Parallel Coordinates—REVISITED

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

SF Data Mining
November 14, 2013

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Outline

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Outline

- What IS parallel coordinates, anyway?

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Outline

- What IS parallel coordinates, anyway?
- SEEMS to be a great tool.

Outline

- What IS parallel coordinates, anyway?
- SEEMS to be a great tool. But has MAJOR problems.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Outline

- What IS parallel coordinates, anyway?
- SEEMS to be a great tool. But has MAJOR problems.
- I will present a novel way to make parallel coordinates usable.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

What IS Parallel Coordinates?

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:
    - Draw a vertical line for each variable ("parallel coords.").

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:
    - Draw a vertical line for each variable ("parallel coords.").
    - For each data point, mark a dot on each vertical line, at the value of that variable for that data point.

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:
  - Draw a vertical line for each variable ("parallel coords.").
  - For each data point, mark a dot on each vertical line, at the value of that variable for that data point.
  - For each data point, "connect the dots."

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:
  - Draw a vertical line for each variable ("parallel coords.").
  - For each data point, mark a dot on each vertical line, at the value of that variable for that data point.
  - For each data point, "connect the dots."
  - Resulting graph: a jagged line for each of your original data point.

What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:
  - Draw a vertical line for each variable (“parallel coords.”).
  - For each data point, mark a dot on each vertical line, at the value of that variable for that data point.
  - For each data point, “connect the dots.”
  - Resulting graph: a jagged line for each of your original data point.
  - Can then try to find relations between variables by looking at line patterns.

What IS Parallel Coordinates?

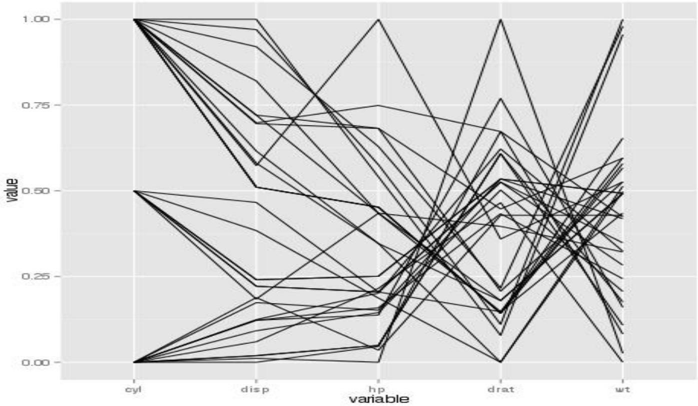
- Attempt to view multidimensional data on 2-dimensiohal screen.
- Simple idea:
  - Draw a vertical line for each variable ("parallel coords.").
  - For each data point, mark a dot on each vertical line, at the value of that variable for that data point.
  - For each data point, "connect the dots."
  - Resulting graph: a jagged line for each of your original data point.
  - Can then try to find relations between variables by looking at line patterns.
  - The operative word is "try."

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

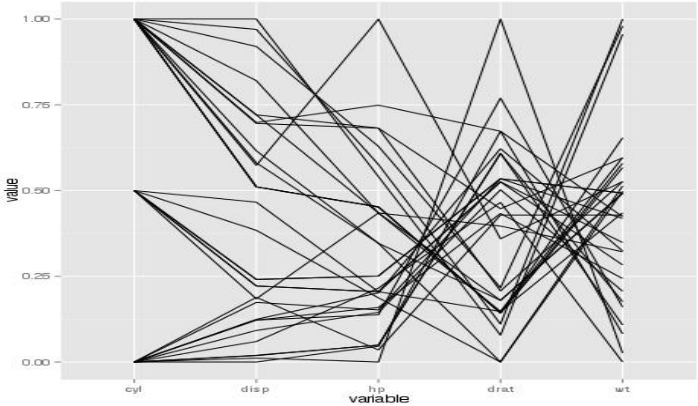
Example: R cars data

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Example: R cars data

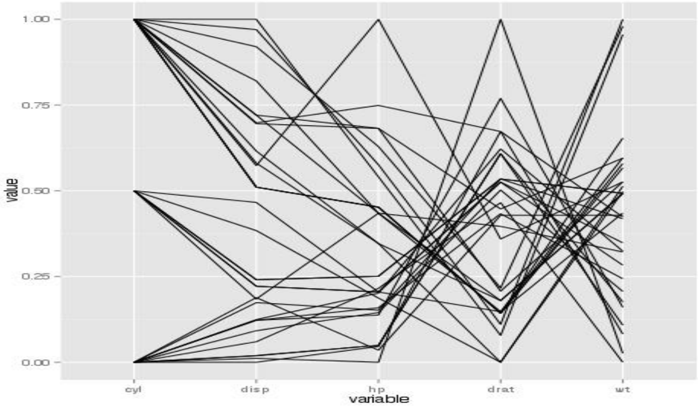


Example: R cars data



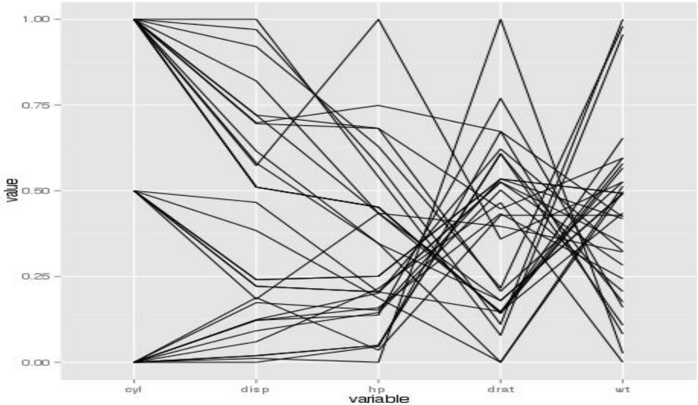
- Each jagged line is one car.

Example: R cars data



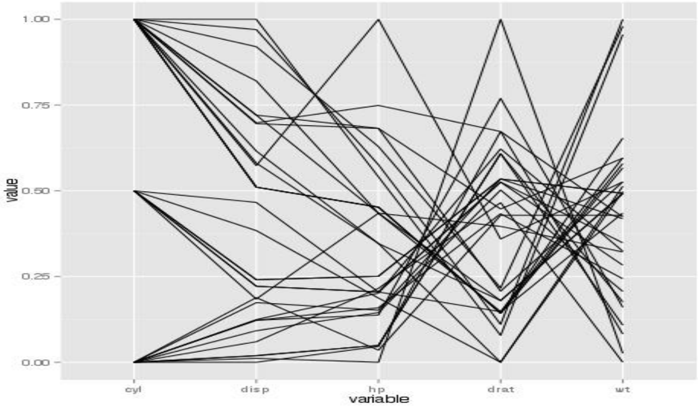
- Each jagged line is one car.
- Vertical axes are the variables, Cyl, Disp, Hp, etc.

Example: R cars data



- Each jagged line is one car.
- Vertical axes are the variables, Cyl, Disp, Hp, etc.
- ALREADY hard to interpret!

Example: R cars data



- Each jagged line is one car.
- Vertical axes are the variables, Cyl, Disp, Hp, etc.
- ALREADY hard to interpret!
- Note: Variables are typically centered and scaled.

Problems

Problems

Hard to interpret, except in “small n, small p” data.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Problems

Hard to interpret, except in "small n, small p" data. (p = number of variables)

Problems

Hard to interpret, except in "small n, small p" data.
(p = number of variables)
Problem 1: Hard to see relation between "far apart" variables

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Problems

Hard to interpret, except in “small n, small p” data.
(p = number of variables)

Problem 1: Hard to see relation between “far apart” variables

Typical solution:

Problems

Hard to interpret, except in "small n, small p" data.
(p = number of variables)

Problem 1: Hard to see relation between "far apart" variables

> *Typical solution:*
> *Allow user to interactively do various permutations of the axes.*

Problems

Hard to interpret, except in “small n, small p” data.
(p = number of variables)

Problem 1: Hard to see relation between “far apart” variables

> *Typical solution:*
> *Allow user to interactively do various permutations of the axes.*

Problem 2: Screen clutter!!!!

Problems

Hard to interpret, except in "small n, small p" data.
(p = number of variables)

Problem 1: Hard to see relation between "far apart" variables

> *Typical solution:*
> *Allow user to interactively do various permutations of the axes.*

Problem 2: Screen clutter!!!!

> *Typical solutions:*

Problems

Hard to interpret, except in “small n, small p” data.
(p = number of variables)

Problem 1: Hard to see relation between “far apart” variables

> *Typical solution:*
> *Allow user to interactively do various permutations of the axes.*

Problem 2: Screen clutter!!!!

> *Typical solutions:*
> *1. α blending (making pixels less dark).*

Problems

Hard to interpret, except in "small n, small p" data.
(p = number of variables)

Problem 1: Hard to see relation between "far apart" variables

Typical solution:
Allow user to interactively do various permutations of the axes.

Problem 2: Screen clutter!!!!

Typical solutions:
1. α blending (making pixels less dark).
2. Plotting line density instead of lines (equiv. to tonal, like α).

Problems

Hard to interpret, except in “small n, small p” data.
(p = number of variables)

Problem 1: Hard to see relation between “far apart” variables

Typical solution:
Allow user to interactively do various permutations of the axes.

Problem 2: Screen clutter!!!!

Typical solutions:
1. α blending (making pixels less dark).
2. Plotting line density instead of lines (equiv. to tonal, like α).
3. Look at random subset of the data.

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Example of Clutter

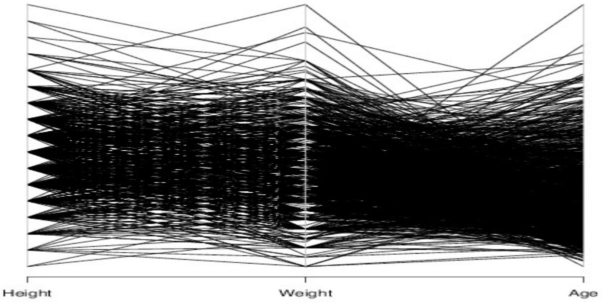
Example of Clutter

Example: Baseball Player data—height, weight, age (courtesy of UCLA Stat. Dept.)

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Example of Clutter

Example: Baseball Player data—height, weight, age (courtesy of UCLA Stat. Dept.)



Another Example of Clutter

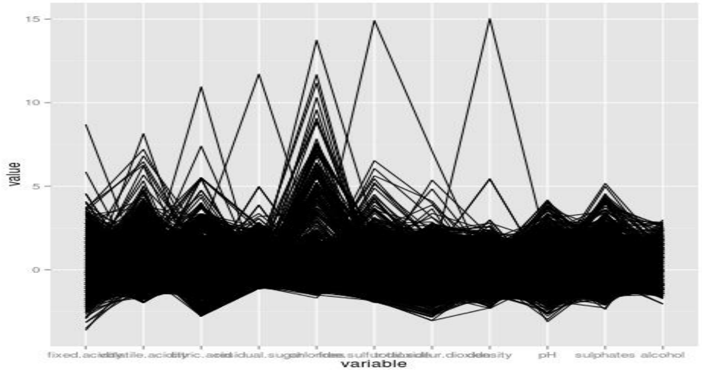
Another Example of Clutter

Example: Wine Quality data—various chemical measures (UCI Repository)

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Another Example of Clutter

Example: Wine Quality data—various chemical measures (UCI Repository)



Alpha Blending May Not Help Much

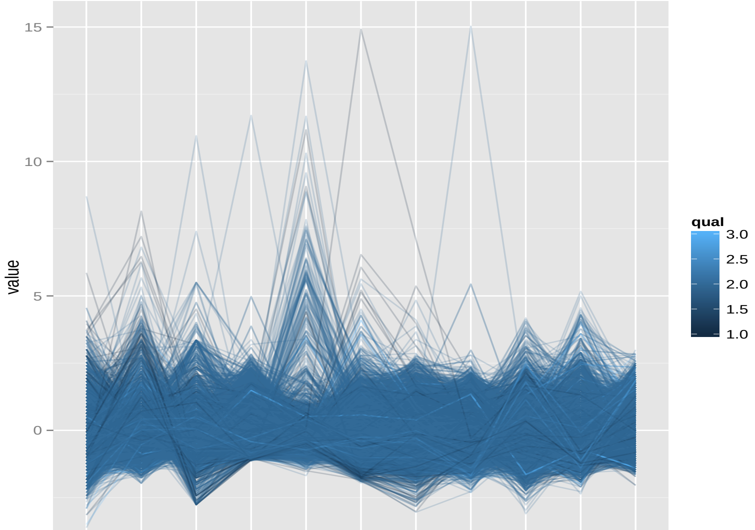
Alpha Blending May Not Help Much

α blending may not help much:

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Alpha Blending May Not Help Much

α blending may not help much:



Yikes!

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Yikes!

Comments:

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Yikes!

Comments:

- Yikes!

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Yikes!

Comments:

- Yikes!
- "Don't let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of cluttered p.c. plots

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Yikes!

Comments:

- Yikes!
- "Don't let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of cluttered p.c. plots
- But it IS intimidating!

Yikes!

Comments:

- Yikes!
- "Don't let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of cluttered p.c. plots
- But it IS intimidating!
- Can TRY to exploit geometric properties, e.g.:

Yikes!

Comments:

- Yikes!
- "Don't let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of cluttered p.c. plots
- But it IS intimidating!
- Can TRY to exploit geometric properties, e.g.:
  - X shape $\Rightarrow$ negative ρ

Yikes!

Comments:

- Yikes!
- "Don't let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of cluttered p.c. plots
- But it IS intimidating!
- Can TRY to exploit geometric properties, e.g.:
    - X shape $\Rightarrow$ negative ρ
    - $<$ shape $\Rightarrow$ positive ρ

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Yikes!

Comments:

- Yikes!
- "Don't let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of cluttered p.c. plots
- But it IS intimidating!
- Can TRY to exploit geometric properties, e.g.:
    - X shape $\Rightarrow$ negative ρ
    - $<$ shape $\Rightarrow$ positive ρ
    - Nice theory, from projective geometry, etc.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Example of Clutter, cont'd.

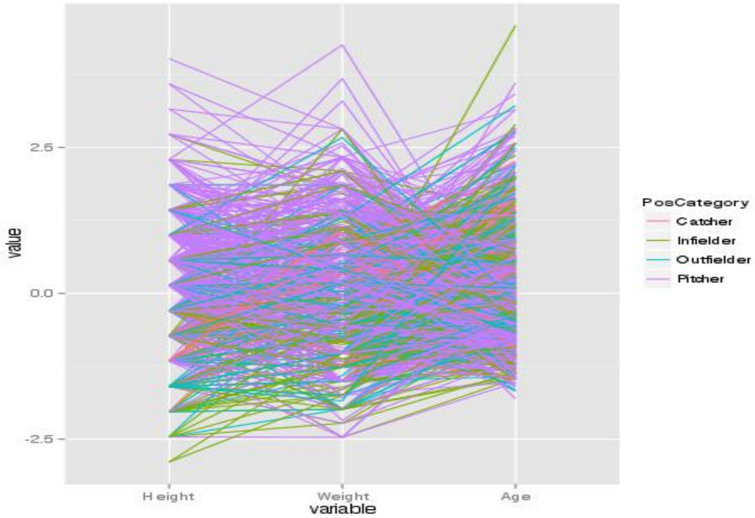
Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Example of Clutter, cont'd.

Grouping by player position doesn't help much:

Example of Clutter, cont'd.

Grouping by player position doesn't help much:



Clutter, cont'd.

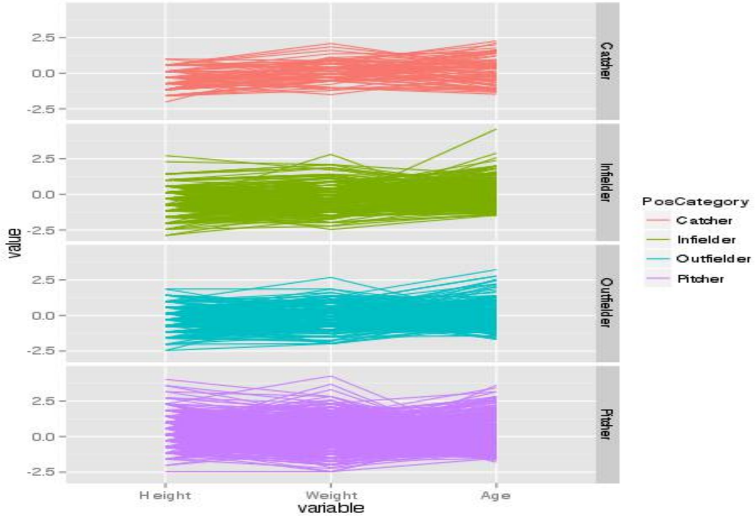
Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Clutter, cont'd.

Grouping by player position doesn't help much—even in lattice display.

Clutter, cont'd.

Grouping by player position doesn't help much—even in lattice display.



Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

My Way

My Way

My approach:

My Way

My approach: **Plot only a few "typical" lines.**

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

My Way

My approach: **Plot only a few "typical" lines.**

- "Typical" means highest estimated multivariate density.

My Way

My approach: **Plot only a few "typical" lines.**

- "Typical" means highest estimated multivariate density.
- No clutter.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

My Way

My approach: **Plot only a few "typical" lines.**

- "Typical" means highest estimated multivariate density.
- No clutter.
- Far-apart variables problem ameliorated.

My Way

My approach: **Plot only a few "typical" lines.**

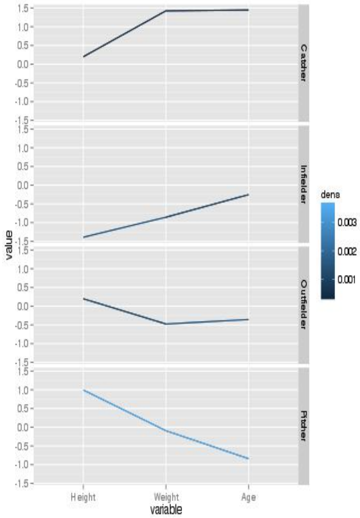
- "Typical" means highest estimated multivariate density.
- No clutter.
- Far-apart variables problem ameliorated.
- (Not related to *parallel coordinate density plots.)*

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Baseball Data, My Way

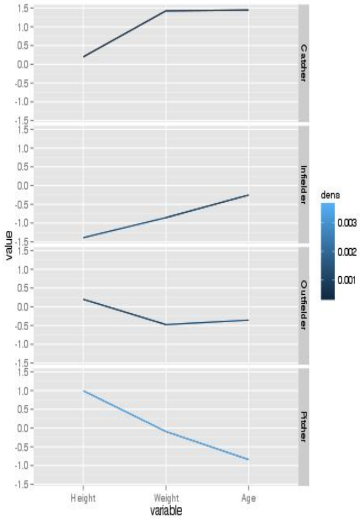
Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Baseball Data, My Way



Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Baseball Data, My Way

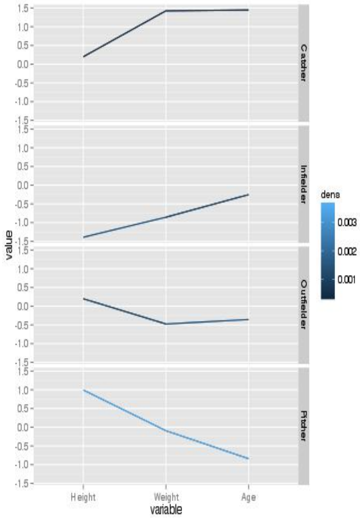


- "The monkeys stand for honesty, Giraffes are insincere, Elephants are kindly but they're dumb"—old Simon & Garfunkel song '

Parallel Coordinates—REVISITED

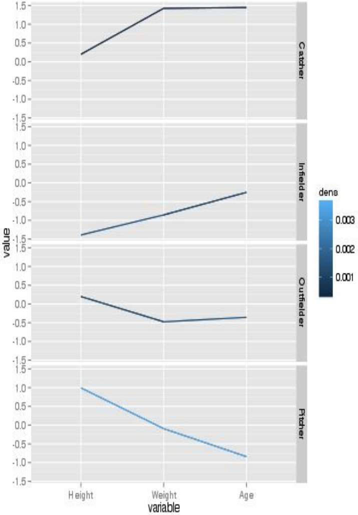
Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Baseball Data, My Way



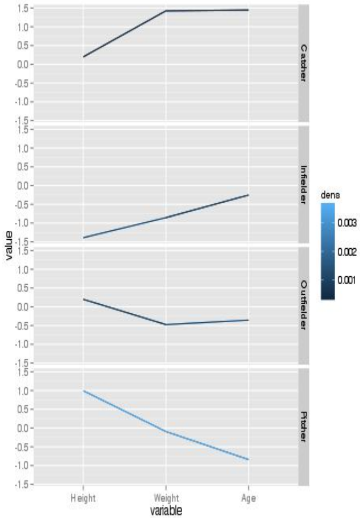
- "The monkeys stand for honesty, Giraffes are insincere, Elephants are kindly but they're dumb"—old Simon & Garfunkel song '
- Pitchers are typically tall, thin, young.

Baseball Data, My Way



- "The monkeys stand for honesty, Giraffes are insincere, Elephants are kindly but they're dumb"—old Simon & Garfunkel song '
- Pitchers are typically tall, thin, young.
- Catchers typically are much heavier, older.

Baseball Data, My Way



- "The monkeys stand for honesty, Giraffes are insincere, Elephants are kindly but they're dumb"—old Simon & Garfunkel song '
- Pitchers are typically tall, thin, young.
- Catchers typically are much heavier, older.
- Infielders typically shorter, thinner, younger.

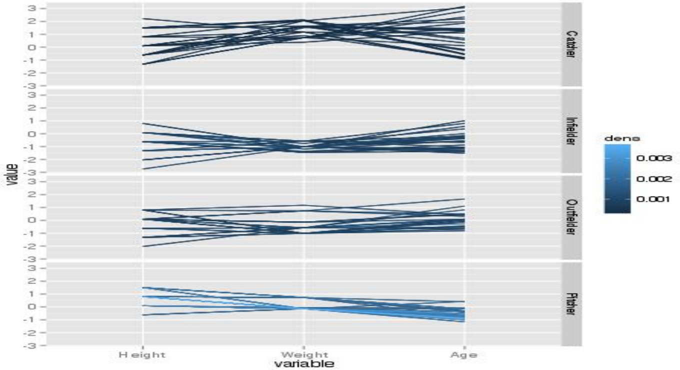
Within-Group Variation

Within-Group Variation

Now look at, say, the 25 most-typical data points in each group, to gauge within-group variation.

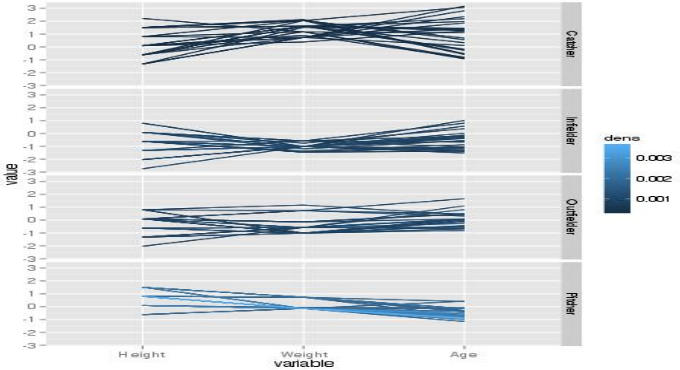
Within-Group Variation

Now look at, say, the 25 most-typical data points in each group, to gauge within-group variation.



Within-Group Variation

Now look at, say, the 25 most-typical data points in each group, to gauge within-group variation.

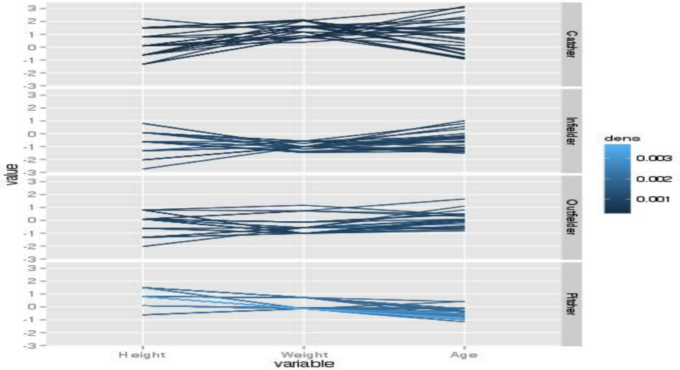


- Pitchers have modest variation in height, very little in weight and age.

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Within-Group Variation

Now look at, say, the 25 most-typical data points in each group, to gauge within-group variation.



- Pitchers have modest variation in height, very little in weight and age.
- Catchers have much more variation.

Cluster Hunting

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Cluster Hunting

- Find local maxima of the density.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Cluster Hunting

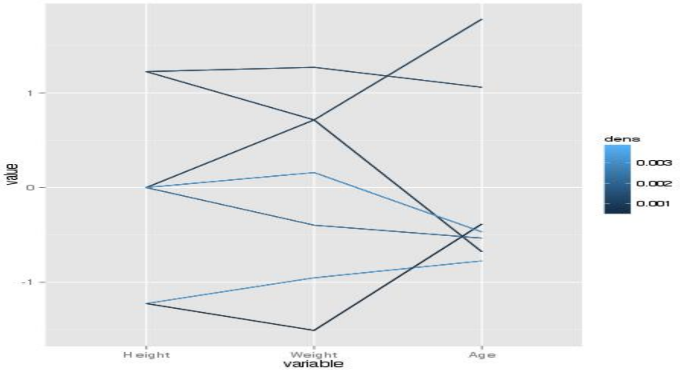
- Find local maxima of the density.
- Pretend we don't know about player position.

Cluster Hunting

- Find local maxima of the density.
- Pretend we don't know about player position. Will the algorithm discover it?

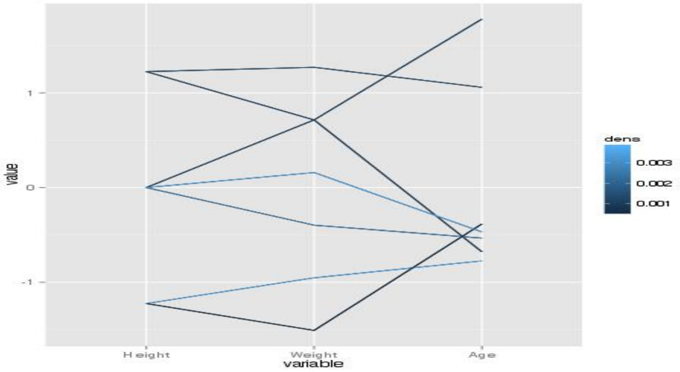
Cluster Hunting

- Find local maxima of the density.
- Pretend we don't know about player position. Will the algorithm discover it?



Cluster Hunting

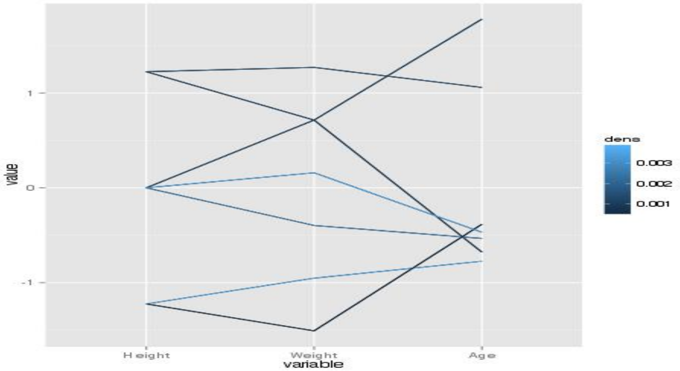
- Find local maxima of the density.
- Pretend we don't know about player position. Will the algorithm discover it?



Suggests 3-7 groups.

Cluster Hunting

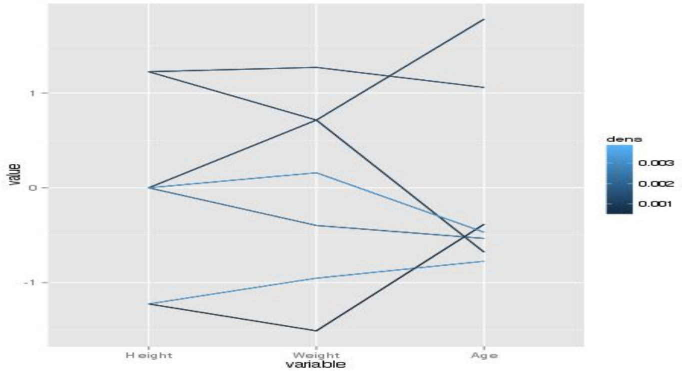
- Find local maxima of the density.
- Pretend we don't know about player position. Will the algorithm discover it?



Suggests 3-7 groups. We have 4 in mind, but there could be subclusters.

Cluster Hunting

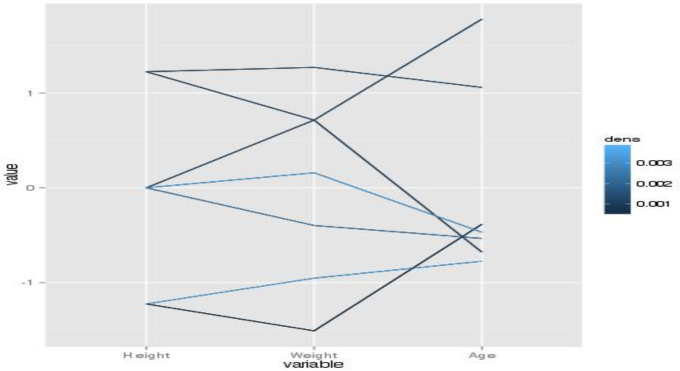
- Find local maxima of the density.
- Pretend we don't know about player position. Will the algorithm discover it?



Suggests 3-7 groups. We have 4 in mind, but there could be subclusters. So the plot is a hint to look more.

Cluster Hunting

- Find local maxima of the density.
- Pretend we don't know about player position. Will the algorithm discover it?



Suggests 3-7 groups. We have 4 in mind, but there could be subclusters. So the plot is a hint to look more. Note: The cluster data points are also printed out, to help find patterns.

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Outlier Hunting

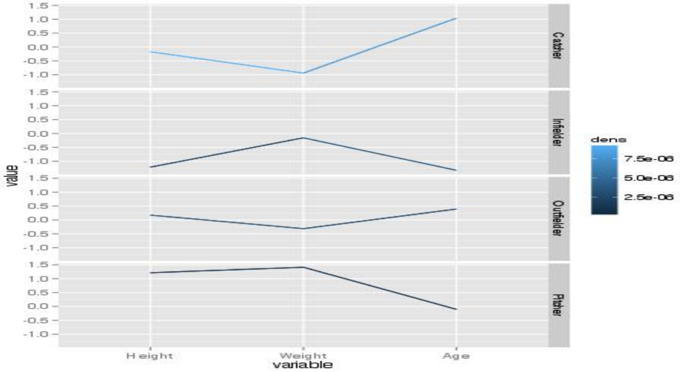
Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Outlier Hunting

To find outliers, find the points having the LOWEST density.

Outlier Hunting

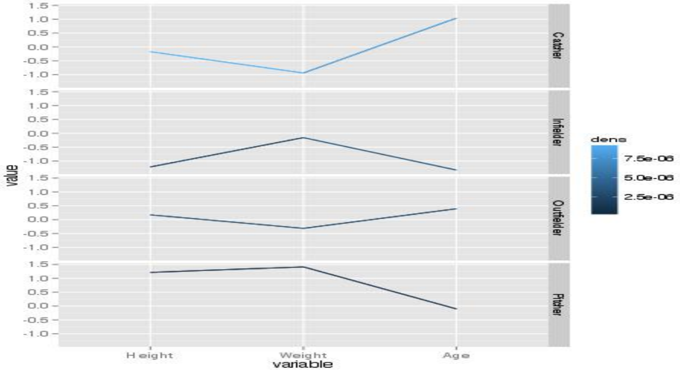
To find outliers, find the points having the LOWEST density.



Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Outlier Hunting

To find outliers, find the points having the LOWEST density.



The unusual ones are thin catchers, fat infielders, very tall/heavy pitchers.

Norm Matloff University of California at Davis (new collaborator: Yingkang Xie)

Computation

Norm Matloff
University of California at Davis
(new collaborator: Yingkang Xie)

Computation

- R package available at http://heather.cs.ucdavis.edu/bdgraphs.html

Computation

- R package available at http://heather.cs.ucdavis.edu/bdgraphs.html
- Use k-NN density estimation.

Computation

- R package available at http://heather.cs.ucdavis.edu/bdgraphs.html
- Use k-NN density estimation.
- Use R's FNN ("fast nearest neighbor") library for some speed.

Computation

- R package available at http://heather.cs.ucdavis.edu/bdgraphs.html
- Use k-NN density estimation.
- Use R's FNN ("fast nearest neighbor") library for some speed.
- Use parallel computing for a lot more speed.