Parallel Coordinates—REVISITED

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

ECS 256
February 20, 2014
Outline

• What is parallel coordinates, anyway?
• SEEMS to be a great tool. But has MAJOR problems.
• We will present a novel way to make parallel coordinates usable.
Outline

• What is parallel coordinates, anyway?
• 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.
• We will present a novel way to make parallel coordinates usable.
What is Parallel Coordinates?
What is Parallel Coordinates?

- An exploratory method aimed at visualizing multidimensional data on 2-dimensional screen.
What is Parallel Coordinates?

- An exploratory method aimed at visualizing multidimensional data on 2-dimensional screen.
- Simple idea:
What is Parallel Coordinates?

• An exploratory method aimed at visualizing multidimensional data on 2-dimensional screen.

• Simple idea:
  • Draw a vertical line for each variable (“parallel coords.”).
What is Parallel Coordinates?

- An exploratory method aimed at visualizing 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?

- An exploratory method aimed at visualizing 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?

• An exploratory method aimed at visualizing 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?

• An exploratory method aimed at visualizing 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.
What is Parallel Coordinates?

- An exploratory method aimed at visualizing 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, mpg, hp, etc.
Example: R cars data

- Each jagged line is one car.
- Vertical axes are the variables: cyl, mpg, hp, etc.
- SEEMS to be a great tool.
Example: R cars data

- Each jagged line is one car.
- Vertical axes are the variables: cyl, mpg, hp, etc.
- SEEMS to be a great tool. (only for 'small n, small p')
Example: R cars data

- Each jagged line is one car.
- Vertical axes are the variables: cyl, mpg, hp, etc.
- SEEMS to be a great tool. (only for 'small n, small p')
- Note: Variables are typically centered and scaled.
Major Problems
Major Problems

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

Hard to interpret, except in “small n, small p” data.
($p =$ number of variables)
Major 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**
Major 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:*
Major 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.
Screen Clutter!!!

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

When the N gets larger, we would have a lot of overplotting.
Screen Clutter!!!

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

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

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

When the N gets larger, we would have a lot of overplotting.
Screen Clutter!!!

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

When the N gets larger, we would have a lot of overplotting.
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)
Another Example of Clutter

Example: Wine Quality data—various chemical measures (UCI Repository)
Major 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.*
Major 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. Hard to see any pattern in a black screen
Major 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. Hard to see any pattern in a black screen**

*Typical solutions:*
Major 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. Hard to see any pattern in a black screen**

*Typical solutions:*
1. Random sampling
Major 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. Hard to see any pattern in a black screen

Typical solutions:
1. Random sampling
2. Brushing (making a subset of data in heavy black with the rest in gray or coloring different groups)
Major 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. Hard to see any pattern in a black screen**

*Typical solutions:*
1. Random sampling
2. Brushing (making a subset of data in heavy black with the rest in gray or coloring different groups)
3. $\alpha$ blending (making pixels less dark or producing a smoothed color density representation of lines).
Major 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. Hard to see any pattern in a black screen

Typical solutions:
1. Random sampling
2. Brushing (making a subset of data in heavy black with the rest in gray or coloring different groups)
3. $\alpha$ blending (making pixels less dark or producing a smoothed color density representation of lines).
Alpha Blending May Not Help
Alpha Blending May Not Help

$\alpha$ blending may not help much:
\( \alpha \) blending may not help much:
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:
Clutter, cont’d.
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.
Yikes!

• Yikes! What should we do with the Big N data?
• "Don’t let the picture intimidate you!"—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of the cluttered p.c. plots
• But it is intimidating!
• We only want the true signal instead of getting lost in the data!
Comments:

• Yikes! What should we do with the Big N data?
• “Don’t let the picture intimidate you!”—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of the cluttered p.c. plots
• But it is intimidating!
• We only want the true signal instead of getting lost in the data!
Yikes!

Comments:

- Yikes! What should we do with the Big N data?
Yikes!

Comments:

- Yikes! What should we do with the Big N data?
- “Don’t let the picture intimidate you!”—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of the cluttered p.c. plots
Comments:

- Yikes! What should we do with the Big N data?
- “Don’t let the picture intimidate you!” — A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of the cluttered p.c. plots
- But it is intimidating!
Comments:

- Yikes! What should we do with the Big N data?
- “Don’t let the picture intimidate you!”—A. Inselberg, one of the pioneers of parallel coordinates, speaking in general of the cluttered p.c. plots
- But it is intimidating!
- We only want the true signal instead of getting lost in the data!
Clutter, cont’d.

Grouping by player position doesn’t help—even in lattice display.
Novel Method
Novel Method

Our approach:

Plot only a few "typical" lines.

• "Typical" means highest estimated multivariate density.
• Plot only a few lines. No screen clutter.
• Far-apart variables problem are ameliorated.
Novel Method

Our approach: **Plot only a few “typical” lines.**
Our approach: **Plot only a few “typical” lines.**

- “Typical” means highest estimated multivariate density.
Our approach: **Plot only a few “typical” lines.**

- “Typical” means highest estimated multivariate density.
- Plot only a few lines. No screen clutter.
Novel Method

Our approach: **Plot only a few “typical” lines.**

- “Typical” means highest estimated multivariate density.
- Plot only a few lines. No screen clutter.
- Far-apart variables problem are ameliorated.
Parallel Coordinates—REVISITED
Norm Matloff
University of California at Davis
(Collaborator: Yingkang Xie)

Novel Method

Our approach: **Plot only a few “typical” lines.**

- “Typical” means highest estimated multivariate density.
- Plot only a few lines. No screen clutter.
- Far-apart variables problem are ameliorated.
Baseball Data, Novel Method
Baseball Data, Novel Method

- 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.
Baseball Data, Novel Method

- “The monkeys stand for honesty, Giraffes are insincere, Elephants are kindly but they’re dumb”—old Simon & Garfunkel song '
“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, Novel Method

- “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, Novel Method

• “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.
Within-Group Variation

Now look at, say, the 10 most-typical data points in each group.

- Pitchers have modest variation in height, little in age.
- Catchers have much more variation in age; they are all heavy.
Within-Group Variation

Now look at, say, the 10 most-typical data points in each group.
Within-Group Variation

Now look at, say, the 10 most-typical data points in each group.

- Pitchers have modest variation in height, little in age.
- Catchers have much more variation in age; they are all heavy.
Within-Group Variation

Now look at, say, the 10 most-typical data points in each group.

- Pitchers have modest variation in height, little in age.
Within-Group Variation

Now look at, say, the 10 most-typical data points in each group.

- Pitchers have modest variation in height, little in age.
- Catchers have much more variation in age; they are all heavy.
Within-Group Variation

Now look at, say, the 10 most-typical data points in each group.

- Pitchers have modest variation in height, little in age.
- Catchers have much more variation in age; they are all heavy.
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.
Cluster Hunting

- Find local maxima of the density.
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?

![Graph showing parallel coordinates plot]

Note: The cluster data points are also printed out, to help find patterns.
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. 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
Outlier Hunting

To find outliers, find the points having the LOWEST density. The unusual ones are old infilder, fat outfielder, also the very heavy pitchers.

Note: The outlier data points are also printed out, to help find outliers.
Outlier Hunting

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

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

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

The unusual ones are old infielder, fat outfielder, also the very heavy pitchers.
Outlier Hunting

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

The unusual ones are old infilder, fat outfielder, also the very heavy pitchers. Note: The outlier data points are also printed out, to help find outliers.
Parallel Coordinates—REVISITED
Norm Matloff
University of California at Davis
(Collaborator: Yingkang Xie)

Recap
• Parallel Coordinates
• Big Data ⇒ Black Screen Issue
• Our novel methods:
  • “Typical” lines
  • Cluster Hunting
  • Outlier Detection
Recap

- Parallel Coordinates
- Big N Data ⇒ Black Screen Issue
Recap

- Parallel Coordinates
- Big N Data ⇒ Black Screen Issue
- Our novel methods:
  - "Typical" lines
  - Cluster Hunting
  - Outlier Detection
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.