TUTORIAL WEEK 10

MLE and MAP

Reminder: Bayes' Rule Notation

$$\underbrace{p(w|D)}_{\text{Posterior}} = \frac{\overbrace{p(D|w)}^{\text{Likelihood}} \cdot \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Marginal Probability of D}}}$$

Reminder: Maximum Likelihood Estimation (MLE)

$$\underbrace{p(w|D)}_{\text{Posterior}} = \frac{\boxed{\overbrace{p(D|w)}^{\text{Likelihood}}} \cdot \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Marginal Probability of D}}}$$

MLE: $\underset{w}{\operatorname{argmax}}\, p(D|w)$

- Based on the *likelihood* term
- Find *w* that makes *D* the highest probability

Reminder: Maximum A Posteriori (MAP)

$$\underbrace{\boxed{p(w|D)}}_{\text{Posterior}} = \frac{\boxed{\overbrace{p(D|w)}^{\text{Likelihood}} \cdot \overbrace{p(w)}^{\text{Prior}}}}{\underbrace{p(D)}_{\text{Marginal Probability of D}}}$$

MAP: $\underset{w}{\operatorname{argmax}}\; p(w|D) \propto \underset{w}{\operatorname{argmax}}\; p(D|w)p(w)$

- Based on the *posterior*
- Find *w* with the highest probability given *D*
  - Through Bayes' Rule we find that this is proportional to the likelihood and the prior

MLE vs MAP

MLE: $\underset{w}{\operatorname{argmax}}\, p(D|w)$

MAP: $\underset{w}{\operatorname{argmax}}\, p(D|w)p(w)$

- Conceptually different
- Mathematically differ only by multiplication by the prior

Reminder: Negative Log Likelihoods (NLL) (2 min)

- Why is log helpful when optimizing functions?

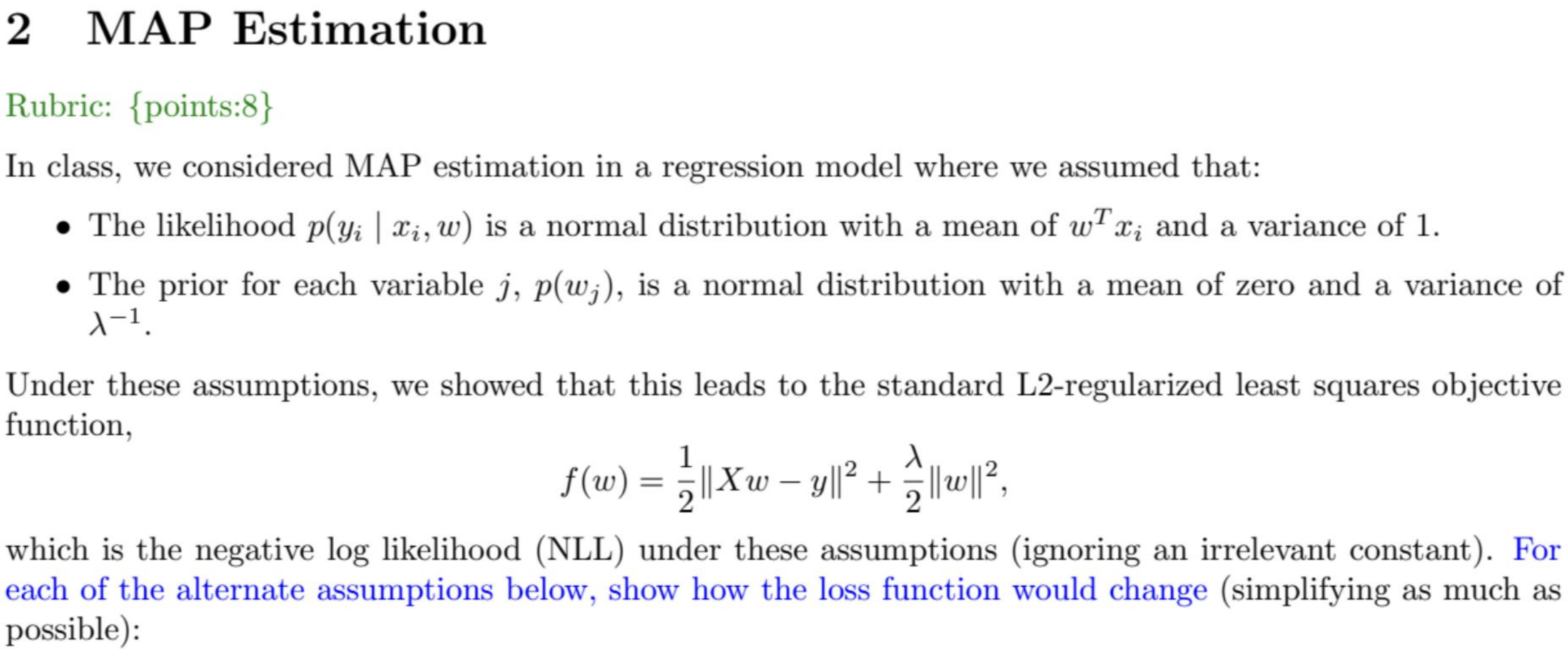
- Why is it mathematically permissible to take the log of a function when we're maximizing it?

- What do we have to do to a maximization problem to make it mathematically equivalent to minimization?

Reminder: Negative Log Likelihoods (NLL) (2 min)

- Why is log helpful when optimizing functions?
  - Answer: It converts products into sums. Since sums are linear under differentiation, this simplifies computing partial derivatives.
- Why is it mathematically permissible to take the log of a function when we're maximizing it?
  - Answer: The logarithm function is monotonic ($a < b \rightarrow \log(a) < \log(b)$)
- What do we have to do to a maximization problem to make it mathematically equivalent to minimization?
  - Answer: Take the negative i.e. $\operatorname{argmax} f(w) = \operatorname{argmin} -f(w)$

MAP Motivation: Assignment 5



2 MAP Estimation

Rubric: {points:8}

In class, we considered MAP estimation in a regression model where we assumed that:

- The likelihood $p(y_i \mid x_i, w)$ is a normal distribution with a mean of $w^T x_i$ and a variance of 1.
- The prior for each variable j, $p(w_j)$, is a normal distribution with a mean of zero and a variance of λ^{-1}.

Under these assumptions, we showed that this leads to the standard L2-regularized least squares objective function,

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2,$$

which is the negative log likelihood (NLL) under these assumptions (ignoring an irrelevant constant). For each of the alternate assumptions below, show how the loss function would change (simplifying as much as possible):

MLE Example (4 mins)

The normal distribution notation is $N(\mu, \sigma^2)$

Given $y_i|x_i, w \sim N(w^T x_i, 1)$, which means:

$$p(y_i|x_i, w) = \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1})$$

- Given this information:
  - 1) Write the likelihood function for N training examples
  - 2) Use the likelihood function to write an expression for the MLE
  - 3) Solve for the MLE (in matrix notation)

MLE Example: Solution

(1)
$$\boxed{p(y|X,w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\cdot 1\cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2\cdot 1}) \qquad (14)}$$

(2)
$$\begin{aligned}
\arg\max_w p(y|X,w) &= \arg\max_w \prod_{i=1}^{N} \frac{1}{\sqrt{2\cdot 1\cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2\cdot 1}) \\
&= \arg\max_w \log(\frac{1}{\sqrt{2\pi}}) + \sum_{i=1}^{N} \log(\exp(-\frac{(w^T x_i - y_i)^2}{2})) \quad \text{(log is monotonic)} \\
&= \arg\max_w -\sum_{i=1}^{N} \frac{(w^T x_i - y_i)^2}{2} \\
&= \arg\min_w \frac{1}{2}\cdot ||Xw - y||_2^2 \quad \text{(negate both sides)}
\end{aligned}$$

(3)
$$= \arg\min_w ||Xw - y||_2^2 \quad \text{(does not change solution)}$$

MAP Example (4 mins)

The normal distribution notation is $N(\mu, \sigma^2)$

Given $y_i | x_i, w \sim N(w^T x_i, 1)$, which means:

$$p(y_i | x_i, w) = \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1})$$

Also assume that our weights are distributed as $w_j \sim N(0, \lambda^{-1})$, i.e.:

$$p(w_j) = \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}})$$

- Given this information:
  - 1) Write the likelihood and prior functions for N training examples
  - 2) Use these functions to write an expression for the MAP
  - 3) Solve for the MAP (in matrix notation)

MAP Example: Solution

(1)

$$p(y|X,w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}) \qquad y_i|x_i, w \sim N(w^T x_i, 1)$$

$$p(w) = \prod_{j=1}^{d} \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}) \qquad w_j \sim N(0, \lambda^{-1})$$

$$\arg\max_{w} p(w|X,y) = \arg\max_{w} p(y|X,w) \cdot p(w)$$

(2)

$$= \arg\max_{w} \log(p(y|X,w)) + \log(\prod_{j=1}^{d} \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}))$$

$$= \arg\max_{w} \log(p(y|X,w)) + \sum_{j=1}^{d} \log(\exp(-\frac{\lambda}{2} w_j^2))$$

MAP Example: Solution

$$\arg\max_w p(w|X,y) = \arg\max_w p(y|X,w)\cdot p(w)$$

$$= \arg\max_w \log(p(y|X,w)) + \log(\prod_{j=1}^{d} \frac{1}{\sqrt{2\cdot\lambda^{-1}\cdot\pi}} \exp(-\frac{(w_j-0)^2}{2\cdot\lambda^{-1}}))$$

$$= \arg\max_w \log(p(y|X,w)) + \sum_{j=1}^{d} \log(\exp(-\frac{\lambda}{2}w_j^2))$$

$$= \arg\min_w -\log(p(y|X,w)) + \sum_{i=1}^{d} \frac{\lambda}{2}w_j^2 \quad \text{(negate both sides)}$$

(3) $$= \arg\min_w \frac{1}{2}||Xw-y||^2 + \frac{\lambda}{2}||w||^2$$

Assignment Suggestions

- Follow the same general procedure
  - *Simplify using operations such as applying the log and negative sign to obtain negative log likelihood*
  - *Show the new loss function*

- General form (for MAP), given: $p(y|X,w) = \ ?$

  $p(w) = \ ?$

- Want to find:

$$\arg\max_{w} P(w|X,y) = \arg\max_{w} p(y|X,w) \cdot p(w)$$

Apply log + negative

$$= \arg\min_{w} \ \text{Loss}$$

MAP Example

$$w_j \sim N(0, \lambda^{-1}) \longrightarrow + \frac{\lambda}{2}||w||^2$$

- MAP loss here differs only by a regularization term
- The regularization term here which we associate with L2 regularization comes from assuming a normal distribution as a prior
- *Bonus*: Is there a prior distribution we can assume that results in a regularization term of 0?

PCA Review

Refresher Question

- PCA always produces a unique solution (True/False)

Refresher Question

- PCA always produces a unique solution (True/False)
- Answer: False. PCA does not produce a unique solution since solutions are equivalent under scalar multiplication.

Refresher Question

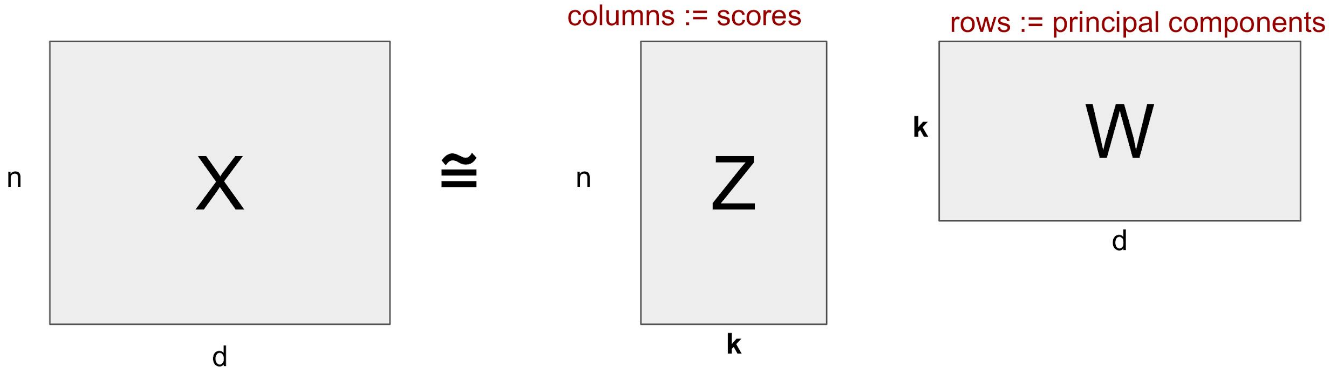
- What modifications do we make to PCA to make our solutions unique?

Refresher Question

- What modifications do we make to PCA to make our solutions unique?
- Answer: Normalization, orthogonality, sequential fitting

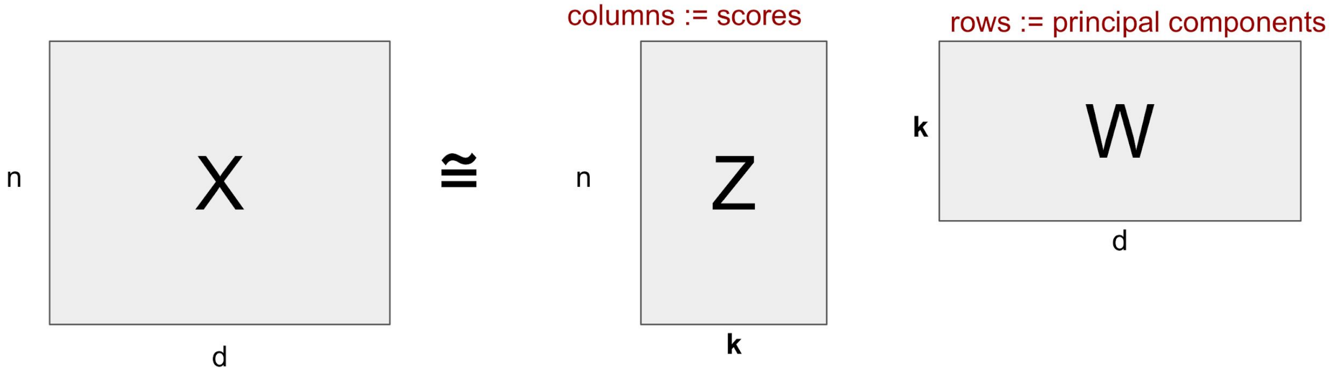
Reminder: PCA

- **P**rincipal **C**omponent **A**nalysis
  - Factor our data into "components" that best describe variance observed in the features
- Data is factored into two matrices:
  - W contains the principal components
    - Each principal component can be thought of as a vector
  - Z contains the scores
    - How much of each component to mix to recover our example
- Thus we are trying to break down our data into a set of vectors that best describe the variance, and the linear combination of those vectors that we have to take to obtain our example in that basis



Reminder: PCA

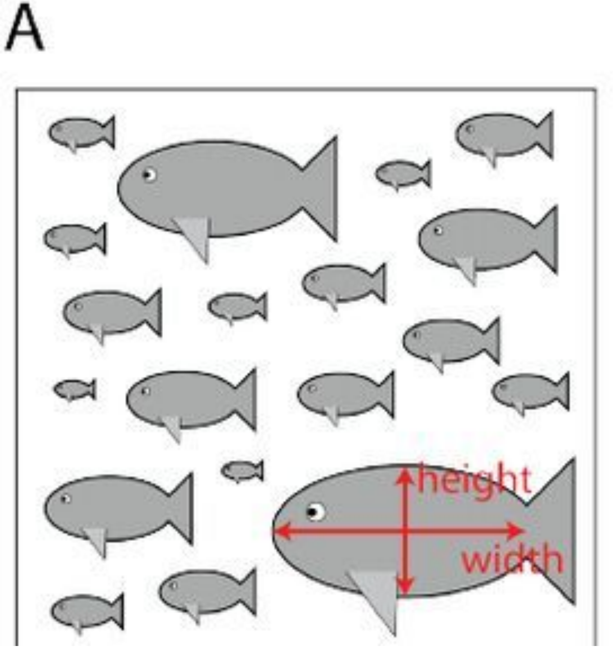
- In particular when k < d, PCA can be thought of as examining our data in a lower dimensional feature space
  - Each example is described by k features in the matrix Z as opposed to d scores in the matrix X
- PCA is a *latent factor model*
  - Output of PCA is a basis that factors our data



PCA Example Questions

PCA Example

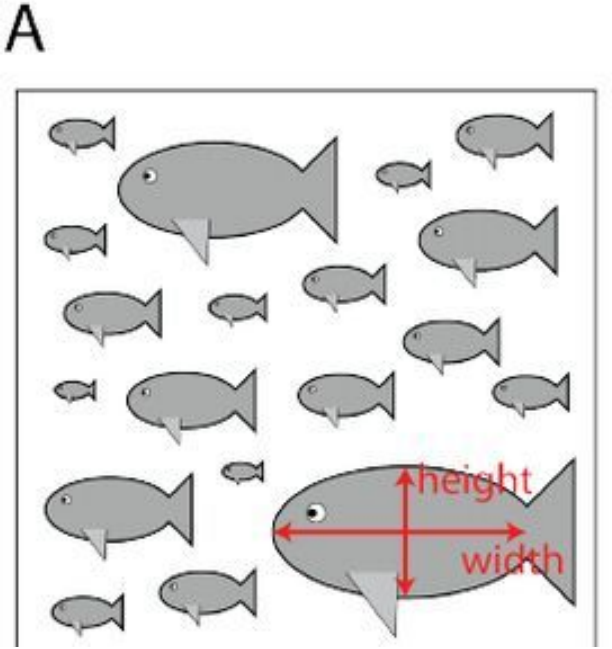
- Suppose we have a population of fish, and our features for each fish consist of its width and height
- Suppose we want to reduce the dimensionality of our data
- How might PCA be applied to this dataset?



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

PCA Example

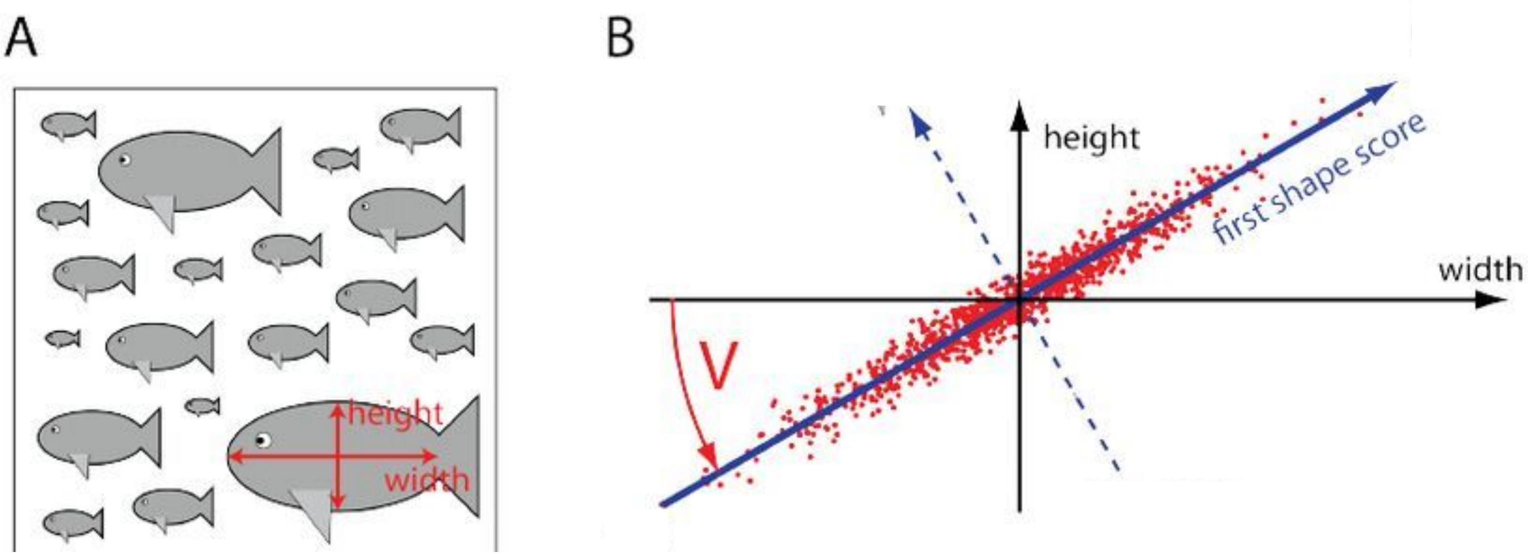
- Suppose we have a population of fish, and our features for each fish consist of its width and height
- Suppose we want to reduce the dimensionality of our data
- How might PCA be applied to this dataset?
- Answer: We might use PCA to describe our fish population in a single dimension rather than 2 (reduce the dimensionality of our data)



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Example: Axis Finding

- If there is a relation (correlation) between height and width, then a large part of the variance of our data can be described by a single number (1 dimension)
- PCA finds the "best representation" in this lower dimension (i.e. reduces dimensionality)



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Reminder: PCA

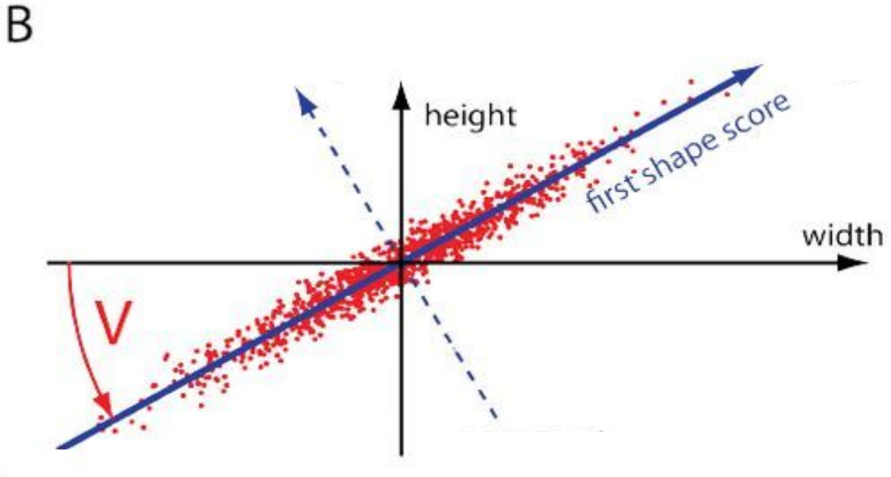
- To clarify what we mean by "best representation":
  - In our original feature space, what are we trying to minimize?

PCA objective:

$$f(Z,W) = \sum_{i=1}^{n} \sum_{j=1}^{2} (w_j z_i - x_{ij})^2$$

nent"

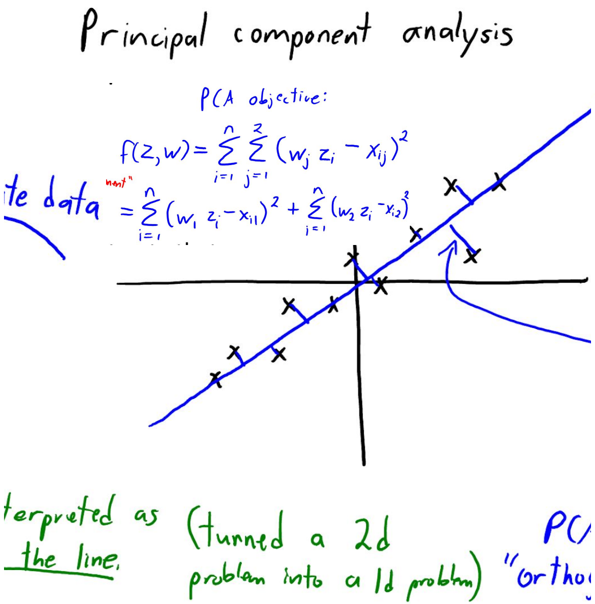
$$= \sum_{i=1}^{n} (w_1 z_i - x_{i1})^2 + \sum_{j=1}^{n} (w_2 z_i - x_{i2})^2$$



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Reminder: PCA

- In our original feature space, what are we trying to minimize?
- Answer: PCA can be interpreted as minimizing the orthogonal distance between our axis and data in the original feature space. So we define "best representation" as being that which

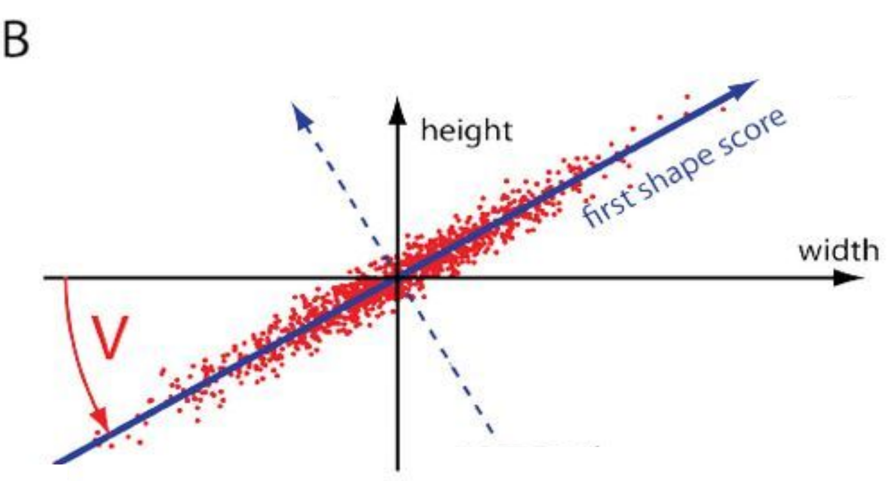


Principal component analysis

PCA objective:

$$f(Z,W) = \sum_{i=1}^{n} \sum_{j=1}^{2} (w_j z_i - x_{ij})^2$$

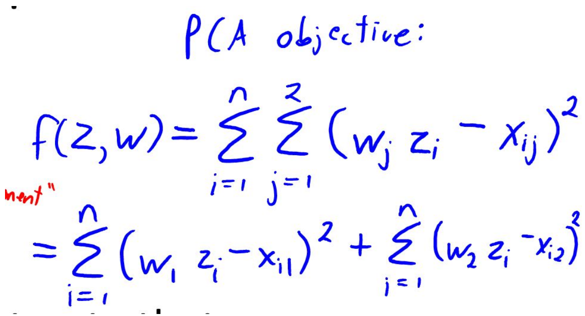
$$= \sum_{i=1}^{n} (w_1 z_i - x_{i1})^2 + \sum_{j=1}^{n} (w_2 z_i - x_{i2})^2$$



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Reminder: PCA

- In our original feature space, what are we trying to minimize?
- Answer 2: Considering the objective function, PCA picks a first PC in the direction that maximizes the variance. This is the statistical interpretation of how PCA is describing the data.

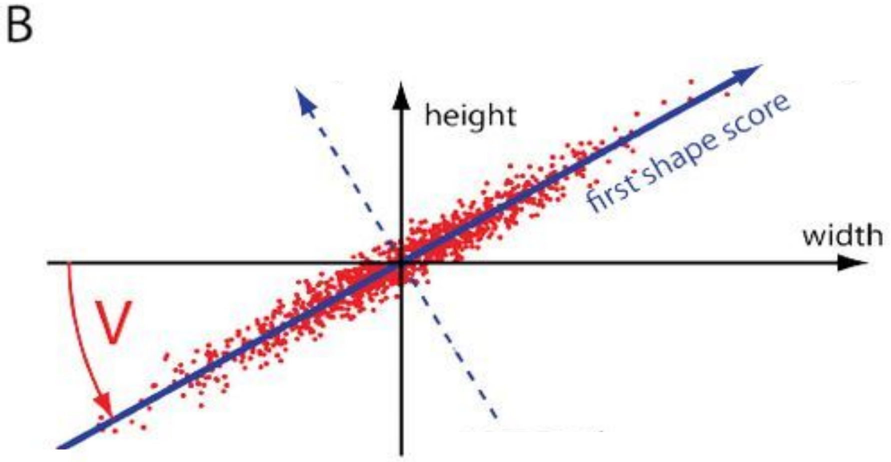


PCA objective:

$$f(Z,w) = \sum_{i=1}^{n} \sum_{j=1}^{2} (w_j z_i - x_{ij})^2$$

ment"

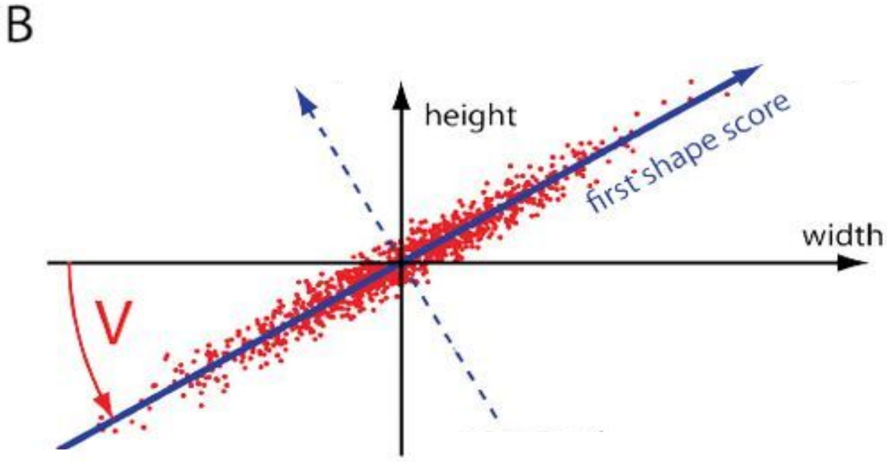
$$= \sum_{i=1}^{n} (w_1 z_i - x_{i1})^2 + \sum_{j=1}^{n} (w_2 z_i - x_{i2})^2$$



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Reminder: PCA

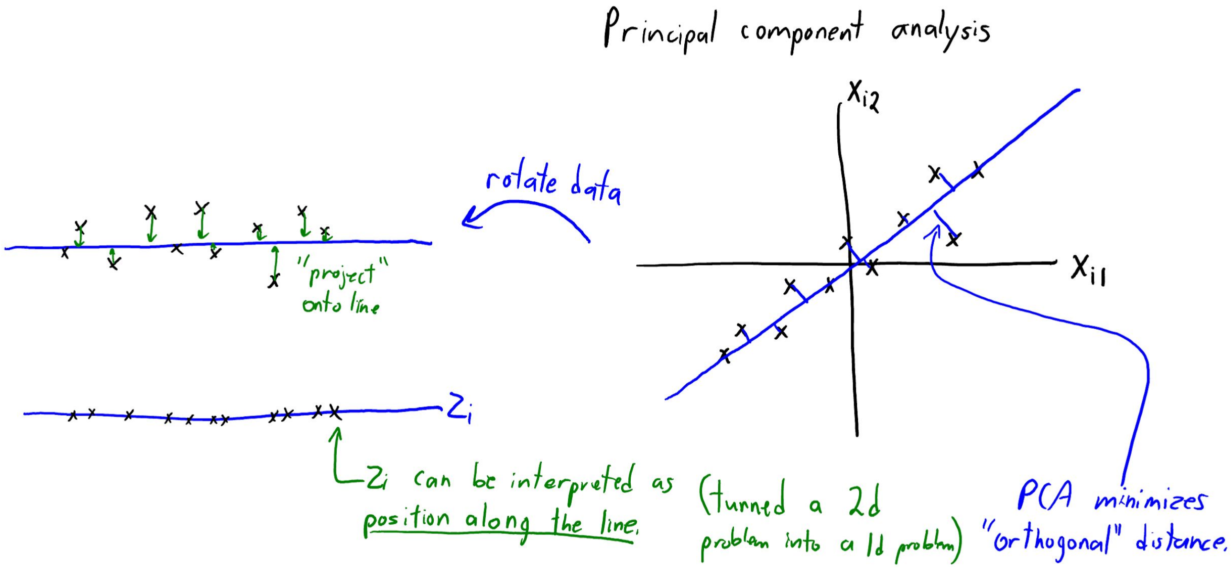
- In this simple example, how might you interpret the first principal component score (zi) of an example?



Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

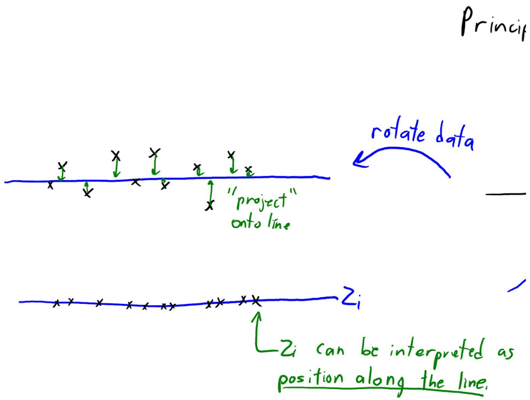
Reminder: PCA

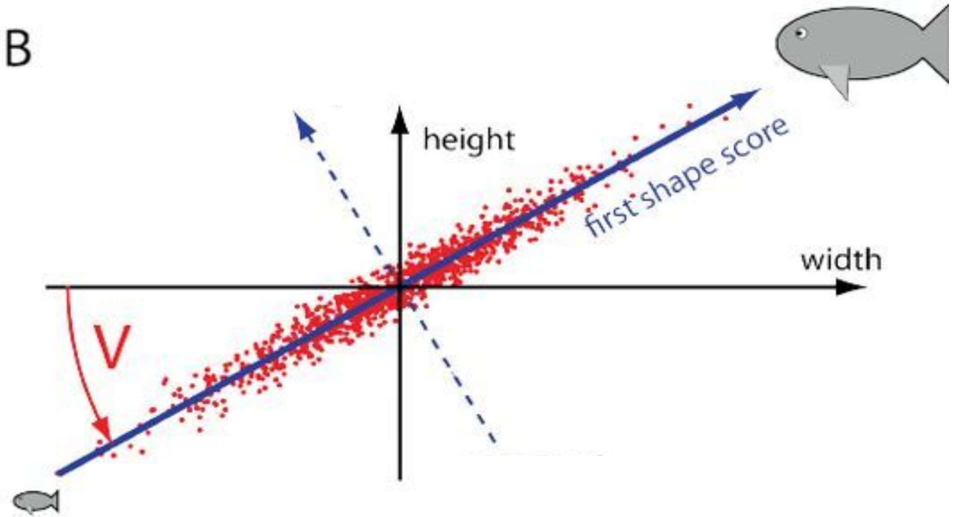
- In this simple example, how might you interpret the first principal component score (zi) of an example?
- Recall that our PC score zi can be thought of as position along a line


Principal component analysis
x_{i2}
x_{i1}
rotate data
"project" onto line
z_i
z_i can be interpreted as position along the line.
(turned a 2d problem into a 1d problem)
PCA minimizes "orthogonal" distance.

Reminder: PCA

- The first PC here describes width and height both increasing linearly with the value of our first PC
  - We can interpret our first PC score as roughly describing "size"
- Fish with large first PC scores will have larger width and height than small first PC scores

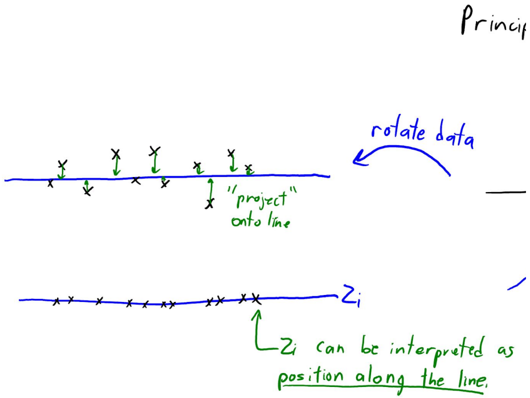


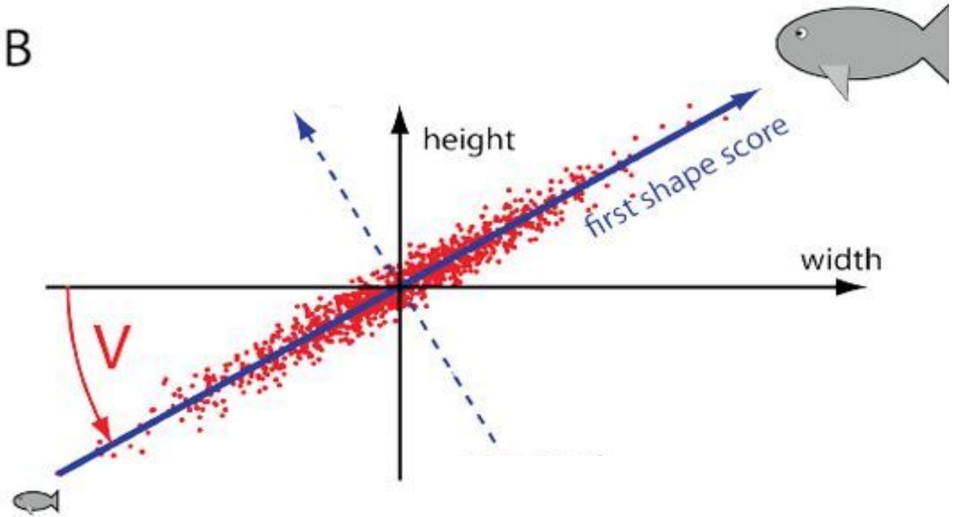


Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Reminder: PCA

- Thus we have gone from describing our data in terms of a width and a height to a single dimension roughly described as size
- Our first PC here reduces the dimensionality well, and so is likely to be useful





Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

Question

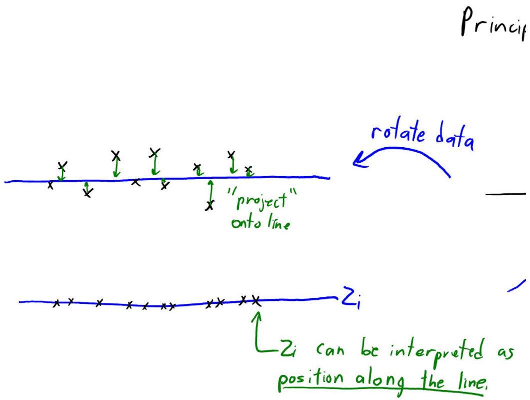
- Could you use PCA if k = 2? What about k > 2?

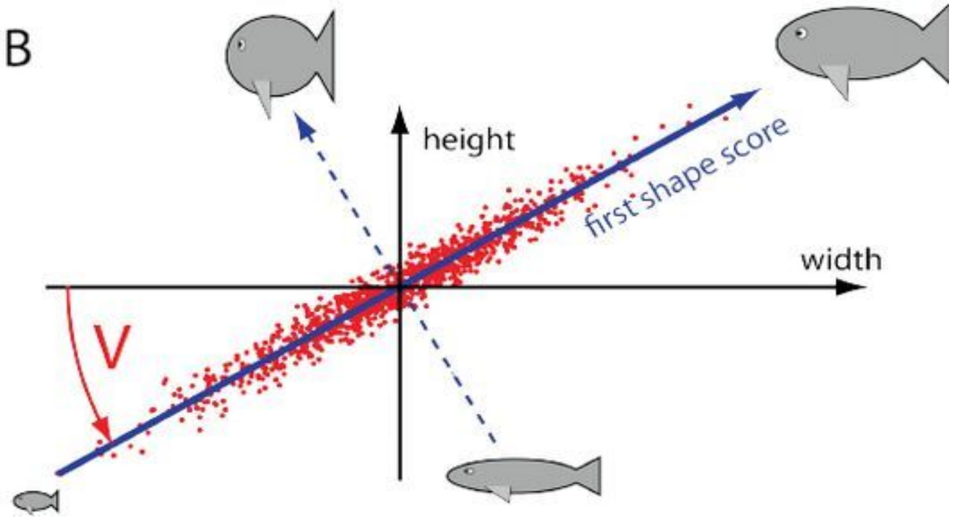
Question

- Could you use PCA if k = d? What about k > d?
- Answer:
  - For k = 2, you can, but any 2D solution will be capable of fully describing our data, and will be a degenerate solution of the PCA objective
  - For k > 2 the components of the solution set z_i will be linearly dependent, and will not provide any added depth

Reminder: PCA

- Consider fitting a second principal component
- If we enforce that the second principal component must be orthogonal to the first, then it will describe the data's variation about the first principal component
- With two PCs we can completely describe the data, and so even though we have defined a unique SVD based solution, there are infinitely many other possible solutions which also minimize the PCA objective in 2D



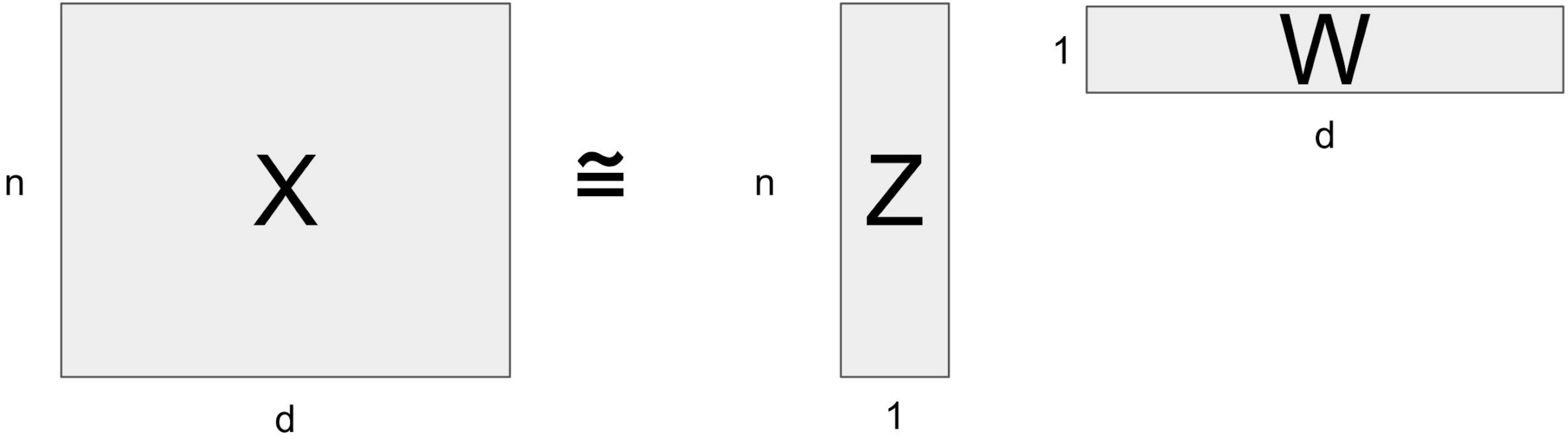


Shape Mode Analysis Exposes Movement Patterns in Biology: Flagella and Flatworms as Case Studies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Illustration-of-principal-component-analysis-A-As-a-minimal-example-we-consider-a_fig1_263968032 [accessed 13 Nov, 2018]

PCA Calculation Questions

K = 1 Case

- We'll consider conceptually the case where k = 1
- How do we find W (aka the first Principal Component)?



bonus!

PCA Computation: SVD

- How do we fit with normalization/orthogonality/sequential-fitting?
  - It can be done with the "singular value decomposition" (SVD).
  - Take CPSC 302.

- 4 lines of Python code:
  - mu = np.mean(X,axis=0)
  - X -= mu
  - U, s, Vh = np.linalg.svd(X)
  - W = Vh[:k]

- Computing Z is cheaper now:

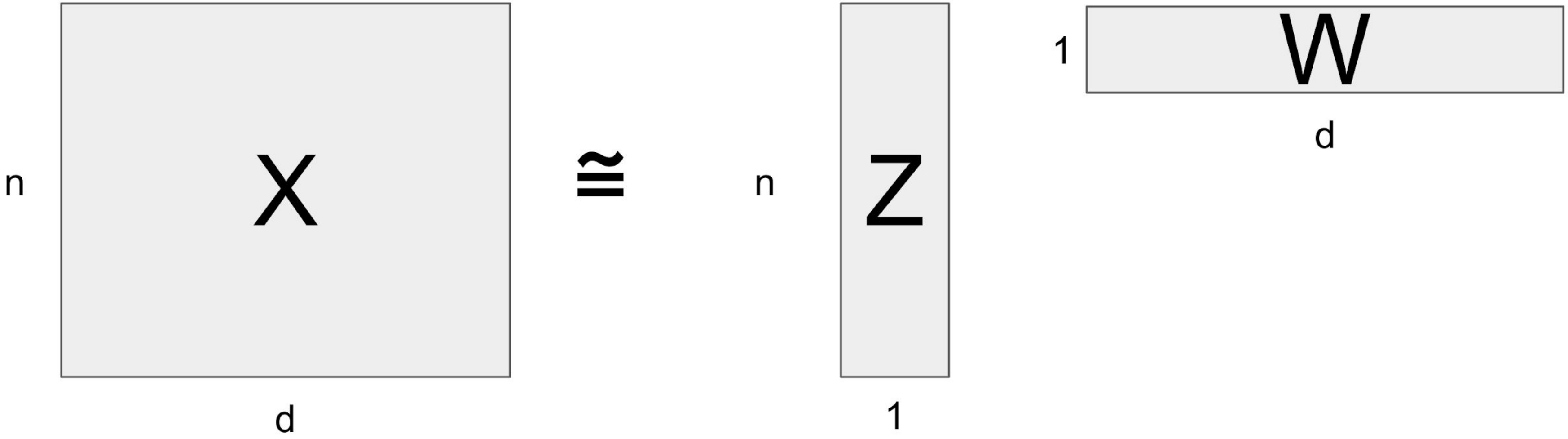
$$Z = XW^T(\underbrace{WW^T}_{})^{-1} = XW^T$$

$$WW^T = \begin{bmatrix} - & w_1 & - \\ - & w_2 & - \\ & \vdots & \\ - & w_k & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ w_1^T & w_2^T & \cdots & w_k^T \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & & \ddots & & 0 \\ \vdots & & & & \\ 0 & \cdots & & 0 & 1 \end{bmatrix} = I$$

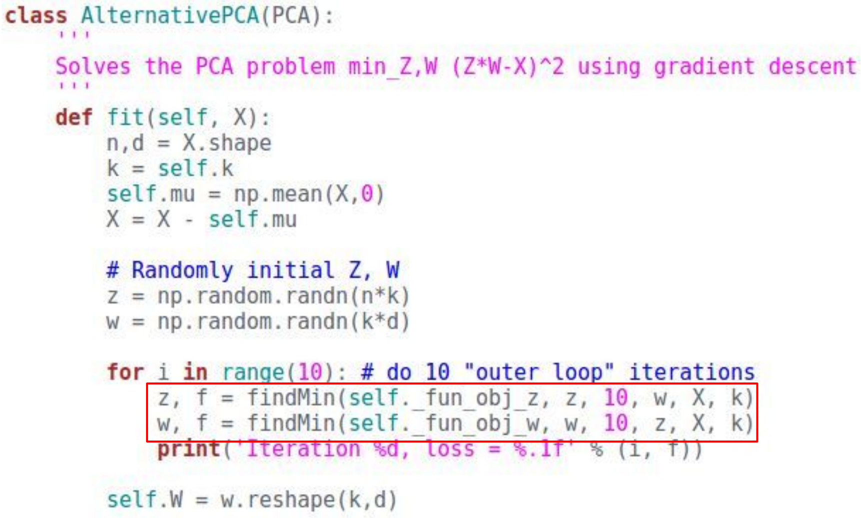
K = 1 Case

- How do we find W (aka the first Principal Component)?
- Answer: Use a minimization algorithm :
  - Alternating Minimization
  - Stochastic Gradient Descent



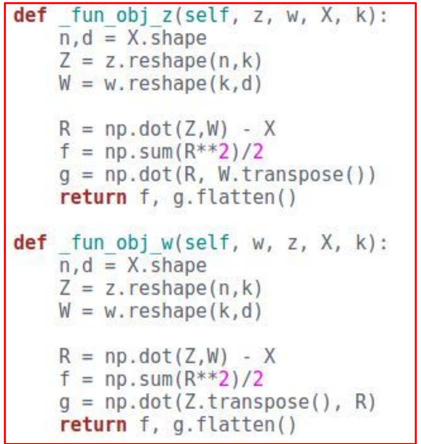
A5 Minimization

- Class from assignment 5 uses alternating minimization and gradient descent
- Alternates computes closed-form gradients with respect to Z and W



```python
class AlternativePCA(PCA):
    '''
    Solves the PCA problem min_Z,W (Z*W-X)^2 using gradient descent
    '''
    def fit(self, X):
        n,d = X.shape
        k = self.k
        self.mu = np.mean(X,0)
        X = X - self.mu

        # Randomly initial Z, W
        z = np.random.randn(n*k)
        w = np.random.randn(k*d)

        for i in range(10): # do 10 "outer loop" iterations
            z, f = findMin(self._fun_obj_z, z, 10, w, X, k)
            w, f = findMin(self._fun_obj_w, w, 10, z, X, k)
            print('Iteration %d, loss = %.1f' % (i, f))

        self.W = w.reshape(k,d)
```

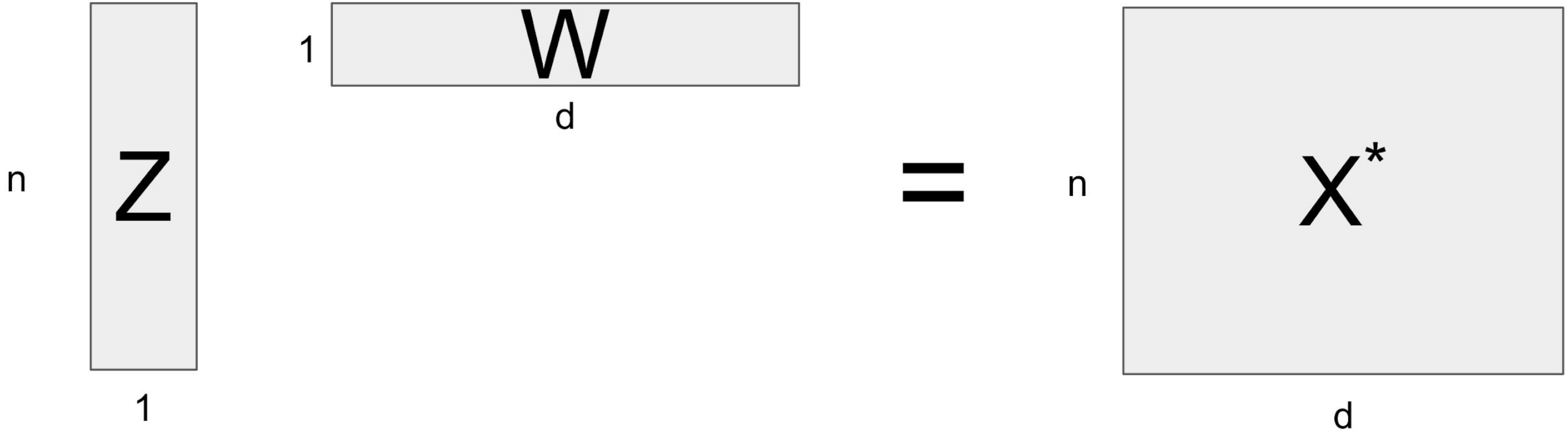
$$\nabla_Z\ f(W,Z)$$
$$\nabla_W\ f(W,Z)$$



```python
def _fun_obj_z(self, z, w, X, k):
    n,d = X.shape
    Z = z.reshape(n,k)
    W = w.reshape(k,d)

    R = np.dot(Z,W) - X
    f = np.sum(R**2)/2
    g = np.dot(R, W.transpose())
    return f, g.flatten()

def _fun_obj_w(self, w, z, X, k):
    n,d = X.shape
    Z = z.reshape(n,k)
    W = w.reshape(k,d)

    R = np.dot(Z,W) - X
    f = np.sum(R**2)/2
    g = np.dot(Z.transpose(), R)
    return f, g.flatten()
```

Reconstruction with k = 1

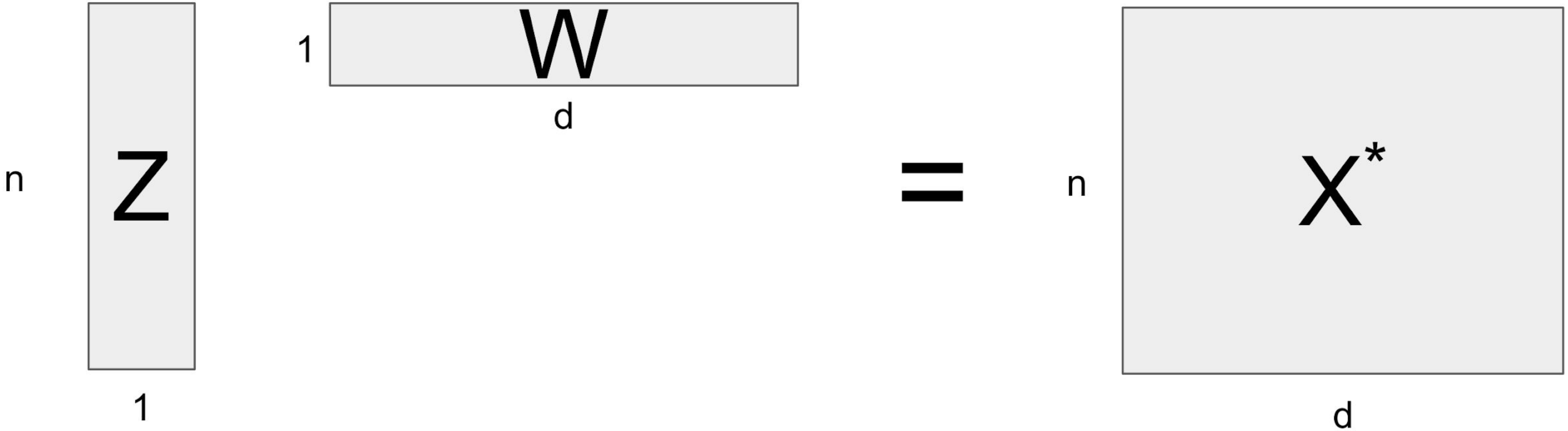
- Suppose we now take our principal components and multiply them out
- Is our output X* the same as X?



Reconstruction with k = 1

- Suppose we now take our principal components and multiply them out
- Is our output X* the same as X?

Answer: If our data is completely describable with a single PC then yes, more generally no, since some information will be lost in the dimensionality reduction, our reconstructed X* will be not quite the same as what we started with.



Reconstruction with k = 1

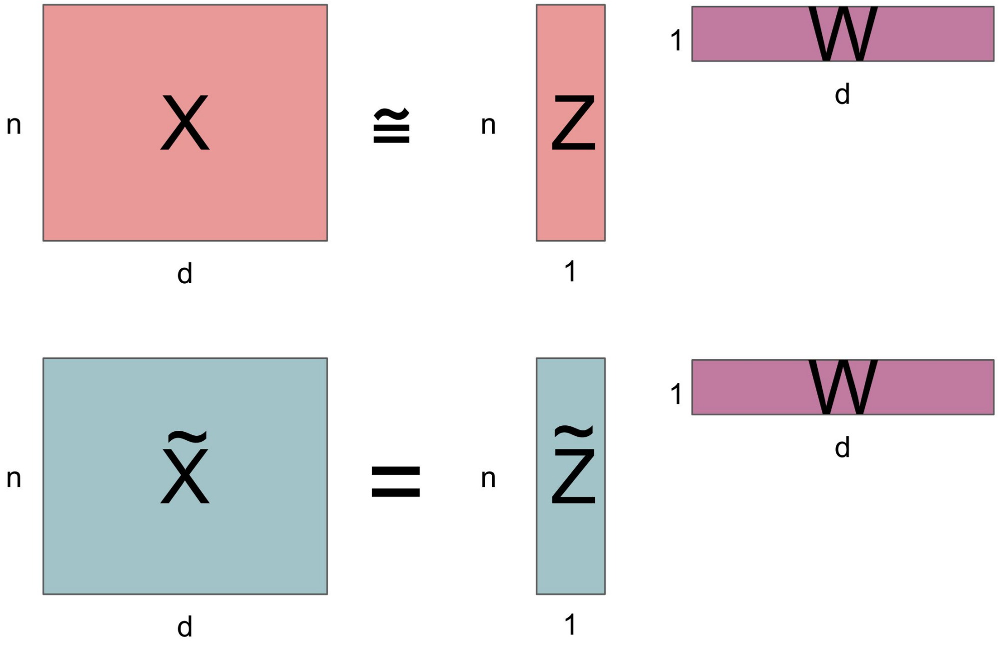
- Refresher: Suppose we train a linear classifier on the features we derived with PCA.
- i) Why might this fact (X =/= X*) be a good thing?
- ii) Why might it be a bad thing?

Reconstruction with k = 1

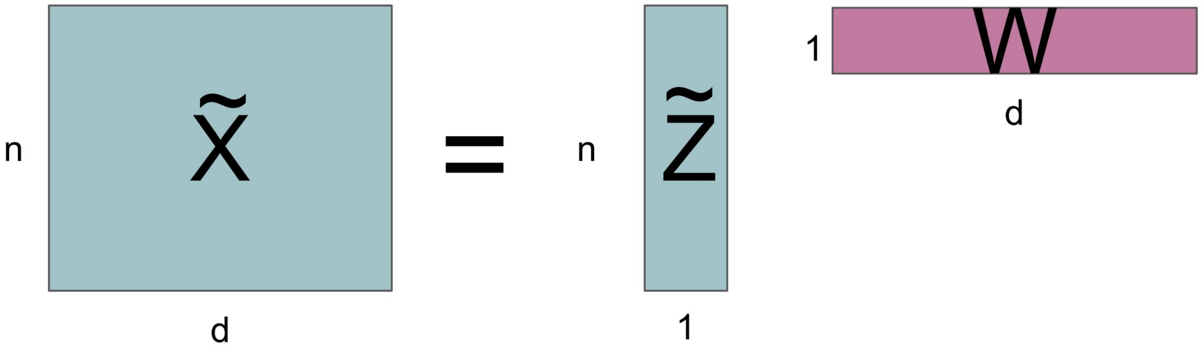
- Refresher: Suppose we train a linear classifier on the features we derived with PCA.
- i) Why might this fact (X =/= X*) be a good thing?
- ii) Why might it be a bad thing?
- Answer:
  - i) We've reduced the complexity of our description of our data while trying to retain the most relevant features. This might reduce overfitting by not giving the model overly specific data about training examples.
  - ii) We may have lost information relevant to our classification task (especially since we're only using k = 1 here)

Transforming Test Data in k = 1 Case

- Consider now our test set $\tilde{X}$
- We would like to factor our test set into the same form as for our training set
- i.e. We would like to describe our test set in terms of the same z1 (same W)
- This requires us to compute the class scores $\tilde{Z}$



Transforming Test Data in k = 1 Case



How do we find $\tilde{Z}$ given W ?

Transforming Test Data in k = 1 Case

- To form Z given W:

$$X = ZW$$

$$XW^T = ZWW^T$$

Rows of W are orthonormal

$$XW^T = ZI_k$$

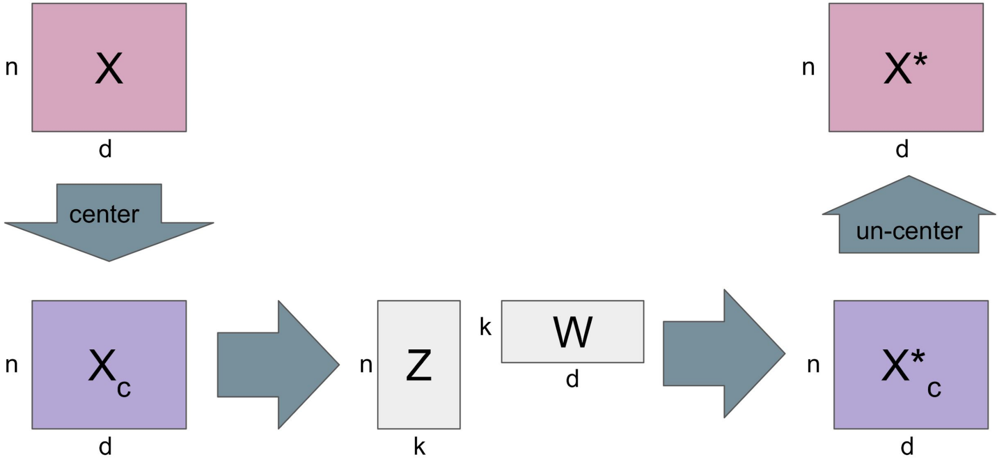
$$XW^T = Z$$

Reminder: PCA with non-centered data

- The PCA objective assumes that the data have features with a mean of 0
- To use PCA on uncentered data, we first subtract the mean of the feature across the training set from each feature

Set $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ (mean of colum 'j')

Replace each x_{ij} with $(x_{ij} - \mu_j)$



Questions