A Convex Hull Peeling Depth Approach to Nonparametric Massive Multivariate Data Analysis with Applications

Hyunsook Lee

hlee@stat.psu.edu

Department of Statistics
The Pennsylvania State University
Outlines

- Convex Hull Peeling (CHP) and Multivariate Data Analysis
  - Definitions on CHP
  - Data Depth (Ordering Multivariate Data)
  - Quantiles and Density Estimation
- Color Magnitude (CM) Diagram and Sloan Digital Sky Survey
- Nonparametric Descriptive Statistics with CHP
  - Multivariate Median
  - Skewness and Kurtosis of a Multivariate Distribution
- Outlier Detection with CHP
  - Level $\alpha$; Shape Distortion; Balloon Plot
- Concluding Remarks
Definitions

Convex Set  A set $C \subseteq \mathbb{R}^d$ is convex if for every two points $x, y \in C$ the whole segment $xy$ is also contained in $C$.

Convex Hull  The convex hull of a set of points $X$ in $\mathbb{R}^d$ is denoted by $CH(X)$, is the intersection of all convex sets in $\mathbb{R}^d$ containing $X$. In algorithms, a convex hull indicates points of a shape invariant minimal subset of $CH(X)$ (vertices, extreme points), connecting these points produces a wrap of $CH(X)$. 

![Diagram showing a set of points and its convex hull]
Convex Hull Peeling

Before

After
Convex Hull Peeling Depth (CHPD)

[CHPD:] For a point \( x \in \mathbb{R}^d \) and the data set \( X = \{X_1, \ldots, X_{N-1}\} \), let \( C_1 = CH\{x, X\} \) and denote a set of its vertices \( V_1 \). We can get \( C_j = C_{j-1} \setminus V_{j-1} \) through CHP until \( x \in V_j \) (\( j = 2, \ldots \)). Then,
\[
CHPD(x) = \frac{\#(\bigcup_{i=1}^{k} V_i)}{N} \text{ for } k \text{ s.t. } k = \min_j \{ j : x \in V_j \} \ ; \text{otherwise CHPD is 0.}
\]

▶ Barnett (1976): Ordering based on Depth
▶ \( \hat{p}^{th} \) quantiles are \( 1 - \hat{p}^{th} \) CHPDs.
▶ Hyper-polygons of \( 1 - \hat{p}^{th} \) depth obtainable from any dimensional data.
▶ QHULL (Barber et. al., 1996) works for general dimensions (http://qhull.org).
▶ Why CHPD...
Challenges in Nonparametric Multivariate Analysis

How to Order Multivariate Data?
How to Order Multivariate Data?

**Ordering Multivariate Data → Data Depth**

- Mahalanobis Depth : Mahalanobis (1936)
- Convex Hull Peeling Depth: Barnett (1976)
- Half Space Depth: Tukey (1975)
- Simplical Depth : Liu (1990)
- Oja Depth : Oja (1983)
- Majority Depth : Singh (1991)

- Ordering is not uniformly defined
(P1) (Affine invariance) \( D(Ax + b; F_{AX+b}) = D(x; F_X) \) for all \( X \) (A nonsingular matrix) holds for any random vector \( X \) in \( R^d \), any \( d \times d \) nonsingular matrix \( A \), and any \( d \)-vector \( b \);

(P2) (Maximality at center) \( D(\theta; F) = \sup_{x \in R^d} D(x; F) \) holds for any \( F \in F \) having center \( \theta \);

(P3) (Monotonicity) for any \( F \in F \) having deepest point \( \theta \), \( D(x; F) \leq D(\theta + \alpha(x - \theta); F) \) holds for \( \alpha \in [0, 1] \); and

(P4) \( D(x; F) \to 0 \) as \( ||x|| \to \infty \), for each \( F \in F \).
Convex Hull Peeling Depth

- affine invariance
- maximality at center
- monotonicity relative to deepest point
- vanishing at infinity

CHPD has these properties and points of smallest depth are possible outliers
Quantile Estimation

- Median: A point(s) left after peeling
  (will show robustness of this estimator later)

- $p^{th}$ Quantile: Level set whose central region contains $\sim 100p\%$ data
  (will define the level set and the central region later)

- No Closed Form; Empirical Process
Empirical Density Estimation

Density Estimation with CHPD on Bivariate Normal Data (McDermott, 2003)

100000 Bivariate Normal Sample

Quantiles={0.99, 0.95, 0.90, 0.80, ..., 0.20, 0.10, 0.05, 0.01}
Lessons and Further Studies

- Sample from a convex distribution (no doughnut shape)
- Works on Massive data
  → Sequential Method
- Without previous knowledge, no model or prior is known to start an analysis. Exploratory data analysis for a large database
- Nonparametric and non-distance based approach
- Where CHP can be applied and how?
  → Multi-color diagram from astronomy, where a plethora of free data archives is available.
Color Magnitude diagram

Two dimensional Color-Color diagram or Celebrated Hertzsprung-Russell diagram (switch)
Two dimensional Color-Color diagram or Celebrated Hertzsprung-Russell diagram (switch)

What if we can see beyond 2 dimensions without bias (projection) Then, 3 or higher dimensional color diagrams might have popularity.
Color Magnitude diagram

Two dimensional Color-Color diagram or Celebrated Hertzsprung-Russell diagram (switch)

What if we can see beyond 2 dimensions without bias (projection) Then, 3 or higher dimensional color diagrams might have popularity.

CHP may assist analyzing multi-color diagrams. Need a suitable data set with colors.
Sloan Digital Sky Survey: SDSS

Commissioned 2000, now Data Release 5 is available.
5 bands; 4 variables (u-g, g-r, r-i, i-z)

- Studies on analyzing astronomical massive data received spotlights recently. http://www.sdss.org

- July, 2005: Data Release Four
  6670 square degrees, 180 million objects
  Available from http://www.sdss.org/dr4
  From SpecPhotoAll with SQL:

- Attributes of photometric data are color indices, $u,b,g,i,z$ along with coordinates.
SQL for SDSS

```sql
select ra, dec, z, psfMag_u, psfMag_g, psfMag_r, psfMag_i, psfMag_z
from SpecPhotoAll
where specclass = 2
```

- Note — 2: galaxies, 3: QSO, 4: HighZ QSO
- Galaxies: 499043
- Quasars: 70204
Multivariate Descriptive Statistics

- CHP Median
- CHP Skewness
- CHP Kurtosis

with bivariate simulated data and SDSS DR4
**Convex Hull Peeling Median (CHPM)**

**Multivariate Median:** the inner most point among data
   → Survey of Multivariate Median (Small, 1990)

**CHPM:** recursive peeling leads to the inner most point(s). The average of these largest depth points is the median of a data set.
Convex Hull Peeling Median (CHPM)

Multivariate Median: the inner most point among data
→ Survey of Multivariate Median (Small, 1990)

CHPM: recursive peeling leads to the inner most point(s). The average of these largest depth points is the median of a data set.

Simulations: Sample from standard bivariate normal distribution

<table>
<thead>
<tr>
<th>n</th>
<th>mean</th>
<th>median</th>
<th>CHPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^4$</td>
<td>(0.001338, -0.02232)</td>
<td>(-0.005305, -0.01643)</td>
<td>(0.000918, -0.010589)</td>
</tr>
<tr>
<td>$10^6$</td>
<td>(0.000072, 0.000114)</td>
<td>(0.001185, -0.000717)</td>
<td>(0.002455, -0.000456)</td>
</tr>
</tbody>
</table>

Sequential CHPM → (0.004741, -0.004111)

Setting for the sequential method: $m=10000$ and $d=0.05$
## Application: Median

<table>
<thead>
<tr>
<th></th>
<th>Quasars</th>
<th>Galaxies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>u-g</td>
<td>g-r</td>
</tr>
<tr>
<td>Mean</td>
<td>0.4619</td>
<td>0.2484</td>
</tr>
<tr>
<td>Median</td>
<td>0.2520</td>
<td>0.1750</td>
</tr>
<tr>
<td>CHPM</td>
<td>0.2530</td>
<td>0.1640</td>
</tr>
<tr>
<td>Mean</td>
<td>1.622</td>
<td>0.9211</td>
</tr>
<tr>
<td>Median</td>
<td>1.680</td>
<td>0.8930</td>
</tr>
<tr>
<td>CHPM</td>
<td>1.790</td>
<td>0.957</td>
</tr>
<tr>
<td>Seq. CHPM</td>
<td>1.772</td>
<td>0.950</td>
</tr>
</tbody>
</table>
Robustness of Convex Hull Peeling Median

Breakdown point of a convex hull peeling median goes to zero as $n \to \infty$ (Donoho, 1982). Outliers are necessarily located at infinity.
Robustness of Convex Hull Peeling Median

Breakdown point of a convex hull peeling median goes to zero as $n \to \infty$ (Donoho, 1982). Outliers are necessarily located at infinity.

Empirical mean square error (EMSE) and Relative Efficiency (RE):

**Model:** $(1 - \epsilon)N((0, 0), \mathbf{I}) + \epsilon N(\cdot, 4\mathbf{I})$

$n = 5000$, $m = 500$, $T_j = \text{(CHPM, Mean)}$

$$EMSE = \frac{1}{m} \sum_{i=1}^{m} ||T_j - \mu||^2$$

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>CHPM</th>
<th>Mean</th>
<th>RE</th>
<th>CHPM</th>
<th>Mean</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.002178</td>
<td>0.000417</td>
<td>0.191689</td>
<td>0.002178</td>
<td>0.000417</td>
<td>0.191689</td>
</tr>
<tr>
<td>0.005</td>
<td>0.0028521</td>
<td>0.001682</td>
<td>0.589961</td>
<td>0.002891</td>
<td>0.005444</td>
<td>1.88291</td>
</tr>
<tr>
<td>0.05</td>
<td>0.016842</td>
<td>0.125522</td>
<td>7.45262</td>
<td>0.017824</td>
<td>0.500610</td>
<td>28.08597</td>
</tr>
<tr>
<td>0.2</td>
<td>0.139215</td>
<td>2.00109</td>
<td>14.37612</td>
<td>0.1435910</td>
<td>8.0017</td>
<td>55.7264</td>
</tr>
</tbody>
</table>
Generalized Quantile Process

EinMahl and Mason (1992)

\[ U_n(t) = \inf \{ \lambda(A) : P_n(A) \geq t, A \in \mathbb{A} \}, 0 < t < 1. \]

- Central Region:
  \[ R_{CH}(t) = \{ x \in \mathbb{R}^d : CHPD(x) \geq t \} \]

- Level Set:
  \[
  B_{CH}(t) = \partial R_{CH}(t) \\
  = \{ x \in \mathbb{R}^d : CHPD(x) = t \}
  \]

- Volume Functional:
  \[ V_{CH}(t) = Volume(R_{CH}(t)) \]

→ One dimensional mapping.
not equi-probability contours, assume smooth convex distributions
Skewness Measure

Let \( x_{j,i} \) be the \( i^{th} \) vertex in a level set \( B_{CH,j} \) comprised by the \( j^{th} \) peeling process. A measure of skewness:

\[
R_j = \frac{\max_i ||x_{j,i} - HPM|| - \min_i ||x_{j,i} - HPM||}{\min_i ||x_{j,i} - HPM||}
\]

Not only a sequence of \( R_j \) visualizes but also quantizes the skewness along depths.

**Denominator for the regularization → affine invariant** \( R_j \)

**symmetric:** flat \( R_j \) along convex hull peels

**skewed:** fluctuating \( R_j \)
Simulation: Skewness Measure

N(0, I)

N(0, Σ)

non normal

\[ \chi^2_{10} \]

\[ R_j \]

\[ j^{\text{th}} \text{convex hull level set} \]

Hyunsook Lee, Department of Statistics, Penn State Univ – p. 22
Application: Skewness Measure (Quasars)
Application: Skewness Measure (Galaxies)
Kurtosis Measure

Quantile (Depth) based Kurtosis:

\[ K_{CH}(r) = \frac{V_{CH}(\frac{1}{2} - \frac{r}{2}) + V_{CH}(\frac{1}{2} + \frac{r}{2}) - 2V_{CH}(\frac{1}{2})}{V_{CH}(\frac{1}{2} - \frac{r}{2}) - V_{CH}(\frac{1}{2} + \frac{r}{2})} \]

Tailweight:

\[ t(r, s) = \frac{V_{CH}(r)}{V_{CH}(s)} \]

for \( 0 < s < r \leq 1 \). Here,
\( V_{CH}(r) \) indicates the volume functional at depth \( r \).
Simulation: Kurtosis Measure (Tailweight)
Application: Kurtosis Measure (Quasars)
Multivariate Outlier Detection

- What are Outliers?
- Detecting Algorithms
  - Level $\alpha$
  - Shape Distortion
  - Balloon Plot
What are Outliers?

Outliers are...

- Cumbersome Observations
- Lead to New Scientific Discoveries
- Improve Models (Robust Statistics)
- ...
- No Clear Objectives but Come Along Often

CHP: Experience and relative Robustness support the Idea of Outlier Detection.

⇒ We need a clear definition on outliers; especially, outliers of the 21st century. And outlier detecting methods.
Outliers are observations....

- Huber (1972): unlikely to belong to the main population.
- Beckman and Cook (1983): surprising and discrepant to the investigator.

Discordant Observations or Contaminants

- Rohlf (1975): somewhat isolated from the main cloud of points.

Yet, somewhat VAGUE!
Some Outlier Detection Methods

**Univariate**: Box-and-Whisker plot, Order statistics, ...
Some Outlier Detection Methods

**Univariate:** Box-and-Whisker plot, Order statistics, ...

**Multivariate:** Mostly bivariate applications
  - Generalized Gap Test (Rolhf, 1975)
  - Bivariate Box Plot (Zani et. al, 1999)
  - Sunburst Plot (Liu et. al., 1999)
  - Bag plot (Miller et. al., 2003)

and **Mahalanobis distance** \( D(x) = (x - \hat{\mu})\hat{\Sigma}^{-1}(x - \hat{\mu}). \)
Some Outlier Detection Methods

**Univariate:** Box-and-Whisker plot, Order statistics, ...

**Multivariate:** Mostly bivariate applications
  ▶ Generalized Gap Test (Rolhf, 1975)
  ▶ Bivariate Box Plot (Zani et. al, 1999)
  ▶ Sunburst Plot (Liu et. al., 1999)
  ▶ Bag plot (Miller et. al., 2003)

and **Mahalanobis distance**

\[ D(x) = (x - \hat{\mu})\hat{\Sigma}^{-1}(x - \hat{\mu}). \]

*Difficulties of multivariate analysis arise from the complexity of ordering multivariate data.*
Quantile Based Outlier Detection

bivariate standard normal

bivariate $t_5$ with $\rho = -0.5$
Contour Shape Changes

- Bivariate Normal Sample
- Outliers Added

- Graphs showing contour shape changes
- Graphs with and without outliers
- Plots of volume against the sum of convex hull levels
- GAP

Hyunsook Lee, Department of Statistics, Penn State Univ – p. 33
A Balloon Plot is obtained by blowing $0.5^{th}$ CHPD polyhedron by 1.5 times (lengthwise). Let $V_{0.5}$ be a set of vertices of $0.5^{th}$ CHPD hull. The balloon for outlier detection is

$$B_{1.5} = \{y_i : y_i = x_i + 1.5(x_i - CHPM), x_i \in V_{0.5}\}.$$ 

In other words, blow the balloon of IQR 1.5 times larger.
Outliers in Quasar Population

Volumes of 1st CH, .01 Depth CH, .05 Depth CH: (474.134, 14.442, 4.353)
Outliers in Galaxy Population

Volumes of 1st CH, .01 Depth CH, .05 Depth CH: (4919.492, 4.310, 1.075)
Discussion on CHP

Convex Hull Peeling is:

- a robust location estimator.
- a tool for descriptive statistics.
  (Skewness and Kurtosis measure.)
- a reasonable approach for detecting multivariate outliers.
- a starter for clustering.

⇒Our methods help to characterize multivariate distributions and identify outlier candidates from multivariate massive data; therefore, the results initiate scientists to study further with less bias.

CHP as Exploratory Data Analysis and Data Mining Tools.
Concluding Remarks

Drawbacks of CHPD

- Limited to moderate dimension data.
- CHPD estimates depths inward not outward.
- Convexity of a data set.
- No population/theoretical counterpart.
Concluding Remarks

Drawbacks of CHPD

- Limited to moderate dimension data.
- CHPD estimates depths inward not outward.
- Convexity of a data set.
- No population/theoretical counterpart.

No assumption on data distribution, Non-distance based, Affine invariant, Applicable to streaming data, Detecting Outliers, Providing Multivariate Descriptive Statistics, Exploratory data analysis.