Statistical Data Mining and Machine Learning
Hilary Term 2016

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
`http://www.stats.ox.ac.uk/~sejdinov/sdmml`

Last time: PCA

PCA

Find an orthogonal basis $\{v_1, v_2, \ldots, v_p\}$ for the data space such that:

- The first principal component (PC) v_1 is the **direction of greatest variance** of data.
- The j-th PC v_j is the **direction orthogonal to** $v_1, v_2, \ldots, v_{j-1}$ **of greatest variance**, for $j = 2, \ldots, p$.

- Eigendecomposition of the sample covariance matrix $S = \frac{1}{n-1}\sum_{i=1}^n x_i x_i^\top$.

$$S = V\Lambda V^\top.$$

  - Λ is a diagonal matrix with eigenvalues (variances along each principal component) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$
  - V is a $p \times p$ orthogonal matrix whose columns are the p eigenvectors of S, i.e. the principal components $v_1, \ldots, v_p$
- Dimensionality reduction by projecting $x_i \in \mathbb{R}^p$ onto first k principal components:

$$z_i = \left[v_1^\top x_i, \ldots, v_k^\top x_i\right]^\top \in \mathbb{R}^k.$$

Eigendecomposition and PCA

$$S = \frac{1}{n-1}\sum_{i=1}^n x_i x_i^\top = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X}.$$

- S is a **real and symmetric** matrix, so there exist p eigenvectors $v_1, \ldots, v_p$ that are pairwise orthogonal and p associated eigenvalues $\lambda_1, \ldots, \lambda_p$ which satisfy the eigenvalue equation $Sv_i = \lambda_i v_i$. In particular, V is an orthogonal matrix:

$$VV^\top = V^\top V = I_p.$$

- S is a **positive-semidefinite** matrix, so the eigenvalues are non-negative:

$$\lambda_i \geq 0, \ \forall i.$$

Why is S symmetric? Why is S positive-semidefinite?
Reminder: A symmetric $p \times p$ matrix R is said to be positive-semidefinite if

$$\forall a \in \mathbb{R}^p, a^\top R a \geq 0.$$

Singular Value Decomposition (SVD)

SVD

Any real-valued $n \times p$ matrix $\mathbf{X}$ can be written as $X = UDV^\top$ where

- U is an $n \times n$ orthogonal matrix: $UU^\top = U^\top U = I_n$
- D is a $n \times p$ matrix with decreasing **non-negative** elements on the diagonal (the singular values) and zero off-diagonal elements.
- V is a $p \times p$ orthogonal matrix: $VV^\top = V^\top V = I_p$

- SVD **always** exists, even for non-square matrices.
- Fast and numerically stable algorithms for SVD are available in most packages.The relevant R command is `svd`.

SVD and PCA

- Let $\mathbf{X} = UDV^\top$ be the SVD of the $n \times p$ data matrix $\mathbf{X}$.
- Note that

$$(n-1)S = \mathbf{X}^\top\mathbf{X} = (UDV^\top)^\top(UDV^\top) = VD^\top U^\top UDV^\top = VD^\top DV^\top,$$

  using orthogonality ($U^\top U = I_n$) of U.
- The eigenvalues of S are thus the diagonal entries of $\Lambda = \frac{1}{n-1}D^\top D$.
- We also have

$$\mathbf{X}\mathbf{X}^\top = (UDV^\top)(UDV^\top)^\top = UDV^\top VD^\top U^\top = UDD^\top U^\top,$$

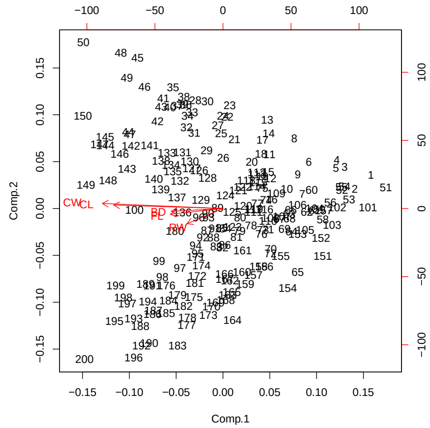
  using orthogonality ($V^\top V = I_p$) of V.

Gram matrix

$\mathbf{B} = \mathbf{X}\mathbf{X}^\top$, $\mathbf{B}_{ij} = x_i^\top x_j$ is called the Gram matrix of dataset $\mathbf{X}$.
$\mathbf{B}$ and $(n-1)S = \mathbf{X}^\top\mathbf{X}$ have the same nonzero eigenvalues, equal to the non-zero squared singular values of $\mathbf{X}$.

Biplots

```
> biplot(Crabs.pca,scale=1)
```



- PCA plots show the data items (rows of $\mathbf{X}$) in the space spanned by PCs.
- **Biplots** allow us to visualize the **original variables** $X^{(1)}, \dots, X^{(p)}$ (corresponding to columns of $\mathbf{X}$) in the same plot.

Biplots

Recall that $X = [X^{(1)}, \dots, X^{(p)}]^\top$ and $\mathbf{X} = UDV^\top$ is the SVD of the data matrix.

- The 'full' PC projection of x_i is the i-th row of UD:

$$z_i = V^\top x_i = D^\top U_i^\top, \text{equivalently: } \mathbf{X}V = UD.$$

- The j-th unit vector $\mathbf{e}_j \in \mathbb{R}^p$ points in the direction of the original variable $X^{(j)}$. Its PC projection η_j is:

$$\eta_j = V^\top \mathbf{e}_j = V_j^\top \qquad \text{(the } j\text{-th row of } V\text{)}$$

- The projection of $\mathbf{e}_j$ indicates the weighting each PC gives to the original variable $X^{(j)}$.
- Dot products between these projections give entries of the data matrix:

$$x_{ij} = \sum_{k=1}^{\min\{n,p\}} U_{ik}D_{kk}V_{jk} = \langle D^\top U_i^\top, V_j^\top \rangle = \langle z_i, \eta_j \rangle.$$

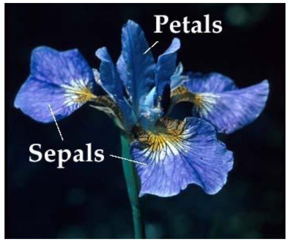
- Biplots focus on the first two PCs and the quality depends on the **proportion of variance explained** by the first two PCs.

Iris Data

50 samples from each of the 3 species of iris: `setosa`, `versicolor`, and `virginica`

Each measuring the length and widths of both sepal and petals

Collected by E. Anderson (1935) and analysed by R.A. Fisher (1936)

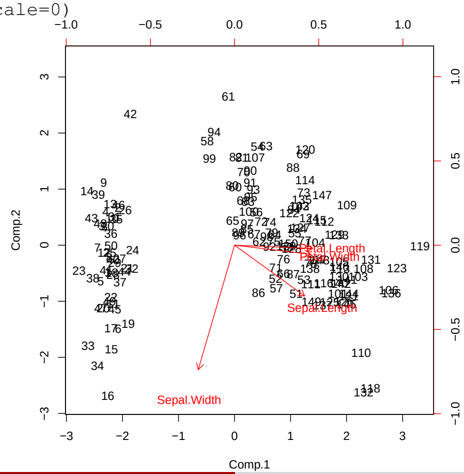


Iris Data

```
> data(iris)
> iris[sample(150,20),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
54           5.5         2.3          4.0         1.3 versicolor
33           5.2         4.1          1.5         0.1     setosa
30           4.7         3.2          1.6         0.2     setosa
73           6.3         2.5          4.9         1.5 versicolor
107          4.9         2.5          4.5         1.7  virginica
4            4.6         3.1          1.5         0.2     setosa
90           5.5         2.5          4.0         1.3 versicolor
83           5.8         2.7          3.9         1.2 versicolor
50           5.0         3.3          1.4         0.2     setosa
92           6.1         3.0          4.6         1.4 versicolor
128          6.1         3.0          4.9         1.8  virginica
57           6.3         3.3          4.7         1.6 versicolor
9            4.4         2.9          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
86           6.0         3.4          4.5         1.6 versicolor
66           6.7         3.1          4.4         1.4 versicolor
85           5.4         3.0          4.5         1.5 versicolor
147          6.3         2.5          5.0         1.9  virginica
8            5.0         3.4          1.5         0.2     setosa
41           5.0         3.5          1.3         0.3     setosa
```


Iris data biplot

```
> iris.pca<-princomp(iris[,-5],cor=TRUE)
> loadings(iris.pca)
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.521 -0.377  0.720  0.261
Sepal.Width  -0.269 -0.923 -0.244 -0.124
Petal.Length  0.580        -0.142 -0.801
Petal.Width   0.565        -0.634  0.524
> biplot(iris.pca,scale=0)
```



Biplots

- There are other projections we can consider for biplots (assuming $p < n$ to simplify notation):

$$x_{ij} = \sum_{k=1}^{p} U_{ik} D_{kk} V_{jk} = \langle D_{1:p,1:p}^\top U_{i,1:p}^\top, V_j^\top \rangle = \langle D_{1:p,1:p}^{1-\alpha} U_{i,1:p}^\top, D_{1:p,1:p}^{\alpha} V_j^\top \rangle.$$

  where $0 \le \alpha \le 1$, i.e., we change representation to

$$\tilde{z}_i = D_{1:p,1:p}^{1-\alpha} U_{i,1:p}^\top, \; \tilde{\eta}_j = D_{1:p,1:p}^{\alpha} V_j^\top$$

- case $\alpha = 1$:
  - Sample covariance of the projected points is:

$$\widehat{\text{Cov}}(\tilde{Z}) = \frac{1}{n-1} U_{1:n,1:p}^\top U_{1:n,1:p} = \frac{1}{n-1} I_p.$$

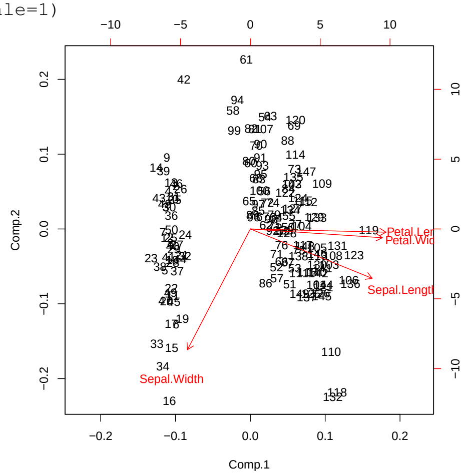
    Projected points are uncorrelated and dimensions are equi-variance.
  - Sample covariance between $X^{(i)}$ and $X^{(j)}$ is:

$$\hat{\mathbb{E}}(X^{(i)} X^{(j)}) = \frac{1}{n-1} \left(V D^\top D V^\top \right)_{i,j} = \frac{1}{n-1} \langle D_{1:p,1:p} V_i^\top, D_{1:p,1:p} V_j^\top \rangle$$

    The angle between the projected variables corresponds to their correlation.

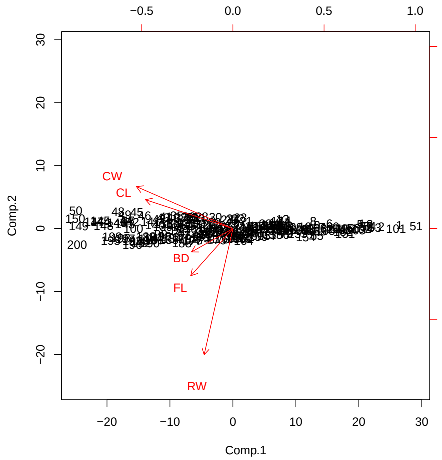
Iris Data biplot - scaled

```
> ?biplot
...
scale: The variables are scaled by lambda ^ scale and the observations
are scaled by lambda ^ (1-scale) where lambda are the singular values
as computed by princomp. (default=1)
...
> biplot(iris.pca,scale=1)
```

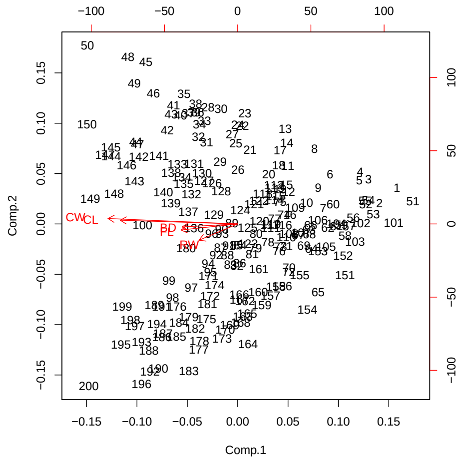


Crabs Data biplots

```
> biplot(Crabs.pca,scale=0)
```




```
> biplot(Crabs.pca,scale=1)
```



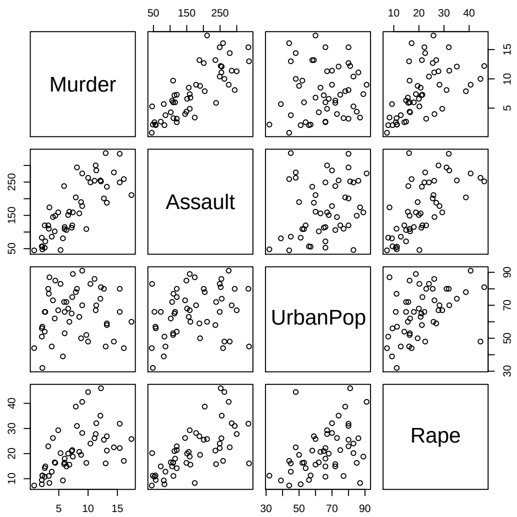
US Arrests Data

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

```
pairs(USArrests)
usarrests.pca <- princomp(USArrests,cor=T)
plot(usarrests.pca)

pairs(predict(usarrests.pca))
biplot(usarrests.pca)
```


US Arrests Data Pairs Plot

```
> pairs(USArrests)
```



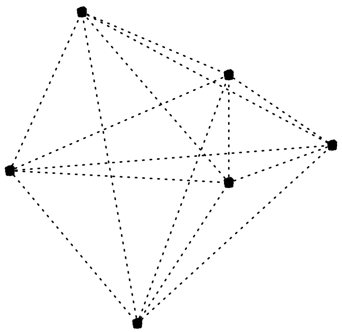
US Arrests Data Biplot

```
> biplot(usarrests.pca)
```

Multidimensional Scaling

Suppose there are n points $\mathbf{X}$ in $\mathbb{R}^p$, but we are only given the $n \times n$ matrix $\mathbf{D}$ of inter-point distances.

Can we reconstruct $\mathbf{X}$?



Multidimensional Scaling

Rigid transformations (translations, rotations and reflections) do not change inter-point distances so cannot recover $\mathbf{X}$ exactly. However $\mathbf{X}$ can be recovered up to these transformations!

- Let $d_{ij} = \|x_i - x_j\|_2$ be the distance between points x_i and x_j.

$$\begin{aligned} d_{ij}^2 &= \|x_i - x_j\|_2^2 \\ &= (x_i - x_j)^\top (x_i - x_j) \\ &= x_i^\top x_i + x_j^\top x_j - 2x_i^\top x_j \end{aligned}$$

- Let $\mathbf{B} = \mathbf{X}\mathbf{X}^\top$ be the $n \times n$ matrix of dot-products, $b_{ij} = x_i^\top x_j$. The above shows that $\mathbf{D}$ can be computed from $\mathbf{B}$.
- Some algebraic exercise shows that $\mathbf{B}$ can be recovered from $\mathbf{D}$ if we assume $\sum_{i=1}^n x_i = 0$.

Multidimensional Scaling

- If we knew $\mathbf{X}$, then SVD gives $\mathbf{X} = UDV^\top$. As $\mathbf{X}$ has rank at most $r = \min(n, p)$, we have at most r non-zero singular values in D and we can assume $U \in \mathbb{R}^{n\times r}$, $D \in \mathbb{R}^{r\times r}$ and $V^\top \in \mathbb{R}^{r\times p}$.
- The eigendecomposition of $\mathbf{B}$ is then:

$$\mathbf{B} = \mathbf{X}\mathbf{X}^\top = UD^2U^\top = U\Lambda U^\top.$$

- This eigendecomposition can be obtained from $\mathbf{B}$ without knowledge of $\mathbf{X}$!
- Let $\tilde{x}_i^\top = U_i\Lambda^{\frac{1}{2}} \in \mathbb{R}^r$. If $r < p$, pad $\tilde{x}_i$ with 0s so that it has length p. Then,

$$\tilde{x}_i^\top \tilde{x}_j = U_i \Lambda U_j^\top = b_{ij} = x_i^\top x_j$$

  and we have found a set of vectors with dot-products given by $\mathbf{B}$, as desired.
- The vectors $\tilde{x}_i$ differs from x_i only via the orthogonal matrix $V^\top$ (recall that $x_i^\top = U_iDV^\top = \tilde{x}_i^\top V^\top$) so are equivalent up to rotation and reflections.

US City Flight Distances

We present a table of flying mileages between 10 American cities, distances calculated from our 2-dimensional world. Using D as the starting point, metric MDS finds a configuration with the same distance matrix.

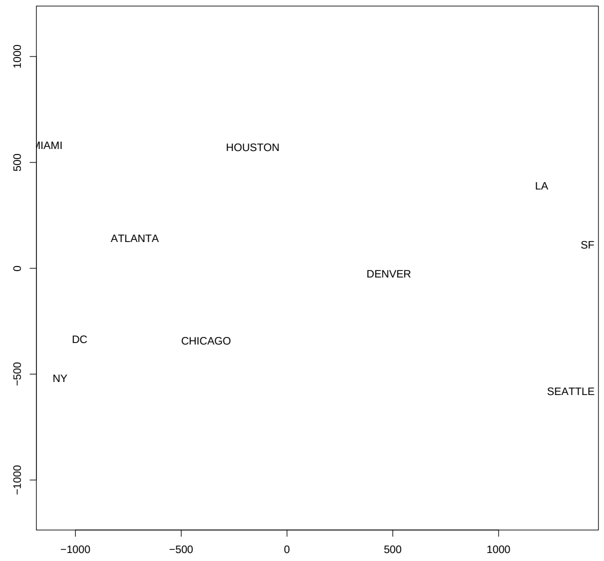
ATLA	CHIG	DENV	HOUS	LA	MIAM	NY	SF	SEAT	DC
0	587	1212	701	1936	604	748	2139	2182	543
587	0	920	940	1745	1188	713	1858	1737	597
1212	920	0	879	831	1726	1631	949	1021	1494
701	940	879	0	1374	968	1420	1645	1891	1220
1936	1745	831	1374	0	2339	2451	347	959	2300
604	1188	1726	968	2339	0	1092	2594	2734	923
748	713	1631	1420	2451	1092	0	2571	2408	205
2139	1858	949	1645	347	2594	2571	0	678	2442
2182	1737	1021	1891	959	2734	2408	678	0	2329
543	597	1494	1220	2300	923	205	2442	2329	0

US City Flight Distances

```
library(MASS)

us <- read.csv("http://www.stats.ox.ac.uk/~sejdinov/sdmml/data/uscities.csv")

## use classical MDS to find lower dimensional views of the data
## recover X in 2 dimensions

us.classical <- cmdscale(d=us,k=2)
plot(us.classical)
text(us.classical,labels=names(us))
```

Visualisation and Dimensionality Reduction Multidimensional Scaling

US City Flight Distances



Visualisation and Dimensionality Reduction Multidimensional Scaling

Lower-dimensional Reconstructions

In classical MDS derivation, we used all eigenvalues in the eigendecomposition of $\mathbf{B}$ to reconstruct

$$\tilde{x}_i = U_i \Lambda^{\frac{1}{2}}.$$

We can use only the largest $k < \min(n, p)$ eigenvalues and eigenvectors in the reconstruction, giving the 'best' k-dimensional view of the data.

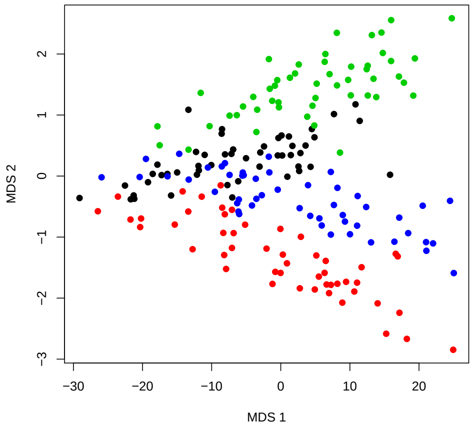
This is analogous to PCA, where only the largest eigenvalues of $\mathbf{X}^\top \mathbf{X}$ are used, and the smallest ones effectively suppressed.

Indeed, PCA and classical MDS are duals and yield effectively the same result.

Visualisation and Dimensionality Reduction Multidimensional Scaling

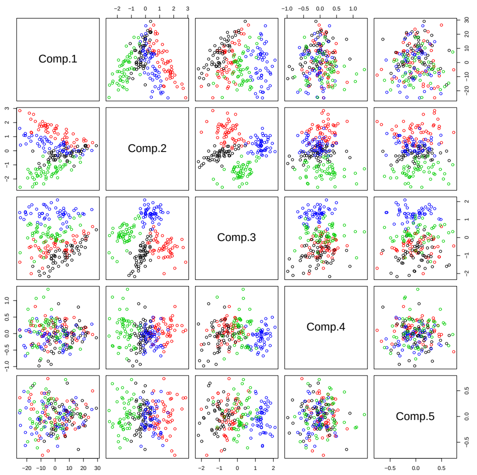
Crabs Data

```
library(MASS)
crabs$spsex=paste(crabs$sp,crabs$sex,sep="")
varnames<-c('FL','RW','CL','CW','BD')
Crabs <- crabs[,varnames]
Crabs.class <- factor(crabs$spsex)
crabsmds <- cmdscale(d= dist(Crabs),k=2)
plot(crabsmds, pch=20, cex=2,col=unclass(Crabs.class))
```



Crabs Data

Compare with previous PCA analysis.
Classical MDS solution corresponds to the first 2 PCs.



Varieties of MDS

Generally, MDS is a class of dimensionality reduction techniques which represents data points $x_1, \dots, x_n \in \mathbb{R}^p$ in a lower-dimensional space $z_1, \dots, z_n \in \mathbb{R}^k$ which tries to preserve inter-point (dis)similarities.

- It requires only the matrix $\mathbf{D}$ of pairwise dissimilarities $d_{ij} = d(x_i, x_j)$. For example, we can use Euclidean distance $d_{ij} = \|x_i - x_j\|_2$, but other dissimilarities are possible.
- MDS finds representations $z_1, \dots, z_n \in \mathbb{R}^k$ such that

$$\|z_i - z_j\|_2 \approx d(x_i, x_j) = d_{ij},$$

  and differences in dissimilarities are measured by the appropriate loss $\Delta(d_{ij}, \|z_i - z_j\|_2)$.
- Goal: Find $\mathbf{Z}$ which minimizes the **stress function**

$$S(\mathbf{Z}) = \sum_{i \neq j} \Delta(d_{ij}, \|z_i - z_j\|_2).$$

Varieties of MDS

- Choices of (dis)similarities and (**stress**) functions lead to different algorithms.
  - **Classical/Torgerson**: preserves inner products instead - **strain function** (`cmdscale`)

$$S(\mathbf{Z}) = \sum_{i \neq j} \left(b_{ij} - \langle z_i - \bar{z}, z_j - \bar{z} \rangle\right)^2$$

  - **Metric Shephard-Kruskal**: preserves distances w.r.t. squared stress

$$S(\mathbf{Z}) = \sum_{i \neq j} \left(d_{ij} - \|z_i - z_j\|_2\right)^2$$

  - **Sammon**: preserves shorter distances more (`sammon`)

$$S(\mathbf{Z}) = \sum_{i \neq j} \frac{\left(d_{ij} - \|z_i - z_j\|_2\right)^2}{d_{ij}}$$

  - **Non-Metric Shephard-Kruskal**: ignores actual distance values, only preserves ranks (`isoMDS`)

$$S(\mathbf{Z}) = \min_{g \text{ increasing}} \frac{\sum_{i \neq j} \left(g(d_{ij}) - \|z_i - z_j\|_2\right)^2}{\sum_{i \neq j} \|z_i - z_j\|_2^2}$$