Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach to the Parallel Coordinates Method for Large Data Sets

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

> JSM 2014 Boston, MA USA August 5, 2014

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• If have k variables, draw k vertical axes.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.

• Very old idea (late 1800s!).

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.
- Very old idea (late 1800s!).
- But only popularized 100 years later.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.
- Very old idea (late 1800s!).
 - But only popularized 100 years later.
- Nice math theory, using affine geometry, aiding practical interpretation;

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.
- Very old idea (late 1800s!).
- But only popularized 100 years later.
- Nice math theory, using affine geometry, aiding practical interpretation; e.g. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*, Alfred Inselberg, Springer, 2009.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.
- Very old idea (late 1800s!).
- But only popularized 100 years later.
- Nice math theory, using affine geometry, aiding practical interpretation; e.g. Parallel Coordinates: Visual Multidimensional Geometry and Its Applications, Alfred Inselberg, Springer, 2009.

• Available in lattice, MASS, GGally etc. -

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- If have k variables, draw k vertical axes. Each data point maps to a polygonal line connecting the value of each variable.
- Very old idea (late 1800s!).
- But only popularized 100 years later.
- Nice math theory, using affine geometry, aiding practical interpretation; e.g. *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*, Alfred Inselberg, Springer, 2009.
- Available in lattice, MASS, GGally etc. but use our approach instead. :-)

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

Example: Height/weight/age data.

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Example

A New Approach to the Parallel Coordinates Method for Large Data Sets

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example: Height/weight/age data.

> d ht wt age 1 71 175 25 2 66 128 36 3 68 162 42 > library(GGally) > p <-+ ggparcoord(d,... > p <- p + annotate(...</pre>

Vertical axes use centered, scaled values.



(日)、

э

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Problems with Parallel Coordinates

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Problems with Parallel Coordinates

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• Highly cluttered, "black screen" problem.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions; see reviews in Heinrich and Weiskopf (IEEE VIS 2014), Zhou *et al* (IEEE-VGTC, 2008);

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions; see reviews in Heinrich and Weiskopf (IEEE VIS 2014), Zhou *et al* (IEEE-VGTC, 2008); e.g. making the lines fainter, or combining them.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions; see reviews in Heinrich and Weiskopf (IEEE VIS 2014), Zhou *et al* (IEEE-VGTC, 2008); e.g. making the lines fainter, or combining them.
- But the larger *n*, the less effective these solutions are,

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions; see reviews in Heinrich and Weiskopf (IEEE VIS 2014), Zhou *et al* (IEEE-VGTC, 2008); e.g. making the lines fainter, or combining them.
- But the larger *n*, the less effective these solutions are, especially with large *p*.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

• Our solution:

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQ@

• Our solution: Plot only a few "typical" lines,

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• Our solution: Plot only a few "typical" lines, based on estimated multivariate density.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with *n*.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with *n*.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with *n*.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)
 - What ht/wt/age combinations are rare?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)
 - What ht/wt/age combinations are rare? (Outlier hunting.)

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)
 - What ht/wt/age combinations are rare? (Outlier hunting.)
 - What ht/wt/age combinations are "locally typical"?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)
 - What ht/wt/age combinations are rare? (Outlier hunting.)
 - What ht/wt/age combinations are "locally typical"? (Cluster hunting.)

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)
 - What ht/wt/age combinations are rare? (Outlier hunting.)
 - What ht/wt/age combinations are "locally typical"? (Cluster hunting.)
 - Bonus: Regression diagnostics.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

A New Approach

- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- Clutter does NOT increase with n.
- Very versatile. E.g., height/weight/age:
 - What ht/wt/age combinations are typical overall? (General analysis.)
 - What ht/wt/age combinations are typical within groups? (Group comparison.)
 - What ht/wt/age combinations are rare? (Outlier hunting.)
 - What ht/wt/age combinations are "locally typical"? (Cluster hunting.)
 - Bonus: Regression diagnostics.
- Implemented in a package freqparcoord on CRAN.

Example: Taxi Data

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com
Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example: Taxi Data

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Example: Taxi data, http://www.theatlantic.com/video/ index/253385/taxi-data-visualization/.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example: Taxi Data

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Example: Taxi data, http://www.theatlantic.com/video/ index/253385/taxi-data-visualization/.

• We used a 100K subsample.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example: Taxi Data

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Example: Taxi data, http://www.theatlantic.com/video/ index/253385/taxi-data-visualization/.

- We used a 100K subsample.
- Consists of **data** and **fare** portions, different variables:

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example: Taxi Data

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Example: Taxi data, http://www.theatlantic.com/video/ index/253385/taxi-data-visualization/.

- We used a 100K subsample.
- Consists of data and fare portions, different variables:
 - data: passenger_count, trip_time_in_secs, trip_distance, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, pickuptime

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Example: Taxi Data

Example: Taxi data, http://www.theatlantic.com/video/ index/253385/taxi-data-visualization/.

- We used a 100K subsample.
- Consists of data and fare portions, different variables:
 - data: passenger_count, trip_time_in_secs, trip_distance, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, pickuptime
 - **fare:** fare_amount, surcharge, mta_tax, tip_amount, tolls_amount, total_amount, cmt, crd (paid with credit card), tippc, booltip (tip, yes or no), pickuptime, daytime

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Outlier Hunting First

Norm Matloff and Yingkang Xie University of California at

Outlier Hunting First p <- freqparcoord(d100, -10, c(8:15), keepidxs=8)</pre>

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Davis e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* mat-

loff.wordpress.com

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Outlier Hunting First p <- freqparcoord(d100,-10,c(8:15),keepidxs=8)</pre>



Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Outliers, cont'd.

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

A New Approach to the Parallel Coordinates Method for Large Data Sets				0
Norm Matloff and Yingkang Xie	> p \$ ×disp[,1	1:14]		
University of California at Davis	plong 74.00399 0.00000 74.00748	plat 40.742107 0.000000 40.703709	dlong -73.94696 -73.96590 -74.07885	dlat 40.81335 40.80481 40.43142
<i>e-mail:</i> mat- loff@cs.ucdavis.e ykxie@ucdavis.ec <i>R/stat blog:</i>	0.00000 0.00000 du, 0.00000 lu -73.88925 -1837.04530 -73.98628	40.783333 40.835121 40.733334 40.769035 0.041667 40.752365	$\begin{array}{r} 0.00000\\ 0.00000\\ -73.94363\\ -73.96226\\ -73.77634\end{array}$	40.79044 40.84693 40.74148 40.75264 40.76774
mat- loff wordpress co	- 73.98028 0.00000	0.000000	0.00000	0.00000

Outliers, cont'd.

A New Approach to the Parallel Coordinates Method for Large Data Sets				Out	liers, cont'd.
Norm Matloff and Yingkang Xie University of California at Davis	> p\$xdisp[,1 plong -74.00399 0.00000 -74.00748	1:14] plat 40.742107 0.000000 40.703709	dlong - 73.94696 - 73.96590 - 74.07885	dlat 40.81335 40.80481 40.43142	
<i>e-mail:</i> mat- loff@cs.ucdavis.e ykxie@ucdavis.e <i>R/stat blog:</i> mat- loff.wordpress.cc	0.00000 0.00000 du, -73.88925 -1837.04530 -73.98628 om 0.00000	40.783333 40.835121 40.733334 40.769035 0.041667 40.752365 0.000000	0.00000 0.00000 -73.94363 -73.96226 -73.77634 0.00000	40.79044 40.84693 40.74148 40.75264 40.76774 40.64601 0.00000	

Bad cases (-1800, 0s) removed (IDs in p\$xdisp but not shown here).

・ロト・日本・モト・モート ヨー うへで

A New Approach to the Parallel Coordinates Method for Large Data Sets				Outl	iers, cont'd.
Norm Matloff and Yingkang Xie University of California at Davis	> p\$xdisp[,1 plong -74.00399 0.00000 -74.00748	1:14] plat 40.742107 0.000000 40.703709	dlong - 73.94696 - 73.96590 - 74.07885	dlat 40.81335 40.80481 40.43142	
<i>e-mail:</i> mat- loff@cs.ucdavis.e ykxie@ucdavis.ec <i>R/stat blog:</i> mat- loff.wordpress.co	0.00000 0.00000 du, -73.88925 -1837.04530 -73.98628 m 0.00000	40.783333 40.835121 40.733334 40.769035 0.041667 40.752365 0.000000	$\begin{array}{c} 0.00000\\ 0.00000\\ -73.94363\\ -73.96226\\ -73.77634\\ 0.00000 \end{array}$	40.79044 40.84693 40.74148 40.75264 40.76774 40.64601 0.00000	

Bad cases (-1800, 0s) removed (IDs in **p\$xdisp** but not shown here). Trip from Altoona, PA to NYC <u>not</u> removed.

A New Approach to the Parallel Coordinates Method for Large Data Sets				Οι
Norm Matloff and Yingkang Xie University of California at Davis	> p\$xdisp[,1 plong -74.00399 0.00000 -74.00748	1:14] plat 40.742107 0.000000 40.703709	dlong - 73.94696 - 73.96590 - 74.07885	dlat 40.81335 40.80481 40.43142
<i>e-mail:</i> mat- loff@cs.ucdavis.ed ykxie@ucdavis.ed <i>R/stat blog:</i> mat- loff.wordpress.col	0.00000 0.00000 u -73.88925 -1837.04530 -73.98628 m 0.00000	40.783333 40.835121 40.733334 40.769035 0.041667 40.752365 0.000000	0.00000 0.00000 -73.94363 -73.96226 -73.77634 0.00000	40.79044 40.84693 40.74148 40.75264 40.76774 40.64601 0.00000

Bad cases (-1800, 0s) removed (IDs in **p\$xdisp** but not shown here). Trip from Altoona, PA to NYC <u>not</u> removed. **Illustrates another advantage of displaying just a few** "typical" cases.

Outliers, cont'd.

・ロト・日本・モート モー うへぐ

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis

p <- freqparcoord(d100,50,c(8:15),keepidxs=8)</pre>

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis





Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis, cont'd.

Norm Matloff and Yingkang Xie University of California at Davis

What do we see?

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis, cont'd.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis, cont'd.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

What do we see?

• Already see at least two clusters, largely differing on pickup/dropoff location and time of day.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

General Analysis, cont'd.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

What do we see?

- Already see at least two clusters, largely differing on pickup/dropoff location and time of day.
- Note there is much more variation in trip time than in trip distance—due to variation in traffic.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Cluster Analysis

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

p <- freqparcoord(d100,1,c(8:15),method="locmax",klm=1000, cls=cl4,keepidxs=15)

Cluster Analysis

Cluster Analysis



A New Approach to the Parallel Coordinates Method for Large Data Sets

Norm Matloff and Yingkang Xie University of California at Davis

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Clustering, cont.d

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Clustering, cont.d

▲ロト ▲冊 ▶ ▲ ヨ ▶ ▲ ヨ ▶ ● の Q @

• We see perhaps 8-9 clusters.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Clustering, cont.d

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

- We see perhaps 8-9 clusters.
- Varying in short vs. long trip distance, pickup/dropoff location, time of day.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Approach to the Parallel Coordinates Method for Large Data Sets

A New

Norm Matloff and Yingkang Xie University of California at Davis

- We see perhaps 8-9 clusters.
- Varying in short vs. long trip distance, pickup/dropoff location, time of day.
- "Changing of the guard," 2 top lines:

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Approach to the Parallel Coordinates Method for Large Data Sets

A New

Norm Matloff and Yingkang Xie University of California at Davis

- We see perhaps 8-9 clusters.
- Varying in short vs. long trip distance, pickup/dropoff location, time of day.
- "Changing of the guard," 2 top lines:
 - Around 1:45 p.m., mid-Manhattan \rightarrow La Guardia Airport.

Norm Matloff and Yingkang Xie University of California at Davis

A New Approach to

the Parallel Coordinates Method for Large Data Sets

- We see perhaps 8-9 clusters.
- Varying in short vs. long trip distance, pickup/dropoff location, time of day.
- "Changing of the guard," 2 top lines:
 - Around 1:45 p.m., mid-Manhattan \rightarrow La Guardia Airport.
 - Around 7:30 p.m., La Guardia Airport \rightarrow mid-Manhattan.

Large Data Sets Norm Matloff and Yingkang Xie University of California at Davis

A New Approach to

the Parallel Coordinates Method for

- We see perhaps 8-9 clusters.
- Varying in short vs. long trip distance, pickup/dropoff location, time of day.
- "Changing of the guard," 2 top lines:
 - Around 1:45 p.m., mid-Manhattan \rightarrow La Guardia Airport.
 - Around 7:30 p.m., La Guardia Airport \rightarrow mid-Manhattan.
 - Good example of the use of viewing variables together, rather than individually.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Group by # of Passengers

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Group by # of Passengers

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ □臣 = の�@

p <- freqparcoord(d100,50,c(9:15),grpvar=8)</pre>

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Group by # of Passengers

p <- freqparcoord(d100,50,c(9:15),grpvar=8)</pre>



Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

of Passengers, cont'd.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

of Passengers, cont'd.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• The 1-passenger trips tend to be earlier in the day, some late.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

of Passengers, cont'd.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

- The 1-passenger trips tend to be earlier in the day, some late.
- The 2-4-passenger trips tend to be later in the day.
Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

of Passengers, cont'd.

- The 1-passenger trips tend to be earlier in the day, some late.
- The 2-4-passenger trips tend to be later in the day.
- The 5-6 passenger trips (families?) more diverse in time.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Credit Card vs. Cash

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Credit Card vs. Cash

▲ロト ▲周ト ▲ヨト ▲ヨト ヨー のくで

p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)</pre>

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Credit Card vs. Cash



p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)</pre>

▲□▶ ▲圖▶ ▲臣▶ ▲臣▶ ―臣 … のへで

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Credit Card vs. Cash



p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)</pre>

Not much difference, e.g. in base fare.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Credit Card vs. Cash

1.0 0.5 -0.0--0.5 dens ane -1.0 e+11 1.0 0.5 -0.0 -0.5 -1.0 fore sur mta tipamt tot cmt tippe booltip ptime variable

p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)</pre>

Not much difference, e.g. in base fare. Some difference in time of day.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Credit Card vs. Cash

1.0 0.5 -0.0--0.5 dens -1.0 alle 1.0 e+11 0.5 0.0 -0.5 -1.0 fore mta tipamt tolls tot cmt tippe booltip ptime variable

p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)</pre>

Not much difference, e.g. in base fare. Some difference in time of day. But stark difference in tips!

・ロト ・ 理 ト ・ ヨ ト ・ ヨ ト

э

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Application: Regression Diagnostics

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Application: Regression Diagnostics

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• Compute *divergences* (<u>not</u> residuals):

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Application: Regression Diagnostics

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

• Compute *divergences* (<u>not</u> residuals):

 $div_i = param_est_i - nonparam_est_i$

• Use freqparcoord() on the divergences,

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Application: Regression Diagnostics

• Compute *divergences* (<u>not</u> residuals):

 $div_i = param_est_i - nonparam_est_i$

• Use **freqparcoord()** on the divergences, to identify regions of predictor space in which there is systematic over- or underestimation of the true regression function.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Application: Regression Diagnostics

• Compute *divergences* (<u>not</u> residuals):

 $div_i = param_est_i - nonparam_est_i$

- Use **freqparcoord()** on the divergences, to identify regions of predictor space in which there is systematic over- or underestimation of the true regression function.
- See our useR! 2014 slides, at http://heather.cs. ucdavis.edu/freqparcoord/UseR2014Slides.pdf.

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Conclusions

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

Norm Matloff and Yingkang Xie University of California at Davis

e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu *R/stat blog:* matloff.wordpress.com

Conclusions

▲ロト ▲周ト ▲ヨト ▲ヨト ヨー のくで

• A new approach to parallel coordinates.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

A New Approach to the Parallel Coordinates Method for Large Data Sets

Norm Matloff and Yingkang Xie University of California at Davis

- A new approach to parallel coordinates.
- Key point: Plots only a few "typical" lines.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Approach to the Parallel Coordinates Method for Large Data Sets

A New

Norm Matloff and Yingkang Xie University of California at Davis

- A new approach to parallel coordinates.
- Key point: Plots only a few "typical" lines.
- This reduces clutter—no increase in clutter as n grows!

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

the Parallel Coordinates Method for Large Data Sets

A New Approach to

Norm Matloff and Yingkang Xie University of California at Davis

- A new approach to parallel coordinates.
- Key point: Plots only a few "typical" lines.
- This reduces clutter—no increase in clutter as n grows!
- Uses: general analysis; group comparison; cluster detection; outlier hunting.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Method for Large Data Sets Norm Matloff and Yingkang Xie University of California at

Davis

A New Approach to

the Parallel

- A new approach to parallel coordinates.
- Key point: Plots only a few "typical" lines.
- This reduces clutter—no increase in clutter as n grows!
- Uses: general analysis; group comparison; cluster detection; outlier hunting.
- Bonus: Regression diagnostics.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Large Data Sets Norm Matloff and Yingkang Xie University of California at Davis

A New Approach to

the Parallel Coordinates Method for

- A new approach to parallel coordinates.
- Key point: Plots only a few "typical" lines.
- This reduces clutter—no increase in clutter as n grows!
- Uses: general analysis; group comparison; cluster detection; outlier hunting.
- Bonus: Regression diagnostics.
- Package freqparcoord on CRAN.

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

Large Data Sets Norm Matloff and Yingkang Xie University of California at Davis

A New Approach to

the Parallel Coordinates Method for

- A new approach to parallel coordinates.
- Key point: Plots only a few "typical" lines.
- This reduces clutter—no increase in clutter as n grows!
- Uses: general analysis; group comparison; cluster detection; outlier hunting.
- Bonus: Regression diagnostics.
- Package freqparcoord on CRAN.
- Location of these slides: http://heather.cs.ucdavis.edu/freqparcoord/ BosSlides.pdf