ell$ in $\Re^d$, $P(\ell \in \mathcal{L}^-_{\mathbf{x}}) \to 1$ and $P(\ell \in \mathcal{L}^+_{\mathbf{x}}) \to 0$, hence $P_{\mathcal{L}^-_{\mathbf{x}}} \to 1$ and $P_{\mathcal{L}^+_{\mathbf{x}}} \to 0$, A demonstration is shown in Fig. 4.

The location estimator $T(D)$ is within $U$, the convex hull of $D$. If the distance between $T(D \cup Q)$ and $T(D)$ is infinity, then the distance between $T(D \cup Q)$ and $U$ is also infinity. Thus suppose $\mathbf{x}^* = T(D \cup Q)$ is infinitely far away from $U$, then the solid angle of $U$ over $\mathbf{x}^*$ is 0, therefore almost surely $\ell \in \mathcal{L}^-_{\mathbf{x}^*}, \forall \ell \in \Re^d$ and $HM(\mathbf{x}^*|D \cup Q) = E_{\mathcal{L}^-_{\mathbf{x}^*}}[HM(\mathbf{x}^{*\ell}|D^{\ell} \cup Q^{\ell})]$. Any movement of finite length from $\mathbf{x}^*$ towards $U$ will increase the one-dimensional mass values $HM(\mathbf{x}^{\ell}|D^{\ell} \cup Q^{\ell}), \forall \ell \in \mathcal{L}^-_{\mathbf{x}}$; thus increase the mass value $HM(\mathbf{x}|D \cup Q)$, which contradicts with the assumption that $HM(\mathbf{x}^*|D \cup Q)$ is the maximum. Therefore $T(D \cup Q)$ can only be finitely far away from $T(D)$ for a contaminating dataset $Q$ of size $n-1$.

Using the same inference as above, any contaminating dataset $Q$ of any size between 1 to $n-1$ combining dataset $D$ of size $n$ can only cause a finite shift of the location estimator $T$. Therefore $\epsilon(T, D) > \frac{n-1}{2n-1}$.

## 5 Relation to other data depth methods

Data depth models data distribution in terms of center-outward ranking rather than density or linear ranking, and it is a means to define multivariate median. Two example data depth definitions and their associated median definitions are given in Tables 2 and 3, respectively. Half-space depth and $L_2$ depth are chosen because the former employs the same half-spaces as in half-space mass; and the latter is another maximally robust method. The definition of half-space mass is also provided for comparison.

It is interesting to note the similarity between half-space mass and half-space depth, i.e., they are both based on the probability mass of half-spaces. The main difference is between taking the expectation or minimum over probability mass of half-spaces. This has led to the improvement of breakdown point and uniqueness of median shown in Table 3.

$L_2$ depth and half-space mass have the same four properties: concavity, unique median which is maximally robust and their distribution extends across dimensions which have zero-

**Table 2** Definitions of half-space mass ($HM(\cdot)$), half-space depth ($HD(\cdot)$) and $L_2$ depth ($L_2D(\cdot)$) with a given dataset $D$

| Depth function | Definition | Equation |
|---|---|---|
| Half-space mass | The expectation of probability mass of all half-spaces covering $\mathbf{x}$ | $HM(\mathbf{x}\|D) = E_{\mathcal{H}(\mathbf{x})}[P_D(H)]$ |
| Half-space depth | The minimum of probability mass of all half-spaces covering $\mathbf{x}$ (Tukey 1975) | $HD(\mathbf{x}\|D) = \min_{H \in \mathcal{H}(\mathbf{x})}[P_D(H)]$ |
| $L_2$ depth | The reciprocal of 1 plus the average of $L_2$ distances between $\mathbf{x}$ and each data point in $D$ (Mosler 2013) | $L_2D(\mathbf{x}\|D) = \left(1 + \frac{1}{\|D\|}\sum_{\mathbf{X}\in D}\|\|\mathbf{x}-\mathbf{X}\|\|_2\right)^{-1}$ |

**Table 3** Medians of half-space mass, half-space depth and $L_2$ depth and their properties

| Depth function | Multivariate median | Breakdown point; median unique? | Extension across dimension | Time complexity |
|---|---|---|---|---|
| Half-space mass | The point $\mathbf{x}$ which has the largest expected probability mass of all half-spaces covering $\mathbf{x}$. | $\frac{1}{2}$; unique | Yes | $O(nt)$ (sample version) $O(\psi t)$ (computation-friendly version) |
| Half-space depth | The point $\mathbf{x}$ which maximizes the minimum probability mass of all half-spaces covering $\mathbf{x}$. | $[1/(1+d), 1/3]$; Not unique (Aloupis 2006) | No | $O(nt)$ [An implementation as in Eq. (8)] |
| $L_2$ depth | The point which minimizes the sum of Euclidean distances to all points in a given data set. | $\frac{1}{2}$; unique (Lopuhaa and Rousseeuw 1991) | Yes | $O(n^2)$ |

volume convex hull. The key difference is the core mechanism: one employs half-space and the other uses distance. The computation without distance calculations leads directly to the advantage of half-space mass in time complexity, as shown in Table 3.

**Implementation**. We implement half-space depth using a technique similar to that used for $\widehat{HM}(\mathbf{x}|D)$. In the same context given in Definition 2, an estimator of half-space depth is defined as follows:

$$\widehat{HD}(\mathbf{x}|D) = \min_{H \in \mathbb{H}(\mathbf{x})}[P_D(H)] \tag{8}$$

We generate $t$ half-spaces, which cover $\mathbf{x}$ and intersect the convex hull of the given dataset, to find the one which gives the minimum probability mass. The implementation is similar to

those shown in Algorithms 1 and 2. The differences are: In training $\widehat{HD}(\mathbf{x}|D)$, $\psi$ must equal to $|D|$ and it is most efficient to set $\lambda = 1$. In the testing phase, $\widehat{HD}(\mathbf{x})$ finds the minimum probability mass of half-spaces, instead of averaging.

The implementation of $L_2$ depth is straightforward: Given a query point $\mathbf{x}$, compute the sum of Euclidean distances to all points in $D$. The output of $L_2D(\mathbf{x}|D)$ is computed as specified in Table 2.

## 6 Applications of half-space mass

We demonstrate the applications of half-space mass in two tasks: anomaly detection and clustering, in the following two subsections.

### 6.1 Anomaly detection

The application of half-space mass to anomaly detection is straightforward since the distribution of half-space mass is concave with center-outward ranking. Once every point in the given dataset is given a score, they can be sorted; and those close to the outer fringe of the distribution, i.e., having low scores, are more likely to be anomalies.

The above property is the same for half-space depth and $L_2$ depth. Thus, all three methods can be directly applied to anomaly detection.

### 6.2 Clustering

We provide a simple algorithm utilizing half-space mass in clustering. This algorithm is designed in a fashion that is similar to the K-means clustering algorithm.

Let $\mathbf{X}_i \in D, i = 1, ..., n$ denote data points in dataset $D$ and $Y_i \in \{1, ..., K\}$ denote the cluster labels, where $K$ is the number of clusters. Let $G_k := \{\mathbf{X}_i \in D : Y_i = k\}$, where $k \in \{1, ..., K\}$, denote the points in the $k$-th group.

The K-mass clustering procedure is given in Algorithm 3. The procedure begins with an initialization that randomly splits the dataset into $K$ equal-size groups. Each iteration consists of two steps. First, data in each group is used to generate a mass distribution $\widehat{HM}$. Second, each point $\mathbf{X}_i$ in the data set is then regrouped based on the mass distributions as follows: $\widehat{HM}$ for each group produces a mass value for $\mathbf{X}_i$; and it is assigned to the group which gives the maximum mass value. We normalise the mass values by the global minimum mass value to give small size groups a better chance to survive the process. The above two steps are iterated until the group labels stay unchanged, between two subsequent iterations, for at least $p$ proportion of the points in the dataset.

K-means clustering algorithm (Jain 2010) is provided in Algorithm 4 for comparison. The K-mass algorithm and the K-means algorithm share the same algorithmic structure. They differ only in the action required in each of the two steps in the iteration process.

Note that when considering K-means as an EM (Expectation-Maximisation) algorithm (Kroese and Chan 2014), K-means implements the expectation step in line 3 and the minimisation step in lines 4–6 in Algorithm 4. Similarly, K-mass implements the expectation step in line 3 and the maximisation step in lines 4–6 in Algorithm 3.

**Algorithm 3**: K-mass clustering algorithm

**input** : $D$ - Dataset; $p$ - proportion of $D$; $K$ - number of clusters
**output**: $\{G_k, k = 1, \ldots, K\}$
1 Initialize: segregate the dataset $D$ into $K$ equal size groups $\{G_k, k = 1, \ldots, K\}$ with hyperplanes of random directions, and $\forall\ X_i \in G_k$, label $Y_i = k$.
2 **while** *labels stay unchanged in* $< p$ *proportion of* $D$ **do**
3 For each group $G_k, k = 1, \ldots, K$, build $\widetilde{HM}(\cdot|G_k)$ to yield $\widetilde{HM_k}(\cdot)$
4 **for** $i = 1, ..., n$ **do**
5 $$Y_i \leftarrow \underset{k\in\{1,\ldots,K\}}{\arg\max} \frac{\widetilde{HM_k}(\mathbf{X}_i)}{\min_{j\in\{1,\ldots,n\}} \widetilde{HM_k}(\mathbf{X}_j)}$$
6 **end**
7 **end**
8 **return** $\{G_k, k = 1, \ldots, K\}$.

**Algorithm 4**: K-means clustering algorithm

**input** : $D$ - Dataset; $p$ - proportion of $D$; $K$ - number of clusters
**output**: $\{G_k, k = 1, \ldots, K\}$
1 Initialize: segregate the dataset $D$ into $K$ equal size groups $\{G_k, k = 1, \ldots, K\}$ with hyperplanes of random directions, and $\forall\ X_i \in G_k$, label $Y_i = k$.
2 **while** *labels stay unchanged in* $< p$ *proportion of* $D$ **do**
3 For each group $G_k, k = 1, \ldots, K$, obtain a group center $\mathbf{C}_k$, by averaging its members.
4 **for** $i = 1, ..., n$ **do**
5 $$Y_i \leftarrow \underset{k\in\{1,\ldots,K\}}{\arg\min} ||\mathbf{X}_i - \mathbf{C}_k||_2$$
6 **end**
7 **end**
8 **return** $\{G_k, k = 1, \ldots, K\}$.

## 7 Empirical evaluations

In this section, we conduct experiments to investigate the advantages of utilizing half-space mass in anomaly detection and clustering, first with artificial data sets and second with real datasets. In both cases, robustness is the key determinant for half-space mass to gain advantage over its contenders.

To simplify notations, we use $HM$ and $HM^*$ hereafter to denote the sample version ($\psi = |D|$) and the computational-friendly version ($\psi \ll |D|$) of half-space mass, respectively. And $HD$ and $L_2D$ denote half-space depth and $L_2$ depth, respectively.

### 7.1 Anomaly detection

In this section, half-space mass, half-space depth and $L_2$ depth are used for anomaly detection. That is, given a dataset, $HM$ is constructed as described in Algorithms 1 and 2; $HD$ and $L_2D$ are constructed as described in Sect. 5. Then, each of the models is used to score each point in the dataset. In all cases, points with low mass/depth scores are more likely to be anomalies. The final ranking of the points is sorted based on the scores produced from each model.

Area under the ROC curve (AUC) is used to measure the detection accuracy of an anomaly detector. $AUC = 1$ indicates that the anomaly detector ranks all anomalies in front of normal points; $AUC = 0.5$ indicates that the anomaly detector is a random ranker. Visualizations are used to show the impact of robustness. When comparing AUC values in the second

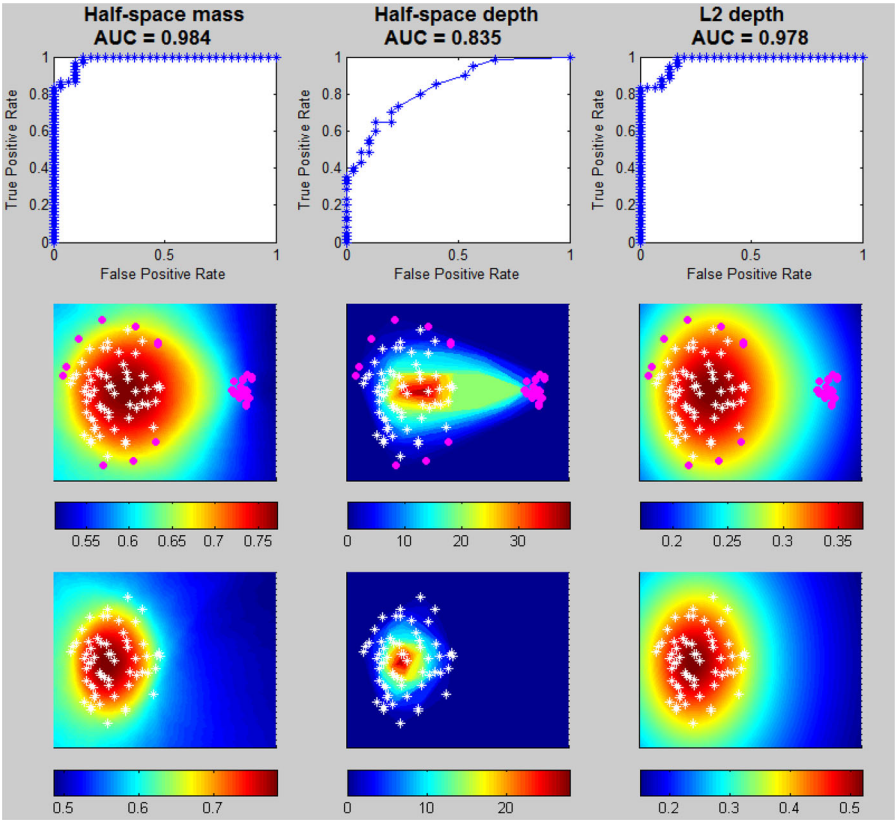


**Fig. 5** Anomaly detection on an artificial dataset, using *HM*, *HD* and $L_2D$. The first row of the plots shows the ROC curves, the second row of the plots shows all the data points and the contour maps, and the third row of the plots shows the normal data points only and the contour maps built with only these normal points. The *white star marker* denotes normal points while the *magenta dot marker* denotes anomalous points. The *color bar* indicates the mass/depth value

experiment, a $t$-test with 5 % significance level is conducted based on AUC values of multiple runs.

The $t$ parameter for both $HM$ and $HD$ is set to 5000 in the experiments, which is sufficiently large since further increase of $t$ observes no noticeable AUC improvement. $L_2$ depth has no parameter setting.

### *7.1.1 Anomaly detection with artificial data*

Here we show the importance of robustness of an anomaly detector in identifying anomalies. An artificial data set with two clusters of data points is generated for the experiment. As shown in Fig. 5, the dataset consists of a cluster of sparse normal points along with a few local anomalies on the left and a dense cluster of anomalies on the right. Center-outward ranking scores are calculated using *HM*, *HD* and $L_2D$.

The AUC results, presented in the first row in Fig. 5, show that both $HM$ and $L_2D$ performed much better than *HD*. In this example, all of the three methods failed to detect

some local anomalies but *HD* failed to detect the anomaly cluster on the right while the other two methods separated the anomaly cluster from the normal points perfectly.

The second row of the plots in Fig. 5 shows the contour maps of mass/depth values when normal points contaminated with noise were used to train the anomaly detectors; and the third row of the plots shows the contour maps when normal data points only were used to train the anomaly detectors.

The contrast between the second row and the third row of the plots is a testament to the impact of robustness. Being maximally robust, the contour maps of *HM* and $L_2 D$ remain centered inside the normal cluster. In contrast, the contour map of *HD* is significantly stretched towards the anomaly cluster. This resulted many clustered anomalies (on the right) being scored with high depth values as equivalent to many normal points; and thus impaired its ability to detect anomalies. Anomalies are contamination to the distribution of normal points. An anomaly detector, which is not robust to contamination, often results in poor ranking outcomes in relation to detecting anomalies. This example shows the impact of contamination has to an anomaly detector which is not robust.

### *7.1.2 Anomaly detection with benchmark datasets*

Here we evaluated the performance of *HM*, $HM^*$, $HD$ and $L_2 D$ in anomaly detection using nine benchmark datasets (Lichman 2013). AUC values and runtime results are shown in Table 4. The figures are the average of 10 runs except for $L_2 D$ which is a deterministic method. Boldface figures in the *HM*, $HM^*$ and $L_2$ columns indicate that the differences are significant compared to $HD$; while boldface figures in the $HD$ column indicate that the differences are significant compared to any of the other methods.

In comparison with $HD$, both $HM$ and $HM^*$ have 7 wins and 2 losses, which is evidence that half-space mass performed better than $HD$ in most datasets.

Note that $HM$ and $L_2 D$ have similar AUC results. This is not surprising since both have the same four properties shown in Table 3.

$HM^*$ using $\psi = 10$ performed comparably with $HM$ in seven out of the nine data sets. This suggests that the performance of $HM^*$ can be further improved by tuning $\psi$.

The major disadvantage of $L_2 D$ is its computational cost. $L_2 D$ ran orders of magnitude slower than the other methods in all data sets, except in the smallest data set with 64 points only. This is because not only $L_2 D$ has a time complexity $O(n^2)$, it also involves distance measures. The freedom from distance measure is an important feature of half-space mass, which makes it much more efficient.

Note that *HD* performed poorly in all three high dimensional datasets. Our investigation suggests that as the number of dimensions increases, an increasing percentage of points will appear at the outer fringe of the convex hull covering the data set. Because *HD* assigns the same lowest depth value to all these points, they are thus unable to be meaningfully ranked. This is the reason why the AUC results of *HD* in these three datasets are close to 0.5, equivalent to random ranking. In a nutshell, *HD* is more prone to the curse of dimensionality than *HM* or $L_2 D$.

*HD* outperformed three other methods in the smtp and covertype datasets. A visualization of the smtp dataset revealed that all anomalous points are located at one corner of the data space close to one normal cluster, as shown in Fig. 6. Being at the corner, *HD* assigned these anomalies with the same lowest score as all points at the outer fringe, while *HM* or $L_2$ would assign them higher scores since they are closer to the center than other fringe points. Had the points located in-between two clusters but had the same distance from the same cluster, *HD*

**Table 4** Anomaly detection performance with the benchmark datasets, where $n$ is data size, $d$ is the number of dimensions, and "ano" is the percentage of anomalies

| Dataset | $n$ | $d$ | ano (%) | AUC | | | | Runtime (second) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $HM$ | $HM^*$ | $HD$ | $L_2$ | $HM$ | $HM^*$ | $HD$ | $L_2$ |
| Mulcross | 262144 | 4 | 10.00 | **1.00** | **1.00** | 0.86 | **1.00** | 30.3 | 26.3 | 30.3 | 2213.0 |
| Satellite | 6435 | 36 | 31.60 | **0.61** | **0.62** | 0.57 | **0.62** | 1.1 | 0.8 | 1.2 | 11.2 |
| Shuttle | 49097 | 9 | 7.15 | **0.99** | **0.99** | 0.92 | **0.99** | 5.4 | 5.3 | 5.2 | 133.5 |
| Smtp | 95156 | 3 | 0.03 | 0.77 | 0.73 | **0.83** | 0.78 | 6.9 | 8.0 | 6.7 | 218.9 |
| Isolet | 7797 | 617 | 3.85 | **0.82** | **0.85** | 0.68 | **0.84** | 24.9 | 13.4 | 25.0 | 229.1 |
| Mfeat | 2000 | 649 | 10.00 | **0.92** | **0.93** | 0.56 | **0.92** | 5.6 | 3.3 | 5.7 | 17.8 |
| Covertype | 286048 | 10 | 0.96 | 0.87 | 0.78 | **0.92** | 0.87 | 45.7 | 35.3 | 44.5 | 5251.3 |
| Http | 567497 | 3 | 0.39 | **1.00** | **1.00** | 0.99 | **1.00** | 55.1 | 57.3 | 54.4 | 7794.4 |
| Dbworld | 64 | 4702 | 45.31 | **0.78** | **0.78** | 0.53 | **0.79** | 2.0 | 2.1 | 2.0 | 0.1 |

Bold values indicate a 5 % significance level difference between $HD$ and the other three methods

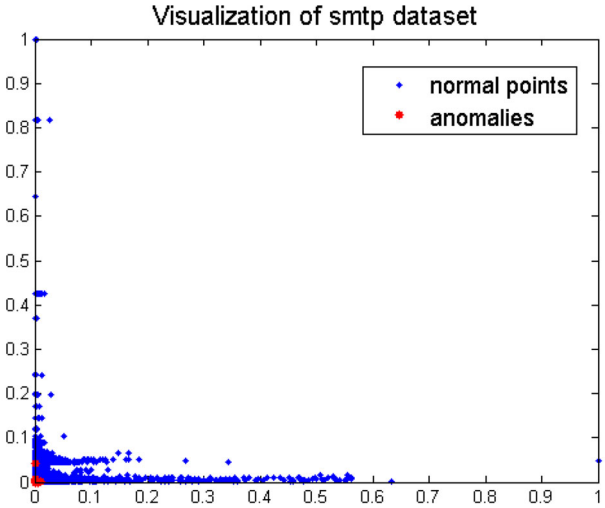


**Fig. 6** Visualization of the smtp dataset projected on the first two dimensions. Since almost all points have very similar values in the third feature, neglecting the third dimension does not affect the point of this visualization. Note that all anomalous points are located at the lower left corner, where dense clusters of normal points are located

would have regarded them as normal points. In other words, *HD* is able to better detect them in this dataset simply because of the special positions the anomalies are placed.[1]

The runtime shown in Table 4 is the sum of training time and testing time. Because the efficiency of the computation-friendly version affects the training process only, Table 5 is provided to show the training and testing time of $HM$ and $HM^*$ separately. With a small subsample size $\psi = 10$, $HM^*$ runs at least two orders of magnitude faster than $HM$ in the training phase in large datasets. Note that in Table 5, the testing time of $HM^*$ is noticeably

[1] We suspect that the result in the covertype dataset is due to the similar reason. But we could not visualize it due to its dimensionality.

**Table 5** The training and testing times of $HM$ and $HM^*$ with subsample size $\psi = 10$

| Dataset | $n$ | $d$ | Training time (second) | | Testing time (second) | |
|---|---|---|---|---|---|---|
| | | | $HM$ | $HM^*$ | $HM$ | $HM^*$ |
| mulcross | 262,144 | 4 | 9.291 | 0.073 | 21.009 | 26.227 |
| satellite | 6435 | 36 | 0.429 | 0.082 | 0.671 | 0.718 |
| shuttle | 49,097 | 9 | 1.545 | 0.073 | 3.855 | 5.227 |
| smtp | 95,156 | 3 | 1.639 | 0.071 | 5.261 | 7.929 |
| isolet | 7797 | 617 | 11.953 | 0.509 | 12.947 | 12.891 |
| mfeat | 2000 | 649 | 2.810 | 0.426 | 2.790 | 2.874 |
| covertype | 286,048 | 10 | 15.632 | 0.080 | 30.068 | 35.220 |
| http | 567,497 | 3 | 17.706 | 0.072 | 37.394 | 57.228 |
| dbworld | 64 | 4702 | 1.315 | 1.370 | 0.685 | 0.730 |

longer than $HM$ for most datasets, while they are theoretically expected to be equal since the amount of computation are exactly the same. Our investigation reveals that this is due to a computational issue of Matlab.[2]

In summary, half-space mass is the best anomaly detectors among the three methods, which has significantly better detection accuracy than $HD$ and runs orders of magnitude faster than $L_2D$.

### 7.2 Clustering

This section reports the empirical evaluation of K-mass in comparison with K-means. The first experiment examines the three scenarios in which K-means is known to have difficulty to find all clusters, i.e., clusters with different sizes, densities and the presence of noise. The second experiment evaluates the clustering performance using eight real data sets (Lichman 2013, Franti et al. 2006).[3]

In every trial using a data set, K-mass or K-means is executed 40 runs and we report the best clustering result. The clustering performance is measured in terms of F-measure, and visualizations of the clustering results are presented where possible in two-dimensional datasets.

K-mass employs $HM^*$ which uses $\psi = 5$ and $t = 2000$ as default in all experiments; it uses $\lambda = 3$ in the first experiment, and $\lambda = 1.6$ in the second experiment. Recall that $\lambda$ controls the size of the convex hull covering the data set. Because the sample size is $\psi = 5$, the convex hull must be enlarged (using $\lambda > 1$) in order to cover points which exist outside the convex hull. For the stopping criterion $p$, both K-mass and K-means use $p = 1$ in the first experiment and search for the best result with $p = 0.98$ and 1 in the second experiment.

[2] When comparing a fixed size vector to a scalar in Matlab, the runtime of such comparison is not constant. It varies significantly depending on the value of the scalar. The closer the scalar is to the median of the numbers in the vector, the longer it takes for the comparison. Because $HM^*$ uses a small subsample for projection, the split points $s_i$ in Algorithm 1 are selected within a narrower range than if the whole dataset was used. Thus $s_i$ lies near the median of the whole dataset more often in $HM^*$ than in $HM$. As a result, the comparisons take significantly longer time in $HM^*$ than in $HM$ in the testing stage. However, this effect is dampened in high dimensional datasets because the high dimensionality makes the range after projection much longer, even for a small subsample. This irregularity will not occur if another programming language is used.

[3] The dim dataset is from Franti et al. (2006) and all other datasets are from Lichman (2013).

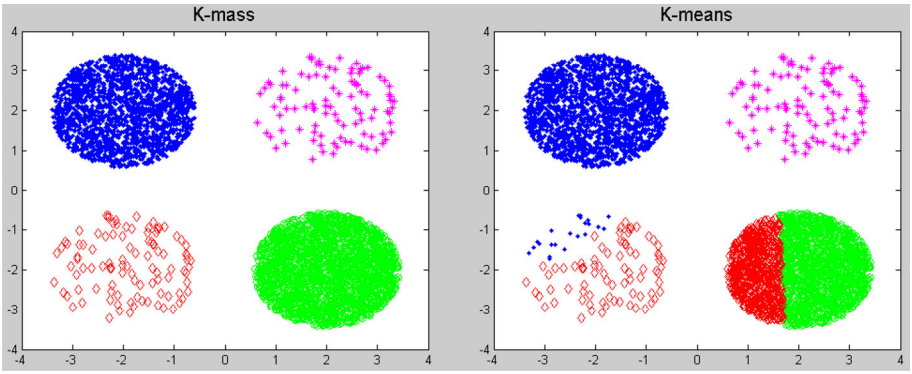


**Fig. 7** Clustering of data groups with different densities. The best converged F-measures are 1 and 0.88 for K-mass and K-means, respectively

### *7.2.1 Clustering with artificial data*

Figures 7, 8 and 9 show the clustering results of K-mass and K-means on three artificial datasets, representing scenarios having clusters with different sizes, densities and the presence of noise, respectively.

In scenario 1, as shown in Fig. 7, the dataset consists of two sparse clusters and two significantly denser clusters. K-mass easily converged to the global optimal result. But K-means converged to a local optimal result which wrongly assigned some points. While it is possible that K-means can converge to the global optimal result if an ideal initialization is generated, this is unlikely because the sparse and dense clusters have different data sizes.

In scenario 2, the four clusters are of equal density but with different data sizes, as shown in Fig. 8. K-mass worked well separating the four clusters; but K-means failed to converge to the global optimum because of its tendency to split half-way between group centers.

Scenario 3 demonstrates the importance of robustness in clustering. The dataset consists of four clusters of equal sizes and density with the presence of noise, scattered around the four clusters. Figure 9 shows that K-mass, in spite of having a F-measure $<1$ because the noise points were assigned to the nearest clusters, was able to separate the four clusters perfectly; while K-means wrongly assigned many points of the four clusters. This is because K-means is not robust against outliers, therefore the group centers could be easily influenced by noise.

In summary, K-mass perfectly separated the four clusters while K-means failed to do so in all three scenarios.

### *7.2.2 Clustering with real datasets*

Table 6 lists the data characteristics as well as the best results of K-mass and K-means in terms of F-measure. K-mass outperforms K-means with 6 wins, 1 draw and 1 loss. K-mass runs slower than K-means because it must train $K$ models at each iteration; and K-mass is expected to need more iterations than K-means in general.

## 8 Discussion

Mass estimation (Ting et al. 2013) was recently proposed as an alternative to density estimation in data modeling. It has significant advantages over density estimation in efficiency

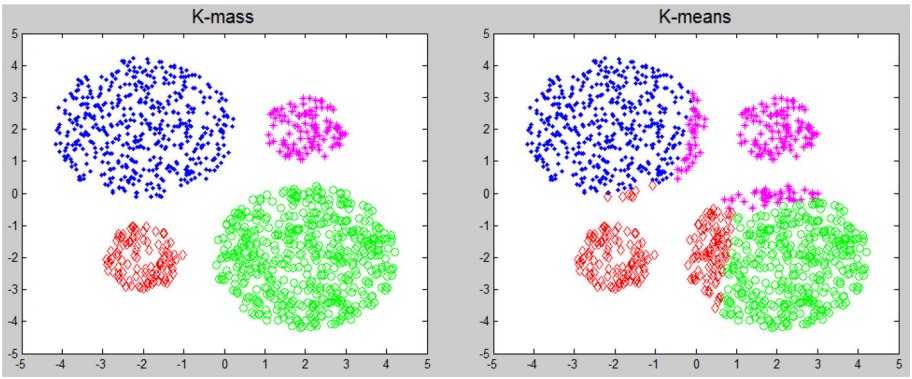


**Fig. 8** Clustering of data groups with same density but different group sizes. The best converged F-measures are 1 and 0.84 for K-mass and K-means, respectively

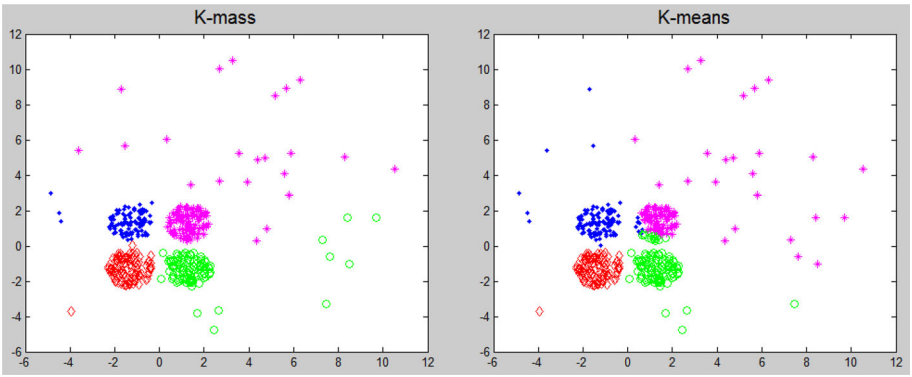


**Fig. 9** Clustering of data groups with the same density and the same group size, with the presence of noise points. The best converged F-measures are 0.89 and 0.84 for K-mass and K-means, respectively

**Table 6** Clustering with real datasets

| Dataset | $n$ | $d$ | $K$ | K-mass | | | | K-means | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Best F | $p$ | time | $l$ | Best F | $p$ | time | $l$ |
| Iris | 150 | 4 | 3 | 0.933 | 1 | 0.40 | 4 | 0.920 | 0.98 | 0.001 | 3 |
| Seeds | 210 | 7 | 3 | 0.923 | 0.98 | 0.53 | 5 | 0.919 | 0.98 | 0.001 | 2 |
| Column | 310 | 6 | 3 | 0.684 | 0.98 | 2.13 | 18 | 0.675 | 0.98 | 0.002 | 4 |
| Banknote | 1372 | 4 | 2 | 0.725 | 0.98 | 0.59 | 4 | 0.602 | 0.98 | 0.012 | 8 |
| Breast | 699 | 9 | 2 | 0.963 | 0.98 | 0.44 | 4 | 0.961 | 0.98 | 0.002 | 2 |
| Dim | 1024 | 1024 | 16 | 1.000 | 1 | 29.16 | 2 | 1.000 | 1 | 0.308 | 2 |
| Wdbc | 569 | 30 | 2 | 0.934 | 0.98 | 0.59 | 5 | 0.929 | 0.98 | 0.004 | 5 |
| Wine | 178 | 13 | 3 | 0.944 | 0.98 | 0.86 | 8 | 0.966 | 1 | 0.002 | 4 |

Best F-measure out of 40 runs. The header "time" means the runtime (in seconds) corresponding to the best F measure and $l$ is the number of iterations before reaching the stopping criterion

and/or efficacy in various data mining tasks such as anomaly detection, clustering, classification and information retrieval (Ting et al. 2013). Despite this success, the formal definition of mass is univariate only and its theoretical analysis is limited to two properties: (i) its mass distribution is concave, and (ii) its maximum mass point is equivalent to median (Ting et al. 2013).

The half-space mass can be viewed as a generalisation of the univariate mass estimation to multi-dimensional spaces, and it has four properties rather than the two revealed previously. The one-dimensional mass estimation is defined as the weighted probability mass (see the details in the Appendix). Half-space splits reduce to binary splits, and the half-space mass reduces to the weighted probability mass in one dimensional space defined in Ting et al. (2013).

The two additional properties of half-space mass, i.e., maximal robustness and extension across dimension, are important in understanding the behaviour of any algorithms designed based on half-space mass, as we have shown in the empirical evaluation section.

The proof for concavity in Lemma 1 made use of the same idea for the concavity proof as presented by Ting et al. (2013). Other ideas in this paper are new.

Ting et al. (2013) also gave a definition of higher level mass estimation, which can be viewed as a localised version of a level-1 mass estimation. We have limited our exposition to level-1 mass estimation in this paper so that we have a direct comparison with data depth and its properties. As a result, it is limited to data modeling with a unimodal distribution having a unique maximum as the median. In datasets which have multi-modal distribution, $HM$ will be outperformed by existing density-based anomaly detectors. We believe that $HM$ can be extended to higher level mass estimation as shown in the one-dimensional case (Ting et al. 2013), which could be regarded as a localized data depth method (Agostinelli and Romanazzi 2011). We will explore higher level mass estimation using half-space mass in the near future.

The successful application of half-space mass in K-mass implies that other data depth methods may also be applicable in K-mass. Our investigation reveals that because half-space depth can only provide its estimations within the convex hull of a given data set (i.e., the lack of the fourth property stated in Sect. 3.4), it could not be applied to K-mass. A K-mass version using $L_2$ depth exhibits a better convergence property than K-mass. However, its performance is in general worse than both K-mass and K-means.[4] Another drawback of $L_2$ depth is that it is very costly to compute in large datasets.

Despite all the advantages of K-mass over K-means shown in this paper, a caveat is in order here: we do not have a proof that K-mass will always converge like K-means.

## 9 Conclusions

This paper makes three key contributions:

First, we propose the first formal definition of half-space mass, which is a significantly improved version of half-space data depth, and it is the only data depth method which is both robust and efficient, as far as we know.

Second, we reveal four theoretical properties of half-space mass: (i) half-space mass is concave in a convex region; (ii) it has a unique median; (iii) the median is maximally robust; and (iv) its estimation extends to higher dimensional space in which training data occupies zero-volume convex hull.

[4] The best F-measure out of 40 runs using $L_2$ depth in clustering with the eight datasets are: 0.947(iris), 0.905(seeds), 0.626(column), 0.595(banknote), 0.939(breast), 1(dim), 0.896(wdbc), 0.943(wine).

Third, we demonstrate applications of half-space mass in two tasks: anomaly detection and clustering. In anomaly detection, it outperforms the popular half-space depth because it is more robust and able to extend across dimensions; and it runs orders of magnitude faster than $L_2$ data depth. In clustering, we introduce K-mass by using half-space mass, instead of a distance function, in the expectation and maximisation steps in K-means. We show that K-mass overcomes three weaknesses of K-means. The maximally robust property of half-space mass contributes directly to these outcomes in both tasks.

**Acknowledgments** This project is partially supported by a grant from the U.S. Air Force Research Laboratory, under agreement # FA2386-13-1-4043, awarded to Kai Ming Ting. It is also partially supported by JSPS KAKENHI Grant Number 25240036, awarded to Takashi Washio. Bo Chen and Gholamreza Haffari are grateful to National ICT Australia (NICTA) for their generous funding, as part of the Machine Learning Collaborative Research Projects. Bo Chen is also supported by a scholarship from the Faculty of Information Technology, Monash University.

## Appendix: One-dimensional mass

This appendix reiterated the one-dimensional mass estimation, as presented by Ting et al. (2010), for ease of comparison with the half-space mass introduced in this paper.

Let $x_1 < x_2 < \cdots < x_{n-1} < x_n$ on the real line, $x_i \in \mathcal{R}$ and $n > 1$. Let $s_i$ be the binary split between $x_i$ and $x_{i+1}$, yielding two non-empty regions having two masses $m_i^L$ and $m_i^R$.

**Definition 4** Mass base function:
$m_i(x)$ as a result of $s_i$, is defined as

$$m_i(x) = \begin{cases} m_i^L & \text{if } x \text{ is on the left of } s_i \\ m_i^R & \text{if } x \text{ is on the right of } s_i \end{cases}$$

Note that $m_i^L = n - m_i^R = i$.

**Definition 5** Mass distribution: $mass(x_a)$ for a point $x_a \in \{x_1, x_2, \cdots, x_{n-1}, x_n\}$ is defined as a summation of a series of mass base functions $m_i(x)$ weighted by $p(s_i)$ over $n-1$ splits as follows, where $p(s_i)$ is the probability of selecting $s_i$.

$$\begin{aligned} mass(x_a) &= \sum_{i=1}^{n-1} m_i(x_a) p(s_i) \\ &= \sum_{i=a}^{n-1} m_i^L p(s_i) + \sum_{j=1}^{a-1} m_j^R p(s_j) \\ &= \sum_{i=a}^{n-1} i p(s_i) + \sum_{j=1}^{a-1} (n-j) p(s_j) \end{aligned}$$

Note that it is defined $\sum_{i=q}^{r} f(i) = 0$, when $r < q$ for any function $f$. $p(s_i)$ can be estimated on the real line as $p(s_i) = (x_{i+1} - x_i)/(x_n - x_1) > 0$, as a result of random selection of splits based on a uniform distribution.

## References


Agostinelli, C., & Romanazzi, M. (2011). Local depth. *Journal of Statistical Planning and Inference*, *141*(2), 817–830.

Aloupis, G. (2006). Geometric measures of data depth. *DIMACS Series in Discrete Math and Theoretical Computer Science*, *72*, 147–158.

Donoho, D. L., & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, *20*(4), 1803–1827.

Dutta, S., Ghosh, A. K., & Chaudhuri, P. (2011). Some intriguing properties of Tukey's half-space depth. *Bernoulli*, *17*(4), 1420–1434.

Franti, P., Virmajoki, O., & Hautamaki, V. (2006). Fast agglomerative clustering using a k-nearest neighbor graph. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(11), 1875–1881.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means, *Pattern Recognition Letters, 31*(8), 651–666.

Kroese, D. P., & Chan, J. C. C. (2014). *Statistical modeling and computation*. New York: Springer.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics*, *27*(3), 783–840.

Lopuhaa, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics*, *19*(1), 229–248. doi:10.1214/aos/1176347978.

Mosler, K. (2013). Depth statistics. In C. Becker, R. Fried, & S. Kuhnt (Eds.), *Robustness and complex data structures. Festschrift in honour of Ursula Gather* (pp. 17–34). Berlin: Springer.

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining* (2nd ed.). Pearson Education, Ltd.

Ting, K. M., Zhou, G.-T., Liu, F., & Tan, J. S.C. (2010). Mass estimation and its applications. In *Proceedings of KDD'10: The 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 989–998)

Ting, K. M., Zhou, G.-T., Liu, F., & Tan, J. S. C. (2013). Mass estimation. Machine Learning, 90(1), 127–160.

Tukey, J. W. (1975). Mathematics and picturing data. *Proceedings of 1975 international congress of mathematics, Vol. 2,* (pp. 523–531).

Zuo, Y., & Serfling, R. (2000). General notion of statistical depth function. *Annals of Statistics*, *28*, 461–482.