Computing Projection Depth and Related Estimators

Yijun Zuo

Michigan State University

May, 2003

0 Outline

- Why data depth

- Projection depth

- Computing issues

- Open problems

Motivation, Order statistics

Motivation

Tukey halfspace depth

Liu simplicial depth

I Why data depth?

- Ordering in $\mathbf{R}^d$ ($d > 1$)

- Order related procedures in $\mathbf{R}^d$

- Other applications

Monotone continuous depth

II Projection Depth

- Outlyingness

$$\mathbf{R}^1: \quad O_1(x, X) = \frac{|x - \mu(X)|}{\sigma(X)}$$

$$\mathbf{R}^d: \quad O_d(x, X) = \sup_{\|u\|=1} O_1(u'x, u'X)$$

[Stahel (1981), Donoho (1982)]

- Projection depth

$$PD(x, X) = \frac{1}{1 + O_d(x, X)}$$

[Liu (1992)]
[Zuo and Serfling (2000abc), Zuo (2000a)]

Bivariate normal projection depth

Uniform over square projection depth

Uniform over triangle projection depth

III Computing Issues

- Outlyingness

- Projection based estimators

Outlyingness

$$O_1(x, X) = \frac{|x - \mathrm{Med}(X)|}{\mathrm{Mad}(X)}$$

$$O_d(x, X) = \sup_{\|u\|=1} O_1(u'x, u'X)$$

$$X = \{X_1, \cdots, X_n\}, \quad X_{(1)} \le \cdots \le X_{(n)}$$

$$\mathrm{Med}(X) = \frac{X_{(\lfloor (n+1)/2 \rfloor)} + X_{(\lfloor (n+2)/2 \rfloor)}}{2}$$

$$\mathrm{Mad}(X) = \mathrm{Med}\{|X_i - \mathrm{Med}(X)|\}$$

$$u'X = \{u'X_1, \cdots, u'X_n\}$$

- Approximate algorithms

  Fix-direction procedure

  Sub-sampling procedure
  [Stahel (1981)]

  Pigeon hole procedure
  [Rousseeuw (1993)]

  Random-direction procedure

- Criticisms

- An exact algorithm ($\mathbf{R}^2$, n odd)

Med sequence (n directions)

Divide $\mathbf{R}^2$ into n angular regions such that within each region the Med of the projected data is the projection of a fixed point to this region

Mad sequence (n directions)

Divide each angular region into n sub-angular regions such that within each of them the Mad of the projected data is the projection of a fixed line segment to this sub-region

Pictures of Med sequence

Pictures of Mad sequence

- An example: Perspiration Data

Individual	X_1 (Sweat rate)	X_2 (Sodium)
1	3.7	48.5
2	5.7	65.1
3	3.8	47.2
4	3.2	53.2
5	3.1	55.5
6	4.6	36.1
7	2.4	24.8
8	7.2	33.1
9	6.7	47.4
10	5.4	54.1
11	3.9	36.9
12	4.5	58.8
13	3.5	27.8
14	4.5	40.2
15	1.5	13.5
16	8.5	56.4
17	4.5	40.2
18	6.5	52.8
19	4.1	44.1

[Johnson and Wichern (2002)]

Scatter plot

PD of perspiration data

Exact	Subsample	Fixed, $10^5 u$
.375349097	.375349097	.375363462
.305747126	.309360731	.305748693
.414516295	.414516295	.414530873
.270949533	.270949533	.270962308
.245614035	.245614035	.24562602
.393545029	.393545029	.393591138
.262133297	.269408451	.262134136
.154569618	.154569618	.154603311
.234636872	.234636872	.234686511
.413690236	.443960827	.413798722
.449288256	.449288256	.449288345
.30121022	.30121022	.301221177
.303121248	.303121248	.303121351
.508064516	.508064516	.50811871
.191923191	.198019081	.191923847
.164594729	.164594729	.164600864
.201177527	.201177527	.201186314
.276771606	.276771606	.276773648
.568047337	.568047337	.568060085

Comparison: PD(approximate)-PD(exact)

fixed 171	fixed 342	fixed 1400	fixed 10^5	SubS 171	EX 342
0185	0140	0013	0000	——	7
0036*	0019*	0000	0000	0036	8
0194*	0142*	0013	0000	——*	4
0148*	0121	0011	0000	——	12
0139	0113	0011	0000	——	14
0221	0044	0008	0000	——	6
0004	0006	0000	0000	0073	13
0057	0027	0006	0000	——	19
0076	0078	0010	0000	——	15
0382*	0251*	0002	0001	0303*	5
0002*	0001	0000	0000	——	3
0168*	0108*	0010	0000	——	10
0002	0001*	0000	0000	——	9
0280	0055	0010	0000	——	2
0003	0004	0000	0000	0061	17
0088	0028	0004	0000	——	18
0129	0071	0008	0000	——	16
0079*	0083	0001	0000	——	11
0219	0123	0011	0000	——	1

Depth plot

Projection depth plot of the data

- Worst case time complexity

  Fixed N directions: $O(Nn)$

  Subsampling: $O(n^3)$

  Exact: $O(n^3)$

  [for all sample points or any one point]

Subsampling in $\mathbf{R}^d$: $O(n^{d+1})$

Exact in $\mathbf{R}^d$: $O((\binom{2(d-1)}{d-1}/d)^2 n^3)$

Slide listing exact algo wctc for d's

G: exact (Hawkins), space shuttle

G: computer scientists, faster exact

B: still relatively slow to me

Projection based estimators

- Stahel-Donoho estimator

$$L(X) = \frac{\sum_i W(O(X_i, X))\ X_i}{\sum_i W(O(X_i, X))}$$

- Projection median

$$PM(X) = \arg \inf_{x \in \mathbf{R}^d} O(x, X)$$

[Tyler (1994)]

[Zuo and Serfling (2000ab), Zuo (2003)]

- Computing

  SD: $O(X_i, X)$

  PM: $O(X_i, X)$, $O(x, X)$
  downhill simplex algor.

- Advantages

  Affine equivariance

  High breakdown point

Finite sample breakdown point

Minimum fraction of 'bad points' in data that can render the estimator useless

- BP($\bar{X}_n, X$) $= 1/n$

- BP(L_{SD}, X) $= \lfloor (n - 2d + 2)/2 \rfloor / n$

  [Donoho (82), Davies (87), Zuo (01)]

- BP(L^*_{SD}, X) $= \lfloor (n - d + 1)/2 \rfloor / n$

  [Tyler(94), Gather and Hilker (97), Zuo (00)]

- BP(PM, X) $= \lfloor (n - d + 2)/2 \rfloor / n$

  [Zuo (03a)]

Open problem related to BP

Under affine equivariance, how high can BP of a location estimator be?

- Answer: $\lfloor (n+1)/2 \rfloor / n$

  Zuo (2003b)

- Exact Computing: n projections

  Zuo (2003b)

Exact in $\mathbf{R}^d$: $O((\binom{2(d-1)}{d-1}/d)n^2)$

Slide listing exact algo wctc for d's

G: exact (Hawkins), space shuttle

G: computer scientists, faster exact

IV Open Problems

- Faster exact algorithms

- Good approximate algorithms