On the Curse of Dimensionality: Can Scagnostics Help?

Kevin Anderson
Theresa Utlaut

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John W. Tukey
Introduction

- One of the products of semiconductor manufacturing is a prodigious amount of data.
- These data are generated and sampled at virtually every one of the ~500 process steps in the process flow.
- Most of these data are happenstance in nature and the subsequent analyses are plagued with the usual problems.

Introduction

Graphical analysis is often the first step in analyses, but the large dimensionality of these data can make the viewing of graphs very time consuming and often unrealistic.
Introduction

A Scatterplot Matrix, variously called a SPLOM, a casement plot, or a draftman’s plot, is a simple graphical technique for visualizing bivariate relationships.

- Invented by John Hartigan in 1975 and popularized by Tukey, Chambers, and others from Bell Labs.
- As Friedman and Stuetzel (2002) noted, these graphs lose their effectiveness when the number of variables is large.
  - Looking for patterns in \( p(p-1)/2 \) scatterplots is impractical when \( p \) is even moderately large.
  - When \( p \) is large, it is difficult to scroll and paint and zoom, keep variables straight, focus on issues, etc.

Introduction

- In the mid-1980’s John and Paul Tukey introduced an exploratory graphical technique called Scagnostics.
  - The word “Scagnostics” is a neologism for the phrase Scatterplot Diagnostics.
  - It is an exploratory graphical technique to help determine notable relationships between two variables.
  - This approach may prove particularly effective when the number of variables is large.
- Interestingly, the Tukeys never published the specifics of the technique despite discussing it in several conferences.
Methods

Wilkinson et al (2005, 2008) have extended the Tukeys’ original idea of Scagnostics by developing measures that are graph-theoretic and computable.

They propose nine scagnostics measures to characterize the scatterplots:

- Outlying
- Skewed
- Sparse
- Clumpy
- Striated
- Convex
- Skinny
- Stringy
- Monotonic

The measures have proven statistical properties, are computable for moderately large data sets, provide insight into the data, and (thanks to Hofmann, et. al. (2006)) are available as a free downloadable package in R.

Graph Theoretic Scagnostics are based on Geometric Graphs

- Convex Hull (the outer edges of a Delaunay triangulation)
- Alpha Hull (a generalization of the convex hull and a subgraph of the Delaunay triangulation)
- Minimum Spanning Tree

The scagnostics measures are (mostly) calculated from areas, perimeters, and lengths of these graphs.

Calculations can be found in the Appendix.

The data are adaptively binned to improve performance.

- Computation time roughly $O(np^2)$.
On the Curse of Dimensionality: Can Scagnostics Help?

Methods

Scagnostics Requirements

- **Consistency**
  - Sample size does not bias the distributions nor impact adaptive binning.

- **Homogeneity**
  - Same scale (0 to 1).
  - Comparable distributions.

- **Sensitivity**
  - Sensitive to the type of variation it intends to represent.

- **Dimensionality**
  - Measures are relatively uncorrelated over real data.

- **Computability**
  - Efficient and scalable.

Wilkinson, et. al. (2008)

Features and Measures on page 9

Each p-choose-2 scatterplot has nine Scagnostics measures identifying it.

Scagnostics Analysis is performed on relationships of variables, not on relationships of observations.

Potentially useful for finding irregularities and regularities.

“Graphs come from data, data don’t come from graphs.”

- Why not? Why not take data wherever you can get it?
Results

**2007MLBaseballPlayers**
- 2007 Major League Baseball players statistics
- 17 variables, some binary, categorical, percents, counts.
- 501 players.

There is utility in analyzing known, familiar relationships first.
- Leaping into the known may generate understanding faster.
- October seems like an appropriate time of the year...
- No proprietary disclosures...an easier leap through the approval loop.

Results

17 variables, 136 scatterplots, 501 observations, many difficulties with high dimensionality when using graphs to discover information.
Results

Possible ways to analyze scagnostics

The Tukeys recommended a recursion: viewing the Scagnostics with a scatterplot matrix, finding interesting relationships to explore further.

But there are many possibilities:
- Univariate
- Multivariate
- Clustering
- Scagnostics over time
- etc.
Results

We are finding the most descriptive information from Clustering the Scagnostics.
Hierarchical and K-Means clustering “discover” many known and interesting facts.

Analysis

- Similar Monotonic relationships between several variables, all of which are intuitive and obvious,
- Outliers in the Triples, Singles, and Stolen Bases plots,
- Clumpiness and stringiness in the League plots,
- Moderate skewness and skinniness in the Position plots,
- etc.
Analysis

K-Means clustering associates certain clusters with descriptive rays in a biplot.

League vs. Position lies far away from all other plots.

Position and League plots are differentiated from other plots.

Results

Parallel Plots of the Cluster Means show the different clusters and detail some interesting characteristics.
Discussion

Possible Utility in the Semiconductor Industry

- Exploratory Data Analysis
- Specific searches for outliers, clumps, etc.
- Monitoring specific relationships over time
  - Speed vs. Leakage
  - Critical Dimensions
  - E-Test vs. Process Parameters

Possible Utility

Clustering the Scagnostics from E-Test data provides some insight into regular and irregular relationships.

These have proven intriguing to Device Engineers.

Monitoring E-Test Scagnostics over time could detect Special Cause variation in scatterplots.
Discussion

Possible Extensions

Multivariate Statistical Process Control of Scagnostics would be an interesting approach for generating OOC signals that would be even more impossible to troubleshoot than the original signals...

Discussion

“Graphs come from data; Data don’t come from graphs.”

There can be a dramatic reduction in dimensionality by evaluating Scagnostics measures.

For instance, a SPLOM of 20 variables generates 190 scatterplots. 65 variables generates 2080.

- Scagnostics can transform all SPLOMs into 36 combinations that are considerably easier to view, paint, zoom, and generate information.

- (It’s still not easy...)
Conclusions

- The Curse of Dimensionality negatively impacts most exploratory data analysis graphics, including scatterplot matrices, sometimes severely.
- John W. Tukey was an amazing statistician (Brillinger, 2002).
- Convenient, free tools make difficult concepts like graph-theoretic scagnostics approachable.
- It’s not always easy to find an appropriate nail when you have a specific hammer...
- Lots of work remains to be done before we can efficiently and effectively apply this technique in the semiconductor industry.

References

5. JMP software version 7.0.2.
“The test of a good procedure is how well it works, not how well it is understood.”

John W. Tukey
Appendix

\[ w = 0.7 + \frac{0.3}{1 + t^2} \quad t = \frac{n}{500} \]
Stabilizing transformation to attenuate the influence of hexagonal binning

\[ \omega = q_{75} + 1.5(q_{75} - q_{25}) \]
Based on MST edge lengths and used to delete outliers before computing other scagnostics

\[ c_{\text{outlying}} = \frac{\text{length}(T_{\text{outlier}})}{\text{length}(T)} \]
Based on MST edge lengths. This is computed before outliers are removed for the other measures.

\[ c_{\text{convex}} = w \frac{\text{area}(A)}{\text{area}(H)} \]
Based on the ratio of the area of alpha hull and area of convex hull

\[ c_{\text{striated}} = \frac{1}{V} \sum_{v \in V} I(\cos \theta_{e(v,a)e(v,b)} < -0.75) \]
Based on a constraint of the number of adjacent edges

\[ q_{\text{skewed}} = \frac{q_{90} - q_{50}}{q_{90} - q_{10}} \]
Based on MST edge lengths

\[ c_{\text{skewed}} = 1 - w(1 - q_{\text{skewed}}) \]
Based on area and perimeter of the alpha hull

\[ c_{\text{skinny}} = 1 - \sqrt{4 \pi \frac{\text{area}(A)}{\text{perimeter}(A)}} \]
Based on a RUT statistic of the MST
Appendix

\[ c_{\text{stringy}} = \frac{\sqrt{2}}{\sqrt{\sum V^2}} \quad \text{Based on counts of the vertices in the MST} \]

\[ c_{\text{sparse}} = WQ_{90} \quad \text{Based on percentiles of MST edge lengths} \]

\[ c_{\text{monotonic}} = r_{\text{spearman}}^2 \quad \text{This is the only scagnostics measure not based on subsets of the Delaunay graphs.} \]