MDS is a linear projection method. It is related to PCA. MDS and PCA can be used for non-linear projection - next lecture.

**Key Idea** of MDS: project to preserve the distances \(|x_i - x_j|\) between the data points \(x_i\).

i.e. \(x_i \rightarrow y_i\) such that \(|x_i - x_j| \approx |y_i - y_j|\)

But the \(y_i\)'s have lower dimension.

This projection constraint is imposed on the dot products \(x_i \cdot x_j \approx y_i \cdot y_j\).

This will imply that \(|x_i - x_j| \approx |y_i - y_j|\).

**Important Property**: we only need to know \(|x_i - x_j| \approx \Delta_{ij}\) in order to calculate \(x_i \cdot x_j\). This will be useful for non-linear applications later. Also, sometimes only \(|x_i - x_j|\) is specified.

Result:

\[
x_i \cdot x_j = \frac{1}{N} \sum_{k} \Delta_{ik}^2 + \frac{1}{2N} \sum_{k} \sum_{j} \Delta_{ik} \Delta_{kj} - \frac{1}{2N} \sum_{k} \Delta_{ik}^2 - \frac{1}{2} \Delta_{ij}^2
\]

provided \(\sum_{k} x_k = 0\) (Subtract \(\frac{1}{N} \sum_{k} x_k\) from data to ensure this)

**Proof**:

\[\Delta_{ij}^2 = |x_i|^2 + |x_j|^2 - 2x_i \cdot x_j\]

Let \(T = \sum_{k} |x_k|^2\), Noth that \(\sum_{k} x_k \cdot x_k = 0 = \sum_{k} x_k x_j\).

Then \(\sum_{k} \Delta_{ik}^2 = N |x_i|^2 + T\), \(\sum_{k} \Delta_{ij}^2 = N |x_j|^2 + T\), \(\sum_{k} \Delta_{ik}^2 = 2NT\).

The result follows by substitution.

Define the Gram matrix:

\[
G_{ij} = x_i \cdot x_j = \frac{1}{N} \sum_{k} \Delta_{ik}^2 + \frac{1}{2N} \sum_{k} \sum_{j} \Delta_{ik} \Delta_{kj} - \frac{1}{2N} \sum_{k} \Delta_{ik}^2 - \frac{1}{2} \Delta_{ij}^2
\]

Define an error function:

\[
\text{err}(y) = \sum_{ij} (G_{ij} - y_i \cdot y_j)^2
\]

\(x_i\) lie in \(d\)-dim space, \(G_{ij}\) is \(N \times N\) matrix

\(y_i\) is a vector in \(d\)-dim space

\(d < C \leq N\)
Minimize $\text{err}(y) = \sum_{i,j} (G_{ij} - y_i y_j^t)^2$.

Do spectral decomposition:

$$G = \sum_{\lambda = 1}^{N} \lambda \mathbf{v}_\lambda \mathbf{v}_\lambda^t$$

$\lambda_1 \geq \ldots \geq \lambda_N \geq 0$

Claim: optimal minimization $\mathbf{v}_\lambda \mathbf{v}_\lambda^t = S \mathbf{S}$

is $y_i = \sqrt{\lambda_i} \mathbf{v}_i$

i.e. $y_i = \langle \sqrt{\lambda_i} \mathbf{v}_i, \sqrt{\lambda_i} \mathbf{v}_i, \ldots, \sqrt{\lambda_i} \mathbf{v}_i \rangle$

$N$-dimension

Proof: $y_i y_j = \sum_{\lambda = 1}^{N} \lambda \mathbf{v}_i^t \mathbf{v}_j = \frac{1}{2} \mathbf{v}_i^T \mathbf{S} \mathbf{v}_j$

This gives $\text{err} = 0$.

We can reduce the dimension by

truncating $y_i$ to $y_i = (\sqrt{\lambda_1} \mathbf{v}_i, \ldots, \sqrt{\lambda_d} \mathbf{v}_i)$ for $d < N$

In this case $G_{ij} = y_i y_j$

$$y_i y_j = \sum_{\lambda = 1}^{N} \lambda \mathbf{v}_i^t \mathbf{v}_j$$

Hence $G_{ij} - y_i y_j = \sum_{\lambda = d+1}^{N} \lambda \mathbf{v}_i^t \mathbf{v}_j$

Claim: $\sum_{i,j} (G_{ij} - y_i y_j)^2 = \sum_{\lambda = d+1}^{N} \lambda^2$

Proof: $\sum_{i,j} \left( \sum_{\lambda = d+1}^{N} \lambda \mathbf{v}_i^t \mathbf{v}_j \mathbf{v}_i \mathbf{v}_j^t \right) = \sum_{\lambda = d+1}^{N} \lambda^2$

$\sum_{\lambda = d+1}^{N} \mathbf{v}_i^t \mathbf{v}_j = S^{\mathbf{S}} \mathbf{v}_i \mathbf{v}_j = S \mathbf{S} \mathbf{v}_i \mathbf{v}_j$

$\sum_{\lambda = d+1}^{N} \lambda^2 = \sum_{\lambda = d+1}^{N} \lambda^2$

Hence we project to $d$-dimension

provided $\sum_{\lambda = d+1}^{N} \lambda^2$ is small, or $\sum_{\lambda = d+1}^{N} \lambda^2$ is small.

$$y_i = (\sqrt{\lambda_1} \mathbf{v}_i, \ldots, \sqrt{\lambda_d} \mathbf{v}_i)$$
(3) Relation between MDS & PCA?
Both linear, both depend on eigenvectors/eigenvalue.

Recall PCA (subtract mean to ensure $\frac{1}{N} \sum X_i = 0$)

$X_p = (X_1, \ldots, X_d)$
the $e$'s are eigenvectors of $K_{ab} = \frac{1}{N} \sum_{c=1}^{N} X_c^T X_c$

MDS $Y_i = (\sqrt{a_1} u_i, \ldots, \sqrt{a_d} u_d)$
the $u$'s are eigenvectors of $G_{ij} = \frac{1}{N} \sum_{c=1}^{N} X_i X_j$

Claim: the eigenvalues of $K$ and $G$ are the same. The eigenvectors are closely related.

Proof: let $X$ be an $N \times D$ matrix, $i = 1 \ldots N$ no. of points
$X = 1 \times D$ space linear
with element $X_{ia}$ a'th component of $i$'th data point

Consider $X^T X$ $N \times N$ matrix, $(X^T X)_{ij} = \frac{1}{N} \sum_{a=1}^{D} X_{ia} X_{ja}$

$X^T X$ $D \times D$ matrix, $(X^T X)_{ab} = \frac{1}{N} \sum_{a=1}^{D} X_{ia} X_{ib}$

Both are square matrices and both are positive definite, so they have positive eigenvalues.

$X^T X$ is used for MDS, $X^T X$ is used for PCA

Suppose $e, \lambda$ are an eigenvector, eigenvalue of $X^T X$

$X^T X e = \lambda e$

So $X^T X e = \lambda X e$

$(X^T X) (\frac{1}{\lambda} X e) = \lambda (\frac{1}{\lambda} X e)$

So $\frac{1}{\lambda} X e$ is an eigenvector (un-normalized)

of $X^T X$ with eigenvalue $\lambda$.

Similarly if $X^T v = \lambda v$
$\frac{1}{\lambda} X^T v$ is an eigenvector (un-normalized)
of $X^T X$ with eigenvalue $\lambda$.

Conclusion: the two matrices have the same eigenvalues and related eigenvectors.
Result: the truncation conditions for MDS and PCA are similar:
\[ \sum_{i=1}^{d} \lambda_i > \text{Threshold} \]

MDS projects \( x_i \) to \( y_i = (\sqrt{\lambda_1} v_{i1}, \ldots, \sqrt{\lambda_d} v_{id}) \)

PCA projects \( x \) to \( y = (x_1, \ldots, x_d) \)

where the \( e \)'s and the \( \nu \)'s are related (see previous page).

Note: The equivalence between the eigenvalues of \( X X^T \) and \( X^T X \) has computational importance.

If \( N < D \), faster to compute eigenvalues/vectors for \( X X^T \), then convert to eigenvalues/vectors of \( X^T X \).

Note: to do PCA we have to know the data \( \{ x_i \} \), but to do MDS we only need to know \( A_{ij} \).

In some applications it is possible to specify \( A_{ij} \) but not the \( \{ x_i \} \).

Deeper Understanding: This relationship between \( X X^T \) and \( X^T X \) can be used to prove SVD:

\[ X = E F \] where \( EE^T = I \) (Identity).

\[ D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_k \end{pmatrix} \] \( EE^T = D \)

This is a generalization of the spectral decomposition:
\[ G = \sum \lambda_a v_a v_a^T \] \( X \) is \( N \times D \) \( N \neq D \)

to any matrix \( X \) \( \rightarrow \) so \( X \) is not square.

It follows that \( X^T X = (E^T D E) (E \Delta E) \)

\[ = E^T \Delta^2 E \]

spectral decomposition \( \rightarrow \) work \( \lambda_1 = d_1^2, \lambda_2 = d_2^2 \ldots \)

and \( X X^T = FDE E^T D F^T = F \Delta^2 F^T \)

So SVD is like the square root of spectral decomposition!