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?

- 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?

- 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?

- 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

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

Example: Height/weight/age data.

Norm Matloff and Yingkang Xie
University of California at Davis

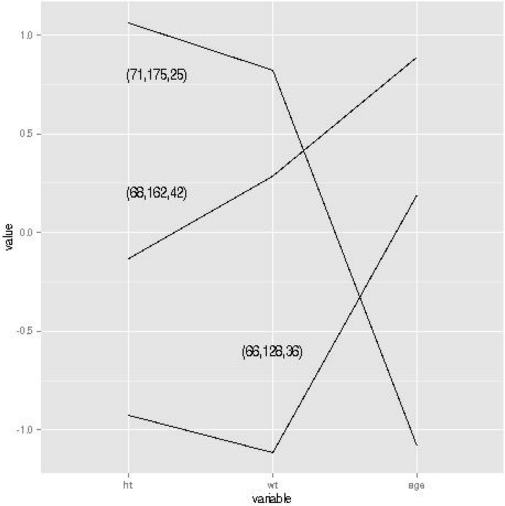
e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu
R/stat blog: matloff.wordpress.com

Example

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(..
...
```

Vertical axes use centered, scaled values.



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

- 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

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

- 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

- 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"?

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.

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

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

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

Outlier Hunting First

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

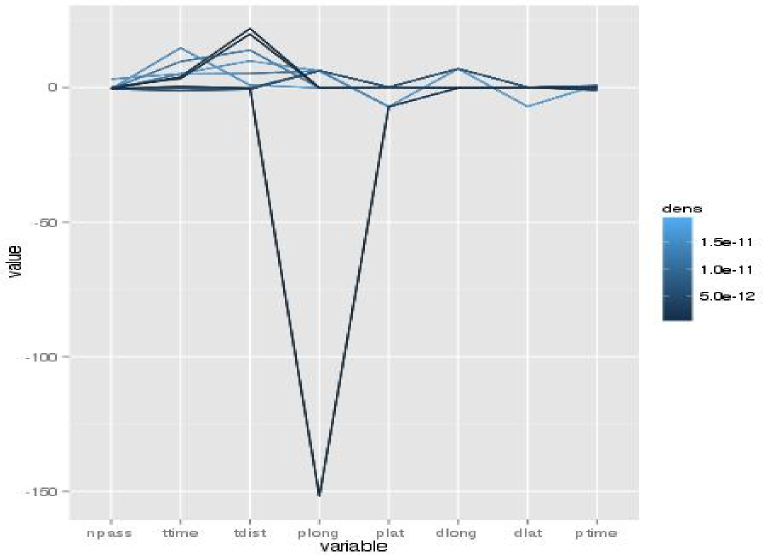
```
p <- freqparcoord(d100,-10,c(8:15),keepidxs=8)
```

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)
```



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.

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.

```
> p$xdisp[,11:14]
        plong      plat     dlong     dlat
    -74.00399 40.742107 -73.94696 40.81335
      0.00000  0.000000 -73.96590 40.80481
    -74.00748 40.703709 -74.07885 40.43142
      0.00000 40.783333   0.00000 40.79044
      0.00000 40.835121   0.00000 40.84693
      0.00000 40.733334   0.00000 40.74148
    -73.88925 40.769035 -73.94363 40.75264
  -1837.04530  0.041667 -73.96226 40.76774
    -73.98628 40.752365 -73.77634 40.64601
      0.00000  0.000000   0.00000  0.00000
```

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.

```
> p$xdisp[,11:14]
        plong       plat     dlong      dlat
    -74.00399  40.742107 -73.94696  40.81335
      0.00000   0.000000 -73.96590  40.80481
    -74.00748  40.703709 -74.07885  40.43142
      0.00000  40.783333   0.00000  40.79044
      0.00000  40.835121   0.00000  40.84693
      0.00000  40.733334   0.00000  40.74148
    -73.88925  40.769035 -73.94363  40.75264
  -1837.04530   0.041667 -73.96226  40.76774
    -73.98628  40.752365 -73.77634  40.64601
      0.00000   0.000000   0.00000   0.00000
```

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

Outliers, cont'd.

```
> p$xdisp[,11:14]
         plong      plat      dlong      dlat
     -74.00399 40.742107  -73.94696  40.81335
       0.00000  0.000000  -73.96590  40.80481
     -74.00748 40.703709  -74.07885  40.43142
       0.00000 40.783333    0.00000  40.79044
       0.00000 40.835121    0.00000  40.84693
       0.00000 40.733334    0.00000  40.74148
     -73.88925 40.769035  -73.94363  40.75264
   -1837.04530  0.041667  -73.96226  40.76774
     -73.98628 40.752365  -73.77634  40.64601
       0.00000  0.000000    0.00000   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.

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.

```
> p$xdisp[,11:14]
       plong       plat     dlong     dlat
   -74.00399  40.742107 -73.94696 40.81335
     0.00000   0.000000 -73.96590 40.80481
   -74.00748  40.703709 -74.07885 40.43142
     0.00000  40.783333   0.00000 40.79044
     0.00000  40.835121   0.00000 40.84693
     0.00000  40.733334   0.00000 40.74148
   -73.88925  40.769035 -73.94363 40.75264
 -1837.04530   0.041667 -73.96226 40.76774
   -73.98628  40.752365 -73.77634 40.64601
     0.00000   0.000000   0.00000  0.00000
```

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

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

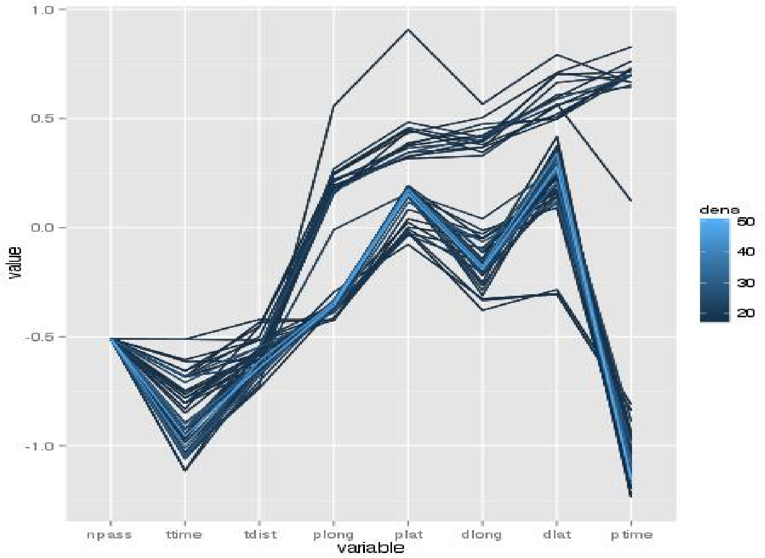
e-mail: matloff@cs.ucdavis.edu, ykxie@ucdavis.edu
R/stat blog: matloff.wordpress.com

General Analysis

```
p <- freqparcoord(d100,50,c(8:15),keepidxs=8)
```

General Analysis

```
p <- freqparcoord(d100,50,c(8:15),keepidxs=8)
```



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

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

General Analysis, cont'd.

What do we see?

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.

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.

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

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

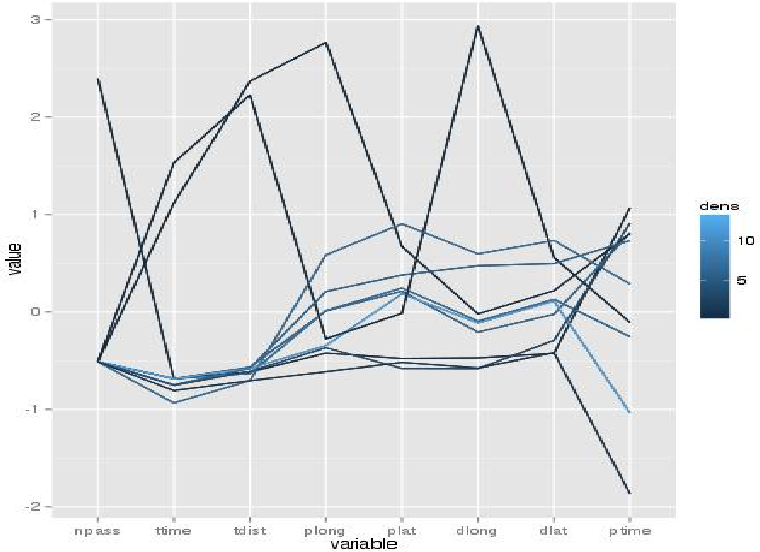
```
p <- freqparcoord(d100,1,c(8:15),method="locmax",klm=1000,
   cls=cl4,keepidxs=15)
```

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

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



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

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

- 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

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

Clustering, cont.d

- 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:

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

- 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

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

Clustering, cont.d

- 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 → La Guardia Airport.
  - Around 7:30 p.m., La Guardia Airport → mid-Manhattan.

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

- 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 → La Guardia Airport.
  - Around 7:30 p.m., La Guardia Airport → 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

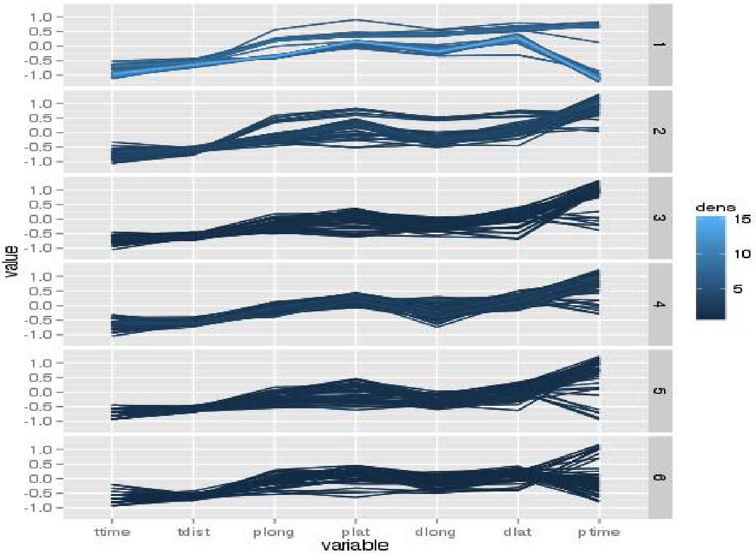
```
p <- freqparcoord(d100,50,c(9:15),grpvar=8)
```

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)
```



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.

- 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.

- 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

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

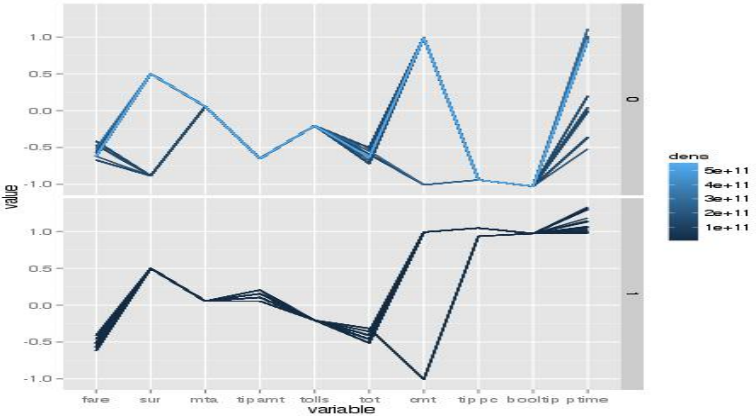
```
p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)
```

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

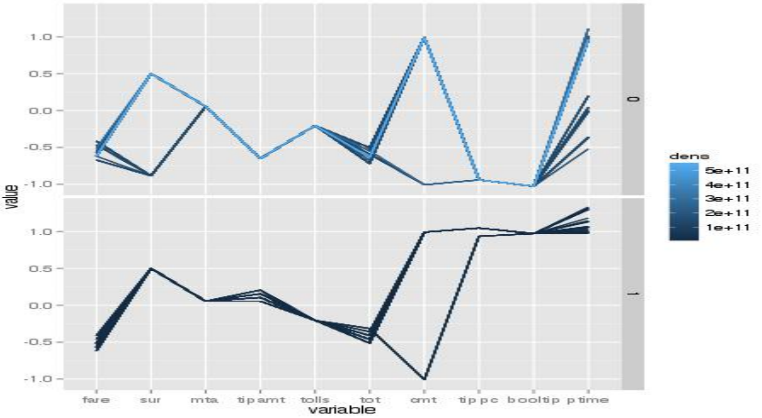
Credit Card vs. Cash

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



Credit Card vs. Cash

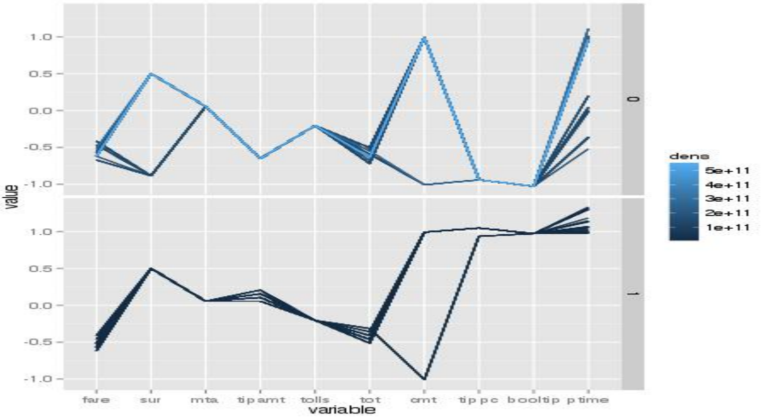
```
p <- freqparcoord(fare100,10,c(6,7,9:12,14:17),grpvar=13)
```



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

Credit Card vs. Cash

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



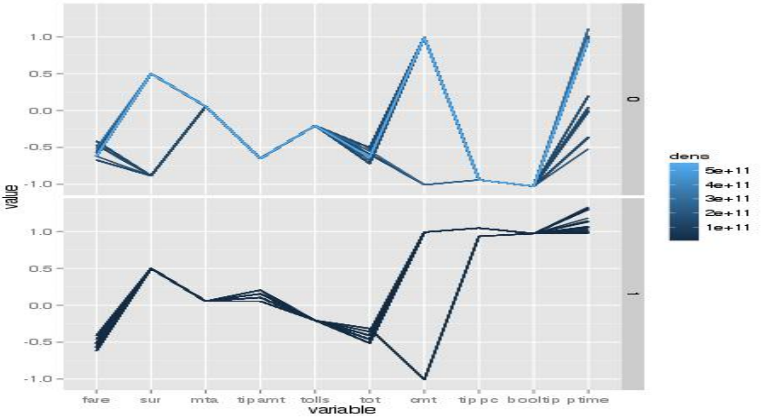
Not much difference, e.g. in base fare. Some difference in time of day.

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

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



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

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

- Compute *divergences* (not 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

- Compute *divergences* (not residuals):

$$div_i = param\_est_i \text{ - } nonparam\_est_i$$

- 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* (not 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* (not residuals):

  $$div_i = param\_est_i \text{ - } nonparam\_est_i$$

- 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

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

- A new approach to 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

Conclusions

- A new approach to parallel coordinates.
- Key point: Plots 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

Conclusions

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

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

- 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.

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

- 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.

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

- 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.

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

- 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`