Unsupervised Learning

Thomas Nichols

Dept. of Statistics & WMG
University of Warwick

Outline

- Data Menu
- Multivariate Data Examples
- Principal Components Analysis
- Multidimensional Scaling

Data Menu: Univariate

- Univariate Data
  - Each unit/subject measured once
- Examples
  - Car fuel efficiency
    - Y_i: MPG of a car model i
    - X_i: Weight of car model i, HP, etc.
  - Factors influencing childhood BMI (cross-sectionally)
    - Y: BMI of child at age 5
    - X: Activity level, dietary factors, SES, etc…

Data Menu: Longitudinal

- Longitudinal Data
  - Each unit/subject measured more than once
    - Usually over time
    - Same variable measured in each instance
  - Usually ‘messy’
    - Differing number of measurements (i.e. drop out)
    - Measurement times differ by subjects
- Examples
  - Factors influencing childhood BMI over time
    - Y: BMI of child at ages 2-7, every ~12 months
      - Exact age varies, spacing not exactly 12 months
    - X: Age in months, Activity level, Dietary factors, SES, etc
      - Measured at same time as BMI

Data Menu: Repeated Measures

- Repeated Measures Data
  - Each unit/subject measured more than once
    - Usually in a single session
    - Same variable measured in each instance
  - Usually ‘neater’
    - Same number of measurements (usually)
    - Measurement over aligned over units/subjects
      - “measurement 1” means same thing for all units
      - But ‘imbalanced’ data can often be accomodated
- Examples
  - Response times in an emotion processing experiment
    - Subjects flashed 50 images of human faces, one at a time
      - M & F, ranging from neutral to angry expressions
      - Must identify gender as “M” or “F”, as quickly as possible
    - Y_{ij}: Response time for subject i for face j
    - X_{ij}: Degree of “anger” in facial expression

Data Menu: Multivariate

- Multivariate Data
  - Each unit/subject measured more than once
    - May be same variable measured in each instance
    - Often completely different variables
  - Must be ‘neat’
    - Same number of measurements per subject
    - Missing data is huge pain for multivariate methods
    - “Neat” repeated measures and longitudinal data is compatible with multivariate methods
  - Often no clear role response/dependent variable
    - Rather, simply want to understand relationship between a ‘bag’ of variables

Multivariate Data Examples

- "Quality of life" scoring of cities/regions
  - For each city/region, measurement of
    - Housing affordability, Crime, Health Care, Transportation, Education, etc…
- Morphometry
  - Lengths of different animal anatomy, plant structure
- Comparison of products
  - Eg. breakfast cereals, each measured on:
    - calories, protein, fat, sodium, fibre, sugars, vitamins
  - No one explanatory variable, just trying to understand how these variables interrelate

Example Data

- US Crime data
  - Arrests per 100,000 residents, by state, in 1973
  - Crimes: Assault, murder, rape
  - Other: Percent population in urban area

	Murder	Assault	Rape	UrbanPop
Alabama	13.2	236	21.2	58
Alaska	10.0	263	44.5	48
Arizona	8.1	294	31.0	80
Arkansas	8.8	190	19.5	50
California	9.0	276	40.6	91
Colorado	7.9	204	38.7	78

Multivariate EDA

- Usual summary…

```
> summary(USArrests)
 Murder          Assault          Rape          UrbanPop
 Min.   : 0.800   Min.   : 45.0   Min.   : 7.30   Min.   :32.00
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:15.07   1st Qu.:54.50
 Median : 7.250   Median :159.0   Median :20.10   Median :66.00
 Mean   : 7.788   Mean   :170.8   Mean   :21.23   Mean   :65.54
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:26.18   3rd Qu.:77.75
 Max.   :17.400   Max.   :337.0   Max.   :46.00   Max.   :91.00
```

  – Tells us nothing about interrelationships

Multivariate EDA

- Covariance

```
> cov(USArrests)
             Murder   Assault      Rape   UrbanPop
Murder    18.970465  291.0624  22.99141   4.386204
Assault  291.062367 6945.1657 519.26906 312.275102
Rape      22.991412  519.2691  87.72916  55.768082
UrbanPop   4.386204  312.2751  55.76808 209.518776
```

  – Mainly shows Assault most variable

- Correlation

```
> cor(USArrests)
             Murder   Assault      Rape   UrbanPop
Murder   1.00000000 0.8018733 0.5635788 0.06957262
Assault  0.80187331 1.0000000 0.6652412 0.25887170
Rape     0.56357883 0.6652412 1.0000000 0.41134124
UrbanPop 0.06957262 0.2588717 0.4113412 1.00000000
```

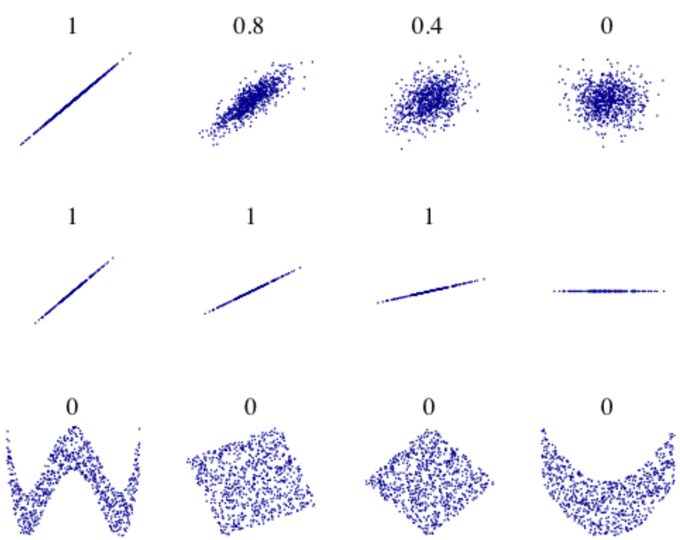
  – Can see Murder and Assault most closely related

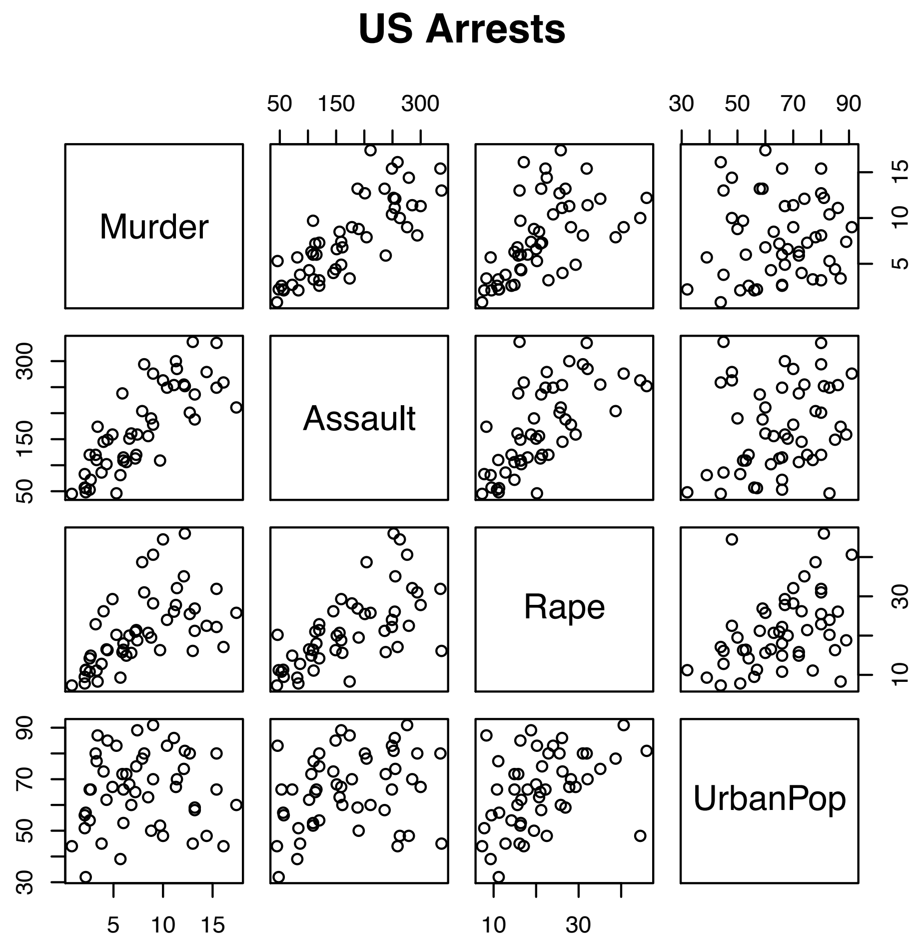
Multivariate EDA

- Scatter plots

```
> pairs(USArrests)
```

  - Essential for gauging strength of linear relationship, role of outliers





Data Reduction with SVD

- SVD is Singular Valuate Decomposition
  - Usually won't need to use directly
- Any matrix X can be decomposed into

  $X = U \Lambda V'$

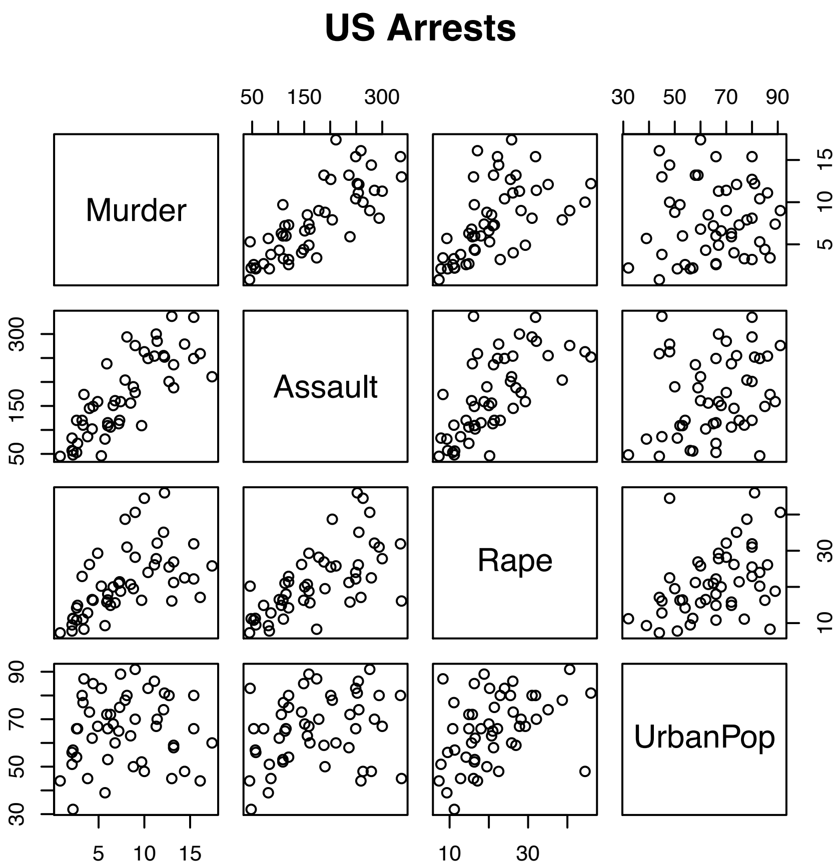
    - Columns of U are left eigenvectors
    - Λ is diagonal matrix of eigenvalues
    - Columns of V are the right eigenvectors

  $X = \Sigma_i \ \lambda_i u_i v_i'$

    - Any matrix X can be written as sum of simpler matrices
    - Weights λ_i determine the importance of each constituent matrix
    - $\lambda_1 u_1 v_1'$ explains most variance, then $\lambda_2 u_2 v_2'$, etc…

Crime Data: Original

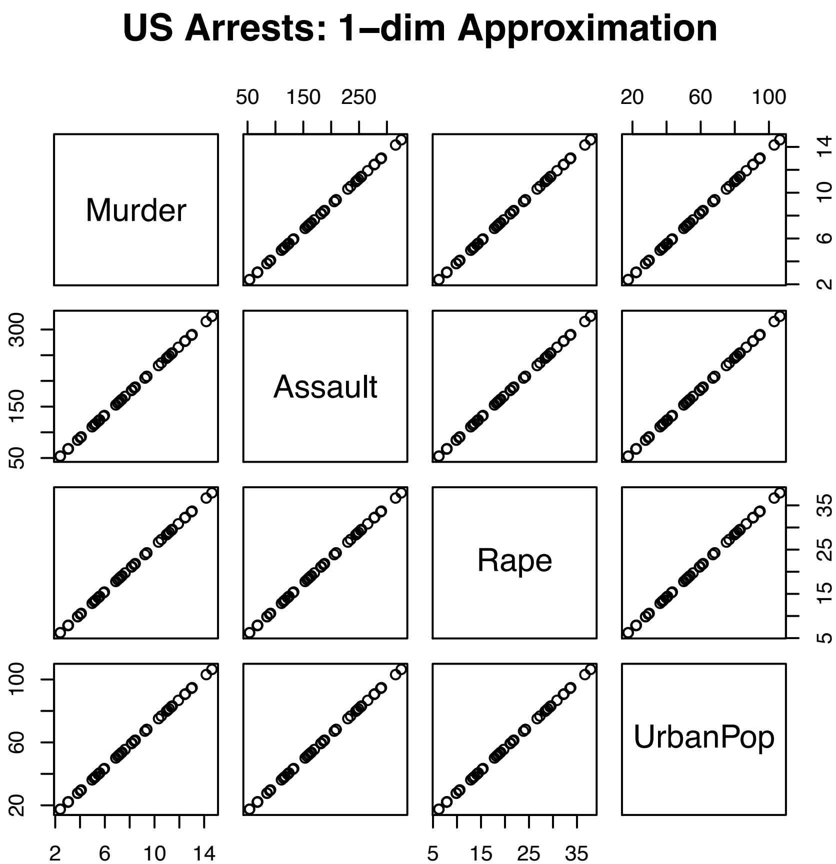
- No approximation



	Murder	Assault	Rape	UrbanPop
Alabama	13.2	236	21.2	58
Alaska	10.0	263	44.5	48
Arizona	8.1	294	31.0	80
Arkansas	8.8	190	19.5	50
California	9.0	276	40.6	91
Colorado	7.9	204	38.7	78

Crime Data: Rank-1 Approximation

- Approximation Data = $\lambda_1 u_1 v_1'$

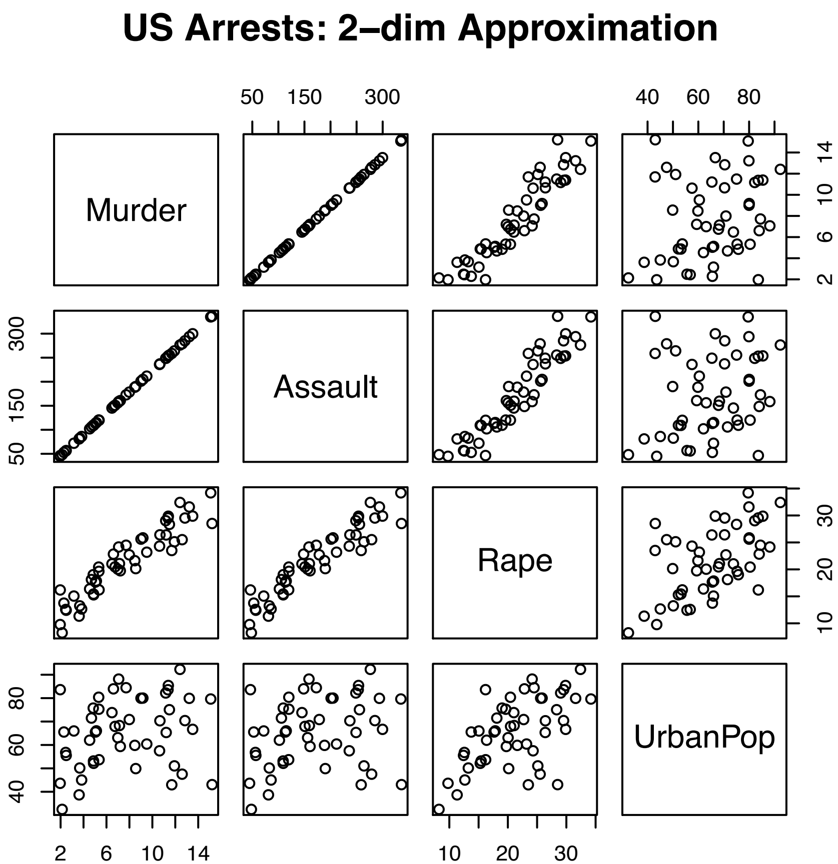


	Murder	Assault	Rape	UrbanPop
Alabama	10.324380	229.8975	26.70182	75.11650
Alaska	11.376607	253.3279	29.42318	82.77212
Arizona	12.969339	288.7940	33.54245	94.36028
Arkansas	8.363235	186.2278	21.62973	60.84791
California	12.439109	276.9871	32.17112	90.50251
Colorado	9.377173	208.8056	24.25207	68.22496

Error:
Total variance unexplained = 10.1%

Crime Data: Rank-2 Approximation

- Approximation Data = $\lambda_1 u_1 v_1' + \lambda_2 u_2 v_2'$

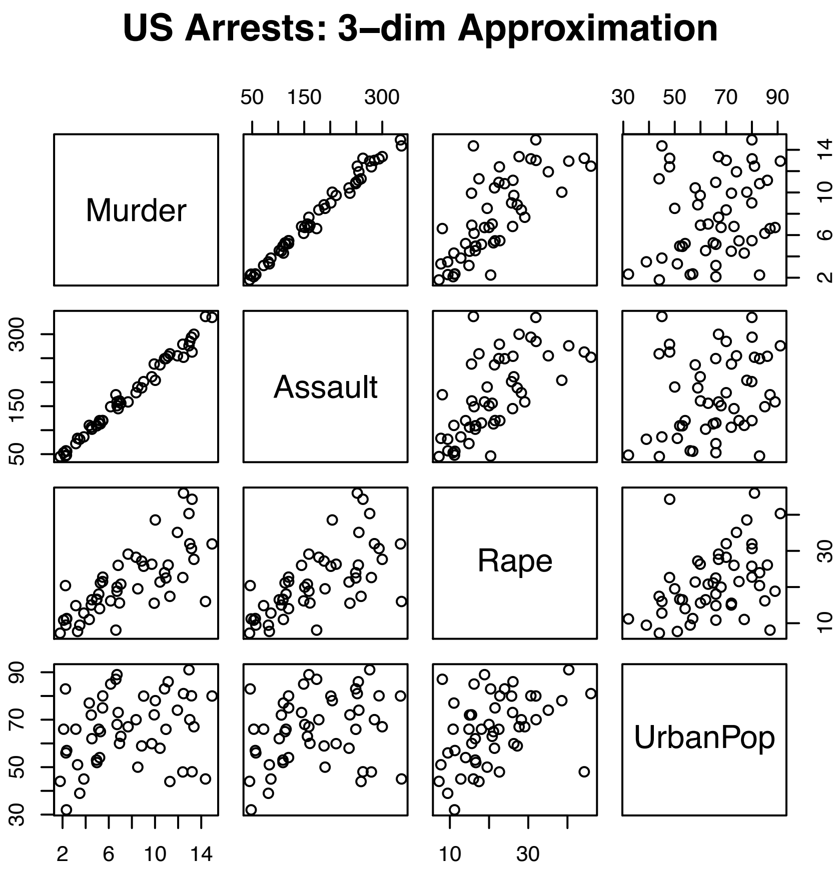


	Murder	Assault	Rape	UrbanPop
Alabama	10.627700	235.9157	24.31363	57.50469
Alaska	11.922791	264.1648	25.12279	51.05874
Arizona	13.218096	293.7296	31.58386	79.91657
Arkansas	8.551822	189.9696	20.14489	49.89789
California	12.408212	276.3741	32.41439	92.29650
Colorado	9.173896	204.7723	25.85258	80.02800

Error:
Variance unexplained = 0.09%

Crime Data: Rank-3 Approximation

- Approximation Data = $\lambda_1 u_1 v_1' + \lambda_2 u_2 v_2' + \lambda_3 u_3 v_3'$

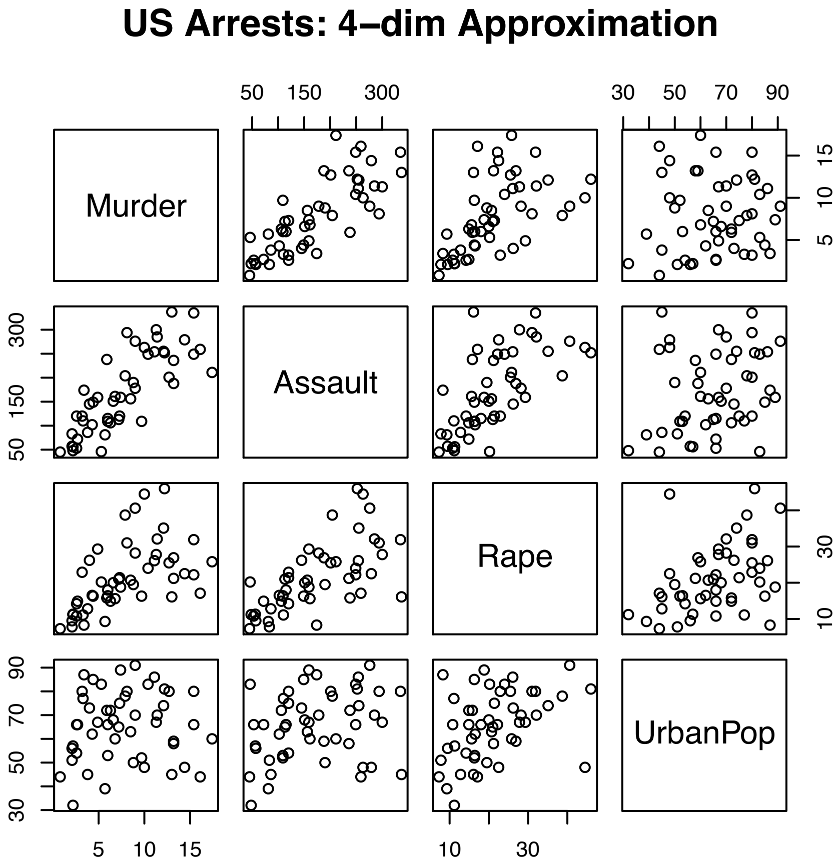


	Murder	Assault	Rape	UrbanPop
Alabama	10.431690	236.1137	21.38775	57.96572
Alaska	13.206333	262.8683	44.28255	48.03970
Arizona	13.156011	293.7923	30.65710	80.06261
Arkansas	8.509937	190.0119	19.51967	49.99641
California	12.938684	275.8382	40.33288	91.04877
Colorado	10.024910	203.9127	38.55589	78.02631

Error:
Variance unexplained = 0.68%

Crime Data: Rank-4 Approximation

- Approximation Data = $\lambda_1 u_1 v_1' + \lambda_2 u_2 v_2' + \lambda_3 u_3 v_3' + \lambda_4 u_4 v_4' = X$



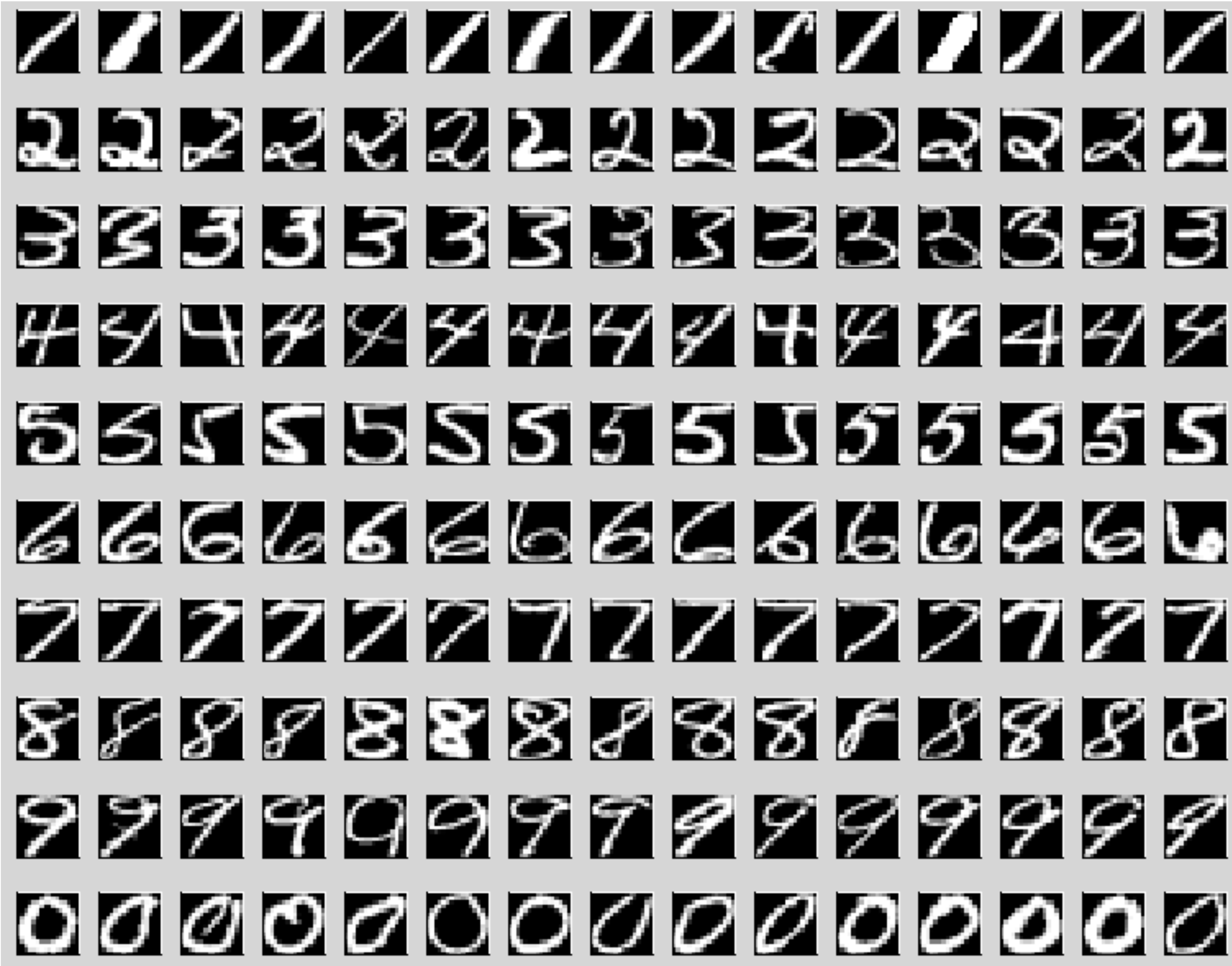
	Murder	Assault	Rape	UrbanPop
Alabama	13.2	236	21.2	58
Alaska	10.0	263	44.5	48
Arizona	8.1	294	31.0	80
Arkansas	8.8	190	19.5	50
California	9.0	276	40.6	91
Colorado	7.9	204	38.7	78

Error:
Variance unexplained = 0%

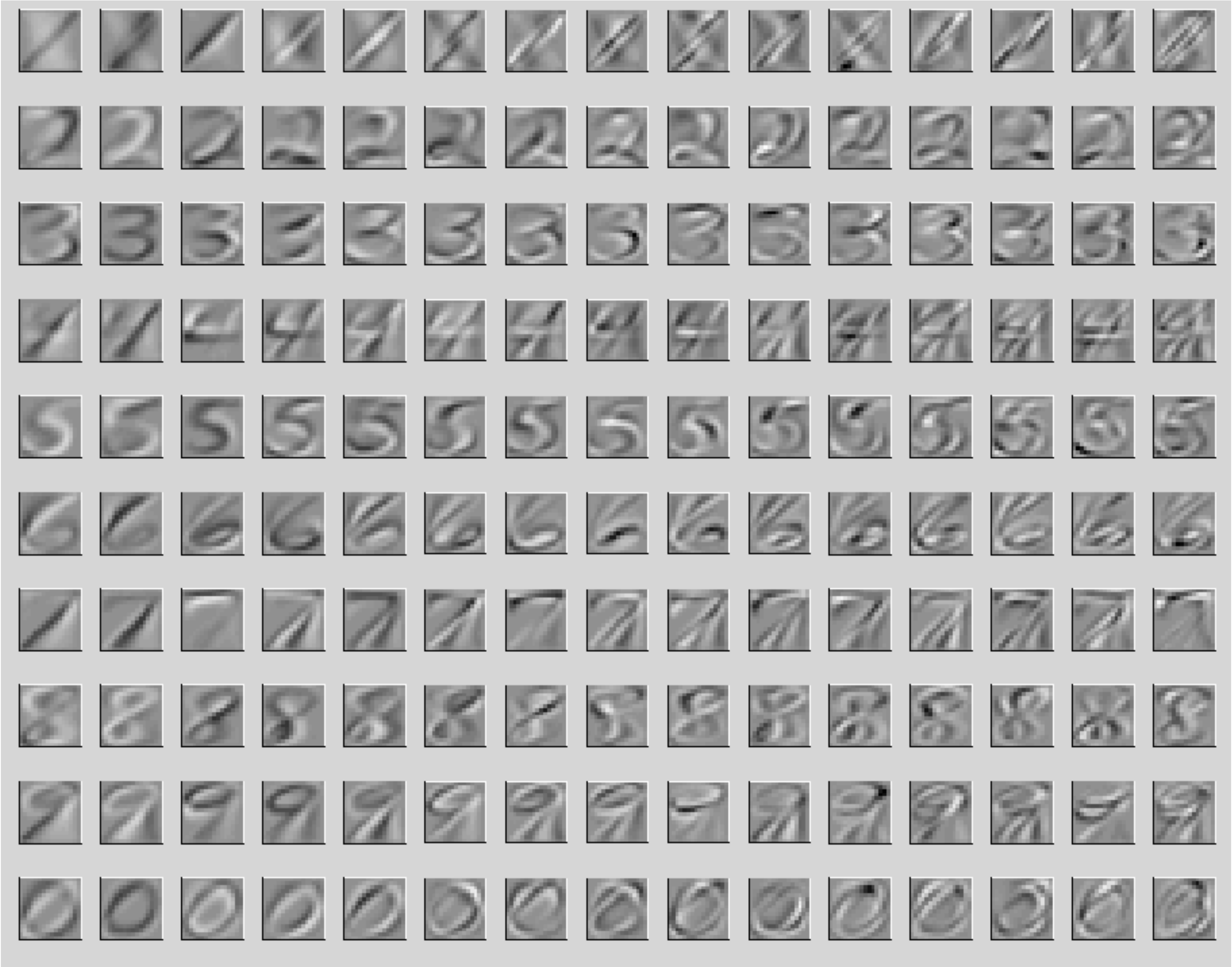
Application: Handwritten Digits

- PCA on handwritten digits
  - Length-256 data vectors (16×16 pixel grayscale images)
  - Full data has 1,100 cases on each of 10 digits
- Data reduction
  - Do we really need 256 dimensions to represent each observation?
  - How many do we need?

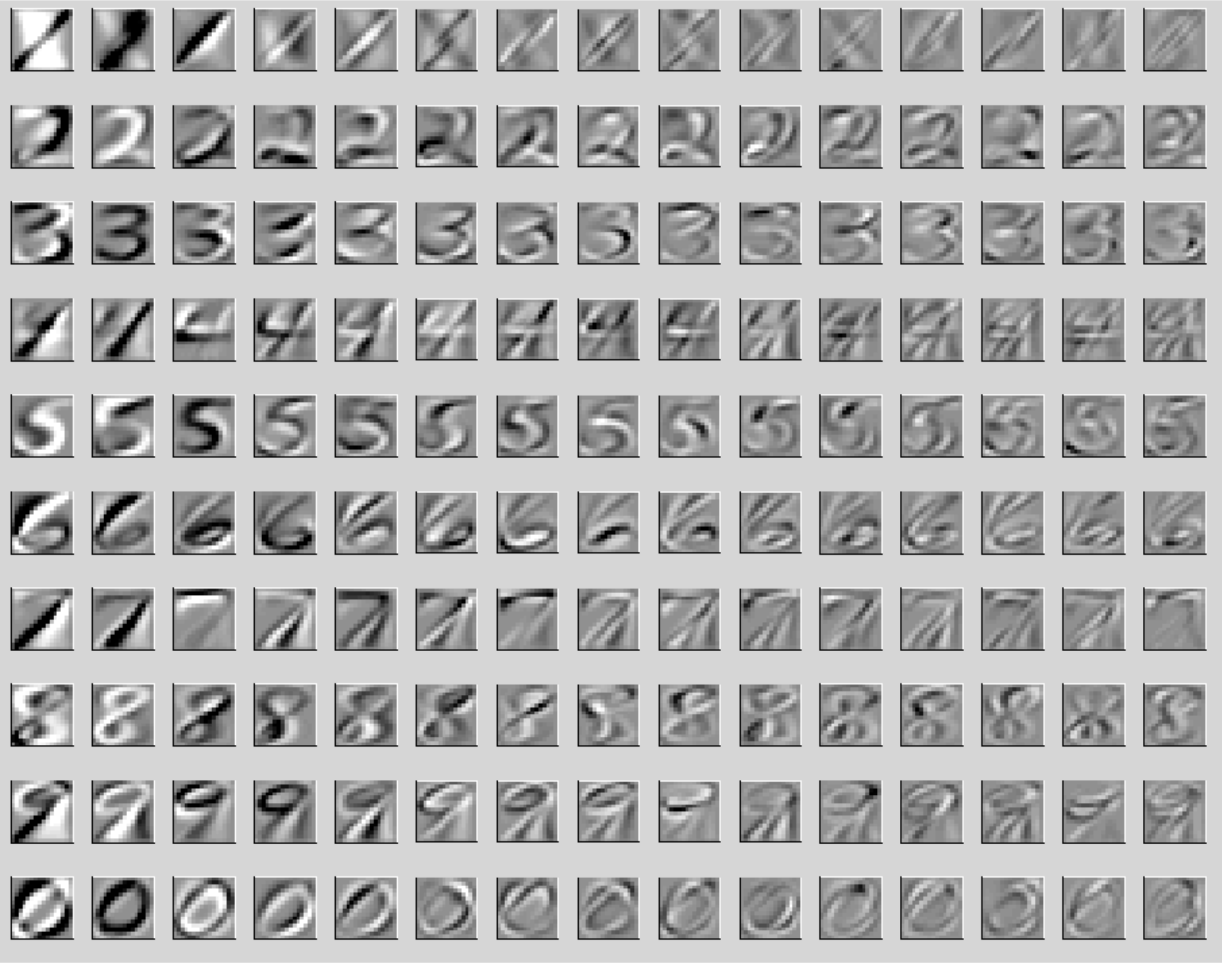
Digits: First 15 cases of 1,100



Digits: First 15 eigenvectors of 1,100

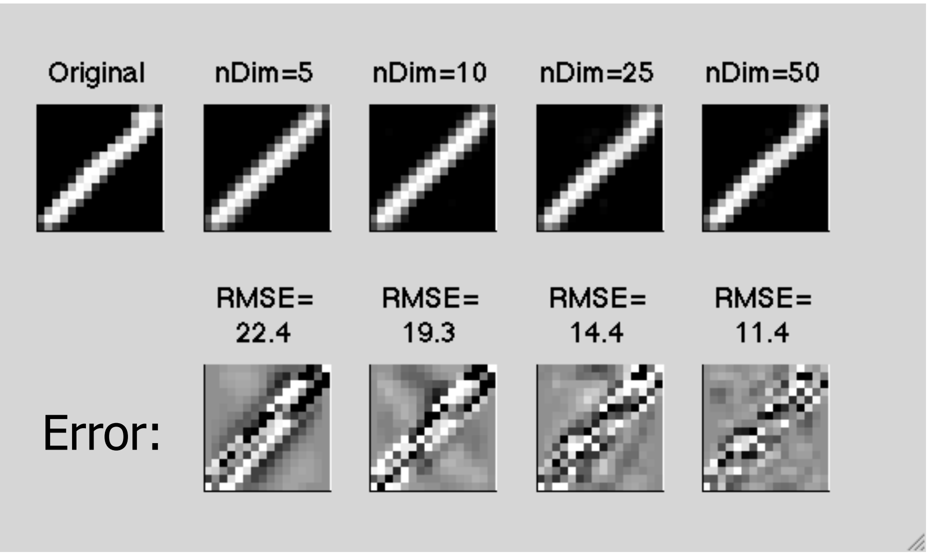


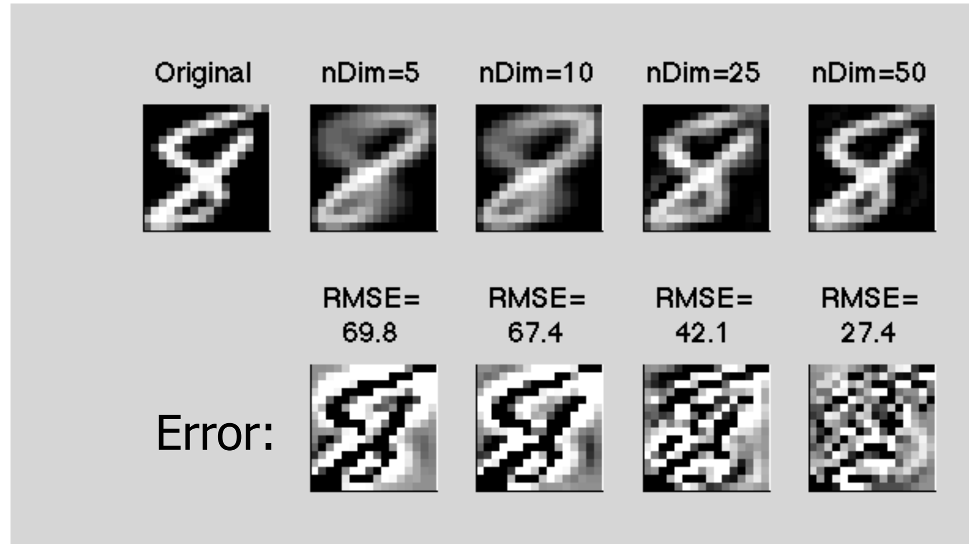
Eigenvectors scaled by $\sqrt{\lambda_j}$

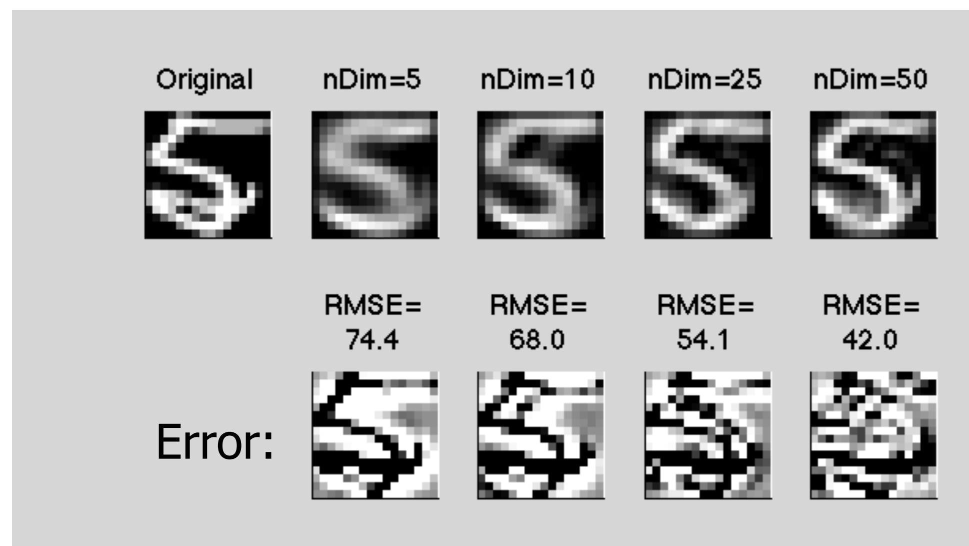


Recall sample covariance of $\mathbf{U}'\mathbf{X}$ for $k=d$?

Approximations of varying k







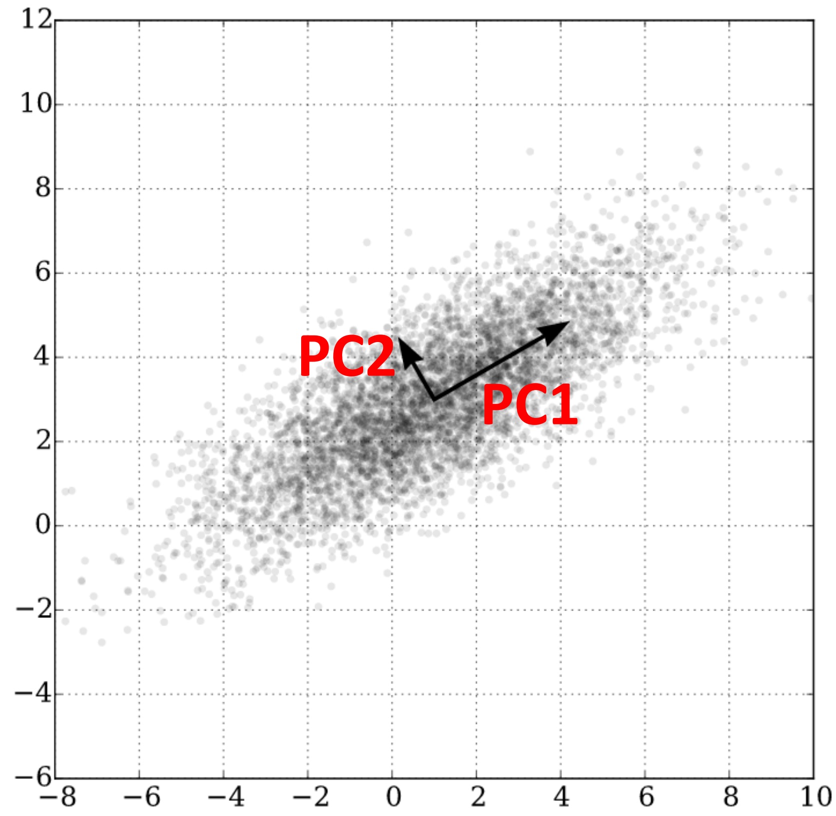
Error images intensity range displayed: [-25, 25]

SVD Redux

- Generally don’t use SVD directly
  - But clearly shows how a data matrix can be summarized
  - That there is ‘latent’ structure that can be exploited

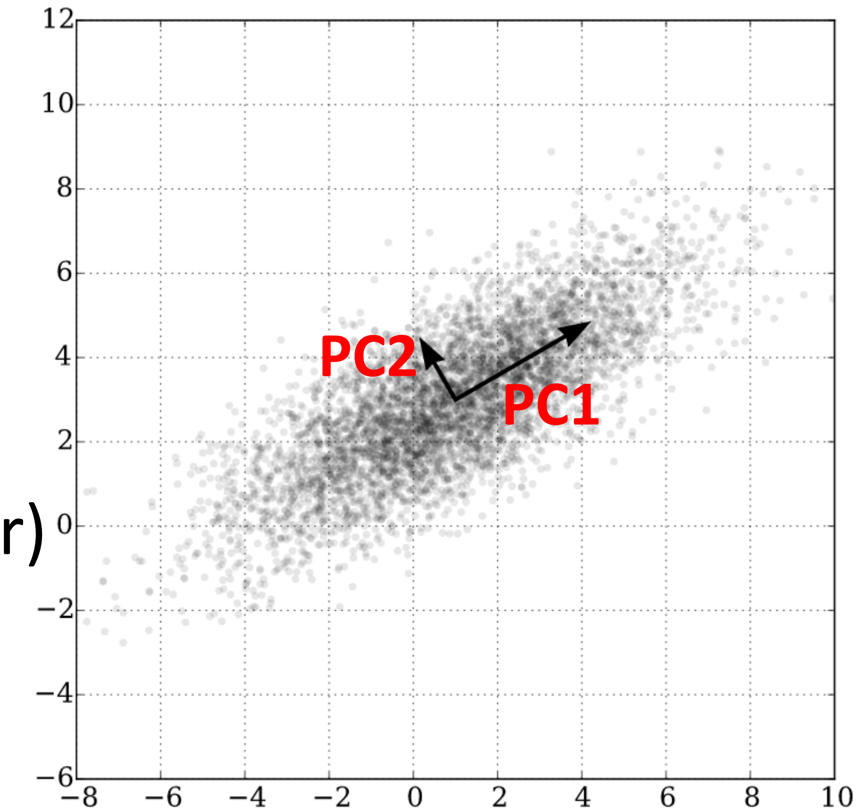
Principal Components Analysis

- Uses covariance or correlation to find the latent structures in the data
- Often don't measure "the right" thing
  - But maybe some 'intrinsic', latent variables exist



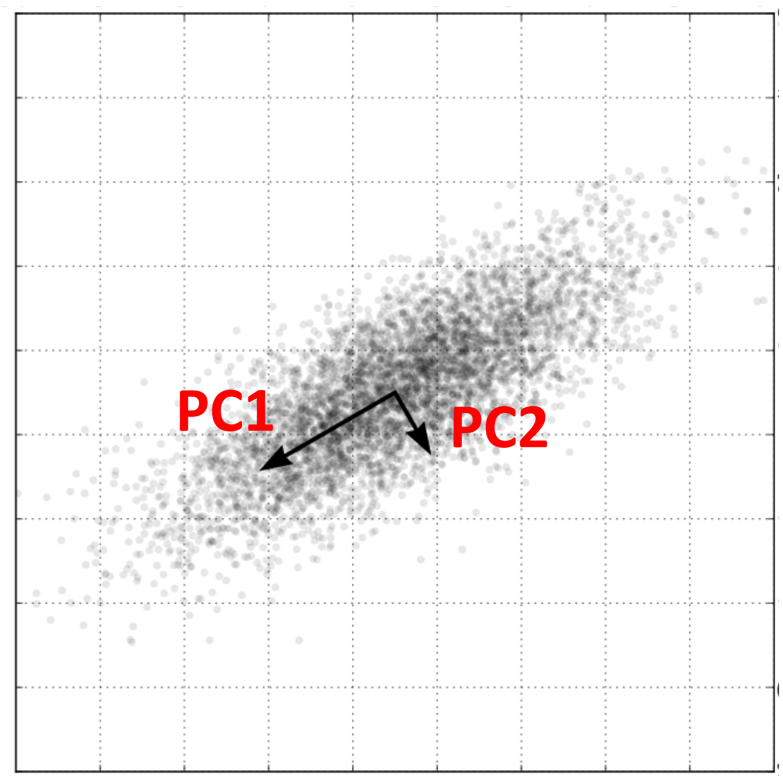
Principal Components Analysis

- Principal Component
  - Variables weights (each a length-$n_{variable}$ vector)
  - First PC is direction in variable space that explains most var.
- Loadings
  - Subject/case weights (each a length n_{subj} vector)
  - How each case “loads” onto the PC



Principal Components Analysis

- Note, sign is arbitrary
  - Sign can flip on the PC
  - Exact same variance explained
- Must interpret PC's with this in mind



PCA in practice

- First, must decide between correlation and covariance
  - Covariance
    - Importance of each variable given by variance
    - As seen in SVD example, one variable can dominate
    - Only makes sense if all variables have equal units, deserve equal weighting despite differences in variance
  - Correlation
    - Use when units not comparable
      - E.g. Arrests per 100,000 residents, vs % pop in urban area
    - Or when comparable units, but variance different
      - E.g. Arrests per 100,000 residents for assault vs murder

PCA on Crime Data

```
> fit <- princomp(USArrests, cor=TRUE)
> summary(fit)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.5748783 0.9948694 0.5971291 0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
```

- First PC accounts for 62% of variance (in correlation, variance normalised data), second 24.7%.
- Interpretation?
  - PC1 – Average
  - PC2 Murder>Pop

```
> loadings(fit) Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
Murder   -0.536  0.418  0.341 -0.649
Assault  -0.583  0.188  0.268  0.743
Rape     -0.543 -0.167 -0.818
UrbanPop -0.278 -0.873  0.378 -0.134
```

PCA on Crime Data

- Loadings
  - Tell you how each unit (state) aligns / weights on a PC

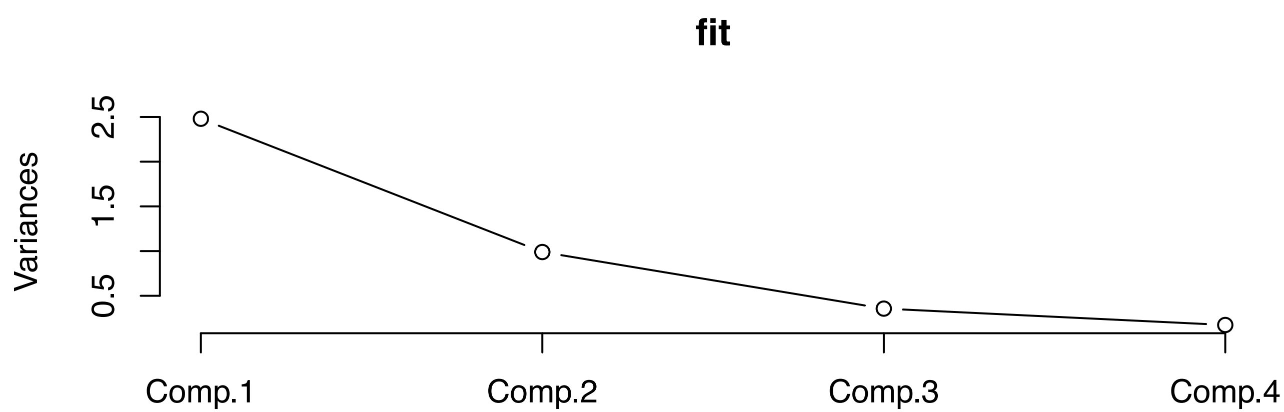
```
> fit$scores

                  Comp.1      Comp.2      Comp.3       Comp.4
Alabama      -0.98556588  1.13339238  0.44426879 -0.156267145
Alaska       -1.95013775  1.07321326 -2.04000333  0.438583440
Arizona      -1.76316354 -0.74595678 -0.05478082  0.834652924
Arkansas      0.14142029  1.11979678 -0.11457369  0.182810896
California   -2.52398013 -1.54293399 -0.59855680  0.341996478
Colorado     -1.51456286 -0.98755509 -1.09500699 -0.001464887
Connecticut   1.35864746 -1.08892789  0.64325757  0.118469414
Delaware     -0.04770931 -0.32535892  0.71863294  0.881977637
```

  - Much easier to visualise a PCA…

PCA on Crime Data

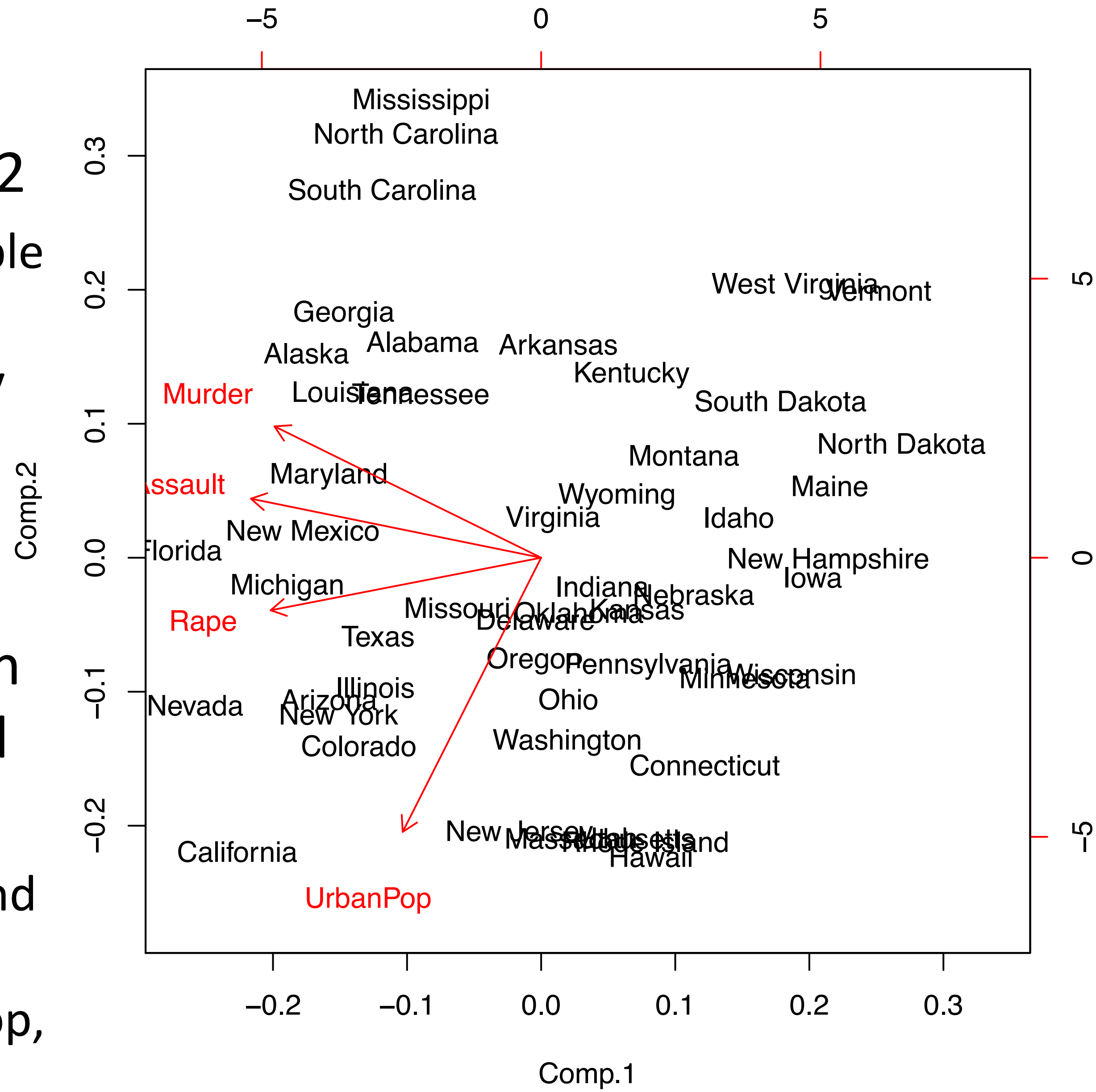
- “Screeplot”, shows variance explained by each PC



  – First two way more important than last one

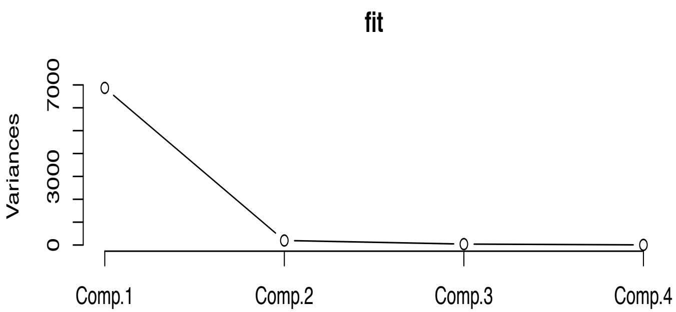
PCA on Crime Data

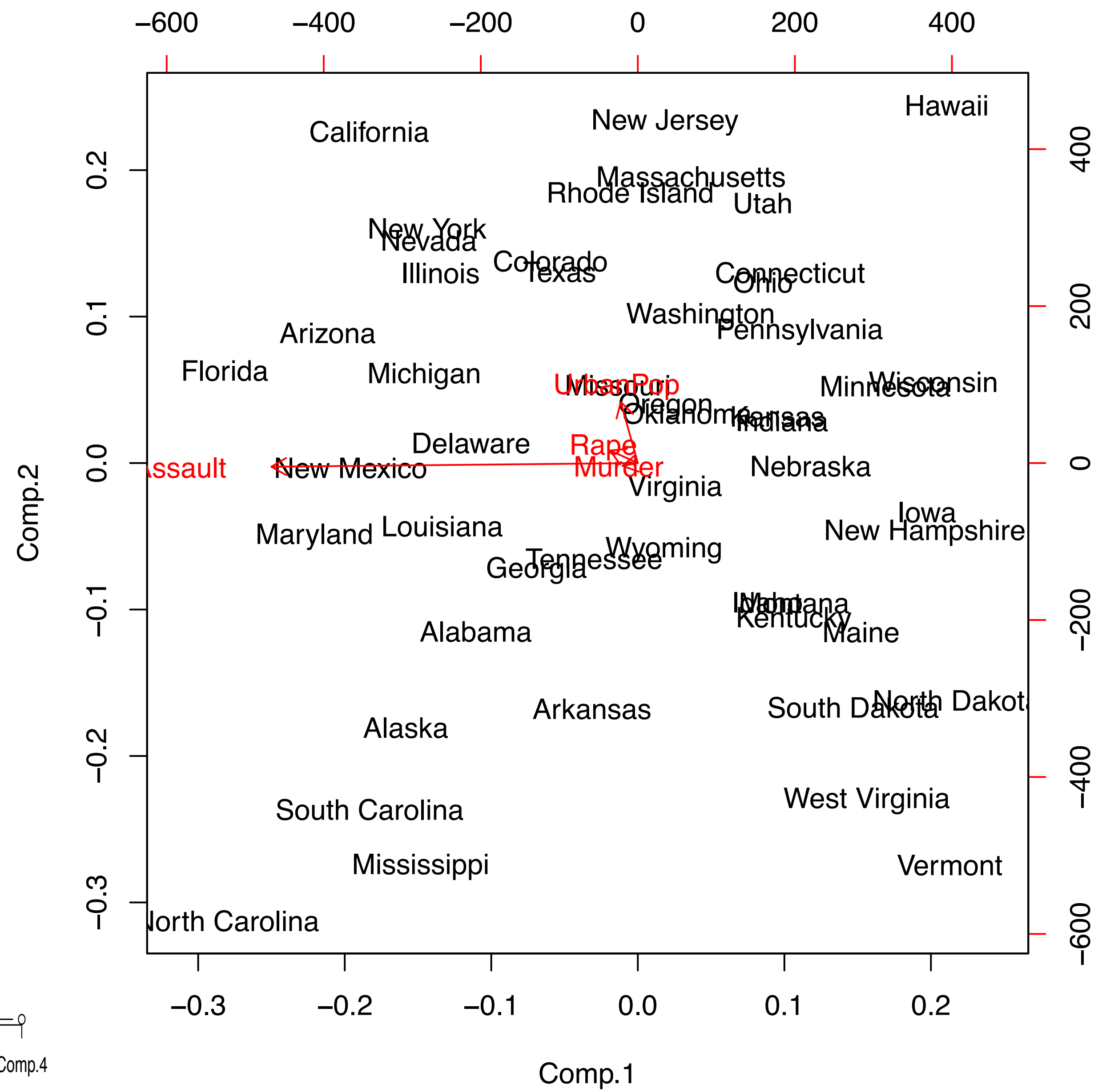
- “biplot”
  - Shows PC1 vs PC2
    - How each variable relates
  - Also shows case/ subject loadings
- Interpretation
  - Urban population is not too related to crime
    - E.g. Maryland and Indiana have similar Urban Pop, but diff crime



PCA on Crime Data

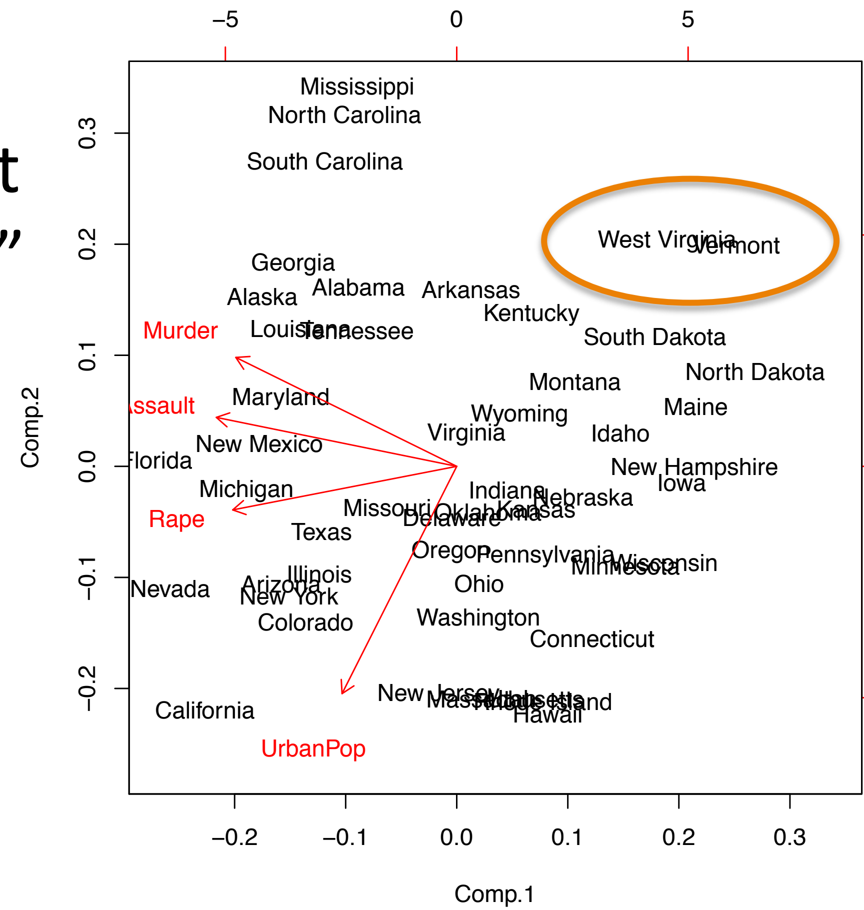
- What if used `cov` instead of `cor`?
  - High-variance `Assault` variable dominates





Multi-Dimensional Scaling Motivation

- The bi-plot is amazing
  - It lets us see how different units are similar in a "PCA" space
    - E.g. West Virginia & Vermont, 2 very different states, by similar…
    - But similar only in terms of PC1 & PC2
- But bi-plot only captures 2 dimensions

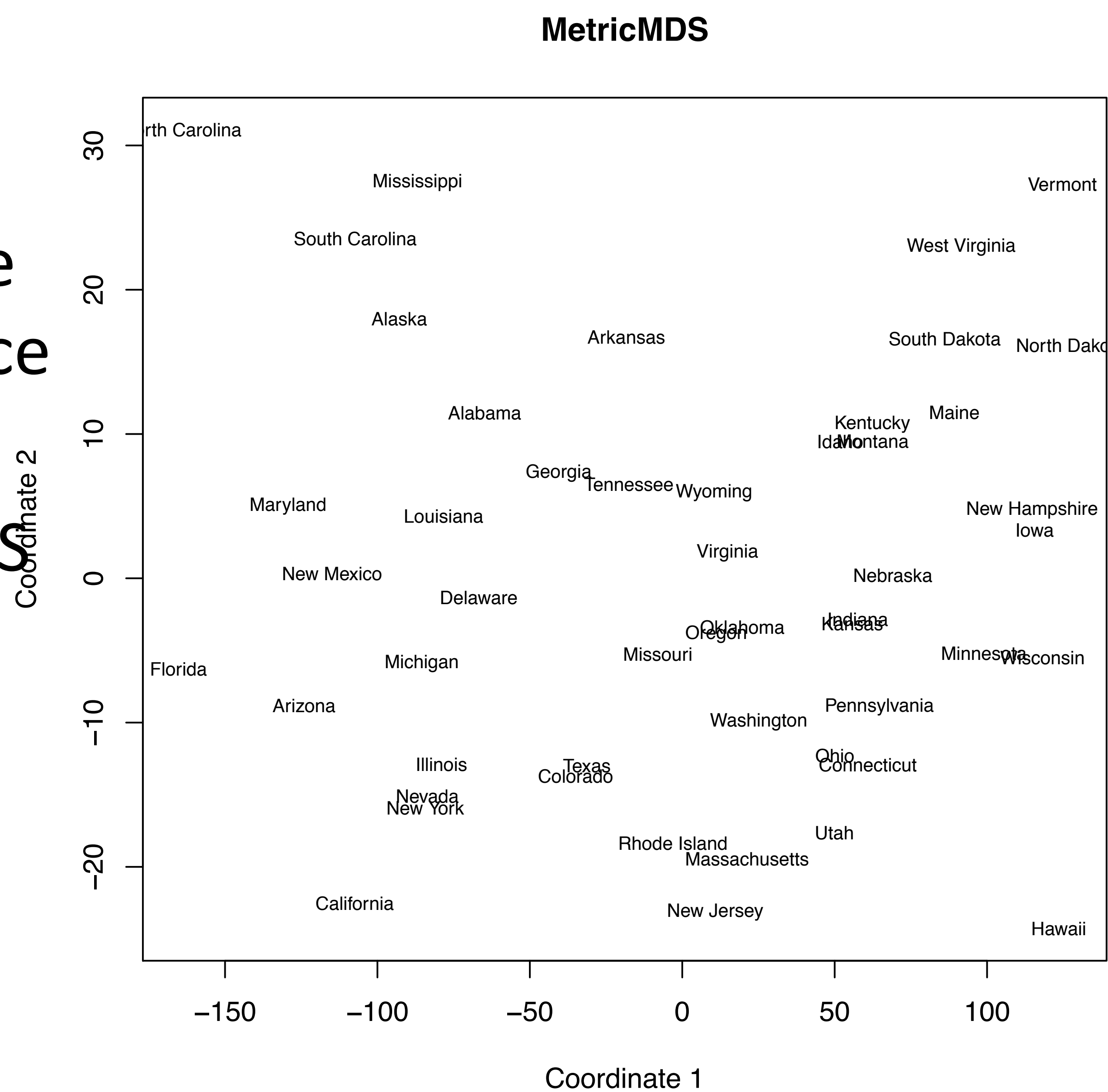


Multi-Dimensional Scaling Motivation

- MDS considers a general notion of distance between each case/unit
  - "Classical MDS" – Distance is Euclidean, like correlation – in full dimensional space
  - "Non-metric MDS" – A monotonic function of Euclidean distance
- Then makes a 2-D picture that accurately as possible captures that distance
  - i.e. distance between, e.g. WV & VT on 2D plane is as similar as possible to WV & VT in 4D variable space

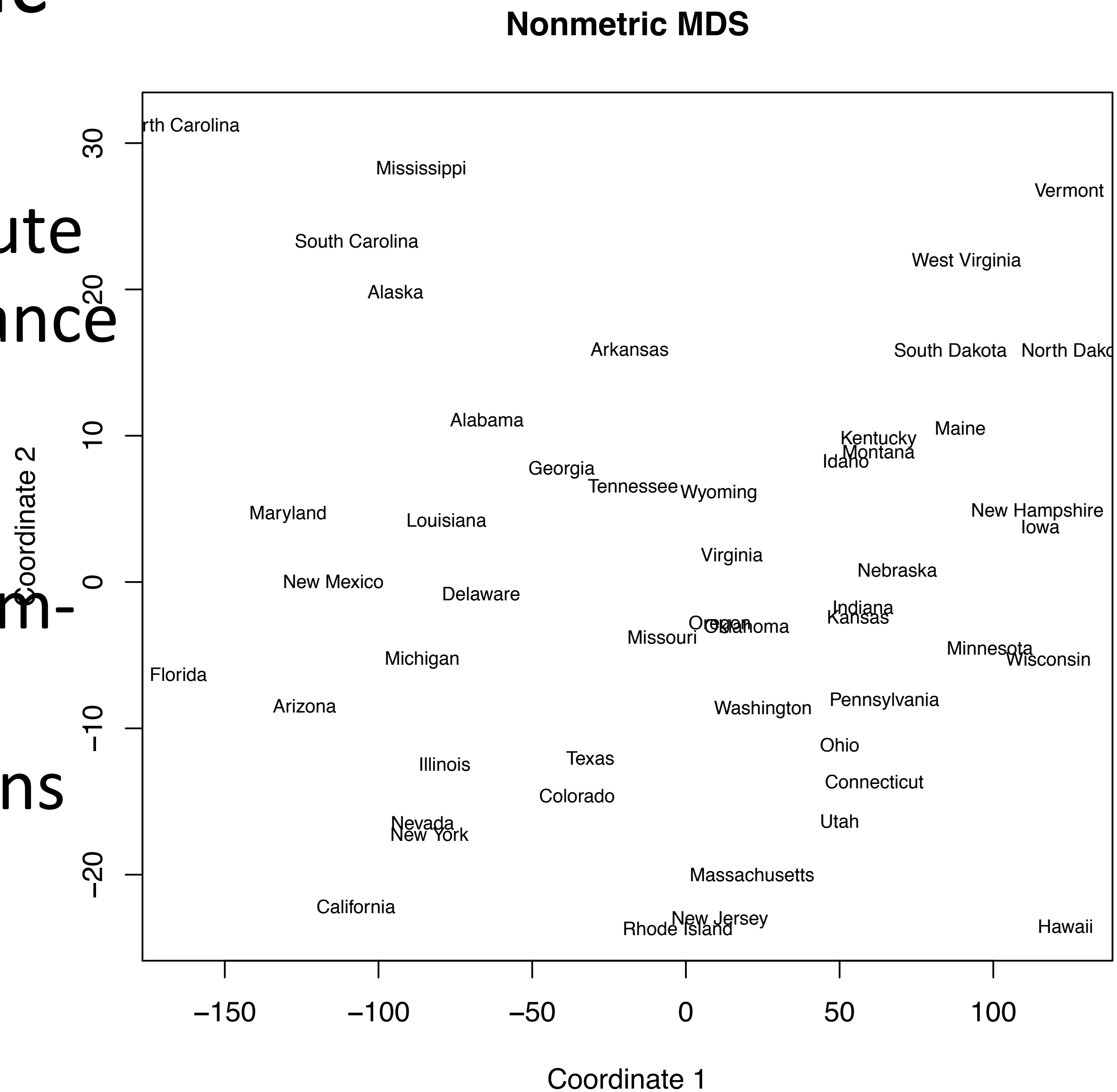
Classical MDS

- First compute 50x50 distance matrix
- Then run MDS for 2 dimensions
- Plot scores



Non-Metric MDS

- First compute 50x50 distance matrix (as before)
- Then run nm-MDS for 2 dimensions
- Plot scores



Conclusions

- Taster of three multivariate methods
  - SVD – For low-level, data reduction
  - PCA – For understanding latent structure
  - MDS – For understanding how different units/ subjects relate in a high-dim space
- Other important tools
  - Factor Analysis
    - Elaborations on PCA
  - Clustering – K-means, Hierachical