Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com
Smoothing

Sometimes we want to smooth relations

Tukey phrased this as

\[ \text{data} = \text{smooth} + \text{rough} \]

The smoothed version should show patterns not evident in raw data
The rough should have no systematic variation

Many of these methods are nonparametric
Some are parametric
We use them to discover, not to confirm

Stephen Stigler (Seven pillars of statistical wisdom, JSM 2014)

Fallacy: Discarding individual level data reduces the amount of information.
Truth: Discarding individual level data, by aggregating or averaging, can increase information.
Smoothing
Smoothing
Smoothing windows (kernels)

uniform: \( f(x) = a : (-w \leq x \leq w), \text{ else } 0 \)

epanechnikov: \( f(x) = a(1 - (x/w))^2 : (-w \leq x \leq w), \text{ else } 0 \)

biweight: \( f(x) = a(1 - (x/w)^2)^2 : (-w \leq x \leq w), \text{ else } 0 \)

triweight: \( f(x) = a(1 - (x/w)^2)^3 : (-w \leq x \leq w), \text{ else } 0 \)

tricube: \( f(x) = a(1 - |x/w|^3)^3 : (-w \leq x \leq w), \text{ else } 0 \)

gaussian: \( f(x) = ae^{- (x/w)^2} \)

cauchy: \( f(x) = a/(b + (x/w)^2) \)
Smoothing

Smoothing Functions

Kernel smoothing
mean
median
mode

Polynomial smoothing
linear regression
quadratic regression
etc.
Smoothing

Bandwidth

Neighborhood

k-nearest neighbor (KNN)

fixed bandwidth
Smoothing

Kernel smoothers

Mean

Median

Mode
Smoothing

Polynomial Smoothers
Spline Regression
Piecewise polynomial regression

Linear Regression
Linear Spline
Smoothing

Spline Regression
Grace Wahba and others
Smoothing

Polynomial smoothers

Loess (Cleveland)

originally called LOWESS (Locally Weighted Scatterplot Smoothing)
renamed Loess (a wind-blown berm)
a robust hybrid of kernel and polynomial regression
Tricube kernel

\[ tricube: f(x) = a(1 - |x/w|^3)^3: (-w \leq x \leq w), \text{else} 0 \]

KNN window

Biweight psi function

Robust linear (quadratic) regression

Copyright © 2016 Leland Wilkinson
Smoothing

Loess
Smoothing

Principal Curves

Hastie and Stuetzle, 1989.

Each point on the curve is the average of points in a window projected onto the space of the curve.

Figure 1. (a) The linear regression line minimizes the sum of squared deviations in the response variable. (b) The principal-component line minimizes the sum of squared deviations in all of the variables. (c) The smooth regression curve minimizes the sum of squared deviations in the response variable, subject to smoothness constraints. (d) The principal curve minimizes the sum of squared deviations in all of the variables, subject to smoothness constraints.
Smoothing Tables

Tukey Median Polish

Model: $y = row + column + rough$

1. Compute the median of each row and record the value to the side of the row. Subtract the row median from each point in that particular row.

2. Compute the median of the row medians, and record the value as the grand effect. Subtract this grand effect from each of the row medians.

3. Compute the median of each column and record the value beneath the column. Subtract the column median from each point in that particular column.

4. Compute the median of the column medians, and add the value to the current grand effect. Subtract this addition to the grand effect from each of the column medians.

5. Repeat steps 1-4 until no major changes occur with the row or column medians.

A Tukey method resurrected for RMA (Robust Microarray Average) analysis of microarrays.
## Smoothing

### Smoothing Tables

**Tukey Median Polish**

Original table

Percentage of married women by country and age

<table>
<thead>
<tr>
<th></th>
<th>X20.24</th>
<th>X25.29</th>
<th>X30.34</th>
<th>X35.34</th>
<th>X40.44</th>
<th>X45.49</th>
<th>X50.54</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>7.5</td>
<td>21.1</td>
<td>37.1</td>
<td>48.4</td>
<td>54.9</td>
<td>57.9</td>
<td>58.4</td>
</tr>
<tr>
<td>China</td>
<td>32.2</td>
<td>77.4</td>
<td>92.6</td>
<td>95.4</td>
<td>95.7</td>
<td>94.6</td>
<td>92.5</td>
</tr>
<tr>
<td>Iceland</td>
<td>3.7</td>
<td>20.8</td>
<td>39.9</td>
<td>51.6</td>
<td>57.6</td>
<td>60.5</td>
<td>63.4</td>
</tr>
<tr>
<td>Japan</td>
<td>9.5</td>
<td>37.1</td>
<td>60.8</td>
<td>69.8</td>
<td>73.2</td>
<td>76.4</td>
<td>79.0</td>
</tr>
<tr>
<td>Mexico</td>
<td>22.2</td>
<td>41.4</td>
<td>53.4</td>
<td>58.8</td>
<td>60.4</td>
<td>61.3</td>
<td>60.0</td>
</tr>
<tr>
<td>Norway</td>
<td>6.3</td>
<td>25.0</td>
<td>43.7</td>
<td>52.0</td>
<td>54.8</td>
<td>56.7</td>
<td>60.7</td>
</tr>
</tbody>
</table>

Jim Albert  http://bayes.bgsu.edu  https://exploreddata.wordpress.com
Smoothing

Smoothing Tables

Tukey Median Polish

Polished table

<table>
<thead>
<tr>
<th></th>
<th>X20.24</th>
<th>X25.29</th>
<th>X30.34</th>
<th>X35.39</th>
<th>X40.44</th>
<th>X45.49</th>
<th>X50.54</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>2.6</td>
<td>21.8</td>
<td>40.8</td>
<td>49.4</td>
<td>52.8</td>
<td>55.1</td>
<td>58.3</td>
</tr>
<tr>
<td>China</td>
<td>45.6</td>
<td>64.8</td>
<td>83.8</td>
<td>92.4</td>
<td>95.8</td>
<td>98.1</td>
<td>101.3</td>
</tr>
<tr>
<td>Iceland</td>
<td>4.8</td>
<td>23.9</td>
<td>43.0</td>
<td>51.6</td>
<td>54.9</td>
<td>57.3</td>
<td>60.4</td>
</tr>
<tr>
<td>Japan</td>
<td>23.0</td>
<td>42.1</td>
<td>61.1</td>
<td>69.8</td>
<td>73.1</td>
<td>75.5</td>
<td>78.6</td>
</tr>
<tr>
<td>Mexico</td>
<td>12.0</td>
<td>31.1</td>
<td>50.1</td>
<td>58.8</td>
<td>62.1</td>
<td>64.5</td>
<td>67.6</td>
</tr>
<tr>
<td>Norway</td>
<td>5.2</td>
<td>24.3</td>
<td>43.4</td>
<td>52.0</td>
<td>55.3</td>
<td>57.7</td>
<td>60.8</td>
</tr>
</tbody>
</table>
## Smoothing

### Smoothing Tables

#### Tukey Median Polish

Residuals

<table>
<thead>
<tr>
<th></th>
<th>X20.24</th>
<th>X25.29</th>
<th>X30.34</th>
<th>X35.39</th>
<th>X40.44</th>
<th>X45.49</th>
<th>X50.54</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>4.9</td>
<td>-0.7</td>
<td>-3.7</td>
<td>-1</td>
<td>2.1</td>
<td>2.8</td>
<td>0.1</td>
</tr>
<tr>
<td>China</td>
<td>-13.4</td>
<td>12.6</td>
<td>8.8</td>
<td>3</td>
<td>-0.1</td>
<td>-3.5</td>
<td>-8.8</td>
</tr>
<tr>
<td>Iceland</td>
<td>-1.1</td>
<td>-3.1</td>
<td>-3.1</td>
<td>0</td>
<td>2.7</td>
<td>3.2</td>
<td>3.0</td>
</tr>
<tr>
<td>Japan</td>
<td>-13.5</td>
<td>-5.0</td>
<td>-0.4</td>
<td>0</td>
<td>0.1</td>
<td>1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>Mexico</td>
<td>10.2</td>
<td>10.3</td>
<td>3.2</td>
<td>0</td>
<td>-1.7</td>
<td>-3.2</td>
<td>-7.6</td>
</tr>
<tr>
<td>Norway</td>
<td>1.1</td>
<td>0.7</td>
<td>0.4</td>
<td>0</td>
<td>-0.5</td>
<td>-1.0</td>
<td>-0.1</td>
</tr>
</tbody>
</table>
Smoothing

Smoothing Tables

Tukey Median Polish

Jim Albert -- https://exploredata.wordpress.com
Smoothing Tables

Conjoint Measurement


This paper refuted the longstanding claim among physicists that *fundamental measurement (concatenation of measures)* was the only admissible measurement foundation for science.

Tukey’s median polish is an implementation of conjoint measurement. Market researchers adopted the idea and called it *conjoint analysis*, but they threw out the baby with the bathwater. They used an ordinary analysis of variance model instead of nonmetric techniques.
Smoothing

Conjoint Measurement

- White
- Male
- Minority
- Female Minority White
- Begin High School
- Graduate School
- Begin College
## Smoothing

**Smoothing Tables**  
**Head Injury Index**  
**Frontal crashes**  
**Cars and Trucks**  
**NHTSA**

![Figure 16.2 Estimated crash head injury criterion (P=passenger, D=driver)](image-url)
Smoothing

**Smoothing Tables**

\[ H = C + M + V + O + T(MV) + MV + MO + VO + OT(MV) + MVO \]

*H*: Head Injury Index  
*C*: constant term (grand mean)  
*M*: Manufacturer  
*V*: Vehicle (car/truck)  
*O*: Occupant (driver/passenger)  
*T*: Model

Be careful if you drive a truck  
Body-on-frame rigid frame rails dangerous  
Don’t absorb shock in head-on collision

*Figure 16.3* Subset model for crash data sorted by estimate
Smoothing

References


Wahba, G. (1990), *Spline Models for Observational Data*, SIAM.