Trouble with the Curve: Automatic Clustering of PITCHf/x Data

Michael A. Pane
mpane@andrew.cmu.edu

Department of Statistics, Carnegie Mellon University

joint work with Samuel L. Ventura, Rebecca C. Steorts, and Andrew C. Thomas

September 21, 2013

- Baseball and Pitcher Background.
- PITCHf/x introduction.
- Automatic Clustering of Pitch Types.
  - Current Methods (MLB-AM and Brooks Baseball).
  - Proposed Methods.
  - Model-Based Clustering with Gaussian Mixture Model.
  - Choosing Correct Number of Pitches (BIC_{adj}).
- Label clusters (Fastball, Curveball, etc.).
- CLUMPD Application
  http://legion.stat.cmu.edu:3838/CLUMPD-server/

(sample 2: p3)

(sample 2: p3)

Baseball and Pitcher Background

- Pitcher's purpose: Make the batter miss or hit poorly.
- Pitches vary in velocity, top-spin, and side-spin.

Spectrum of Pitches:

	Fastball	Change-Up	Slider	Curveball	...
Speed	Fastest	med	med	low	...
Movement	Low	med-low	med-high	high	...

Different pitchers throw different combinations of pitch types

- Pitchers throw different sets of pitch types depending on their role on the team, arm strength, ability, etc.
- Example: starting pitcher versus relief pitcher.
  - Barry Zito (Starting Pitcher) throws a four-seam fastball, sinker, changeup, curveball, and slider.
  - Craig Kimbrel (Relief Pitcher) throws a four-seam fastball and curveball.

Pitch type is unknown to batter

- Pitcher's team determines what pitch type will be thrown.
- Batter doesn't know what type of pitch will be thrown.
- No official record of pitch type thrown.

Identifying pitch types

- If each pitch type is known, we can improve measurement of pitcher and batter performance, predict future injury, and analyze other baseball research questions.
- Identify pitch types with velocity, side-spin, and top-spin.

PITCHf/x and Data

- PITCHf/x:
  - A system for recording data on pitches thrown.
  - PITCHf/x used by Major League Baseball since 2006.
- 30+ variables: velocity, release point, acceleration, etc.
- 2008 – 2013: 1000+ pitchers (100 – 15,000 pitches each)
- Back/side spin derived from PITCHf/x data (Nathan 2007).

Pitcher	Start Speed (mph)	Top Spin (rps)	Side Spin (rps)	Label	...
Barry Zito	89.70	-84.59	56.17	Four-seam	...
Barry Zito	70.80	50.39	-50.50	Curveball	...
Tim Wakefield	75.20	-107.19	50.23	Four-seam	...
Tim Wakefield	75.30	-113.89	46.10	Four-seam	...

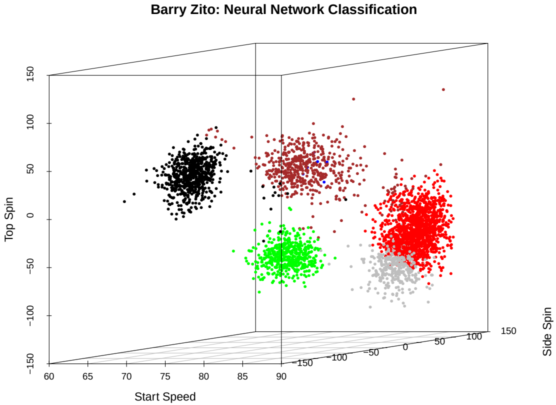
How to automatically identify all pitch types?

1. Identify groups of pitches with similar characteristics using features of the PITCHf/x database.

2. Label each group with a pitch type (e.g. four-seam fastball).

MLB Current Method: Neural Networks Classification

- MLB uses proprietary labeled dataset and classification.
    - Labeled dataset not publicly available, and may be inaccurate.



Pitch Name	Four-Seam	Two Seam	Cutter	Changeup	Curveball	Slider
Color	Red	Grey	Blue	Green	**Black**	Brown

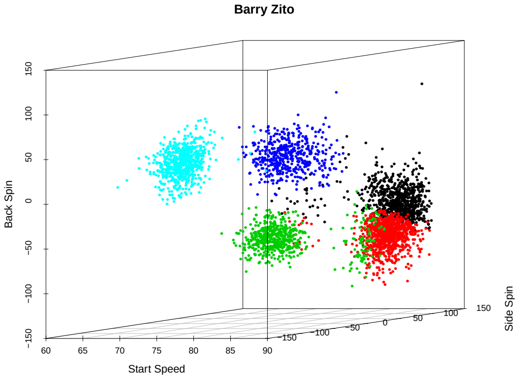
Identify groups of pitches with similar characteristics.

- Possible solution: Unsupervised learning (clustering)
  - k-means
  - hierarchical clustering
  - model-based clustering with a Gaussian mixture model (MBC)
- Two-step approach:
  - Cluster pitches for each individual pitcher.
    - Three variables: velocity, top-spin, side-spin.
    - Adapts to pitcher specific characteristics.
    - Choose number of pitch types (clusters) for each pitcher.
  - Develop algorithm to label clusters.

k-means

Let $x_1, \ldots, x_n \in \mathbb{R}^3$ and $C_1, \ldots, C_K$ clusters with μ_k for each cluster.

$$argmin \sum_{k=1}^{K} \sum_{i \in C_k} ||\bar{x}_i - \mu_k||^2$$

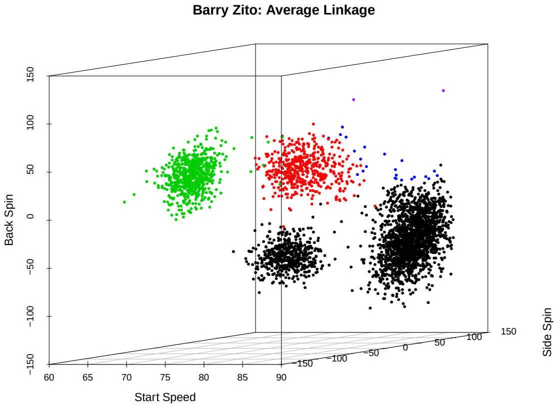


4-Seam Fastball	2-Seam Fastball	Changeup	Slider	Curveball
Black	Red	Green	Blue	Light Blue

Average Linkage (out-performs complete and single)

Let N represents the number of observations in clusters A and B, and d represents the individual pairwise dissimilarities. The distance between clusters A and B:

$$dist(A, B) = \frac{1}{N_A N_B} \sum_{i \varepsilon A \ i' \varepsilon B} d_{ii'},$$



Model–Based Clustering with Gaussian mixture model

A multivariate Gaussian model for each pitcher profile is intuitive.

- Each pitch has a mean value for desired speed and spin.
- The resulting pitches are noisy, both in the pitchers delivery and due to other external factors, such as wind.
- The resulting noisy pattern forms a hyper-ellipsoid.

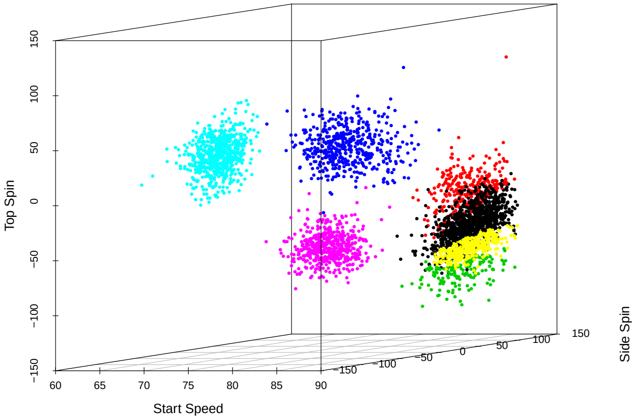
$$y_i|c_i, \mu_k, \Sigma_k \sim N_3(\mu_k, \Sigma_k) \qquad f(y;K) = \sum_{k=1}^{K} f_k(y_i|c_i)\pi(k)$$

$$\mathrm{BIC(K)} = -2\log(\hat{f}(Y;K)) + g(K,d)\cdot\log(n)$$

where $\hat{f}(Y)$ is the likelihood for K compoments, and $g(K,d)\cdot\log(n)$ is the penalty term.

Model-Based Clustering with BIC

MBC Barry Zito: Original BIC



Choosing number of pitch types (clusters)

- Prior knowledge: clustering variables should be uncorrelated.
  - Velocity, side and top-spin should be uncorrelated within clusters.

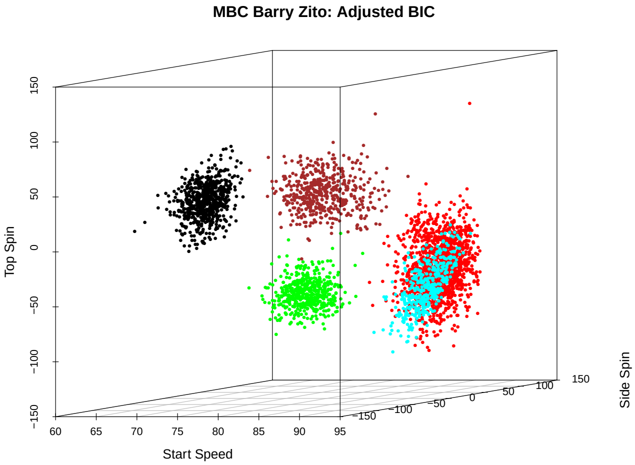
- We develop BIC_{adj}: Penalizes for high intra-cluster correlation.

$$BIC_{\text{adj}}(K) = BIC(K) + \lambda * \sum_{k=1}^{K} \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} log|r_{kij}|$$

  - K is the number of clusters, d is the number of variables, and r is correlation.
  - λ chosen via cross-validation (2010 as training data, 2011 as test data).

Model-Based Clustering with BIC_{adj}

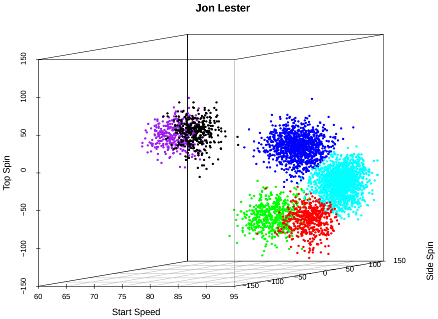
▸ CLUMPD

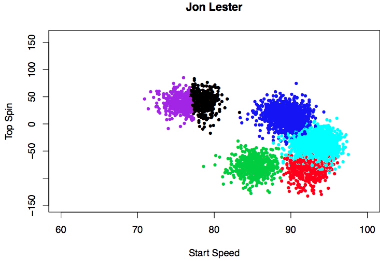


Pitch Name	Four-Seam	Sinker	Changeup	Curveball	Slider
Color	Red	Light Blue	Green	**Black**	Brown

Result of MBC: Identify pitch evolution across time

▸ CLUMPD





Pitch Name	4-Seam Fastball	Sinker	Cutter	Changeup	Curveball 2010	Curveball 2011
Color	Red	Light Blue	Blue	Green	**Black**	Purple

Comparing BIC and $BIC_{\mathbf{adj}}$

- Used both criterions on all pitchers (1051 pitchers).
- Randomly select 50 pitchers:
  - All 50 cases BIC_{adj} out-performs BIC based on visual inspection.
  - In 46 of 50 pitchers, BIC chooses the maximum allowed number of clusters.
- BIC_{adj} out-performs BIC in this application.

Develop Labeling System for Clusters

Original Method:

- Heuristic decision tree algorithm to label clusters with typical pitch types (Fastball, Curveball, etc.)

New Method:

- Split each clustering space into 8 groups and label cluster based on where they fall.
  - Labels clusters off of pitch characteristics, not pitcher intent.
  - **Types of pitches:**
    Fast Rise (Fastball), Slow Drop (Curveball), Slow Left (Slider), etc.
  - Feedback and suggestions?

CLUMPD Application

▸ CLUMPD

Conclusions

- New criterion for choosing the number of clusters.
  - BIC_{adj} factors in intra-cluster correlation structure.
- New method for MLB pitch type clustering and classification.
- BIC_{adj} and MBC are intuitive models for PITCHf/x data.
- Pitch type labeling system.
- Developed pitch classification application that updates daily.

Current and Future Work

- Will be available on FanGraphs.
- Currently fine-tuning and updating CLUMPD method and application.
- Explore new baseball applications using clustering results.

Contact Information:

Email: mpane@andrew.cmu.edu
Version of paper: http://repository.cmu.edu/hsshonors/
CLUMPD Prototype: http://legion.stat.cmu.edu:3838/

- Try out application. Email me if you have any questions or suggestions.