Binary Classification Using Some Notions of Kernelized Data Depths

Kaushik Thakkar

Indian Institute of Technology Kanpur

June 8, 2024

Under the supervision of
Dr. Subhajit Dutta

Data depth

Data depth is a non-parametric approach that measures the relative position of a point with respect to the given data cloud (or, a probability distribution). A simple example is

$$D(x, F_n) = \min\{F_n(x), 1 - F_n(x)\}.$$

$F_n(x)$ is the proportion of data points that are on the left of x (the empirical distribution function). Note: $\max\limits_{x \in \mathbb{R}} D(x, F_n) = \frac{1}{2}$.

$D(x, F_n) = \min\{\frac{6}{10}, \frac{4}{10}\} = \frac{2}{5}$; $D(x, F_n) = \min\{\frac{2}{10}, \frac{8}{10}\} = \frac{1}{5}$

Red point is more close to the centre as compared to green point.

Desirable properties of a depth function

Let $D(.,.) : \mathbb{R}^d \times \mathcal{F} \to [0,1]$ be a bounded, non-negative depth function. Ideally, D should satisfy:

- **Affine invariant**: $D(A\mathbf{x} + \mathbf{b}, F_{A\mathbf{X}+\mathbf{b}}) = D(\mathbf{x}, F_{\mathbf{X}})$.
- **Maximality at center**: $D(\boldsymbol{\theta}, F) = \sup_{\mathbf{x} \in \mathbb{R}^d} D(\mathbf{x}, F)$ for any F with center $\boldsymbol{\theta} \in \mathbb{R}^d$.
- **Monotonicity relative to deepest point**: For symmetric F (with centre at $\boldsymbol{\theta}$), $D(\mathbf{x}, F) \leq D(\alpha \cdot \mathbf{x} + (1-\alpha) \cdot \boldsymbol{\theta}, F)$, for $\alpha \in [0,1]$.
- **Vanishing at infinity**: $D(\mathbf{x}, F) \to 0$ as $\|\mathbf{x}\|_2 \to \infty$.

Half-space depth (HD)

Half-space depth at $\mathbf{x} \in \mathbb{R}^d$ w.r.t the distribution F is defined to be

$$HD(\mathbf{x}; F) = \inf_{H}\{P(H) : \text{H is a closed half-space in } \mathbb{R}^d \text{ and } \mathbf{x} \in H\}. \quad (1)$$

Alternatively,

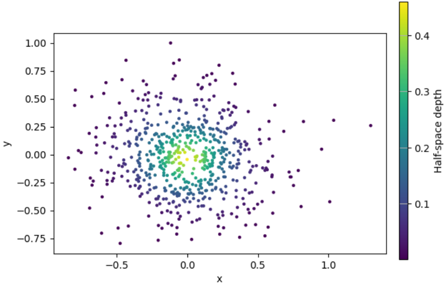
$$HD(\mathbf{x}; F) = \inf_{\mathbf{u} \in \mathbb{S}(0,1)} \mathbb{P}(\langle \mathbf{u}, \mathbf{X} \rangle \geq \langle \mathbf{u}, \mathbf{x} \rangle) \text{ for } \mathbf{X} \sim F. \quad (2)$$

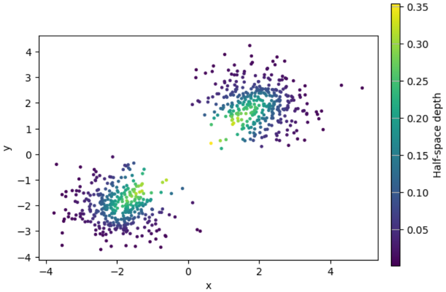
A sample version of HD is given by:

$$HD(\mathbf{x}; F_n) = \inf_{\mathbf{u} \in \mathbb{S}(0,1)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\langle \mathbf{u}, \mathbf{x_i} \rangle \geq \langle \mathbf{u}, \mathbf{x} \rangle\right), \quad (3)$$

where $\mathbb{I}$ is the indicator function.

HD for data with convex and non-convex support





Left: Data is simulated from $N_2((0,0)^T, I_2)$ (convex support).

Right: Data is simulated from $\frac{1}{2}N_2((2,2)^T, I_2) + \frac{1}{2}N_2((-2,-2)^T, I_2)$ (non-convex support).

Spatial depth (SPD)

Spatial depth at $\mathbf{x} \in \mathbb{R}^d$ w.r.t the distribution F is defined as:

$$SPD(\mathbf{x}; F) = 1 - \|E_F[S(\mathbf{Y} - \mathbf{x})]\|_2 \text{ for } \mathbf{Y} \sim F \ , \tag{4}$$

where

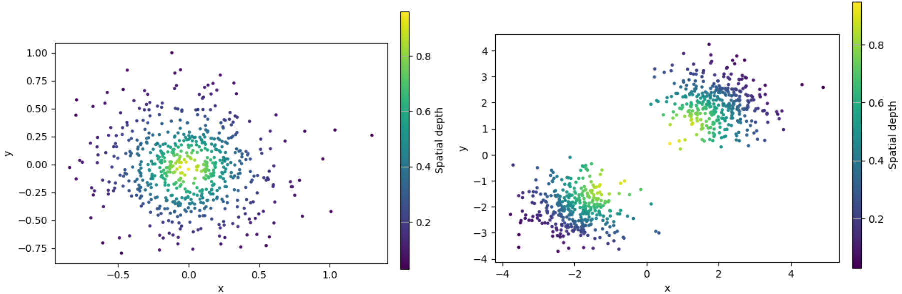
$$S(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|_2} & \text{if } \mathbf{x} \neq \mathbf{0} \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0} \ . \end{cases}$$

The sample version is:

$$SPD(\mathbf{x}; F_n) = 1 - \frac{1}{|\mathscr{X} \cup \mathbf{x}| - 1} \left\| \sum_{\mathbf{y} \in \mathscr{X}} S(\mathbf{y} - \mathbf{x}) \right\|_2 \ , \tag{5}$$

where $\mathscr{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ is the sample and $|\mathscr{X} \cup \mathbf{x}|$ denotes cardinality of the union $\mathscr{X} \cup \mathbf{x}$.

SPD for data with convex and non-convex support



Left: Data is simulated from $N_2((0,0)^T, I_2)$ (convex support).

Right: Data is simulated from $\frac{1}{2}N_2((2,2)^T, I_2) + \frac{1}{2}N_2((-2,-2)^T, I_2)$ (non-convex support).

HD and SPD for non-convex support

HD and SPD cannot adapt to non-convex supports.

This is because they are based on Euclidean scalar products i.e. linear projections.

Note that,

$$HD(\mathbf{x}; F_n) = \inf_{\mathbf{u} \in \mathbb{S}(0,1)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\langle \mathbf{u}, \mathbf{x_i} - \mathbf{x} \rangle \geq 0\right) \text{ and}$$

$$SPD(\mathbf{x}; F_n) = 1 - \frac{1}{|\mathscr{X} \cup \mathbf{x}| - 1} \left\| \sum_{\mathbf{y} \in \mathscr{X}} S(\mathbf{y} - \mathbf{x}) \right\|_2 \text{ where}$$

$$\left\| \sum_{\mathbf{y} \in \mathscr{X}} S(\mathbf{y} - \mathbf{x}) \right\|_2 = \Big(\sum_{\mathbf{z}, \mathbf{y} \in \mathscr{X}} \frac{\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{z} \rangle}{\delta_\kappa(\mathbf{x}, \mathbf{y}) \times \delta_\kappa(\mathbf{x}, \mathbf{z})} \Big)^{\frac{1}{2}} ,$$

$$\delta_\kappa(\mathbf{x}, \mathbf{y}) = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle - 2 \times \langle \mathbf{x}, \mathbf{y} \rangle} \ .$$

Kernel trick

Let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be a positive definite kernel ,i.e.,

$$\sum_{i,j=1}^{n} c_i c_j \kappa(\mathbf{x_i}, \mathbf{x_j}) \geq 0 \ \forall \ c_i \in \mathbb{R} \text{ and } \mathbf{x_i} \in \mathbb{R}^d \text{ for } 1 \leq i \leq n \ .$$

A positive definite kernel κ, implicitly defines an embedding map

$$\phi : \mathbf{x} \in \mathbb{R}^d \to \phi(\mathbf{x}) \in \mathbb{F}$$

via an inner product in the feature space $\mathbb{F}$. Thus, we have

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \ .$$

In HD and SPD, features appear as **inner products**. These inner products are replaced by kernels evaluated on features in the induced space.

Kernelized depths: Sphere depth (SD)

Kernelized version of HD is given by:

$$SD^r(\mathbf{z}, F) \stackrel{\text{def}}{=} \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \mathbb{P}(\kappa(\mathbf{c}, \mathbf{X}) \geq \kappa(\mathbf{c}, \mathbf{z})) \ , \tag{6}$$

where $\mathbf{X} \sim F$ and $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a positive definite kernel with $r > 0$.

After some modification, the sphere depth can be expressed as:

$$SD^r_s(\mathbf{z}, F) = \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \mathbb{E}\left[sig_s(r^2 - \|\mathbf{X} - c\|_2^2)\right] \ . \tag{7}$$

The sample version of sphere depth is given by:

$$SD^r_s(\mathbf{z}, F_n) = \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \frac{1}{n} \sum_{i=1}^{n} \left[sig_s(r^2 - \|\mathbf{x_i} - c\|_2^2)\right] \text{ where } sig_s : x \to \frac{1}{1 + e^{-x/s}} \ . \tag{8}$$

Kernelized Spatial depth (KSPD)

Kernelized spatial depth is the kernelized version of SD and is defined as:

$$KSPD(\mathbf{x}, F) = 1 - \|E[S(\phi(\mathbf{Y}) - \phi(\mathbf{x}))]\|_2 \ , \tag{9}$$

where $\phi : \mathbf{x} \in \mathbb{R}^d \to \phi(\mathbf{x}) \in \mathbb{F}$ (here, $\mathbb{F}$ is the induced feature space).

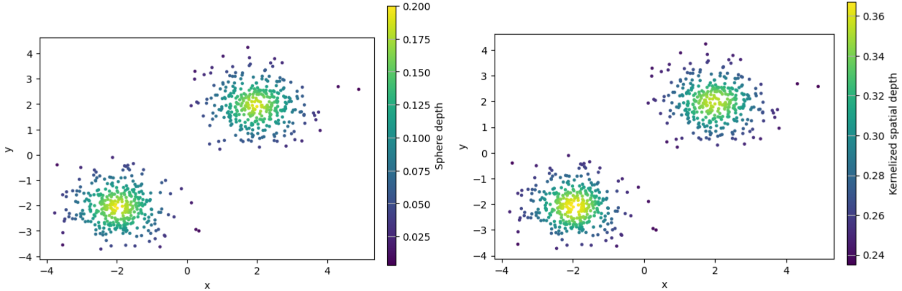
The sample version is given by:

$$KSPD_\kappa(\mathbf{x}, F_n) = 1 - \frac{1}{|\mathscr{X} \cup \mathbf{x}| - 1} \left(\sum_{\mathbf{z},\mathbf{y} \in \mathscr{X}} \frac{\kappa(\mathbf{x},\mathbf{x}) + \kappa(\mathbf{y},\mathbf{z}) - \kappa(\mathbf{x},\mathbf{y}) - \kappa(\mathbf{x},\mathbf{z})}{\delta_\kappa(\mathbf{x},\mathbf{y}) \times \delta_\kappa(\mathbf{x},\mathbf{z})} \right)^{\frac{1}{2}} \ , \tag{10}$$

where $\delta_\kappa(\mathbf{x}, \mathbf{y}) = \sqrt{\kappa(\mathbf{x}, \mathbf{x}) + \kappa(\mathbf{y}, \mathbf{y}) - 2 \times \kappa(\mathbf{x}, \mathbf{y})}$,

$\mathscr{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ and $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(- \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2} \right)$.

SD and KSPD for data with non-convex support



SD (left) and KSPD (right) are able to capture the bimodal structure.

Using the kernel trick, the data projections are non-linear and can be adapted for distributions with non-convex supports.

Classification problem

Problem of interest: A supervised classification problem with 2 competing classes.

Let $\mathbf{X_1}, \cdots, \mathbf{X_n} \stackrel{i.i.d}{\sim} F_0$ and $\mathbf{Y_1}, \cdots, \mathbf{Y_m} \stackrel{i.i.d}{\sim} F_1$.

Task: To construct a decision rule for classifying an unlabeled observation $\mathbf{X}$ to one of these 2 classes.

We used the idea of data depth for this purpose.

Transformed $d-$dimensional data into $2-$dimensional data ($\mathbb{R}^d \rightarrow [0, 1]^2$) using depth values calculated for each observation with respect to two class subsets.

Trained multiple classifiers on the transformed dataset.

Classifiers

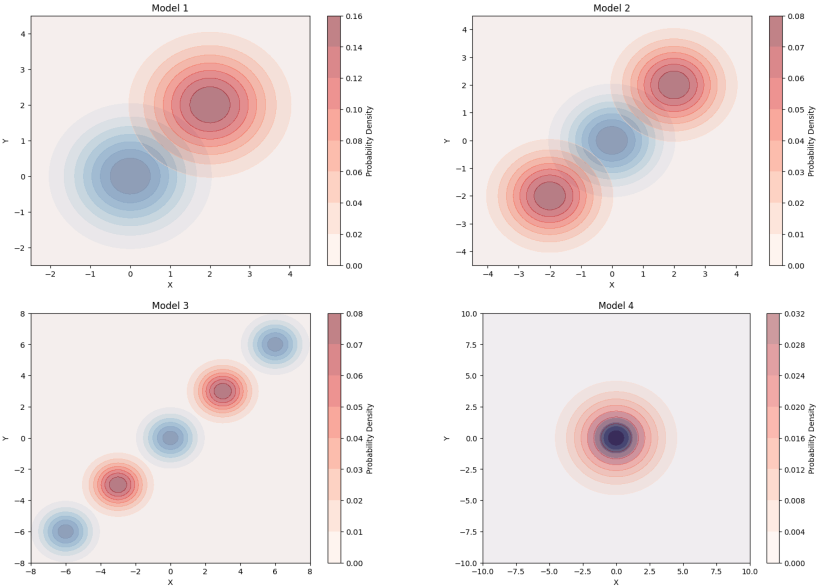
- Support vector machines with Radial Basis Function kernel (SVM-RBF).
- K-nearest neighbors (KNN).
- Generalized additive models (GAM) with logistic link. The model is given by

$$\log\left[\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}\right] = \alpha + f_1(X_1) + \cdots + f_p(X_p) \; ,$$

where $\mu(\mathbf{x}) = Pr(Y = 1|\mathbf{X} = \mathbf{x})$ and each of the functions f_j are unspecified smooth functions for $1 \leq j \leq p$.

Non parametric form of f_j ensures flexibility, while the additive nature ensures interpretability.

Simulation Models



Results-I

Model 1:

dim	Bayes	On $\mathbb{R}^d$	SD	KSPD
5	1.28 (0.59)	1.75 (0.77)	3.47 (0.85)	4.20 (2.70)
10	0.12 (0.13)	3.55 (2.68)	0.48 (0.58)	6.43 (0.12)

Table: 100 $\times$ Mean misclassification (standard error) using SVM-RBF.

Model 2:

dim	Bayes	On $\mathbb{R}^d$	SD	KSPD
5	2.00 (0.55)	3.05 (0.74)	10.80 (2.93)	12.47 (6.37)
10	0.10 (0.12)	7.80 (3.89)	7.97 (5.20)	6.00 (5.32)

Table: 100 $\times$ Mean misclassification (standard error) using SVM-RBF.

Results-II

Model 3:

dim	Bayes	On $\mathbb{R}^d$	SD	KSPD
5	0.10 (0.12)	0.83 (0.54)	4.43 (4.14)	5.60 (8.46)
10	0.00 (0.00)	0.08 (0.11)	13.93 (6.67)	8.38 (9.00)

Table: 100 $\times$ Mean misclassification (standard error) using GAM.

Model 4:

dim	Bayes	On $\mathbb{R}^d$	SD	KSPD
5	26.15 (2.00)	34.50 (3.26)	36.35 (3.25)	39.75 (4.38)
10	17.80 (1.63)	29.75 (1.82)	36.44 (3.40)	36.44 (3.40)

Table: 100 $\times$ Mean misclassification (standard error) using GAM.

Thank you!

References

Arturo Castellanos, Pavlo Mozharovskyi, Florence d'Alché Buc, and Hicham Janati.
Fast kernel half-space depth for data with non-convex supports.
2023.

Yixin Chen, Xin Dang, Hanxiang Peng, and Bart H. L.
Outlier detection with the kernelized spatial depth function.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, 2009.

Luc Devroye, László Györfi, and Gábor Lugosi.
A Probablistic Theory of Pattern Recognition.
01 1996.

Subhajit Dutta, Soham Sarkar, and Anil K. Ghosh.
Multi-scale classification using localized spatial depth.
Journal of Machine Learning Research, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning.
Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Robert Serfling and Yijun Zuo.
General notions of statistical depth function.
The Annals of Statistics, 28(2), 2000.

Appendix: Bayes classifier

The Bayes classifier classifies a test point ($\mathbf{x} \in \mathbb{R}^d$) based on the product of the likelihood and the prior probability. It has smallest probability of misclassification.

Let $C(\mathbf{x})$ be the Bayes classifier. Then

$$\begin{aligned} C(\mathbf{x}) &= \arg \max_{r \in \{0,1\}} \mathbb{P}(Y = r | \mathbf{X} = \mathbf{x}) \\ &= \arg \max_{r \in \{0,1\}} \underbrace{\mathbb{P}(\mathbf{X} = \mathbf{x} | Y = r)}_{\text{likelihood}} \cdot \underbrace{\mathbb{P}(Y = r)}_{\text{prior}} \end{aligned}$$

With simulated data the prior and likelihood are known hence the Bayes misclassification rate is used as benchmark for comparison.

SD-I

SD is defined as:

$$SD^r(\mathbf{z}, \mathbf{F}) \stackrel{\text{def}}{=} \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \mathbb{P}(\kappa(\mathbf{c}, \mathbf{X}) \geq \kappa(\mathbf{c}, \mathbf{z})) \text{ where, } \mathbf{X} \sim F \tag{1}$$

The kernelized depth of equation (1) has two main intuitive advantages:

1. data projections are non-linear and can be adapted for distributions with non-convex support,
2. additional parameter r provides a flexible lever to control the depth's sensitivity depending on the data.

SD-II

If κ is the Gaussian kernel $\kappa(\mathbf{x}, \mathbf{y}) \overset{def}{=} \exp(-\gamma\|\mathbf{x} - \mathbf{y}\|_2^2)$ with $\gamma > 0$, then $SD^r(\mathbf{z}, \mathbf{F})$ can be written as:

$$SD^r(\mathbf{z}, \mathbf{F}) = \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \mathbb{P}_{\mathbf{X}}(\mathbb{B}(\mathbf{c}, r)) = \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \mathbb{E}\left[\, \mathbb{I}\{r^2 - \|\mathbf{X} - c\|_2^2 \geq 0\}\right]$$

Justification: For any $\mathbf{x}$, $\mathbf{z} \in \mathbb{R}^d$ and $\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)$:

$\kappa(\mathbf{c}, \mathbf{x}) \geq \kappa(\mathbf{c}, \mathbf{z}) \iff e^{-\gamma\|\mathbf{x}-\mathbf{c}\|_2^2} \geq e^{-\gamma r^2} \iff \|\mathbf{x} - \mathbf{c}\|_2 \leq r$.

Therefore, $\mathbb{P}(\kappa(\mathbf{c}, \mathbf{X}) \geq \kappa(\mathbf{c}, \mathbf{z})) = \mathbb{P}_{\mathbf{X}}(\mathbb{B}(\mathbf{c}, r))$.

To make the depth function differentiable the indicator function is replaced by the sigmoid function and SD is given by:

$$SD_s^r(\mathbf{z}, F) = \inf_{\mathbf{c} \in \mathbb{S}(\mathbf{z}, r)} \mathbb{E}\left[sig_s(r^2 - \|\mathbf{X} - c\|_2^2)\right] .$$

Algorithm to compute SD

Algorithm 1 Riemannian gradient descent for computing SD

Input: $\mathbf{z}, \mathbf{X_{1:n}}, \mathrm{tol}, \alpha$
Result: $l \stackrel{\mathrm{def}}{=} SD_s^r(\mathbf{z}|\mathbf{X_{1:n}})$
Initialize $\mathbf{u} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}$
$\mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|_2}$
$l \leftarrow L(u)$
for $i = 1$ to n_iter **do**
    $v \leftarrow -\nabla_u \mathbf{L}(u)$
    $v \leftarrow v - \langle v, u \rangle u$
    $v \leftarrow \frac{v}{\|v\|_2}$
    $u \leftarrow \cos(\alpha)u + \sin(\alpha)v$
    $l' \leftarrow \mathbf{L}(u)$
    $\mathrm{dist} \leftarrow |l' - l|$
    if $l' > l$ **then**
        $\alpha \leftarrow \frac{\alpha}{2}$
    else if $\mathrm{dist} < \mathrm{tol}$ **then**
        break
    else
        $l \leftarrow l'$
    end if
end for
Return l

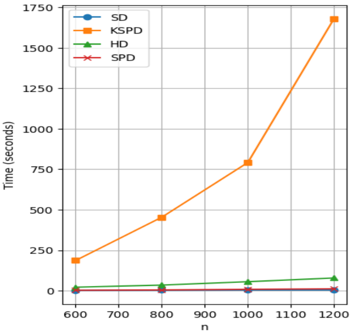
Algorithm to compute KSPD

Algorithm 2 Computation of KSPD

for every pair of $\mathbf{x_i}$ and $\mathbf{x_j}$ in $\mathbf{X}_{1:n}$ **do**
    $K_{ij} = \kappa(\mathbf{x_i}, \mathbf{x_j})$
end for
given input $\mathbf{x}$
for every observation $\mathbf{x_i}$ in $\mathbf{X}_{1:n}$ **do**
    $\zeta_i = \kappa(\mathbf{x}, \mathbf{x_i})$
    $\delta_i = \sqrt{\kappa(\mathbf{x}, \mathbf{x}) + K_{ii} - 2\zeta_i}$
    if $\delta_i = 0$ **then**
        $z_i = 0$
    else
        $z_i = \frac{1}{\delta_i}$
    end if
end for
for every pair of $\mathbf{x_i}$ and $\mathbf{x_j}$ in $\mathbf{X}_{1:n}$ **do**
    $\tilde{K_{ij}} = \kappa(\mathbf{x}, \mathbf{x}) + K_{ij} - \zeta_i - \zeta_j$
end for
$D_\kappa(\mathbf{x}, \mathbf{X}_{1:n}) = 1 - \frac{1}{|\mathcal{X} \cup \{\mathbf{x}\}| - 1} \sqrt{\mathbf{z}^T \tilde{K} \mathbf{z}}$

Computation time for SD and KSPD for varying n

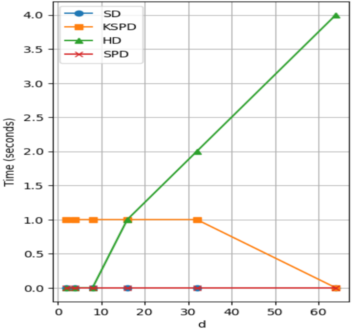
Computation time for varying sample sizes:



The time computation is performed on data simulated from $N_2((0,0)^T, I_2)$.

Computation time for SD and KSPD for varying d

Computation time for varying dimensions:



The time computation is performed on data simulated from $N_2(\mathbf{0}_d, I_d)$. with n = 200.