

Dimensionality Reduction

先进计算与数字工程研究中心
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning



Lecture 11 Dimensionality Reduction

Chizhi Chris ZHANG

zhangchizhi@ciomp.ac.cn

Advanced Computing and Digital Technology Research Center

University of Chinese Academy of Sciences

Autumn 2026

Lecture plan and time allocation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Topic	Minutes
Principal component analysis	45
Global structure	45
Local structure	45

Coordinates, distances and neighborhood relations

Compute PCA; recover coordinates from distances; examine local structure.

The numerical steps are shown on the slides.



High-dimensional neighborhoods require much more coverage

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Four two-dimensional observations will remain visible throughout the PCA calculation.

A 28×28 image has 784 pixel features; a sensor record may contain 500 measurements. Many features move together.

Check. A two-dimensional picture can look clean while discarding a feature needed for classification.

For a uniform unit cube, an axis-aligned box of side length a occupies fraction a^d of the total volume.

$$a^d = 0.1 \quad \Rightarrow \quad a = 0.1^{1/d}.$$

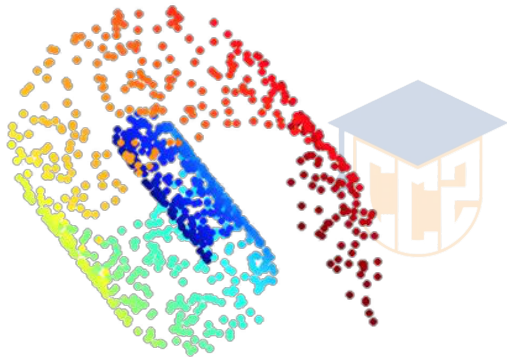
Dimensions d	1	10	100
Required side length a	0.100	0.794	0.977



A curved sheet can use fewer coordinates than its surroundings

Intrinsic dimension: the number of independent factors needed to move along the data.

The rolled sheet occupies three-dimensional space but has two coordinates along its surface. This distinguishes the number of recorded coordinates from the number of local degrees of freedom.



State what the embedding should preserve

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

No map into two dimensions can preserve every relation in arbitrary high-dimensional data.

Method	Relation to preserve
PCA	Variance after linear projection
Kernel PCA	Variance in implicit feature space
MDS	Given pairwise distances
Isomap	Global distances along neighborhood-graph paths
LLE	Local weights for reconstruction from neighbors
t-SNE, UMAP	Local neighborhoods under the selected objective



First calculate PCA on four points, then examine distances and neighborhoods. Visual separation does not establish classification performance.

The covariance matrix is built from centered rows

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

PCA describes variation around the mean, not distance from the original origin.

$$X = \begin{bmatrix} 4 & 3 \\ 3 & 4 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \bar{x} = (2, 2), \quad X_c = X - \mathbf{1}\bar{x}^\top = \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}.$$

Each diagonal averages a squared centered feature. Each off-diagonal averages the product of the two centered features.

$$(2, 1), (1, 2), (-1, -2), (-2, -1).$$
$$S = \frac{1}{4} \left(2 \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} + 2 \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \right) = \frac{1}{4} \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix} = \begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix}.$$

Scaling changes the question that PCA answers

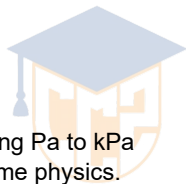
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Using $1/(n-1)$ rescales every eigenvalue but leaves principal directions unchanged.

$$X_c = \begin{bmatrix} 2 & 1 \\ 1 & 2 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}, \quad S = \frac{1}{n} X_c^\top X_c = \begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix}, \quad n = 4.$$

$$S_{11} = (4 + 1 + 1 + 4)/4 = 2.5 \quad S_{12} = (2 + 2 + 2 + 2)/4 = 2$$

Temperature variance is 4 in degrees Celsius; pressure variance is 10,000 in Pa. Converting Pa to kPa divides its variance by 10^6 . Raw PCA changes although the observations describe the same physics.



A principal-component score is a position along an axis

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

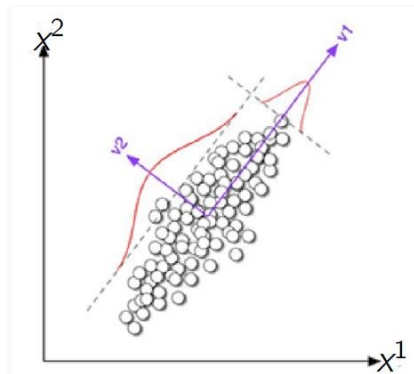
Replacing w by $-w$ flips all scores, but produces the same reconstructed point.

$$\mathbf{z} = \mathbf{w}^\top \mathbf{x}_c, \quad \|\mathbf{w}\| = 1.$$

$$\mathbf{w} = (1, 1)/\sqrt{2} \quad \mathbf{x}_c = (2, 1)$$

$$\mathbf{z} = 3/\sqrt{2} \approx 2.121$$

$$\hat{\mathbf{x}}_c = \mathbf{z}\mathbf{w} = (1.5, 1.5)$$



PCA chooses the direction with the largest projected variance

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

The first direction captures the largest spread of centered observations.

$$\text{Var}(z) = \frac{1}{n} \sum_i (w^\top x_{c,i})^2 = w^\top S w, \quad \max_{\|w\|=1} w^\top S w.$$

$$S = \begin{bmatrix} 2.5 & 2 \\ 2 & 2.5 \end{bmatrix}$$

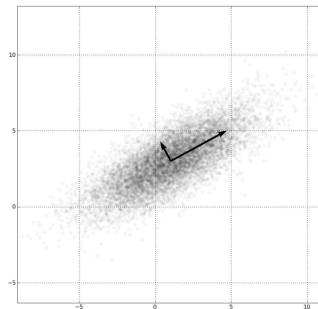
Direction w	Projected variance	Interpretation
$(1, 0)$	2.5	Original first feature
$(1, 1)/\sqrt{2}$	4.5	Features increase together
$(1, -1)/\sqrt{2}$	0.5	Features change oppositely



Compare the spread of two projected directions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Project the centered points onto a unit direction. The long axis retains more spread than the short axis. Comparing unit directions prevents a longer arrow from winning merely by scale.



The variance objective leads to an eigenvector

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

w : a direction unchanged by multiplication with covariance, except for scale.

$$\mathcal{L}(w, \lambda) = w^\top S w - \lambda(w^\top w - 1), \quad \nabla_w \mathcal{L} = 2S w - 2\lambda w = 0.$$

$$S w = \lambda w, \quad w^\top S w = \lambda \quad (\|w\| = 1).$$



An eigenvalue gives an equation for the direction

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Equal eigenvalues would allow multiple equally good choices of principal axes.

$$\det(\mathbf{S} - \lambda \mathbf{I}) = (2.5 - \lambda)^2 - 4 = 0 \quad \Rightarrow$$
$$\lambda_1 = 4.5, \lambda_2 = 0.5.$$

$$\mathbf{w}_1 = \frac{1}{\sqrt{2}}(1, 1)^\top, \quad \mathbf{w}_2 = \frac{1}{\sqrt{2}}(1, -1)^\top.$$

$$\text{tr}(\mathbf{S}) = 5 = 4.5 + 0.5 \quad 4.5/5 = 90\%$$

$$\mathbf{S}\mathbf{w}_1 = 4.5\mathbf{w}_1 \quad \mathbf{w}_1^\top \mathbf{w}_2 = 0$$

Changing the sign reverses coordinates along the same axis.

$$\lambda_1 = 4.5 \quad \mathbf{w} = (\mathbf{a}, \mathbf{b})^\top$$

$$\mathbf{S}\mathbf{w} = 4.5\mathbf{w}$$

$$2.5\mathbf{a} + 2\mathbf{b} = 4.5\mathbf{a} \Rightarrow 2\mathbf{b} = 2\mathbf{a} \Rightarrow \mathbf{b} = \mathbf{a}.$$

$$\mathbf{a}^2 + \mathbf{b}^2 = 1 \Rightarrow 2\mathbf{a}^2 = 1 \Rightarrow \mathbf{a} = \pm 1/\sqrt{2}.$$

$$\mathbf{w}_1 = (1, 1)^\top / \sqrt{2} \quad \mathbf{w}_2 = (1, -1)^\top / \sqrt{2}$$

$$\mathbf{S}\mathbf{w}_1 = \frac{1}{\sqrt{2}}(4.5, 4.5)^\top = 4.5\mathbf{w}_1.$$

Projection and reconstruction complete the calculation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Add the original training mean only when decoding back to original coordinates.

Worked calculation

x_c	$z = x_c^\top w_1$	$\hat{x}_c = zw_1$	$\ x_c - \hat{x}_c\ ^2$
(2, 1)	$3/\sqrt{2}$	(1.5, 1.5)	0.5
(1, 2)	$3/\sqrt{2}$	(1.5, 1.5)	0.5
(-1, -2)	$-3/\sqrt{2}$	(-1.5, -1.5)	0.5
(-2, -1)	$-3/\sqrt{2}$	(-1.5, -1.5)	0.5

$$Z = X_c W_m, \quad \hat{X} = Z W_m^\top + \mathbf{1} \bar{x}^\top.$$

Maximum variance and minimum reconstruction error agree

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Orthogonal projection splits squared length into retained and discarded parts.

$$\|\mathbf{x}_c\|^2 = \|\mathbf{W}_m^\top \mathbf{x}_c\|^2 + \|\mathbf{x}_c - \mathbf{W}_m \mathbf{W}_m^\top \mathbf{x}_c\|^2.$$

$$\frac{1}{n} \|\mathbf{X}_c - \hat{\mathbf{X}}_c\|_F^2 = \text{tr}(\mathbf{S}) - \sum_{j=1}^m \lambda_j = \sum_{j=m+1}^d \lambda_j.$$

$$5 - 4.5 = 0.5$$



Worked review: interpret PCA without relying on a picture

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For covariance $S = \text{diag}(9, 1, .25)$, total variance is 10.25. The first axis retains $9/10.25 \approx 87.8\%$ and discards variance 1.25. A class signal in a discarded direction can still be lost.

$$\text{Retained fraction} = \frac{9}{10.25} \approx 87.8\%, \quad \text{Mean reconstruction error} = 1 + 0.25 = 1.25.$$

Check. If label information lies in the third coordinate, retaining only the first can discard it. PCA preserves variance; use labeled validation to check classification.



SVD computes PCA directly from centered observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For the same centered data, SVD factors the matrix directly. Here Σ contains singular values, not the covariance matrix.

Rows of $X_c \in \mathbb{R}^{n \times d}$ are observations. Columns of V are feature-space principal directions.

$$X_c = U \Sigma V^T, \quad S = \frac{1}{n} V \Sigma^T \Sigma V^T.$$

$$X_c \in \mathbb{R}^{n \times d} \quad \lambda_j = \sigma_j^2 / n$$

$$Z = X_c V_m = U_m \Sigma_m$$



The SVD numbers must agree with the PCA numbers

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

$\|\mathbf{X}_c\|_F^2 = \sigma_1^2 + \sigma_2^2 = 20$; dividing by four gives total variance five.

$n = 4$

$$\sigma_1 = \sqrt{n\lambda_1} = \sqrt{18} \approx 4.243, \quad \sigma_2 = \sqrt{2} \approx 1.414.$$

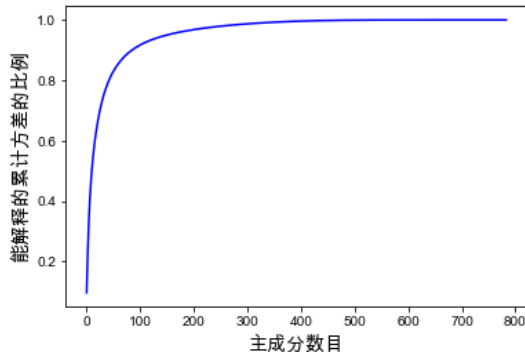
Quantity	Component 1	Component 2
variance σ_j^2/n	4.5	0.5
Variance fraction	0.90	0.10
Feature direction	$(1, 1)/\sqrt{2}$	$(1, -1)/\sqrt{2}$

$$\|\mathbf{X}_c\|_F^2 = \sigma_1^2 + \sigma_2^2 = 20$$



Choose a dimension using cumulative variance

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



One component retains 60% with error 4; two retain 90% with error 1; three retain everything with error 0.

The curve gives cumulative retained variance against the number of components. State the required fraction first, then find the smallest qualifying dimension. Retained variance alone does not establish retained label information.

$$R(m) = \frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^d \lambda_j}, \quad E(m) = \sum_{j=m+1}^d \lambda_j.$$

A small sensor signal can matter more than 99.99 percent of variance

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two sensors describe a machine. Sensor 1 has large background variation; sensor 2 carries a small fault signal. Use these four noise-free teaching records, on the displayed scales.

State	Sensor 1	Sensor 2
Normal	-10	-.1
Fault	-10	+.1
Normal	+10	-.1
Fault	+10	+.1

$$\bar{\mathbf{x}} = (0, 0), \quad \mathbf{S} = \begin{bmatrix} 100 & 0 \\ 0 & .01 \end{bmatrix}.$$

The first component is sensor 1. It retains $100/(100 + .01) \approx 99.99\%$ of variance.

Can one component still detect the fault?

At score -10 , normal and fault coincide. The same happens at $+10$. The discarded sensor separates the classes by its sign.

Condition to check

Retained variance measures reconstruction, while fault detection needs label information.

Even if sensor 2 had within-class noise SD $.01$, its mean gap $.2$ would be 20 such SDs. Small total variation can carry a strong class signal.

Fit preprocessing and PCA inside the training split

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

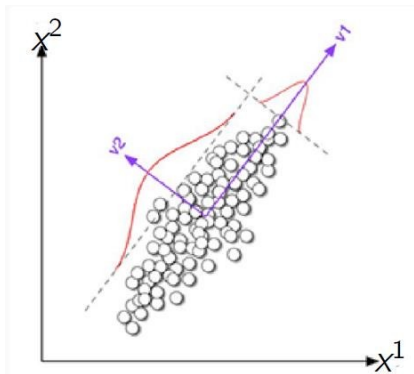
Split labeled observations into training, validation and test sets before fitting any transformation.

$$\mathbf{z} = \mathbf{W}_m^\top [(\mathbf{x} - \bar{\mathbf{x}})/\mathbf{s}]$$

Stage	Use the training coordinate system	Permitted role
Training	Fit mean, scale and PCA axes	Estimate the representation and classifier.
Validation	Transform using those saved quantities	Choose dimension and model settings.
Final test	Apply the finalized complete pipeline	Report independent performance.

Read the principal axes rather than just the scatter plot

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



PCA rotates to orthogonal axes and then discards selected coordinates.



A flat subspace cannot always unfold a curved structure

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

PCA maps x to $W_m^\top x_c$: every output coordinate is a linear combination of the original features.

Check. A nonlinear method can reveal useful structure, but its geometry depends on its kernel or neighborhood choices.

$$x_a = (0, 1), \quad x_b = (0, -1), \quad z = x_1 : \quad z_a = z_b = 0.$$

In the original circle

The upper and lower points are different locations.

After projection onto the horizontal axis

Both become zero. A collision in the display can be information loss, rather than identical observations.

Kernel PCA reuses the kernel trick from SVM

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Perform linear PCA on centered $\phi(x)$ without constructing every feature-space coordinate.

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle.$$



Center the kernel matrix before eigendecomposition

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

For normalized eigenvector v_j with $K_c v_j = \eta_j v_j$ and $\eta_j > 0$, the training score vector is $\sqrt{\eta_j} v_j$.

$$H = I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top, \quad K_c = H K H.$$

$$(K_c)_{ij} = K_{ij} - \bar{K}_i - \bar{K}_j + \bar{K}...$$

$$K_c v_j = \eta_j v_j \quad \eta_j > 0$$

$$K = \begin{bmatrix} 1 & 3 \\ 3 & 9 \end{bmatrix}, \quad H = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad K_c = H K H = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$



Kernel centering on two observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Subtract training means

$$K = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}.$$

Both row means and column means are 3.
The grand mean is also 3.

$$(K_c)_{11} = 4 - 3 - 3 + 3 = 1,$$

$$(K_c)_{12} = 2 - 3 - 3 + 3 = -1.$$

Extract one coordinate

$$K_c = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

$$v = \frac{1}{\sqrt{2}}(1, -1)^T, \quad K_c v = 2v.$$

The score vector is

$$z = \sqrt{2} v = (1, -1)^T.$$

Its average is zero. No explicit mapped feature vector was needed.

A new point must use training-set kernel centering

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

x_i , row means, total kernel mean and eigenvectors all come from training data.

$$k_{c,i}(x) = k(x_i, x) - \frac{1}{n} \sum_l k(x_l, x) - \frac{1}{n} \sum_l K_{il} + \frac{1}{n^2} \sum_{l,r} K_{lr}.$$

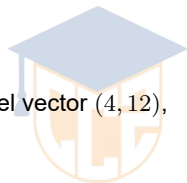
$$\mathbf{z}_j(x) = \frac{\mathbf{v}_j^\top k_c(x)}{\sqrt{\eta_j}}.$$

Training kernel row means are 2 and 6, and the grand mean is 4. The new point has kernel vector (4, 12), with mean 8:

$$k_c(4) = (4 - 8 - 2 + 4, 12 - 8 - 6 + 4) = (-2, 2).$$

$$\eta = 2, \quad \mathbf{v} = (-1, 1)^\top / \sqrt{2} \quad \Rightarrow \quad \mathbf{z}(4) = \mathbf{v}^\top k_c(4) / \sqrt{2} = 2.$$

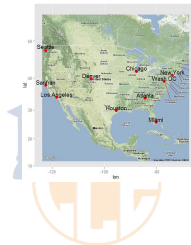
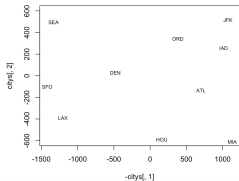
This agrees with linear PCA's $4 - \bar{x} = 4 - 2 = 2$ agrees.



A map can be reconstructed from distances alone

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Input distances in the source table's units: ATL–ORD 587; LAX–SFO 347; JFK–IAD 205.



Double centering removes unknown squared lengths

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

PCA starts from centered coordinates; classical MDS starts from distances and recovers centered coordinates.

$$D_{ij}^{(2)} = d_{ij}^2, \quad H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top, \quad B = -\frac{1}{2} H D^{(2)} H.$$

$$B_{ij} = \mathbf{z}_i^\top \mathbf{z}_j \quad B = V \Lambda V^\top$$

$$Z = V_m \Lambda_m^{1/2}$$

Assume centered coordinates $\sum_i \mathbf{z}_i = 0$ and write $m = \frac{1}{n} \sum_i \|\mathbf{z}_i\|^2$. Expanding each squared distance separates coordinate lengths from inner products. Double centering removes those length terms.

$$d_{ij}^2 = \|\mathbf{z}_i\|^2 + \|\mathbf{z}_j\|^2 - 2\mathbf{z}_i^\top \mathbf{z}_j,$$

$$\overline{d_{i\cdot}^2} = \|\mathbf{z}_i\|^2 + m, \quad \overline{d_{\cdot j}^2} = \|\mathbf{z}_j\|^2 + m,$$

$$\overline{d_{\cdot\cdot}^2} = 2m.$$

$$d_{ij}^2 - \overline{d_{i\cdot}^2} - \overline{d_{\cdot j}^2} + \overline{d_{\cdot\cdot}^2} = -2\mathbf{z}_i^\top \mathbf{z}_j.$$

Thus multiplication by $-1/2$ gives the inner-product matrix B .

Worked MDS: recover three points on a line

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Three unknown positions have pairwise distances $d_{12} = 1$, $d_{23} = 1$, $d_{13} = 2$.

Square and center the distances

$$D^{(2)} = \begin{bmatrix} 0 & 1 & 4 \\ 1 & 0 & 1 \\ 4 & 1 & 0 \end{bmatrix}.$$

Row means: $(5/3, 2/3, 5/3)$. Grand mean: $4/3$.

$$B_{11} = -\frac{1}{2}(0 - 5/3 - 5/3 + 4/3) = 1.$$

$$B_{12} = -\frac{1}{2}(1 - 5/3 - 2/3 + 4/3) = 0.$$

Recover coordinates

$$B = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

$$\lambda = 2, \quad \mathbf{v} = \frac{1}{\sqrt{2}}(-1, 0, 1)^T.$$

$$\mathbf{z} = \sqrt{\lambda} \mathbf{v} = (-1, 0, 1)^T.$$

Recovered distances are 1, 1 and 2. Translation or reversal gives the same distance information.

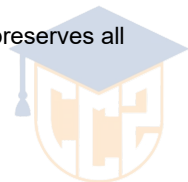
Worked review: choose coordinates from what you are given

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Metric MDS directly minimizes a distance-stress objective, usually by iterative optimization.

$$\text{Stress}(Z) = \sum_{i < j} (d_{ij} - \|z_i - z_j\|)^2.$$

Centering limits rank; square the distances; evaluate a labeled validation task; reflection preserves all distances.



Two nearby dots may still require a long walk

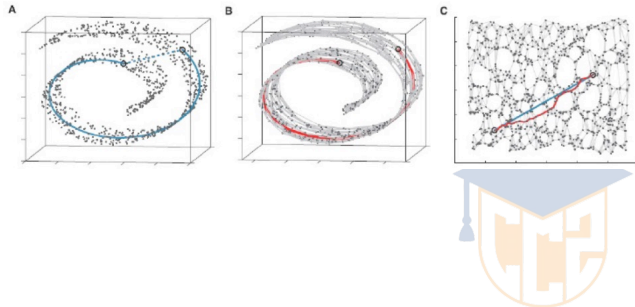
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Walk on the paper, or through the air?

Imagine a route drawn on this rolled sheet. Points on neighboring turns can look close through the air but require a long walk along the sheet.

Isomap estimates the walk using short local edges. A cross-layer edge invents a shortcut.

Read the figure as three stages: rolled sheet, neighborhood graph, unfolded coordinates. The next slide calculates one route.



One false shortcut changes the distances used by Isomap

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Four landmarks on one route

The supplied local graph has edges $A-B$, $B-C$, $C-D$, each of length 1. These are schematic graph lengths.

$$d_G(A, D) = 1 + 1 + 1 = 3.$$

Suppose a false cross-layer edge $A-D$ of length .2 is added.

$$d_G(A, D) = \min(3, .2) = .2.$$

MDS now receives .2 for a route that should have distance 3. Later eigendecomposition cannot recover the missing distinction.

How Isomap constructs coordinates

1. Connect nearby points using k neighbors or a radius ε .
2. Compute shortest-path distances D_G on that graph.
3. Apply classical MDS to D_G .

Too few edges may disconnect the graph. Too many may bridge different layers.

Select a meaningful scale and metric, then inspect graph connectivity and possible shortcuts.

LLE preserves a local interpolation weight

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Find weights that reconstruct each point from its local neighbors.

$$x_i \approx \sum_{j \in \mathcal{N}_i} w_{ij} x_j, \quad \sum_{j \in \mathcal{N}_i} w_{ij} = 1.$$

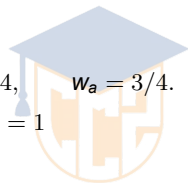
In the same original two-dimensional example, all second coordinates are zero. First coordinates are 0, 2, with the target at .5. Weights satisfy

$$w_a + w_b = 1.$$

$$w_a + w_b = 1$$

$$.5 = w_a(0) + w_b(2) \Rightarrow w_b = 1/4, \quad w_a = 3/4.$$

$$z_a = 0, z_b = 4 \quad z_i = \frac{3}{4}z_a + \frac{1}{4}z_b = 1$$



Zero reconstruction error can collapse every point to one place

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A zero-error answer can be useless

For the preceding weights $w_a = 3/4$, $w_b = 1/4$, put all embedded points at 5:

$$\mathbf{z}_i = \mathbf{z}_a = \mathbf{z}_b = 5.$$

$$\mathbf{z}_i - \frac{3}{4}\mathbf{z}_a - \frac{1}{4}\mathbf{z}_b = 5 - 3.75 - 1.25 = 0.$$

Putting all points at 0 also gives zero error.
Neither map distinguishes any observations.

Centering removes a common offset; fixing scale excludes the collapsed zero map.

Keep useful global coordinates

W is $n \times n$, Z is $n \times m$, and $M = (I - W)^T(I - W)$. The input-space weights stay fixed.

$$\min_Z \sum_i \left\| \mathbf{z}_i - \sum_j w_{ij} \mathbf{z}_j \right\|^2 = \min_Z \text{tr}(\mathbf{Z}^T \mathbf{M} \mathbf{Z}),$$
$$\mathbf{1}^T \mathbf{Z} = 0, \quad \mathbf{Z}^T \mathbf{Z} = I.$$

Use the nonconstant low-eigenvalue directions of M .
Check disconnected components and extra zero modes.

Laplacian eigenmaps retain smooth nonconstant coordinates

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A large edge weight penalizes placing its endpoints far apart.

$$A_{ij} = \text{similarity of connected points}, \quad D_{ii} = \sum_j A_{ij}, \quad L = D - A.$$

$$\mathbf{z}^\top L \mathbf{z} = \frac{1}{2} \sum_{i,j} A_{ij} (\mathbf{z}_i - \mathbf{z}_j)^2.$$

For this three-node chain, $D = \text{diag}(1, 2, 1)$. Both displayed directions have $\mathbf{z}^\top D \mathbf{z} = 1$ and $\mathbf{z}^\top D \mathbf{1} = 0$. Their graph energies are 1 and 2, so the first is smoother under equal normalization.

$$\min \mathbf{z}^\top L \mathbf{z} \quad \Rightarrow \quad L \mathbf{z} = \lambda D \mathbf{z}.$$

Normalized direction	Graph energy	Result
$\mathbf{z} = (-1, 0, 1)^\top / \sqrt{2}$	$\frac{1}{2} + \frac{1}{2} = 1$	Smoother
$\mathbf{z} = (1, -1, 1)^\top / 2$	$1 + 1 = 2$	Less smooth

Without fixed normalization, shrinking all coordinates reduces energy without improving the representation.

Neighbor probabilities from three points

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Use input positions $x = (0, 1, 2)$ and $\sigma_i = 1$ for this example. A point does not count itself as its neighbor.

Start at point 0

Distances to 1 and 2 are 1 and 2. Weights are $e^{-1/2} = .6065$ and $e^{-2} = .1353$.

$$p_{1|0} = \frac{.6065}{.6065 + .1353} = .8176,$$

$$p_{2|0} = .1824.$$

Nearer points receive more weight, but both weights must share one denominator.

Repeat and symmetrize

Pair	$p_{j i}, p_{i j}$	p_{ij}
0-1	.8176, .5	.2196
0-2	.1824, .1824	.0608
1-2	.5, .8176	.2196

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \quad n = 3.$$

Both directions together sum to 1. Next compare these probabilities with a low-dimensional display.



t-SNE compares probabilities of being neighbors

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The preceding three-point example calculated p : input-space neighbor probabilities. q describes neighbors in the displayed coordinates z . The objective compares these two probability tables. Set $p_{ii} = q_{ii} = 0$.

$$p_{j|i} = \frac{\exp[-\|x_i - x_j\|^2 / (2\sigma_i^2)]}{\sum_{k \neq i} \exp[-\|x_i - x_k\|^2 / (2\sigma_i^2)]}, \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$
$$q_{ij} = \frac{(1 + \|z_i - z_j\|^2)^{-1}}{\sum_{a \neq b} (1 + \|z_a - z_b\|^2)^{-1}}, \quad \min_z \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

$$p_{ii} = q_{ii} = 0$$



A misplaced neighbor has a measurable cost

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Keep the preceding input probabilities: $p_{01} = p_{12} = .2196$, $p_{02} = .0608$ and reverse pairs equal.

A display with the middle point misplaced

Display coordinates are $(z_0, z_1, z_2) = (0, 3, 1)$.

The unnormalized pair weights are

$$u_{01} = 1/(1 + 9) = .1,$$

$$u_{02} = 1/(1 + 1) = .5, \quad u_{12} = 1/(1 + 4) = .2.$$

The ordered-pair total is $2(.1 + .5 + .2) = 1.6$.

Normalize and compare

$$q_{01} = .0625, \quad q_{02} = .3125, \quad q_{12} = .125.$$

The original near pair 0–1 is now far apart; the original far pair 0–2 is now close.

$$C = \sum_{i \neq j} p_{ij} \log(p_{ij}/q_{ij}).$$

Using unrounded p , the cost is about 0.6003. A display $(0, 1, 2)$ costs about .0079.

Read a t-SNE digit map cautiously

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Many moderately distant neighbors in high dimensions cannot all fit at comparable distances around one point in 2-D.

Return to original samples and compare neighborhoods across settings.



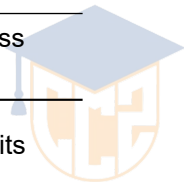
A local display needs checks for two kinds of neighbor error

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

UMAP and t-SNE use neighborhood objectives. A useful display should be checked against the original observations.

Error	What changed in the display?	Check
False neighbor	An originally distant point moved close.	Trustworthiness
Missing neighbor	An originally close point moved far away.	Continuity

Neither method guarantees that every plotted distance or cluster area preserves its original-space meaning. Compare settings and return to the samples.



Choose a method by the relation you need

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The objective determines what the low-dimensional coordinates mean.

Purpose	Starting method	Evidence to check
Linear compression	PCA or SVD	Reconstruction error and retained variance
Distance map	MDS or Isomap	Distance distortion and graph connectivity
Local exploration	LLE, graph methods or t-SNE	Neighbor retention and stability across settings
Support classification	Select representation inside training	Independent labeled-validation performance

When a method fails, first check whether the input relationships match the task.

Can one representation answer two different questions?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Question

We have 500 features, 2000 training rows and 300 labeled validation rows. We need fault classification and a 2-D exploratory display.

Evidence and reasoning

Fit scaling and PCA only on training rows; choose the retained dimension using the labeled validation task. Make a separate 2-D neighborhood display for exploration and inspect original examples. Compare classification performance before and after reduction.

Check. Retained variance and visually separated islands do not by themselves establish preserved fault information.