Lecture 16

Supervised principal component analysis

Which directions should we keep when the goal is prediction?

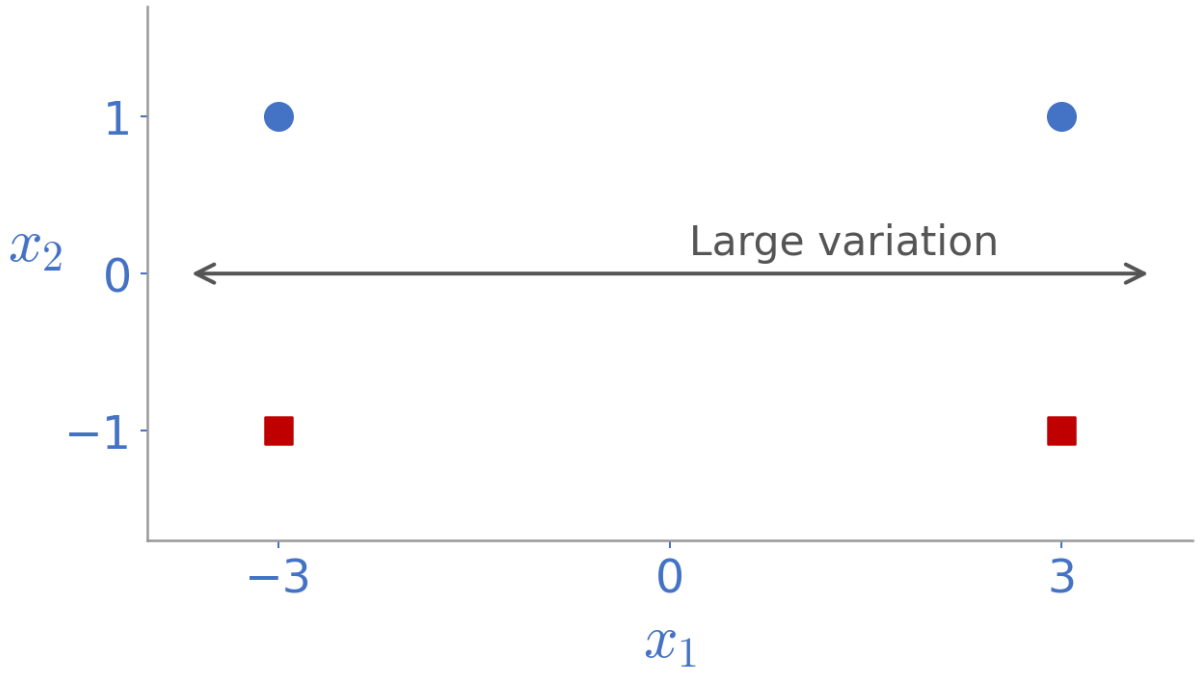
Use labels to guide the projection.

Measure dependence with HSIC, then solve an eigenvalue problem.

Variance can miss the label

PCA keeps directions with large variation.

The direction that separates the classes can have much less variation.



PCA keeps the horizontal direction.

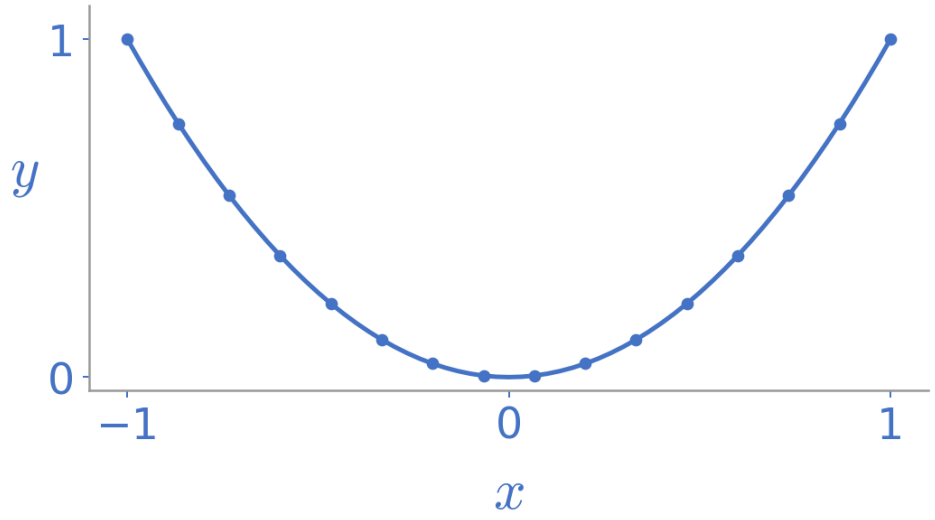
The vertical direction keeps the class information.

HSIC measures dependence

Hilbert-Schmidt independence criterion (HSIC)

Zero correlation does not imply independence.

$$y = x^2, \quad x \sim \text{Unif}[-1, 1] \quad \Longrightarrow \quad \text{Cov}(x, y) = 0$$



The response is completely determined by the input.

Kernel features can reveal this dependence.

MMD compares two distributions

First map observations to a feature space, then compare the mean vectors.

$$\boldsymbol{\mu}_p = \mathbb{E}_{\mathbf{x}\sim p}[\boldsymbol{\phi}(\mathbf{x})], \qquad \boldsymbol{\mu}_q = \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\phi}(\mathbf{x})]$$

$$\mathrm{MMD}_k^2(p,q) = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\|_{\mathcal{H}}^2$$

The kernel determines which differences in the distributions are visible.

Why “maximum mean discrepancy”?

$$f_{\mathbf{v}}(\mathbf{x}) = \langle \mathbf{v}, \boldsymbol{\phi}(\mathbf{x}) \rangle_{\mathcal{H}}, \qquad \|\mathbf{v}\|_{\mathcal{H}} \leq 1$$

$$\mathbb{E}_p[f_{\mathbf{v}}(\mathbf{x})] - \mathbb{E}_q[f_{\mathbf{v}}(\mathbf{x})] = \langle \mathbf{v}, \boldsymbol{\mu}_p - \boldsymbol{\mu}_q \rangle_{\mathcal{H}}$$

$$\max_{\|\mathbf{v}\|_{\mathcal{H}} \leq 1} \left| \langle \mathbf{v}, \boldsymbol{\mu}_p - \boldsymbol{\mu}_q \rangle_{\mathcal{H}} \right| = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_q\|_{\mathcal{H}}$$

Choose the function whose mean differs most between the two distributions.

Equal means can hide a difference

Each distribution places equal probability on its two values shown below.

$$p : \{-1, 1\}, \qquad q : \{-2, 2\}, \qquad \mathbb{E}_p[x] = \mathbb{E}_q[x] = 0$$

$$\boldsymbol{\phi}(x) = (x, x^2)^\mathsf{T} \implies \boldsymbol{\mu}_p = (0, 1)^\mathsf{T}, \quad \boldsymbol{\mu}_q = (0, 4)^\mathsf{T}$$

$$\mathrm{MMD}_k^2(p, q) = 0^2 + (1 - 4)^2 = 9$$

MMD can be computed using kernels

$$k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle_{\mathcal{H}}$$

$$\mathrm{MMD}_k^2(p, q) = \langle \boldsymbol{\mu}_p, \boldsymbol{\mu}_p \rangle + \langle \boldsymbol{\mu}_q, \boldsymbol{\mu}_q \rangle - 2\langle \boldsymbol{\mu}_p, \boldsymbol{\mu}_q \rangle$$

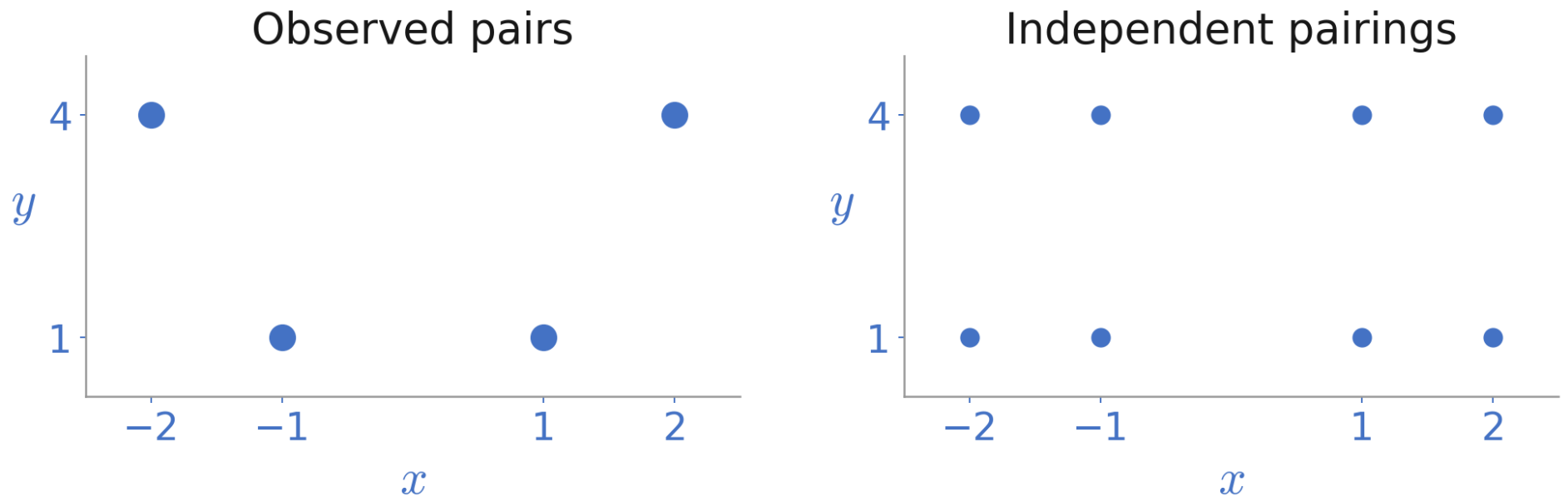
$$= \mathbb{E}_{p,p}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}_{q,q}[k(\mathbf{v}, \mathbf{v}')] - 2\mathbb{E}_{p,q}[k(\mathbf{x}, \mathbf{v})]$$

All draws in each expectation are independent. No explicit feature map is needed.

Independence is a distribution comparison

$$\mathbf{x} \perp y \quad \Longleftrightarrow \quad p(\mathbf{x}, y) = p(\mathbf{x})p(y)$$

Compare the observed pairings with pairings formed independently.



HSIC is MMD for this comparison

Use a kernel that compares both parts of a pair.

$$\kappa\big((\mathbf{x}, y), (\mathbf{x}', y')\big) = k(\mathbf{x}, \mathbf{x}')\, b(y, y')$$

$$\mathrm{HSIC}(\mathbf{x}, y) = \mathrm{MMD}^2_\kappa(p(\mathbf{x}, y),\ p(\mathbf{x})p(y))$$

With sufficiently rich kernels, such as Gaussian kernels, population HSIC is zero exactly at independence.

HSIC as squared cross-covariance

For finite feature maps, the joint feature is an outer product.

$$\boldsymbol{\mu}_\phi = \mathbb{E}[\boldsymbol{\phi}(\mathbf{x})], \qquad \boldsymbol{\mu}_\psi = \mathbb{E}[\boldsymbol{\psi}(y)]$$

$$C = \underbrace{\mathbb{E}[\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\psi}(y)^\mathsf{T}]}_{\text{joint mean}} - \underbrace{\boldsymbol{\mu}_\phi \boldsymbol{\mu}_\psi^\mathsf{T}}_{\text{independent mean}}$$

$$= \mathbb{E}\big[(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\mu}_\phi)(\boldsymbol{\psi}(y) - \boldsymbol{\mu}_\psi)^\mathsf{T}\big]$$

$$\mathrm{HSIC}(\mathbf{x}, y) = \|C\|_\mathrm{F}^2 = \sum_{a,b} c_{ab}^2$$

Centering with one matrix

$$X \in \mathbb{R}^{d \times n}, \qquad H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad \mathbf{1} = (1, \ldots, 1)^\top$$

$$XH = X - \frac{1}{n}X\mathbf{1}\mathbf{1}^\top = X - \bar{\mathbf{x}}\mathbf{1}^\top$$

$$XH = [\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}], \qquad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

$$H^\top = H, \qquad H^2 = H$$

Sample cross-covariance in feature space

$$\Phi = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}(\mathbf{x}_n)], \qquad \Psi = [\boldsymbol{\psi}(y_1), \ldots, \boldsymbol{\psi}(y_n)]$$

$$\widehat{C} = \frac{1}{n-1}(\Phi H)(\Psi H)^{\mathsf{T}} = \frac{1}{n-1}\Phi H \Psi^{\mathsf{T}}$$

$$K = \Phi^{\mathsf{T}}\Phi, \qquad B = \Psi^{\mathsf{T}}\Psi, \qquad K, B \in \mathbb{R}^{n \times n}$$

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \qquad b_{ij} = b(y_i, y_j)$$

Deriving the empirical HSIC formula

$$\widehat{\mathrm{HSIC}} = \|\widehat{C}\|_{\mathrm{F}}^2 = \mathrm{tr}(\widehat{C}\widehat{C}^{\mathsf{T}})$$

$$= \frac{1}{(n-1)^2}\,\mathrm{tr}\big(\Phi H \Psi^{\mathsf{T}} \Psi H \Phi^{\mathsf{T}}\big)$$

$$= \frac{1}{(n-1)^2}\,\mathrm{tr}\big(\Phi^{\mathsf{T}} \Phi H \Psi^{\mathsf{T}} \Psi H\big)$$

$$= \frac{1}{(n-1)^2}\,\mathrm{tr}(KHBH)$$

What a large HSIC score means

$$K_c = HKH, \qquad B_c = HBH$$

$$(n-1)^2 \widehat{\mathrm{HSIC}} = \mathrm{tr}(K_c B_c) = \sum_{i,j} k_{c,ij}\, b_{c,ij}$$

The score is large when the centered similarity patterns agree.

For classification, we want projected similarities to agree with class similarities.

The supervised PCA projection

$$X \in \mathbb{R}^{d \times n}, \quad X_c = XH, \quad \mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$$

$$U \in \mathbb{R}^{q \times d}, \qquad Z = UX_c \in \mathbb{R}^{q \times n}, \qquad q < d$$

$$\mathbf{z}_i = U(\mathbf{x}_i - \bar{\mathbf{x}})$$

Choose a projection whose output has large HSIC with the labels.

The label kernel defines supervision

For classification, observations in the same class are similar.

$$b(y_i, y_j) = \begin{cases} 1, & y_i = y_j, \\ 0, & y_i \neq y_j. \end{cases}$$

For regression, nearby response values can be treated as similar.

$$b(y_i, y_j) = \exp\left[-\frac{(y_i - y_j)^2}{2\sigma_y^2}\right], \qquad \sigma_y > 0$$

The kernel of the projected observations

Use a linear kernel on the projected observations.

$$K_Z = Z^\top Z = (UX_c)^\top (UX_c)$$

$$= X_c^\top U^\top U X_c = HX^\top U^\top U X H$$

$$(n-1)^2 \widehat{\mathrm{HSIC}}(Z, \mathbf{y}) = \mathrm{tr}(K_Z H B H)$$

Moving the projection outside the trace

$$\mathrm{tr}(K_Z H B H) = \mathrm{tr}(H X^\mathsf{T} U^\mathsf{T} U X H H B H)$$

$$= \mathrm{tr}(H X^\mathsf{T} U^\mathsf{T} U X H B H) \qquad (H^2 = H)$$

$$= \mathrm{tr}(U X H B H H X^\mathsf{T} U^\mathsf{T}) \qquad \text{(cycle the trace)}$$

$$= \mathrm{tr}(U X H B H X^\mathsf{T} U^\mathsf{T})$$

Fixing the scale of the projection

$$A = XHBHX^\mathsf{T} \in \mathbb{R}^{d \times d}, \qquad h(U) = \mathrm{tr}(UAU^\mathsf{T})$$

$h(tU) = t^2 h(U)$ so the scale must be constrained

$$\max_U \mathrm{tr}(UAU^\mathsf{T}) \qquad \text{subject to} \qquad UU^\mathsf{T} = I_q$$

The rows are unit projection directions, orthogonal to one another.

One direction: the eigenvalue equation

$$\max_{\mathbf{u}} \mathbf{u}^T A \mathbf{u} \qquad \text{subject to} \quad \mathbf{u}^T \mathbf{u} = 1$$

$$\mathcal{L}(\mathbf{u}, \lambda) = \mathbf{u}^T A \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1)$$

$$\nabla_{\mathbf{u}} \mathcal{L} = 2A\mathbf{u} - 2\lambda \mathbf{u} = \mathbf{0} \quad \Longrightarrow \quad A\mathbf{u} = \lambda \mathbf{u}$$

$$\mathbf{u}^T A \mathbf{u} = \lambda \mathbf{u}^T \mathbf{u} = \lambda \quad \Longrightarrow \quad \text{choose the largest eigenvalue}$$

Several directions: the leading eigenvectors

$$A\mathbf{u}_j = \lambda_j \mathbf{u}_j, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

$$U = \begin{bmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_q^\top \end{bmatrix}, \qquad UU^\top = I_q$$

$$\mathrm{tr}(UAU^\top) = \sum_{j=1}^{q} \mathbf{u}_j^\top A\mathbf{u}_j = \sum_{j=1}^{q} \lambda_j$$

Choose orthonormal eigenvectors corresponding to the largest eigenvalues.

A small example: data and labels

$$X = \begin{bmatrix} -3 & 3 & -3 & 3 \\ -1 & -1 & 1 & 1 \end{bmatrix}, \qquad \mathbf{y} = (0, 0, 1, 1)^\mathsf{T}$$

The columns are observations. Each feature already has mean zero.

$$B = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad XH = X$$

A small example: forming the matrix

$$A = XHBHX^\top = XBX^\top$$

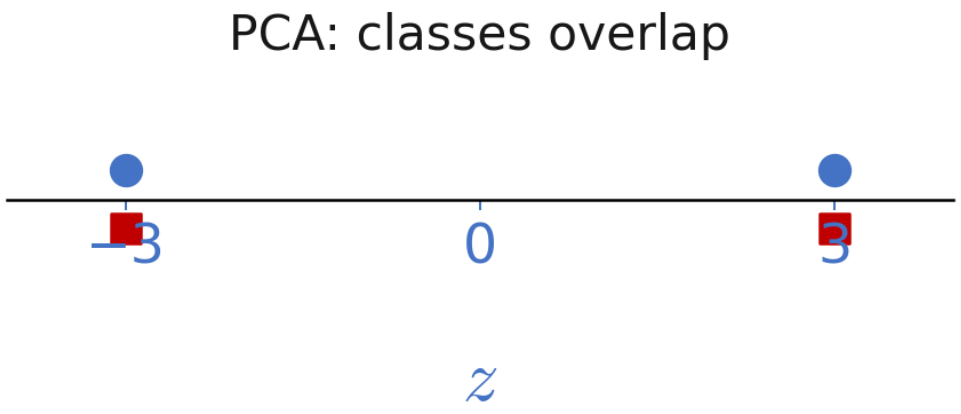
$$XB = \begin{bmatrix} -3 & 3 & -3 & 3 \\ -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 2 \end{bmatrix}$$

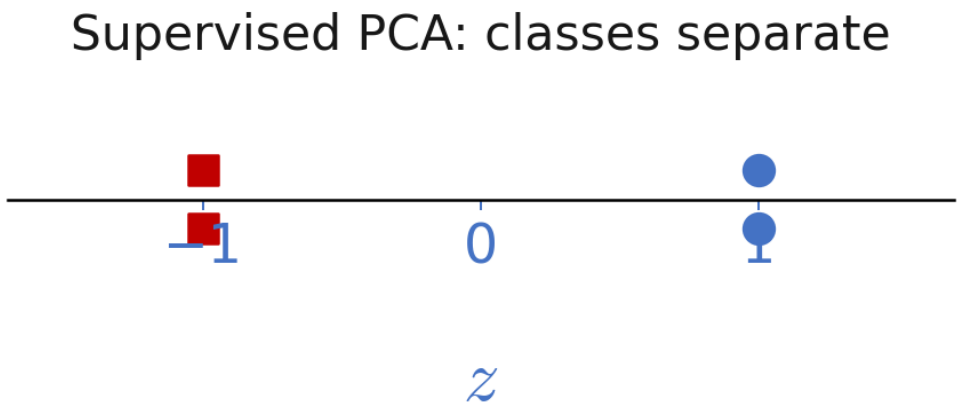
$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 2 \end{bmatrix} \begin{bmatrix} -3 & -1 \\ 3 & -1 \\ -3 & 1 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 8 \end{bmatrix}$$

The two projections keep different information

$$XX^\top = \begin{bmatrix} 36 & 0 \\ 0 & 4 \end{bmatrix} \implies U_{\text{PCA}} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 8 \end{bmatrix} \implies U_{\text{SPCA}} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$





Vertical offsets only separate overlapping markers.

Ordinary PCA is a special case

$$B = I_n \quad \Longrightarrow \quad A = XHI_nHX^\top = XHX^\top$$

$$= X_cX_c^\top = (n-1)S, \qquad S = \frac{1}{n-1}X_cX_c^\top$$

The leading eigenvectors are those of the usual sample covariance matrix.

Using the identity matrix removes label information from the objective.

Training and predicting

Training: center the data, form the label kernel, and find the leading eigenvectors.

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i, \qquad A = XHBHX^\top, \qquad Z = U(X - \bar{\mathbf{x}}\mathbf{1}^\top)$$

Fit a classifier or regression model using the projected training observations.

$$\mathbf{z}_{\text{new}} = U(\mathbf{x}_{\text{new}} - \bar{\mathbf{x}}), \qquad \widehat{y}_{\text{new}} = g(\mathbf{z}_{\text{new}})$$

The connection to metric learning

$$\|\mathbf{z}_i - \mathbf{z}_j\|^2 = \|U(\mathbf{x}_i - \mathbf{x}_j)\|^2$$

$$= (\mathbf{x}_i - \mathbf{x}_j)^\mathsf{T} U^\mathsf{T} U (\mathbf{x}_i - \mathbf{x}_j)$$

$$M = U^\mathsf{T} U \succeq 0, \qquad \text{rank}(M) = q$$

Supervised PCA learns a projection, which also defines a distance for k-NN.

What this linear formulation can miss

The eigenvalue solution uses a linear kernel on the projected observations.

With the same-class label kernel, it uses differences between class means.

Classes with equal means but different spreads can therefore give zero score.

Nonlinear input features can reveal differences that a linear projection misses.

From a dependence measure to a projection

MMD compares distributions through their feature means.

HSIC applies that comparison to the joint distribution and independent marginals.

$$\max_{UU^\top = I_q} \operatorname{tr}(UXHBHX^\top U^\top)$$

Supervised PCA keeps the directions that make the projected data most dependent on the labels.

Lecture 17