Combinatorial Fusion Analysis

Naomi Ephraim

Mentor: Dr. Frank Hsu

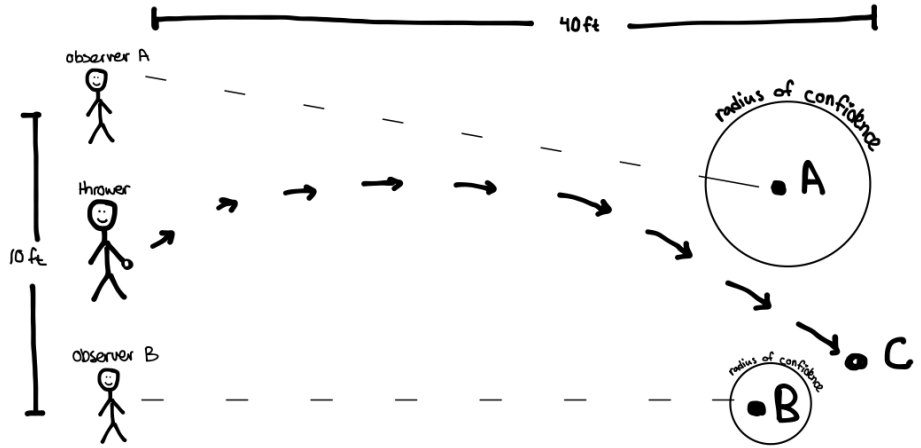
Dimacs REU

nephrai1@jhu.edu

July 18, 2014

Problem Statement

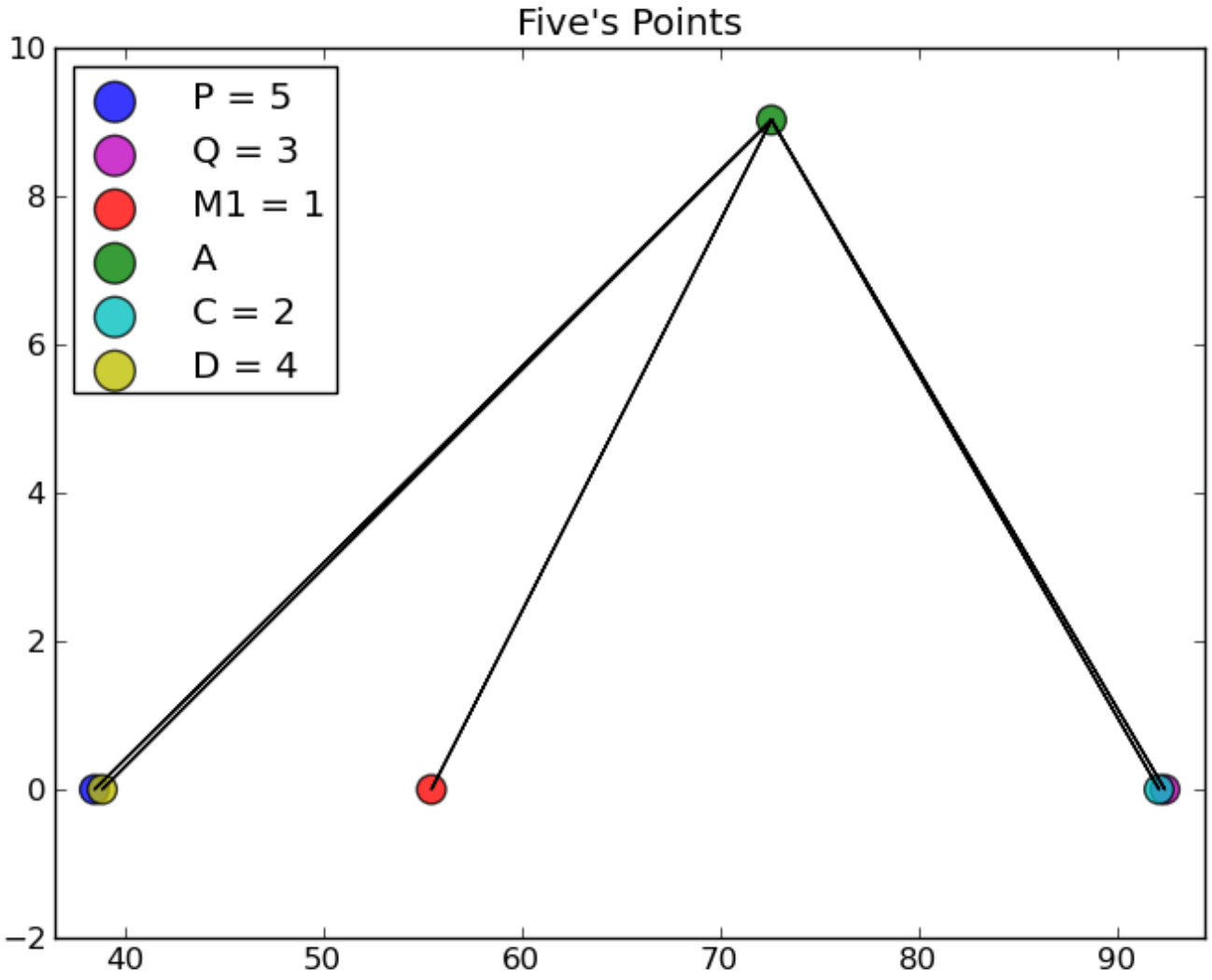
- Combinatorial Fusion Analysis is a method to combine data in order to improve the estimation of another parameter.
- We have considered the following ways to combine data:
  1. Statistical averaging
  2. Score combination
  3. Rank combination
- I explored these methods while trying to answer the question: Are two heads better than one?
- In the problem I worked on, I tried to form a model for an experiment with a target and two observers.



Creating a Model

- I began with eight trials of the same experiment.
- Each trial has a target point and two estimations of the target.
- My model redistributes the data points in each graph.
- It also calculates the rank and score combination, and provides statistics on the closest point to the target.
- In order to visualize the data, I generated scatterplots of each trial.

Scatterplot Example



Distance and Confidence

- There are two main paths from this point:
  - Exploring the relationship between confidence and combination by score and rank.
  - Exploring the relationship between correlation and diversity.
- I began by testing the results of different confidence intervals and seeing whether the score or rank combination was more effective.
- I used a method called bootstrapping, which is a way to generate data that is consistent with the inherent distribution of the original data set.

Bootstrapping

Example (Bootstrapping)

1. Choose a target (A) and two estimations of the target (P, Q).
2. Generate a set of random confidence intervals between 0 and some maximum number.
3. Resample the generated set a given number of times.
4. Calculate a specific statistic for each trial.

Bootstrapping Results

- The result of 4000 trials using four different starting points were the following:
  - P was the closest to A 4.4% of the time.
  - Q was the closest to A 4.4% of the time.
  - M was the closest to A 66.1% of the time.
  - C (score combination) was the closest to A 20.3% of the time.
  - D (rank combination) was the closest to A 4.83% of the time.
- When I looked at the confidence intervals, in the 4000 trials the average confidence interval difference for the rank combination was 3.46, and the average confidence interval difference for the score combination was 8.41.

Hypothesis

Rank combination is more effective than score combination if the two confidence intervals are closer to one another.

Correlation versus Diversity

Pearson's Correlation Coefficient

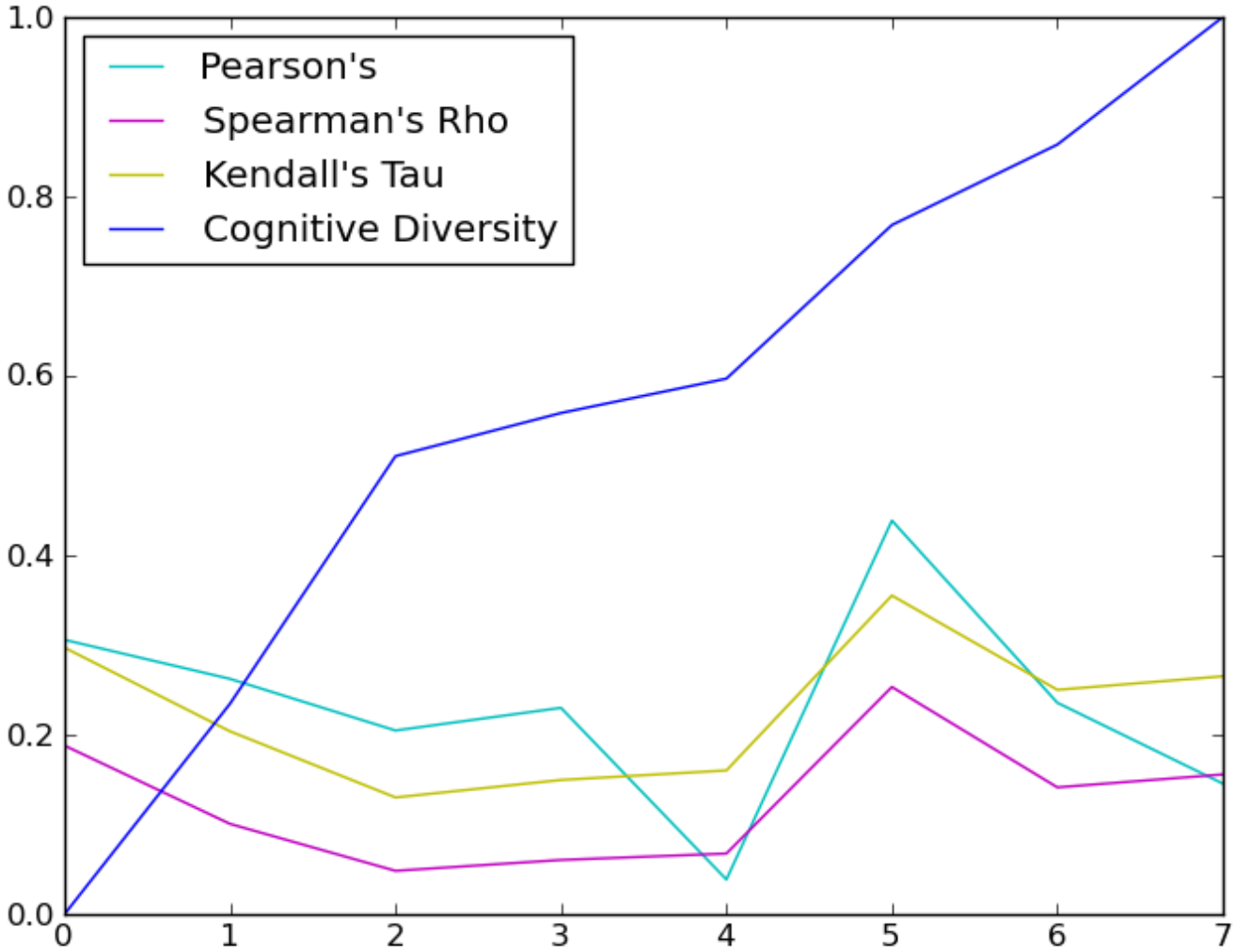
$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}}$$

Pearson's Correlation Coefficient is a measure of the linear relationship between two data sets. It ranges from -1 to 1, and is -1 when the two sets have total negative correlation, 0 when the two sets have no correlation, and 1 when the two sets have total positive correlation.

Cognitive Diversity

Let s_A be a score function, where $s_A : A \to \mathbb{R}$ and let r_A be a rank function, where $r_A : A \to \mathbb{N}$ for some data set A. Then, define $f_A = s_A \circ r_A$. Now, for two data sets A and B, the cognitive diversity between them is defined as $\int_0^n f_A(x) - f_B(x)dx$.

Correlation Graph



Correlation versus Diversity

- The second idea was to look at the relationship between correlation and diversity.
- Generally, the cognitive diversity between two sets is at its highest when the correlation is at its lowest.
- After analyzing many data sets, the rank combination and score combination seem to be better approximations than the original estimates if the estimates are highly correlated.
- Trials were only analyzed when the x-coordinate of the target was between the x-coordinates of the two estimates; in all other cases, the combination proved to be worse than the original estimates.

Conclusion

- We have studied several approaches for data fusion.
- Applying the bootstrap method was my approach to this problem.
- I am currently analyzing more cases and generating data to arrive at a final conclusion.
- Future ideas that I have looked into include using a least square estimator or other known statistical estimators.