Nonparametric Tolerance Regions Based on Multivariate Spacings

Jun Li
(jun.li@ucr.edu)
Department of Statistics
University of California, Riverside

Joint work with Prof. Regina Y. Liu (Rutgers University)
Outline

• Data depth and its induced center-outward ordering of multivariate data
• Multivariate spacings based on data depth
• Multivariate tolerance regions based on multivariate spacings
• Simulation studies
• Concluding remarks
Background Material on Data Depth

Data Depth:

A measure of the “depth” or “centrality” of a given point w.r.t. a multivariate data cloud or its underlying distribution

Given sample: \(X = \{X_1, \cdots, X_n\} \rightarrow F;\)

• **Simplicial Depth** (Liu 1990)

\[D_{F_n}(x) = \frac{1}{\binom{n}{3}} \sum_{\ast} I(x \in \Delta(X_i, X_j, X_k))\]

* Larger \(D_{F_n}(x) \leftrightarrow \) deeper (or more central)
* Smaller \(D_{F_n}(x) \leftrightarrow \) more outlying

\[D_F(x) = P_F(x \in \Delta(X_1, X_2, X_3))\]
\( x \in \mathbb{R}^d \)

\[
D_{F_n}(x) = \frac{1}{n} \binom{n}{d+1} \sum_{*} I(x \in S[X_{i_1}, \ldots, X_{i_{d+1}}])
\]

\[
D_{F}(x) = P_{F}(x \in S[X_1, \ldots, X_{d+1}])
\]
Given sample: \( X = \{X_1, \ldots, X_n\} \rightarrow F; \)

Compute \( D_n(X_1), D_n(X_2), \ldots, D_n(X_n) \)

\[ \Rightarrow D_n(X_{[1]}) \geq D_n(X_{[2]}) \geq \cdots \geq D_n(X_{[n]}) \]

\[ \Rightarrow X_{[1]}, X_{[2]}, \ldots, X_{[n]} \]

the deepest (the most central) point

**Data Depth**

**Center Outward Ordering of Multivariate Data**

Order statistics w/ **center-outward** ordering (ranking)

* smaller rank \( \iff \) deeper (or more central)

* higher rank \( \iff \) more outlying
Figure: 500 sample points from the standard bivariate normal
“+” is the deepest point
Figure: 500 sample points from the bivariate exponential
"+" is the deepest point
1) **Half-space depth** (Tukey (1975))

$$D_{F_n}(x) = \min \left( \# \{i : u^T X_i \geq u^T x\}, \# \{i : u^T X_i \leq u^T x\} \right) / n$$

$$= \min \left( \# \{i : X_i \in H, \text{ and } x \in H\} \right) / n$$

$$D_F(x) = \inf_{H \in P_F(H)} \{x \in H\} \quad H \text{ is a closed halfspace in } \mathbb{R}^d$$

$$= \min_{|u|=1} \left\{ P_F(u^T X \geq u^T x), P_F(u^T X \leq u^T x) \right\}$$
2) **Mahalanobis Depth** (Mahalanobis (1936))

\[
D_F(x) = [1 + (x - \mu)' \Sigma^{-1} (x - \mu)]^{-1}
\]

\[
D_{F_n}(x) = [1 + (x - \bar{X})' S^{-1} (x - \bar{X})]^{-1}
\]

3) **Projection Depth** (Stahel (1981), Donoho (1982))

\[
D_F(x) = \left[ 1 + \sup_{||u||=1} \frac{u'x - \text{Med}(u'X)}{\text{MAD}(u'X)} \right]^{-1}
\]

(See Liu, Parelius and Singh (1999), Zuo and Serfling (2000))
Data depth and its induced center outward ordering of multivariate data

• Usefulness:

  - *Characterize distributions*: descriptive statistics
    (location “center”, scale, skewness, kurtosis, depth contours, quantiles, …) (Liu, Parelius and Singh (1999), …)

  - *Statistical Inference*: sample comparisons (DD-plot), confidence regions, testing, … (Liu and Singh (1997), Li and Liu (2004), Yeh and Singh (1997) …)

  - *Applications*: classification, multivariate control charts, regression, … (Ghosh and Chaudhuri (2005), Cui et. al. (2008), Liu (1995), Rousseeuw and Hubert (1999), Liu et. al. (2004)…)

• *Yield a systematic nonparametric inference scheme* …
- With $\mu$ and $\sigma$ known, $[\mu - z_{(1+\beta)/2} \cdot \sigma, \mu + z_{(1+\beta)/2} \cdot \sigma]$ covers $100\% \beta$ of the distribution, i.e.,

$$P_F \left( X \in [\mu - z_{(1+\beta)/2} \cdot \sigma, \mu + z_{(1+\beta)/2} \cdot \sigma] \right) = \beta$$

where $z_{(1+\beta)/2}$ is the $(1 + \beta) / 2$-quantile of $N(0,1)$.

- With $\mu$ and $\sigma$ unknown, we want to find $[\bar{X} - k_1 \cdot s, \bar{X} + k_1 \cdot s]$, s.t.,

$$P\left( P_F \left( X \in [\bar{X} - k_1 \cdot s, \bar{X} + k_1 \cdot s] \right) \geq \beta \right) = \gamma$$

or $[\bar{X} - k_2 \cdot s, \bar{X} + k_2 \cdot s]$, s.t.,

$$E\left( P_F \left( X \in [\bar{X} - k_1 \cdot s, \bar{X} + k_1 \cdot s] \right) \right) = \beta$$

The above $[\bar{X} - k_1 \cdot s, \bar{X} + k_1 \cdot s]$ is called $\beta$-content tolerance interval for $X$ at confidence level $\gamma$, and $[\bar{X} - k_2 \cdot s, \bar{X} + k_2 \cdot s]$ is called $\beta$-expectation tolerance interval.
Definition:

Let \( \{X_1, \ldots, X_n\} \rightarrow F \in \mathbb{R}^d, d \geq 1 \). \( S(X_1, \ldots, X_n) \) is called

- \( \beta \) - content tolerance region at confidence level \( \gamma \) if
  \[
P \left( P_F \left( X \in S(X_1, \ldots, X_n) \right) \geq \beta \right) = \gamma
\]

- \( \beta \) - expectation tolerance region if
  \[
  E \left( P_F \left( X \in S(X_1, \ldots, X_n) \right) \right) = \beta
  \]
**Univariate case: (d=1)**

**Parametric:** *(Wald and Wolfowitz (1946), Wallis (1951))*

\[
\{ X_1, \ldots, X_n \} \sim N(\mu, \sigma^2) \\
S(X_1, \ldots, X_n) = \left[ \bar{X} - k \cdot s, \bar{X} + k \cdot s \right]
\]

**Nonparametric:** *(Wilks (1941))*

Define
\[
S(X_1, \ldots, X_n) = \left( X_{[r]}, X_{[n-r+1]} \right), \quad r < \frac{n + 1}{2}
\]
**Result:** (Wilks (1941))

\[ P_F \left( X \in S(X_1, \ldots, X_n) \right) = P_F \left( X \in \left( X_r, X_{n-r+1} \right) \right) \sim \text{Beta}(n - 2r + 1, 2r) \]

- \( S(X_1, \ldots, X_n) = \left( X_r, X_{n-r+1} \right) \) is \( \beta \)-content tolerance region at confidence level \( \gamma \) if \( r \) is determined by
  \[ P \left( \text{Beta}(n - 2r + 1, 2r) \geq \beta \right) = \gamma \]

- \( S(X_1, \ldots, X_n) = \left( X_r, X_{n-r+1} \right) \) is \( \beta \)-expectation tolerance region if \( r \) is determined by
  \[ E \left( \text{Beta}(n - 2r + 1, 2r) \right) = \beta \]
Multivariate case: \( (d>1) \)

- Wald (1943): adapt Wilks’s method to each coordinate
  
  Tolerance region: hyperrectangles.

- Tukey (1947): “statistically equivalent blocks”
  
  Depends on the ordering function.

- Charterjee and Patra (1980):
  
  - nonparametric density estimation
  - depends on density estimation and smoothing method
  - overly conservative

- Di Bucchianico, Einmahl and Mushkudiani (2001):
  
  - empirical process theory
  - pre-specify the shape of the tolerance region, ellipsoid, hyperrectangles, etc.
Potential problem:
Univariate Spacings

**Definition:**

\[ X_1 < X_2 < \cdots < X_n \]

Define \( L_i = (X_{i-1}, X_i), \ i = 1, \ldots, n + 1 \), with \( X_0 = a, X_{n+1} = b \).

**Uniform spacings**

Define \( D_i = P_F(X \in L_i), \ i = 1, \ldots, n+1 \), then

(i) \( D_1 + D_2 + \cdots + D_{n+1} = 1 \)

(ii) the density function of \((D_1, D_2, \ldots, D_{n+1})\) is

\[ f(d_1, d_2, \ldots, d_{n+1}) = \begin{cases} n! & \text{if } d_i \geq 0 \text{ and } d_1 + d_2 + \cdots + d_{n+1} = 1 \\ 0 & \text{otherwise.} \end{cases} \]

\((D_1, D_2, \ldots, D_{n+1})\) has the same distribution as the uniform spacings,

i.e., \( D_i \overset{D}{=} Y_{[i]} - Y_{[i-1]} \), where the \( Y_{[i]} \) are the ordered sample from \( U[0,1] \).
Multivariate Spacings

\{X_1, \cdots, X_n\} \rightarrow F \in \mathbb{R}^d, \ d \geq 2

Define \( Z_i = D_F(X_i), i = 1, \ldots, n, \)

\[
\begin{align*}
Z_{[1]} &\geq Z_{[2]} \geq \cdots \geq Z_{[n]} \\
\downarrow &\quad \downarrow \quad \cdots \quad \downarrow \\
X_{[1]}, X_{[2]}, \cdots, X_{[n]}
\end{align*}
\]

Multivariate spacings

\( MS_i = \left\{ X : Z_{[i]} \leq D_F(X) \leq Z_{[i-1]} \right\}, i = 1, \ldots, n + 1 \)

with \( Z_{[0]} = \sup_x \{D_F(x)\}, \ Z_{[n+1]} = 0. \)
Example:

\[ X_1, \ldots, X_5 \sim \text{Norm} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix} \right) \]
• **Multivariate spacings**

\[ MS_i = \left\{ X : Z_{[i]} \leq D_F (X) \leq Z_{[i-1]} \right\}, i = 1, \ldots, n + 1 \]

with \( Z_{[0]} = \sup_x \{ D_F (x) \}, \ Z_{[n+1]} = 0. \)

• **Property:**

Define \( T_i = P_F (X \in MS_i), \ i = 1, \ldots, n + 1, \) then \( (T_1, T_2, \ldots, T_{n+1}) \) follow the same distribution as the uniform spacings.

• **Sample version:**

\[ \hat{MS}_i = \left\{ X : \hat{Z}_{[i]} \leq D_{F_n} (X) \leq \hat{Z}_{[i-1]} \right\}, i = 1, \ldots, n + 1 \]

where \( \hat{Z}_{[1]} \geq \hat{Z}_{[2]} \geq \cdots \geq \hat{Z}_{[n]}, \ \hat{Z}_i = D_{F_n} (X_i), i = 1, \ldots, n. \)
Recall Wilks's univariate tolerance region:

\[ S( X_1, \ldots, X_n ) = \left( X_{[r]}, X_{[n-r+1]} \right) = \bigcup_{i=r+1}^{n-r+1} L_i, \]

Multivariate Tolerance Region
Based on Multivariate Spacings
Multivariate tolerance region:

- \( F \) is known

\[
O_{Z_{[r_n]}} = \bigcup_{i=1}^{r_n} MS_i = \left\{ X : D_F(X) \geq Z_{[r_n]} \right\}.
\]

**Theorem:**

\[
P_F \left( X \in O_{Z_{[r_n]}} \right) \sim Beta(r_n, n + 1 - r_n)
\]

- \( O_{Z_{[r_n]}} = \bigcup_{i=1}^{r_n} MS_i \) is \( \beta \) - content tolerance region at confidence level \( \gamma \) if \( r_n \) is determined by

\[
P \left( Beta(r_n, n - r_n + 1) \geq \beta \right) = \gamma
\]

- \( O_{Z_{[r_n]}} = \bigcup_{i=1}^{r_n} MS_i \) is \( \beta \) - expectation tolerance region if \( r \) is determined by

\[
E \left( Beta(r_n, n - r_n + 1) \right) = \beta
\]
• F is unknown

\[
O^n_{\hat{Z}_{\{r_n\}}} = \bigcup_{i=1}^{r_n} \hat{M} S_i = \left\{ X : D_{F_n}(X) \geq \hat{Z}_{\{r_n\}} \right\}
\]

**Theorem:**

Under proper conditions, if \( \frac{r_n}{n+1} \to \beta \), then

\[
\lim_{n \to \infty} E\left( P_F\left( O^n_{\hat{Z}_{\{r_n\}}} \right) \right) = \beta
\]
**Theorem:** Under proper conditions, for any $\varepsilon > 0$, if
\[
\sqrt{n}\left(\frac{r_1^n}{n} - (\beta + \varepsilon)\right) \rightarrow z_\gamma \sqrt{\beta (1 - \beta)}
\]
\[
\sqrt{n}\left(\frac{r_2^n}{n} - (\beta - \varepsilon)\right) \rightarrow z_\gamma \sqrt{\beta (1 - \beta)}
\]
Then
\[
\lim_{n \to \infty} P\left(P_F\left(O_{\hat{Z}_{[r_1^n]}}^n\right) \geq \beta\right) \geq \gamma
\]
\[
\lim_{n \to \infty} P\left(P_F\left(O_{\hat{Z}_{[r_2^n]}}^n\right) \geq \beta\right) \leq \gamma
\]

**Remark:** In practice, we may take $\varepsilon = 0$ and calculate $r_n$ by solving
\[
r_n = n \beta + z_\gamma \sqrt{n \beta (1 - \beta)}
\]
If $r_n$ is not an integer, we use $\lfloor r_n \rfloor$ or $\lceil r_n \rceil$, depending on which of following is closer to $\gamma$,
\[
P\left(Beta(\lfloor r_n \rfloor, n - \lfloor r_n \rfloor + 1) \geq \beta\right) \text{ and } P\left(Beta(\lceil r_n \rceil, n - \lceil r_n \rceil + 1) \geq \beta\right)
\]
Asymptotic Minimum Property: (Chatterjee and Patra (1980))

Definition:
A sequence of $\beta$-content tolerance regions $S_n$ is called asymptotically minimal if

$$\lambda(S_n \Delta R_{f, \beta})^p \to 0 \quad \text{as} \quad n \to \infty$$

where $R_{f, \beta}$ is minimal among all the sets which satisfy

$$P_F(R_{f, \beta}) = \beta$$

Theorem:
Under proper conditions, for elliptical distributions,

$$\lambda(O_{\hat{Z}_{[n]}} \Delta R_{f, \beta})^p \to 0 \quad \text{as} \quad n \to \infty.$$  

Therefore, the proposed tolerance regions are asymptotically minimal.
Simulation studies

\[ \beta = 0.90, \gamma = 0.95 \]

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Bivariate Normal</th>
<th>Bivariate Cauchy</th>
<th>Bivariate Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=300</td>
<td>( \hat{\gamma} )</td>
<td>0.954</td>
<td>0.963</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>0.90131</td>
<td>0.90036</td>
<td>0.90043</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Bivariate Normal</th>
<th>Bivariate Cauchy</th>
<th>Bivariate Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=1000</td>
<td>( \hat{\gamma} )</td>
<td>0.949</td>
<td>0.9615</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>0.9000486</td>
<td>0.9006111</td>
<td>0.899852</td>
</tr>
</tbody>
</table>
(a) Bivariate normal

(b) Bivariate exponential
Summary

- Nonparametric multivariate tolerance region based on multivariate spacings
  - Completely nonparametric
  - Multi-dimensional generalization of Wilks’s methods
  - Completely data driven and reflects the underlying geometric structure of the data

Reference:
Thank you!
Simulation studies

\( \beta = .90, \gamma = .95 \)

<table>
<thead>
<tr>
<th>n=300</th>
<th>F</th>
<th>Bivariate Normal</th>
<th>Bivariate Cauchy</th>
<th>Bivariate Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\gamma} )</td>
<td>.954</td>
<td>.963</td>
<td>.941</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>.90131 (0.877)</td>
<td>.90036 (0.862)</td>
<td>.90043 (0.885)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n=1000</th>
<th>F</th>
<th>Bivariate Normal</th>
<th>Bivariate Cauchy</th>
<th>Bivariate Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\gamma} )</td>
<td>.949</td>
<td>.9615</td>
<td>.943</td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>.9000486 (0.887)</td>
<td>.9006111 (0.863)</td>
<td>.899852 (0.890)</td>
</tr>
</tbody>
</table>

- Results in () are those reported in Di Bucchianico, Einmahl and Mushkudiani (2001, *Annals of Statistics*).
Figure: Depth contours for the bivariate exponential by using the Mahalanobis depth
Figure: 500 sample points from the bivariate exponential
“+” is the deepest point
Figure: Depth contours for the standard bivariate normal by using the Mahalanobis depth
Figure: 500 sample points from the standard bivariate normal. "+" is the deepest point.