Statistical Depth Functions

In the multivariate and functional setting

Satvik Saha

Supervised by

Dr. Anirvan Chakraborty

Master's Thesis

Department of Mathematics and Statistics
Indian Institute of Science Education and Research, Kolkata

March 2024

ABSTRACT

Dedication

Declaration

Acknowledgments

Contents

Chapter 1

Introduction

1.1 Centrality vs Density

1.2 Nonparametric procedures

Chapter 2

Multivariate Data

2.1 Depth contours

The following definitions are adapted from Liu et al., 1999.

Definition 2.1.1. The contour of depth t is the set $\{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, F) = t\}$.

Definition 2.1.2. The region enclosed by the contour of depth t is the set

$$R_F(t) = \{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}, F) > t\}. \tag{2.1.1}$$

Definition 2.1.3. The p-th central region is the set

$$C_F(p) = \bigcap_t \{R_F(t) : P_F(R_F(t)) \geq p\}. \tag{2.1.2}$$

Definition 2.1.4. The p-th level contour, or center-outward contour surface, is the set $Q_F(p) = \partial C_F(p)$.

2.2 Monge-Kantorovich Depth

2.3 Depth-Depth plots

Definition 2.3.1 (DD plot)**.** Let F, G be two distributions on $\mathbb{R}^d$, and let D be a depth function. The Depth-Depth plot, also known as the DD plot, of F and G is given by

$$\mathrm{DD}(F, G) = \{(D(\boldsymbol{z}, F), D(\boldsymbol{z}, G)) : \boldsymbol{z} \in \mathbb{R}^d\}. \tag{2.3.1}$$

Remark. The above definition generalizes naturally to involve more than two distributions on $\mathbb{R}^d$.

When the depth function D only takes values in $[0,1]$, the DD plot is a subset of $[0,1]^2$ and hence easily visualized. Clearly when $F = G$, the corresponding DD plot is confined to the diagonal $\{(t,t) : t \in [0,1]\}$. However, when $d \geq 2$ and F, G are absolutely continuous, $\mathrm{DD}(F,G)$ has non-zero area (Lebesgue measure) when $F \neq G$. Assuming that D is affine invariant, Liu et al. (1999) propose this area as an affine invariant measure of the discrepancy between F and G.

If the distributions F, G are unknown, we may use data $\mathscr{D}_F = \{\boldsymbol{x}_i\}$ and $\mathscr{D}_G = \{\boldsymbol{y}_j\}$ where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} F$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \overset{\text{iid}}{\sim} G$, then construct empirical distributions $\hat{F}_n$ and $\hat{G}_m$. With this, we may examine the empirical DD plot

$$\mathrm{DD}(\hat{F}_n, \hat{G}_n) = \{(D(\boldsymbol{z}, \hat{F}_n), D(\boldsymbol{z}, \hat{G}_m)) : \boldsymbol{z} \in \mathscr{D}_F \cup \mathscr{D}_G\}. \tag{2.3.2}$$

DD plots can be used as a diagnostic tool to detect differences in location, scale, skewness, and kurtosis between two multivariate distributions (Liu et al., 1999).

1. If the same point $\boldsymbol{z}_0$ achieves maximum depths with respect to both distributions F and G, this indicates that $\boldsymbol{z}_0$ is their common center.
2. Suppose that F and G have the same center. If the points in $\mathrm{DD}(\hat{F}_n, \hat{G}_m)$ arch above the diagonal, i.e. the bulk of points are deeper in G than in F, this indicates that F has a greater spread than G.

2.4 Testing

2.5 Classification

The k-class classification task involves assigning an observation $\boldsymbol{x}$ to one of k populations, described by distributions F_i for $i \in \mathbb{N}_k$. The populations may also be associated with prior probabilities π_i.

Definition 2.5.1 (Classifier)**.** A classifier is a map $\hat{\imath} \colon \mathbb{R}^d \to \mathbb{N}_k$.

Example 2.5.2 (Bayes classifier)**.** Suppose that the population densities f_i for each $i \in \mathbb{N}_k$ are known. The Bayes classifier assigns $\boldsymbol{x}$ to the $\hat{\imath}_B$-th population where

$$\hat{\imath}_B(\boldsymbol{x}) = \underset{i \in \mathbb{N}_k}{\arg\max}\ \pi_i f_i(\boldsymbol{x}). \tag{2.5.1}$$

One way of measuring the performance of a classifier (given the population distributions and their priors) is by measuring its average misclassification rate.

Definition 2.5.3. The average misclassification rate of a classifier $\hat{\imath}$ is given by

$$\Delta(\hat{\imath}) = \sum_{i=1}^{k} \pi_i P(\hat{\imath}(\boldsymbol{X}) \neq i \mid \boldsymbol{X} \sim F_i). \tag{2.5.2}$$

Proposition 2.5.4. *The Bayes classifier has the lowest possible average misclassification rate. This is known as the optimal Bayes risk, denoted* Δ_B.

The simplest depth based classifier is the maximum depth classifier (Ghosh & Chaudhuri, 2005).

Example 2.5.5 (Maximum depth classifier)**.** Suppose that the prior probabilities π_i are equal. The maximum depth classifier $\hat{\imath}_D$ for a choice of depth function D is described by

$$\hat{\imath}_D(\boldsymbol{x}) \;=\; \underset{i\in\mathbb{N}_k}{\arg\max}\; D(\boldsymbol{x}, F_i). \tag{2.5.3}$$

In practice, instead of having direct access to the population distributions F_i, we have typically deal with labeled training data

$$\mathscr{D} \;=\; \{(\boldsymbol{x}_{ij}, i)\} \subset \mathbb{R}^d \times \mathbb{N}_k, \tag{2.5.4}$$

where $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i} \overset{\text{iid}}{\sim} F_i$ for each $i \in \mathbb{N}_k$. The empirical maximum depth classifier simply replaces the population distributions F_i with their empirical counterparts $\hat{F}_i$ determined by $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$. Thus, it is given by

$$\hat{\imath}_{D,\mathscr{D}}(\boldsymbol{x}) \;=\; \underset{i\in\mathbb{N}_k}{\arg\max}\; D(\boldsymbol{x}, \hat{F}_i). \tag{2.5.5}$$

Under certain restrictions, this classifier becomes asymptotically optimal in the following sense.

Theorem 2.5.6 (Ghosh and Chaudhuri, 2005)**.** *Suppose that the population density functions f_i are elliptically symmetric, with $f_i(\boldsymbol{x}) = g(\boldsymbol{x} - \boldsymbol{\mu}_i)$ for parameters $\boldsymbol{\mu}_i$ and a density function g such that $g(k\boldsymbol{x}) \le g(\boldsymbol{x})$ for every $\boldsymbol{x}$ and $k > 1$. Further suppose that the priors on the populations are equal, and the depth function D is one of HD, SD, MJD, PD. Then, $\Delta(\hat{\imath}_{D,\mathscr{D}}) \to \Delta_B$ as $\min\{n_1, \ldots, n_k\} \to \infty$.*

Note that this result deals with elliptic population densities differing only in location. Relax this assumption, and instead suppose that f_i are elliptic of the form

$$f_i(\boldsymbol{x}) = c_i|\Sigma|^{-1/2} h_i\left((\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right) \tag{2.5.6}$$

for strictly decreasing h_i, and that the depths can be expressed as $D(\cdot, F_i) = l_i(f_i(\cdot))$ for strictly increasing functions l_i. It follows that the Bayes decision rule can be reformulated as

$$\pi_i f_i(\boldsymbol{x}) > \pi_j f_j(\boldsymbol{x}) \iff D(\boldsymbol{x}, F_i) > r_{ij}(D(\boldsymbol{x}, F_j)) \tag{2.5.7}$$

for some real increasing function r_{ij}. Using this observation, the DD classifier (Li et al., 2012) picks separating functions r_{ij} which best classify the training data $\mathscr{D}$.

Definition 2.5.7. The empirical misclassification rate of a classifier $\hat{\imath}$, with respect to data $\mathscr{D}$, is given by

$$\Delta_{\mathscr{D}}(\hat{\imath}) \;=\; \sum_{i=1}^{k} \frac{\pi_i}{n_i} \sum_{j=1}^{n_i} \mathbf{1}(\hat{\imath}(\boldsymbol{x}_{ij}) \neq i). \tag{2.5.8}$$

Definition 2.5.8 (DD classifier)**.** Suppose that $k = 2$, that D is a depth function, and that $r\colon [0,1] \to [0,1]$ is an increasing function. The DD classifier $\hat{\imath}_{D,r}$ is given by

$$\hat{\imath}_{D,r}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } D(\boldsymbol{x}, F_2) \leq r(D(\boldsymbol{x}, F_1)), \\ 2, & \text{if } D(\boldsymbol{x}, F_2) > r(D(\boldsymbol{x}, F_1)). \end{cases} \tag{2.5.9}$$

The empirical DD classifier $\hat{\imath}_{D,\hat{r},\mathcal{D}}$ replaces F_i by their empirical counterparts $\hat{F}_i$. Here, the separating curve $\hat{r}$ is chosen from a family Γ so as to minimize the empirical misclassification rate, i.e.

$$\hat{r} = \arg\min_{r' \in \Gamma} \Delta_{\mathcal{D}}(\hat{\imath}_{D,r',\mathcal{D}}). \tag{2.5.10}$$

Remark. The maximum depth classifier $\hat{\imath}_D$ is simply the DD classifier $\hat{\imath}_{D,\mathrm{id}}$, where $\mathrm{id}(x) = x$.

Li et al. (2012) show that under certain restrictions, the empirical DD classifier is asymptotically equivalent to the Bayes rule. We give one such instance below.

Lemma 2.5.9. *Suppose that the following conditions hold.*

1. Γ *is the class of polynomial functions on* $[0,1]$.
2. *The depth functions* $D(\cdot, F_i)$ *are continuous.*
3. *As* $\min\{n_1, n_2\} \to \infty$, *we have for each* $i \in \mathbb{N}_2$,

$$\sup_{\boldsymbol{z} \in \mathbb{R}^d} |D(\boldsymbol{z}, \hat{F}_i) - D(\boldsymbol{z}, F_i)| \xrightarrow{a.s.} 0. \tag{2.5.11}$$

4. *For each* $i \in \mathbb{N}_2$, *the distributions* F_i *are elliptical and satisfy for all* $\delta \in \mathbb{R}$

$$P(D(\boldsymbol{Z}, F_i) = \delta \mid \boldsymbol{Z} \sim F_i) = 0. \tag{2.5.12}$$

Then, $\Delta(\hat{\imath}_{D,\hat{r},\mathcal{D}}) \to \Delta_B$ *as* $\min\{n_1, n_2\} \to \infty$.

In all the depth based classifiers we have seen so far, the classification rule depends on the observation $\boldsymbol{x}$ only through the depths $D(\boldsymbol{x}, F_i)$. Thus, we are motivated to define the following transformation from $\mathbb{R}^d$ to a depth feature space.

Definition 2.5.10. The depth feature vector $\boldsymbol{x}^D$ of an observation $\boldsymbol{x}$, with respect to the population distributions F_i and a choice of depth function D, is defined as

$$\boldsymbol{x}^D = (D(\boldsymbol{x}, F_1), \ldots, D(\boldsymbol{x}, F_k)). \tag{2.5.13}$$

Remark. The graph

$$\mathrm{DD}(F_1, \ldots, F_k) = \{\boldsymbol{x}^D : \boldsymbol{x} \in \mathbb{R}^d\} \tag{2.5.14}$$

is the analogue of the DD plot, with k distributions.

Assuming that the depth function D only takes values in $[0,1]$, the map $\boldsymbol{x} \mapsto \boldsymbol{x}^D$ takes values in $[0,1]^k$, regardless of the dimensionality of the original vector $\boldsymbol{x}$. With this, the maximum depth classification rule can be expressed as

$$\hat{\imath}_D(\boldsymbol{x}) = i \iff \boldsymbol{x}^D \in R_i^D = \{\boldsymbol{y} \in [0,1]^k : y_i = \max_j y_j\}. \tag{2.5.15}$$

Indeed, any partition of the unit cube $[0,1]^k$ into k decision regions R_i^D gives rise to a depth based classifier. The DD classifier achieves this by using an increasing separating function r to partition $[0,1]^2$. Furthermore, $r \in \Gamma$ is chosen so as to best separate the training data $\mathscr{D}$ transformed into the depth feature space. However, we can in principle use the transformed training data

$$\mathscr{D}^D = \{(\boldsymbol{x}_{ij}^D, i)\} \subset [0,1]^k \times \mathbb{N}_k \tag{2.5.16}$$

along with any multivariate classification algorithm (LDA, QDA, kNN, GLM, etc) to devise suitable decision regions. This is the basis of the DD^G classifier (Cuesta-Albertos et al., 2017).

2.6 Clustering

2.7 Outlier detection

Chapter 3

Functional Data

3.1 Classification

3.2 Clustering

3.3 Outlier detection

3.4 Partially Observed Functional Data

Chapter 4

Local Depth Functions

4.1 Regression using Local Depth Regions

Chapter 5

Conclusion

Bibliography

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

Chakraborty, A., & Chaudhuri, P. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, *42*(3), 1203–1231.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, *45*(1), 223–256.

Cuesta-Albertos, J. A., Febrero-Bande, M., & Oviedo de la Fuente, M. (2017). The DD^G-classifier in the functional setting. *TEST*, *26*(1), 119–142.

Cuesta-Albertos, J. A., & Nieto-Reyes, A. (2008). The Tukey and the random Tukey depths characterize discrete distributions. *Journal of Multivariate Analysis*, *99*(10), 2304–2311.

Dai, W., & Genton, M. G. (2018). An outlyingness matrix for multivariate functional data classification. *Statistica Sinica*, *28*(4), 2435–2454.

Elías, A., Jiménez, R., Paganoni, A. M., & Sangalli, L. M. (2023). Integrated depths for partially observed functional data. *Journal of Computational and Graphical Statistics*, *32*(2), 341–352.

Elías, A., Jiménez, R., & Shang, H. L. (2022). On projection methods for functional time series forecasting. *Journal of Multivariate Analysis*, *189*, 104890.

Elías, A., Jiménez, R., & Shang, H. L. (2023). Depth-based reconstruction method for incomplete functional data. *Computational Statistics*, *38*(3), 1507–1535.

Fraiman, R., Liu, R. Y., & Meloche, J. (1997). Multivariate Density Estimation by Probing Depth. *Lecture Notes-Monograph Series*, *31*, 415–430.

Ghosh, A. K., & Chaudhuri, P. (2005). On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, *32*(2), 327–350.

Gijbels, I., & Nagy, S. (2017). On a General Definition of Depth for Functional Data. *Statistical Science*, *32*(4), 630–639.

Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth [Special Issue on Multivariate Methods in Genomic Data Analysis]. *Journal of Multivariate Analysis*, *90*(1), 67–89.

Kneip, A., & Liebl, D. (2020). On the optimal reconstruction of partially observed functional data. *The Annals of Statistics*, *28*(3), 1692–1717.

Kong, L., & Zuo, Y. (2010). Smooth depth contours characterize the underlying distribution. *Journal of Multivariate Analysis*, *101*(9), 2222–2226.

Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *77*(4), 777–801.

Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-Classifier: Nonparametric Classification Procedure Based on DD-Plot. *Journal of the American Statistical Association*, *107*(498), 737–753.

Liu, R. Y. (1990). On a Notion of Data Depth Based on Random Simplices. *The Annals of Statistics*, *18*(1), 405–414.

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), 783–858.

Liu, R. Y., & Singh, K. (1993). A Quality Index Based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association*, *88*, 252–260.

López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, *104*(486), 718–734.

Mosler, K., & Mozharovskyi, P. (2022). Choosing Among Notions of Multivariate Depth Statistics. *Statistical Science*, *37*(3), 348–368.

Nagy, S. (2017). Monotonicity properties of spatial depth. *Statistics & Probability Letters*, *129*, 373–378.

Nagy, S. (2021). Halfspace depth does not characterize probability distributions. *Statistical Papers*, *62*(3), 1135–1139.

Nieto-Reyes, A., & Battey, H. (2016). A Topologically Valid Definition of Depth for Functional Data. *Statistical Science*, *31*(1), 61–79.

Shi, X., Zhang, Y., & Fu, Y. (2023). Two-sample tests based on data depth. *Entropy*, *25*(2).

Villani, C. (2003). *Topics in optimal transportation*. American Mathematical Society.

Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, *28*(2), 461–482.