Parallel Coordinates—REVISITED

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

Bay Area R Users Group
November 12, 2013
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.
Outline

- What IS parallel coordinates, anyway?
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.
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.
What IS Parallel Coordinates?

• Attempt to view multidimensional data on 2-dimensional 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."
What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensional screen.
What IS Parallel Coordinates?

• Attempt to view multidimensional data on 2-dimensional 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-dimensional 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-dimensional 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-dimensional 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-dimensional 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."
What IS Parallel Coordinates?

- Attempt to view multidimensional data on 2-dimensional 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."
Example: R cars data
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.
Parallel Coordinates—
REVISITED

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:
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 α).
3. Look at random subset of the data.
Problems

Hard to interpret, except in “small n, small p” data.
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
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. ($p = \text{number of variables}$)

**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. ($p = \text{number of variables}$)

**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:*
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. \( \alpha \) blending (making pixels less dark).
Problems

Hard to interpret, except in “small n, small p” data. ($p = \text{number of variables}$)

**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. $\alpha$ blending (making pixels less dark).
2. Plotting line density instead of lines (equiv. to $\alpha$).
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. \( \alpha \) blending (making pixels less dark).
2. Plotting line density instead of lines (equiv. to \( \alpha \)).
3. Look at random subset of the data.
Example of Clutter
Example of Clutter

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

Example: Wine Quality data—various chemical measures (UCI Repository)
Alpha Blending May Not Help
Alpha Blending May Not Help

$\alpha$ blending may not help:
Alpha Blending May Not Help

$\alpha$ blending may not help:
Yikes!

• 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:
  • X shape ⇒ negative ρ
  • < shape ⇒ positive ρ
• Nice theory, from projective geometry, etc.
Comments:

- 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
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!
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:
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:
  - $\times$ shape $\Rightarrow$ negative $\rho$
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:
  - $X$ shape $\Rightarrow$ negative $\rho$
  - $<$ shape $\Rightarrow$ positive $\rho$
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:
  - $\times$ shape $\Rightarrow$ negative $\rho$
  - $<$ shape $\Rightarrow$ positive $\rho$
  - Nice theory, form projective geometry, etc.
Example of Clutter, cont’d.
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:
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.)
My approach:
My approach: Plot only a few “typical” lines.
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.
My approach: **Plot only a few “typical” lines.**

- “Typical” means highest estimated multivariate density.
- No clutter.
- Far-apart variables problem ameliorated.
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.*)
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.)
Parallel Coordinates—REVISITED

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

Baseball Data, My Way
Baseball Data, My Way

- "The monkeys stand for honesty, giraffes are insincere, elephants are kindly but they're dumb"—Simon & Garfunkel song
- Pitchers are typically tall, thin, young.
- Catchers typically are much heavier, older.
- Infielders typically shorter, thinner, younger.
Baseball Data, My Way

- “The monkeys stand for honesty, giraffes are insincere, elephants are kindly but they’re dumb”—Simon & Garfunkel song
Baseball Data, My Way

- “The monkeys stand for honesty, giraffes are insincere, elephants are kindly but they’re dumb”—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”—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”—Simon & Garfunkel song
- Pitchers are typically tall, thin, young.
- Catchers typically are much heavier, older.
- Infielders typically shorter, thinner, younger.
Example: UCI Wine Quality Data
Example: UCI Wine Quality Data

Variables are chemical measures, plus a quality variable.
Example: UCI Wine Quality Data

Variables are chemical measures, plus a quality variable. I broke quality into 3 broad ranges.
Example: UCI Wine Quality Data

Variables are chemical measures, plus a quality variable. I broke quality into 3 broad ranges.
Example: UCI Wine Quality Data

Variables are chemical measures, plus a quality variable. I broke quality into 3 broad ranges.

Need domain expert to interpret, but distinction between quality groups is clear.
How Many to Plot?

We plot "a few of the most typical lines."

• My software asks the user to define "a few," in an argument $m$.

• Above plots had $m = 1$.

• $m$ smaller $\Rightarrow$ higher "typicalness"

• $m$ larger $\Rightarrow$ better chance to see $\rho$'s among the most typical data points
How Many to Plot?

We plot “a few of the most typical lines.”
How Many to Plot?

We plot “a few of the most typical lines.”

• My software asks the user to define “a few,” in an argument $m$. 

How Many to Plot?

We plot “a few of the most typical lines.”

• My software asks the user to define “a few,” in an argument $m$.

• Above plots had $m = 1$. 
How Many to Plot?

We plot “a few of the most typical lines.”

- My software asks the user to define “a few,” in an argument $m$.
- Above plots had $m = 1$.
- $m$ smaller $\Rightarrow$ higher “typicalness”
How Many to Plot?

We plot “a few of the most typical lines.”

- My software asks the user to define “a few,” in an argument $m$.
- Above plots had $m = 1$.
- $m$ smaller $\Rightarrow$ higher “typicalness”
- $m$ larger $\Rightarrow$ better chance to see $\rho$’s among the most typical data points
More Features
More features of the software (currently experimental):
More Features

More features of the software (currently experimental):

- Cluster hunting.
More features of the software (currently experimental):

- Cluster hunting.
  Instead of finding global max of density, plot jagged line at each local maximum
More features of the software (currently experimental):

- Cluster hunting. Instead of finding global max of density, plot jagged line at each local maximum
- Outlier hunting.
More features of the software (currently experimental):

- Cluster hunting. Instead of finding global max of density, plot jagged line at each local maximum
- Outlier hunting. Plot the least typical lines.
Computation
Computation

Computation

- Use k-NN density estimation.
Computation

- Use k-NN density estimation.
- Use R’s FNN ("fast nearest neighbor") library for some speed.
Computation

- Use k-NN density estimation.
- Use R’s FNN (“fast nearest neighbor”) library for some speed.
- Use parallel computing for a lot more speed.