# Project Groups

So far, I got the following groups:

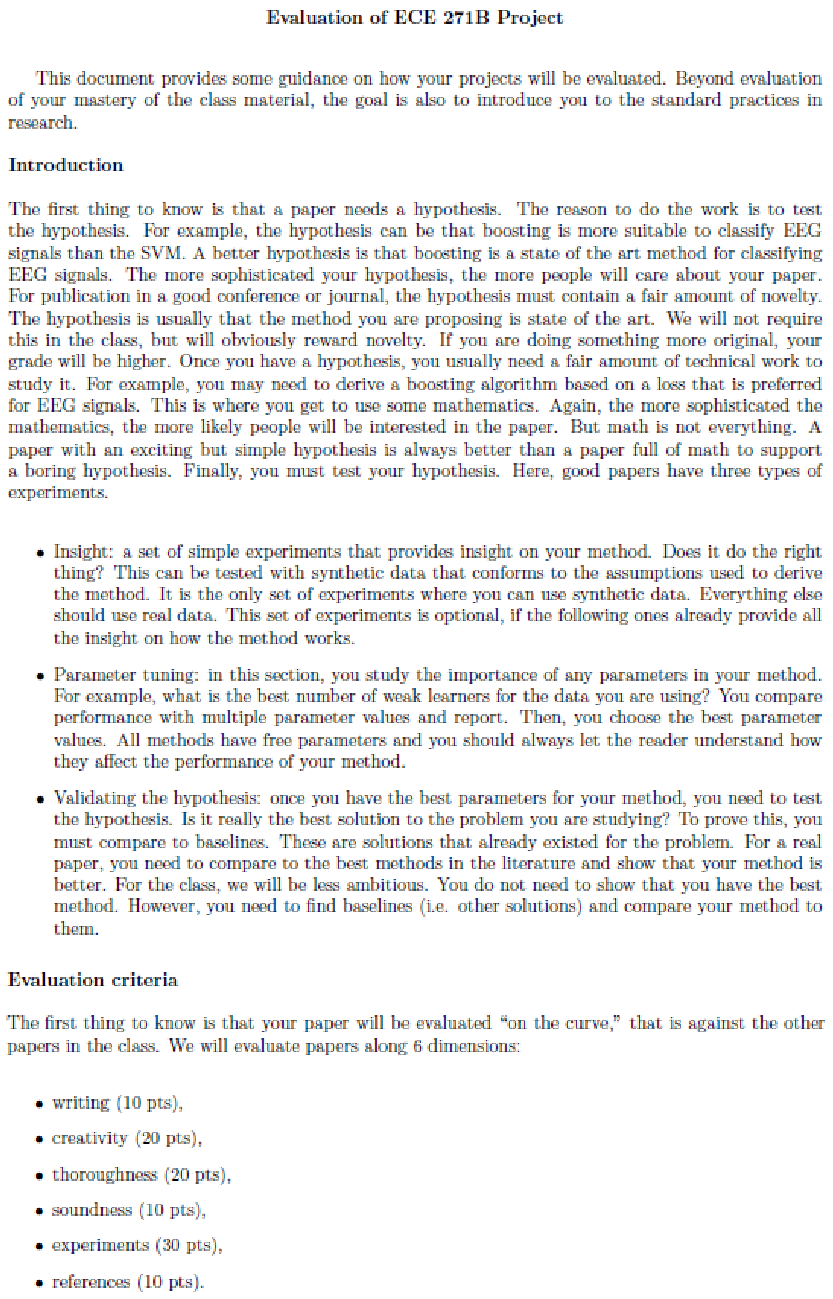
1. **Hussain**, Tanvir; **Lewis**, Cameron; **Villamar**, Sandra
2. **Dong**, Meng; **Long**, Jianzhi; **Wen**, Bo; **Zhang**, Haochen
3. **Chen**, Yuzhao; **Li**, Zonghuan; **Song**, Yuze; **Yan**, Ge
4. **Li**, Jiayuan; **Xiao**, Nan; **Yu**, Nancy; **Zhou**, Pei
5. **Li**, Zheng; **Tao**; Jianyu; **Yang**, Fengqi
6. **Bian**, Xintong; **Jiang**, Yufan; **Wu**, Qiyao
7. **Chen**, Yongxing; **Yao**, Yanzhi; **Zhang**, Canwei
8. **Nukala**, Kishore; **Pulleti**, Sai; **Vaidyula**, Srikar
9. **Baluja**, Michael; **Cao**, Fangning; **Huff**, Mikael; **Shen**, Xuyang
10. **Arun**, Aditya; **Long**, Heyang; **Peng**, Haonan
11. **Cowin**, Samuel; **Hanna**, Aaron; **Liao**, Albert; **Mandadi**, Sumega
12. **Jia**, Yichen; **Jiang**, Zhiyun; **Li**, Zhuofan
13. **Dandu**, Murali; **Daru**, Srinivas; **Pamidi**, Sri
14. **Huang**, Yen-Ting; **Wang**, Shi; **Wang**, Tzu-Kao
15. **Chen**, Luobin; **Feng**, Ruining; **Wu**, Ximei; **Xu**, Haoran
16. **Chen**, Rex; **Liang**, Youwei; **Zheng**, Xinran
17. **Aguilar**, Matthew; **Millhiser**, Jacob; **O'Boyle**, John; **Sharpless**, Will
18. **Wang**, Haoyu; **Wang**, Jiawei; **Zhang**, Yuwei
19. **Chen**, Yinbo; **Di**, Zonglin; **Mu**, Jiteng
20. **Chowdhury**, Debalina; **He**, Scott; **Ye**, Yiheng
21. **Lin**, Wei-Ru; **Ru**, Liyang; **Zhang**, Shaohua

If you haven't sent me the composition of your group, please send me an email: mvasconcelos@eng.ucsd.edu with the group members.

**If I don't hear from you by Monday, 1/24 @ 11:59pm, I will assume that you are not doing the class for credit and will not be doing a project.** Keep in mind that the Project Proposal is due 2/1.

# Project Evaluation

(posted on Canvas)



### Evaluation of ECE 271B Project

This document provides some guidance on how your projects will be evaluated. Beyond evaluation of your mastery of the class material, the goal is also to introduce you to the standard practices in research.

#### Introduction

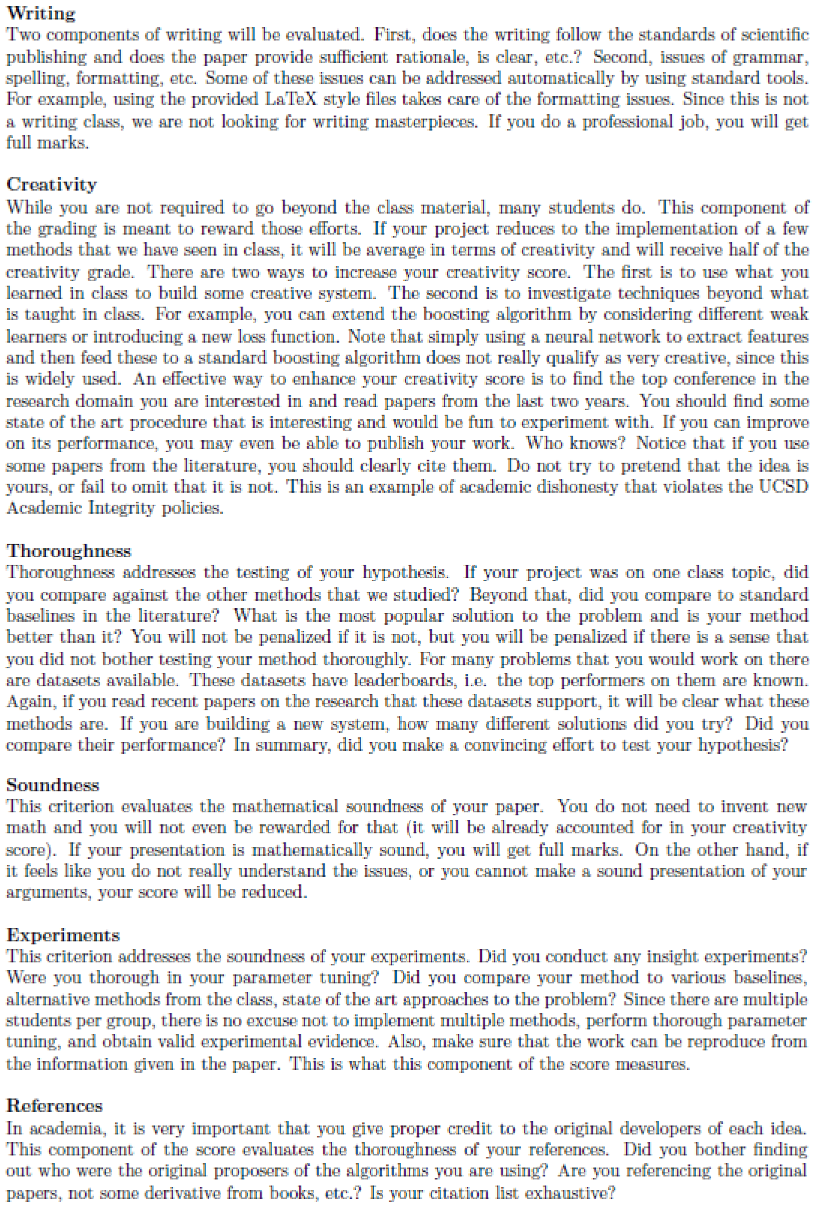
The first thing to know is that a paper needs a hypothesis. The reason to do the work is to test the hypothesis. For example, the hypothesis can be that boosting is more suitable to classify EEG signals than the SVM. A better hypothesis is that boosting is a state of the art method for classifying EEG signals. The more sophisticated your hypothesis, the more people will care about your paper. For publication in a good conference or journal, the hypothesis must contain a fair amount of novelty. The hypothesis is usually that the method you are proposing is state of the art. We will not require this in the class, but will obviously reward novelty. If you are doing something more original, your grade will be higher. Once you have a hypothesis, you usually need a fair amount of technical work to study it. For example, you may need to derive a boosting algorithm based on a loss that is preferred for EEG signals. This is where you get to use some mathematics. Again, the more sophisticated the mathematics, the more likely people will be interested in the paper. But math is not everything. A paper with an exciting but simple hypothesis is always better than a paper full of math to support a boring hypothesis. Finally, you must test your hypothesis. Here, good papers have three types of experiments.

- Insight: a set of simple experiments that provides insight on your method. Does it do the right thing? This can be tested with synthetic data that conforms to the assumptions used to derive the method. It is the only set of experiments where you can use synthetic data. Everything else should use real data. This set of experiments is optional, if the following ones already provide all the insight on how the method works.
- Parameter tuning: in this section, you study the importance of any parameters in your method. For example, what is the best number of weak learners for the data you are using? You compare performance with multiple parameter values and report. Then, you choose the best parameter values. All methods have free parameters and you should always let the reader understand how they affect the performance of your method.
- Validating the hypothesis: once you have the best parameters for your method, you need to test the hypothesis. Is it really the best solution to the problem you are studying? To prove this, you must compare to baselines. These are solutions that already existed for the problem. For a real paper, you need to compare to the best methods in the literature and show that your method is better. For the class, we will be less ambitious. You do not need to show that you have the best method. However, you need to find baselines (i.e. other solutions) and compare your method to them.

#### Evaluation criteria

The first thing to know is that your paper will be evaluated "on the curve," that is against the other papers in the class. We will evaluate papers along 6 dimensions:

- writing (10 pts),
- creativity (20 pts),
- thoroughness (20 pts),
- soundness (10 pts),
- experiments (30 pts),
- references (10 pts).



**Writing**
Two components of writing will be evaluated. First, does the writing follow the standards of scientific publishing and does the paper provide sufficient rationale, is clear, etc.? Second, issues of grammar, spelling, formatting, etc. Some of these issues can be addressed automatically by using standard tools. For example, using the provided LaTeX style files takes care of the formatting issues. Since this is not a writing class, we are not looking for writing masterpieces. If you do a professional job, you will get full marks.

**Creativity**
While you are not required to go beyond the class material, many students do. This component of the grading is meant to reward those efforts. If your project reduces to the implementation of a few methods that we have seen in class, it will be average in terms of creativity and will receive half of the creativity grade. There are two ways to increase your creativity score. The first is to use what you learned in class to build some creative system. The second is to investigate techniques beyond what is taught in class. For example, you can extend the boosting algorithm by considering different weak learners or introducing a new loss function. Note that simply using a neural network to extract features and then feed these to a standard boosting algorithm does not really qualify as very creative, since this is widely used. An effective way to enhance your creativity score is to find the top conference in the research domain you are interested in and read papers from the last two years. You should find some state of the art procedure that is interesting and would be fun to experiment with. If you can improve on its performance, you may even be able to publish your work. Who knows? Notice that if you use some papers from the literature, you should clearly cite them. Do not try to pretend that the idea is yours, or fail to omit that it is not. This is an example of academic dishonesty that violates the UCSD Academic Integrity policies.

**Thoroughness**
Thoroughness addresses the testing of your hypothesis. If your project was on one class topic, did you compare against the other methods that we studied? Beyond that, did you compare to standard baselines in the literature? What is the most popular solution to the problem and is your method better than it? You will not be penalized if it is not, but you will be penalized if there is a sense that you did not bother testing your method thoroughly. For many problems that you would work on there are datasets available. These datasets have leaderboards, i.e. the top performers on them are known. Again, if you read recent papers on the research that these datasets support, it will be clear what these methods are. If you are building a new system, how many different solutions did you try? Did you compare their performance? In summary, did you make a convincing effort to test your hypothesis?

**Soundness**
This criterion evaluates the mathematical soundness of your paper. You do not need to invent new math and you will not even be rewarded for that (it will be already accounted for in your creativity score). If your presentation is mathematically sound, you will get full marks. On the other hand, if it feels like you do not really understand the issues, or you cannot make a sound presentation of your arguments, your score will be reduced.

**Experiments**
This criterion addresses the soundness of your experiments. Did you conduct any insight experiments? Were you thorough in your parameter tuning? Did you compare your method to various baselines, alternative methods from the class, state of the art approaches to the problem? Since there are multiple students per group, there is no excuse not to implement multiple methods, perform thorough parameter tuning, and obtain valid experimental evidence. Also, make sure that the work can be reproduce from the information given in the paper. This is what this component of the score measures.

**References**
In academia, it is very important that you give proper credit to the original developers of each idea. This component of the score evaluates the thoroughness of your references. Did you bother finding out who were the original proposers of the algorithms you are using? Are you referencing the original papers, not some derivative from books, etc.? Is your citation list exhaustive?

# ECE 271B – Winter 2022

# The Rayleigh Quotient

Disclaimer:
This class will be recorded
and made available to students asynchronously.

Manuela Vasconcelos

ECE Department, UCSD

# Plan for today

- PCA and LDA (quick review)
- **LDA**: a special case of the **Rayleigh quotient**
- invertibility and regularization: **RDA**
- **Rayleigh quotient** as **unified formulation** for PCA, LDA, and RDA
- **dual form** of the Rayleigh quotient

# The Role of the Mean

- the mean of the entire data is a function of the coordinate system
  - if $\mathbf{X}$ has mean $\boldsymbol{\mu}$, then $\mathbf{X} - \boldsymbol{\mu}$ has mean $\mathbf{0}$
- we can always make the data have **zero−mean** by **centering**
  - if

$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix} \quad \text{and} \quad \boxed{\mathbf{X}_c^T = \left(\mathbf{I} - \frac{1}{n}\,\mathbf{1}\mathbf{1}^T\right)\mathbf{X}^T}$$

then $\mathbf{X}_c$ has **zero mean**

- we can assume that $\mathbf{X}$ is zero mean **without** loss of generality

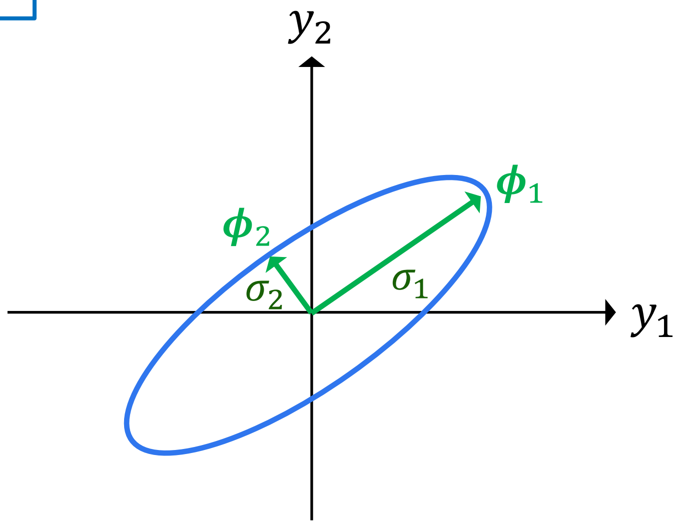
# Principal Component Analysis (Learning)

- If $\mathbf{y}$ is Gaussian with covariance $\mathbf{\Sigma}$, the equiprobability contours

$$\boxed{\mathbf{y}^T \mathbf{\Sigma}^{-1} \mathbf{y} = K}$$

are the ellipses whose

- principal components $\boldsymbol{\phi}_i$ are the eigenvectors of $\mathbf{\Sigma}$
- principal lengths $\sigma_i$ are the eigenvalues of $\mathbf{\Sigma}$



- by detecting **small eigenvalues**, we can eliminate dimensions that have little variance

- this is PCA

# PCA by SVD

- computation of PCA by SVD

$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}$$

given $\mathbf{X}$ with one example per column

1) create the centered data−matrix

$$\boxed{\mathbf{X}_c^T = \left(\mathbf{I} - \frac{1}{n}\,\mathbf{1}\mathbf{1}^T\right)\mathbf{X}^T}$$

2) compute its SVD

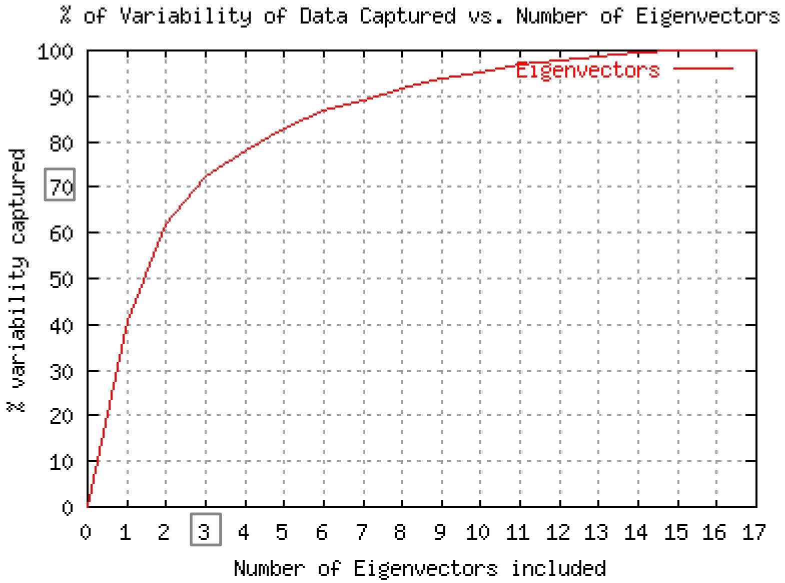
$$\boxed{\mathbf{X}_c^T = \mathbf{M}\,\mathbf{\Pi}\,\mathbf{N}^T}$$

3) principal components are columns of $\mathbf{N}$, eigenvalues are

$$\boxed{\sigma_i = \frac{1}{n}\pi_i^2}$$

# Principal Component Analysis

- a natural measure is to pick the eigenvectors that **explain $p\%$ of the data variability** → can be done by plotting the ratio $r_k$ as a function of $k$



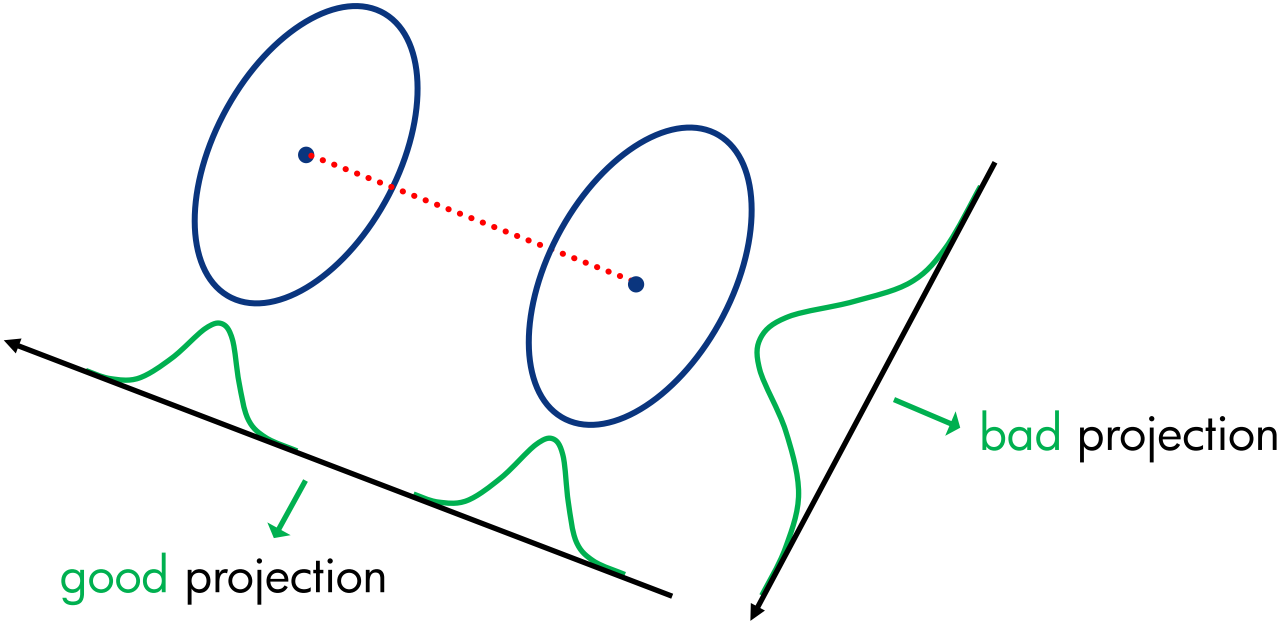
$$r_k = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2}$$

e.g. we need 3 eigenvectors to cover 70% of the variability of this dataset

- However, PCA is <u>not</u> optimal for classification
  - the **discriminant dimensions** could be **<u>thrown out</u>**
  - it is not hard to construct examples where PCA is the <u>worst</u> possible thing we could do (see last lecture example)

# Fischer’s Linear Discriminant

- Find the line $z = \mathbf{w}^T \mathbf{x}$ (i.e. the direction $\mathbf{w}$) that **best separate** the two classes



$$\mathbf{w}^* = \max_{\mathbf{w}} \frac{\left(E_{Z|Y}[z|y=1] - E_{Z|Y}[z|y=0]\right)^2}{\mathrm{var}[z|y=1] + \mathrm{var}[z|y=0]}$$

measures separation between class means

measures variability inside the classes

# Linear Discriminant Analysis

- this can be written as

$$\boxed{J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}}$$

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T$$

$$\mathbf{S}_W = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$$

**between** class scatter

measures separation between class means

**within** class scatter

measures variability inside the classes

- solution of

$$\boxed{\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = K}$$

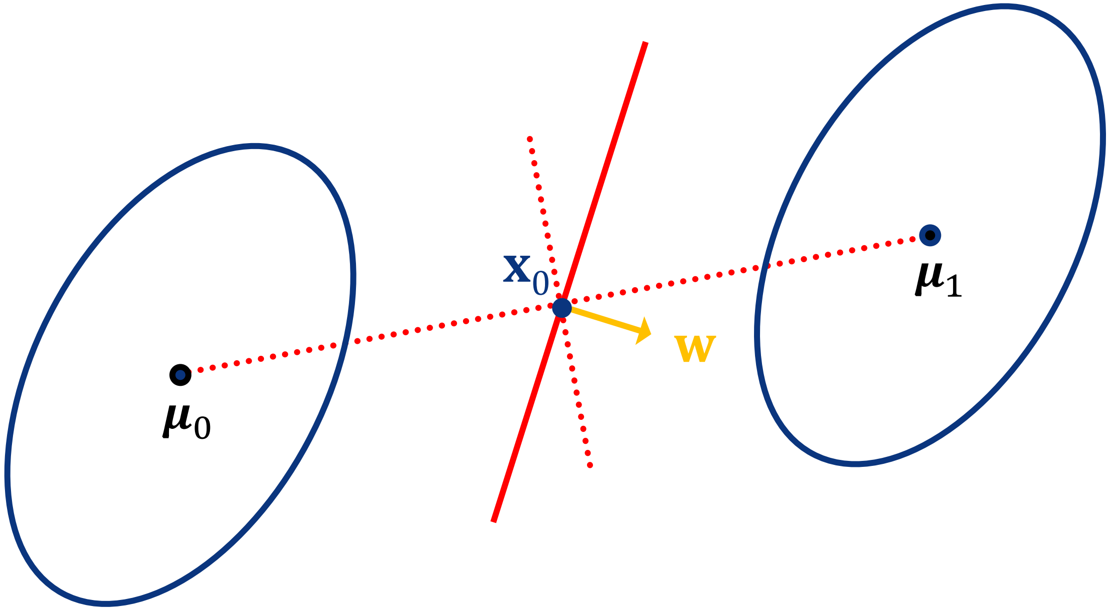
is

$$\boxed{\mathbf{w}^* = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = (\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$$

# Linear Discriminant Analysis

- note that you have seen this before
  - for a classification problem with Gaussian classes of equal covariance $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0$, the BDR boundary is the plane of normal

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$



Gaussian classes, equal covariance $\mathbf{\Sigma}$

  - if $\Sigma_1 = \Sigma_0$, this is also the LDA solution

# Linear Discriminant Analysis

- this gives two different interpretations of LDA

1. classical Fisher interpretation, which is just about separating data (assumes no probability model)

2. Bayes decision rule

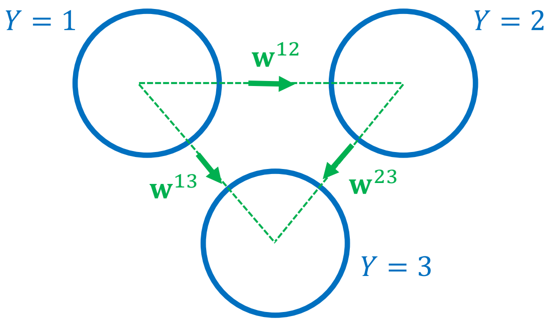
$$\boxed{i^* = \arg\max_i P(y = i|\mathbf{x})}$$

after approximating the data by two Gaussians with equal covariance

- hence, LDA is usually better than PCA
  - it explicitly optimizes discrimination
- but not necessarily good enough
  - if the Gaussian approximation is a poor one

# Linear Discriminant Analysis

- what if there are more than two classes?
  - you simply compute the discriminants $\mathbf{w}^{ij}$ between all pairs of classes $i$ and $j$
  - e.g. for $C = 3$ classes



$$\boxed{\mathbf{w}^{ij} = \left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)}$$

  - note that there are $C(C-1)/2$ different pairs

- this **constrains** the dimension of LDA
  - you **cannot** simply pick what you want
  - the dimension after dimensionality reduction is $\boxed{C(C-1)/2}$

# PCA + LDA

- the main difficulty of LDA is that computation of the **linear discriminant**

  $$\mathbf{w}^* = (\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$$

  requires **matrix inversion**

- the inversion can be very error prone when the matrix $\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1$ is close to singular

- this happens when $\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1$ has eigenvalues close to zero

- to avoid this problem it can be useful to adopt a **two–step** solution

  1. use PCA to eliminate the dimensions of small eigenvalues
  2. project into the remaining dimensions and apply LDA

- this is known as "PCA + LDA"

# The Rayleigh Quotient

- it turns out that the maximization of the Rayleigh quotient

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$\mathbf{S}_B, \mathbf{S}_W$ symmetric positive−semidefinite

appears in <u>many</u> problems in engineering and pattern recognition

- we have already seen that this is equivalent to

$$\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = K$$

and can be solved using Lagrange optimization

$$\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = K$$

# The Rayleigh Quotient

- define the Lagrangian

$$\boxed{L = \mathbf{w}^T \mathbf{S}_B \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{S}_W \mathbf{w} - K)}$$

- maximize with respect to $\mathbf{w}$

$$\nabla_{\mathbf{w}} L = 2(\mathbf{S}_B - \lambda \mathbf{S}_W)\mathbf{w} = 0$$

to obtain the solution

$$\boxed{\mathbf{S}_B \mathbf{w} = \lambda\, \mathbf{S}_W \mathbf{w}}$$

- this is a generalized eigenvalue problem that you can solve using any eigenvalue routine

- which eigenvalue?

# The Rayleigh Quotient

- recall that we want

$$\boxed{\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = K}$$

  and the optimal $\mathbf{w}^*$ satisfies

$$\boxed{\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}}$$

- hence,

$$\boxed{(\mathbf{w}^*)^T \mathbf{S}_B \mathbf{w}^* = \lambda (\mathbf{w}^*)^T \mathbf{S}_W \mathbf{w}^* = \lambda K}$$

  which is maximum for the **largest eigenvalue**

- in summary, we need the generalized eigenvector $\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$ of largest eigenvalue

# The Rayleigh Quotient

$\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w}$ subject to $\mathbf{w}^T \mathbf{S}_W \mathbf{w} = K$

optimal $\mathbf{w}^*$ satisfies $\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$

- case 1: $S_W$ invertible
  - simplifies to a **standard eigenvalue problem**

$$\boxed{\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}}$$

  - $\mathbf{w}$ is the eigenvector of largest eigenvalue of $\mathbf{S}_W^{-1} \mathbf{S}_B$

- case 2: $S_W$ not−invertible
  - this case is more **problematic** since the cost can be unbounded
  - to see this, recall that a non−invertible $\mathbf{S}_W$ has a null−space $N(\mathbf{S}_W)$
  - consider $\mathbf{w} = \mathbf{w}_r + \mathbf{w}_n$, $\mathbf{w}_r$ in the row−space of $\mathbf{S}_W$ and $\mathbf{w}_n \in N(\mathbf{S}_W)$

$$\mathbf{w}^T \mathbf{S}_W \mathbf{w} = (\mathbf{w}_r + \mathbf{w}_n)^T \mathbf{S}_W (\mathbf{w}_r + \mathbf{w}_n) = (\mathbf{w}_r + \mathbf{w}_n)^T \mathbf{S}_W \mathbf{w}_r = \mathbf{w}_r^T \mathbf{S}_W \mathbf{w}_r$$

and the constraint holds for any $\mathbf{w}_n$

Recall:
$\mathbf{w}_n \in N(\mathbf{S}_W) \Leftrightarrow \mathbf{S}_W \mathbf{w}_n = 0$

# The Rayleigh Quotient

$$\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = K$$

optimal $\mathbf{w}^*$ satisfies $\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$

- on the other hand,

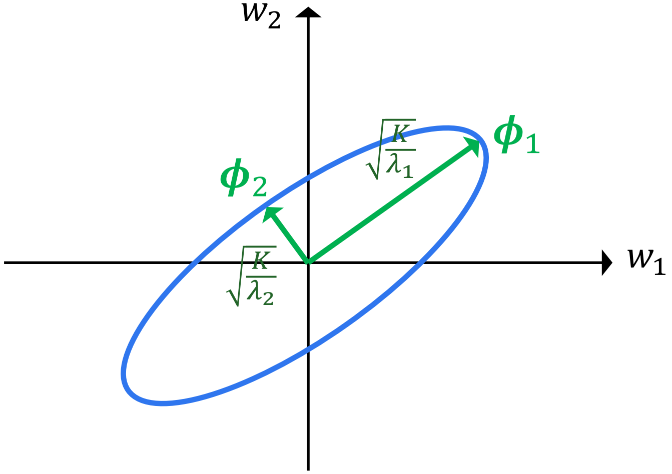
$$\begin{aligned} \mathbf{w}^T \mathbf{S}_B \mathbf{w} &= (\mathbf{w}_r + \mathbf{w}_n)^T \mathbf{S}_B (\mathbf{w}_r + \mathbf{w}_n) \\ &= \underbrace{\mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_r}_{\geq 0} + 2\, \mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_n + \underbrace{\mathbf{w}_n^T \mathbf{S}_B \mathbf{w}_n}_{\geq 0} \end{aligned}$$

because $\mathbf{S}_B$ is positive–definite

- hence, if there is a pair $(\mathbf{w}_r, \mathbf{w}_n)$ such that $\mathbf{w}_r^T \mathbf{S}_B \mathbf{w}_n > 0$
  - we can make the cost arbitrarily large
  - by simply scaling up the null–space component $\mathbf{w}_n$
- this can also be seen geometrically

$$\max_{\mathbf{w}} \ \mathbf{w}^T\mathbf{S}_B\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T\mathbf{S}_W\mathbf{w} = K$$

# The Rayleigh Quotient



- recall that

$$\boxed{\mathbf{w}^T\mathbf{S}_W\mathbf{w} = K \quad \text{with} \quad \mathbf{S}_W = \mathbf{\Phi\Lambda\Phi}^T}$$

  are the ellipses whose

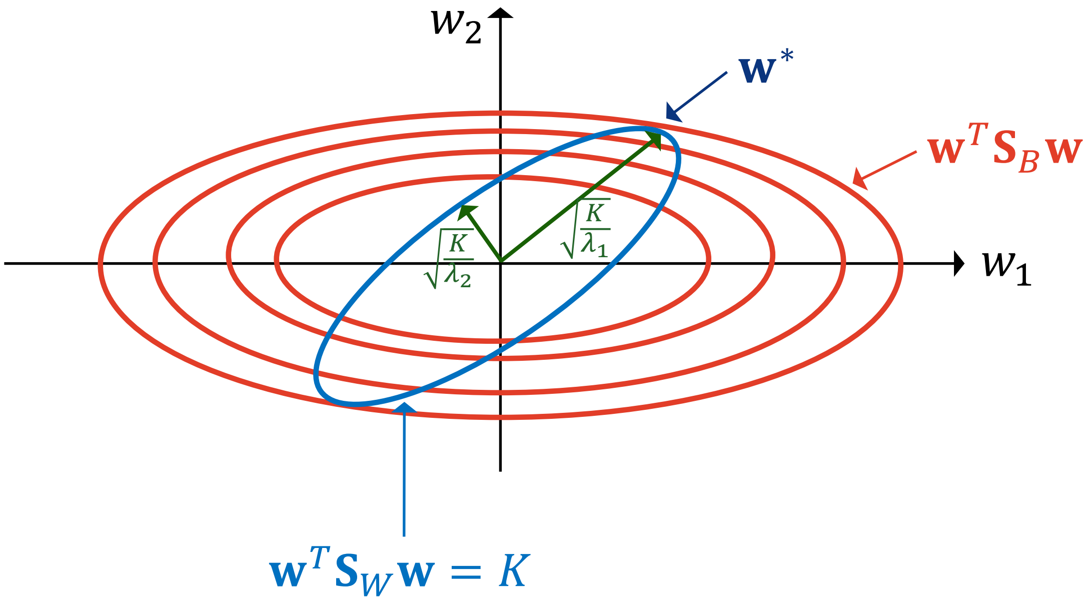
  - principal components $\boldsymbol{\phi}_i$ are the eigenvectors of $\mathbf{S}_W$
  - principal lengths are $\sqrt{K/\lambda_i}$, where $\lambda_i$ are the eigenvalues of $\mathbf{S}_W$

- when the eigenvalues go to **zero**, the ellipses **blow up**
- consider the picture of the optimization problem

$$\boxed{\max_{\mathbf{w}} \ \mathbf{w}^T\mathbf{S}_B\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T\mathbf{S}_W\mathbf{w} = K}$$

# The Rayleigh Quotient

$$\max_{\mathbf{w}} \boxed{\mathbf{w}^T\mathbf{S}_B\mathbf{w}} \quad \text{subject to} \quad \boxed{\mathbf{w}^T\mathbf{S}_W\mathbf{w} = K}$$


$w_2$
$w_1$
$\mathbf{w}^*$
$\mathbf{w}^T\mathbf{S}_B\mathbf{w}$
$\sqrt{\frac{K}{\lambda_1}}$
$\sqrt{\frac{K}{\lambda_2}}$
$\mathbf{w}^T\mathbf{S}_W\mathbf{w} = K$

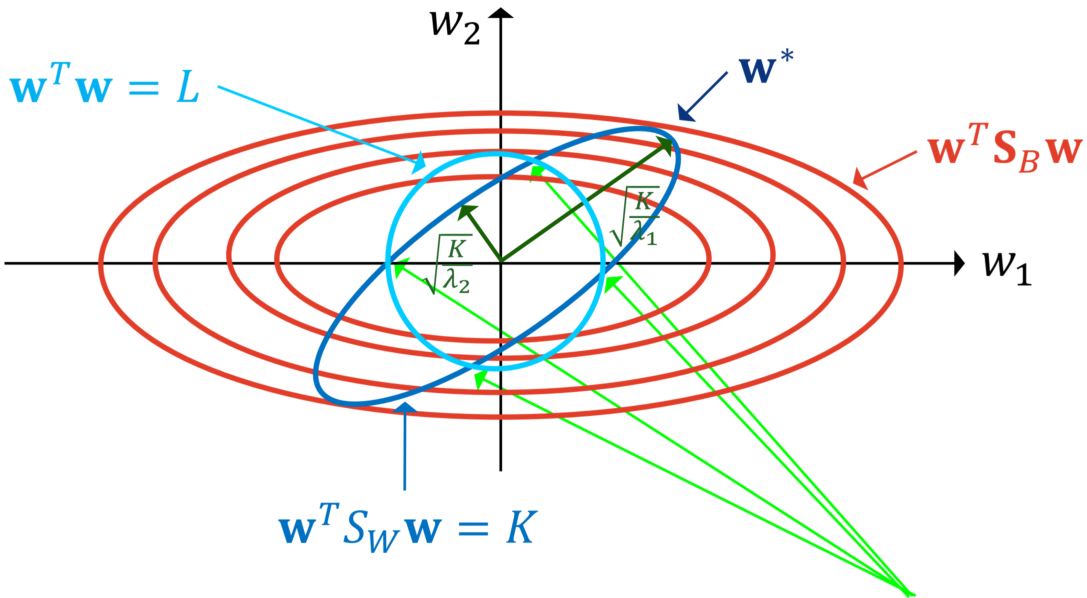
- the optimal solution $\mathbf{w}^*$ is where the outer red ellipse (**cost**) touches the blue ellipse (**constraint**)
  - in this example, as $\lambda_1$ goes to zero, $\|\mathbf{w}^*\|$ and the cost go to infinity

# The Rayleigh Quotient

- how do we avoid this problem?
  - we introduce another **constraint**

normalization constraint

$$\max_{\mathbf{w}} \boxed{\mathbf{w}^T \mathbf{S}_B \mathbf{w}} \quad \text{subject to} \quad \boxed{\mathbf{w}^T \mathbf{S}_W \mathbf{w} = K} \quad \boxed{\|\mathbf{w}\|^2 = L}$$


$w_2$
$w_1$
$\mathbf{w}^T\mathbf{w} = L$
$\mathbf{w}^*$
$\mathbf{w}^T\mathbf{S}_B\mathbf{w}$
$\sqrt{\frac{K}{\lambda_1}}$
$\sqrt{\frac{K}{\lambda_2}}$
$\mathbf{w}^T S_W \mathbf{w} = K$

- **restricts the set of possible solutions** to these points (surfaces in high–dimensional case)

$$\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} = K$$

$$(\mathbf{S}_B - \lambda \mathbf{S}_W)\mathbf{w} = 0$$

# The Rayleigh Quotient

- The Lagrangian is now

$$\boxed{L = \mathbf{w}^T \mathbf{S}_B \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{S}_W \mathbf{w} - K) - \beta(\mathbf{w}^T \mathbf{w} - L)}$$

- and the solution satisfies

$$\nabla_{\mathbf{w}} L = 2(\mathbf{S}_B - \lambda \mathbf{S}_w - \beta \mathbf{I})\mathbf{w} = 0$$

- or

$$(\mathbf{S}_B - \lambda[\mathbf{S}_W + \gamma \mathbf{I}])\mathbf{w} = 0, \qquad \gamma = \beta/\lambda$$

- but this is **exactly** the solution of the original problem with $\mathbf{S}_W + \gamma \mathbf{I}$ instead of $\mathbf{S}_W$

$$\boxed{\max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{S}_B \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T [\mathbf{S}_W + \gamma \mathbf{I}] \mathbf{w} = K}$$

# The Rayleigh Quotient

- adding the constraint is equivalent to maximizing the regularized Rayleigh quotient

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T [\mathbf{S}_W + \gamma \mathbf{I}] \mathbf{w}}$$

$\mathbf{S}_B, \mathbf{S}_W$ symmetric positive−semidefinite

- what does this accomplish?
  - note that

$$\begin{aligned} \mathbf{S}_W = \mathbf{\Phi\Lambda\Phi}^{\boldsymbol{T}} \;\Rightarrow\; \mathbf{S}_W + \gamma\mathbf{I} &= \mathbf{\Phi\Lambda\Phi}^{\boldsymbol{T}} + \gamma\mathbf{\Phi I \Phi}^{\boldsymbol{T}} \\ &= \mathbf{\Phi}[\mathbf{\Lambda} + \gamma\mathbf{I}]\mathbf{\Phi}^{\boldsymbol{T}} \end{aligned}$$

  - this makes all eigenvalues positive
  - the matrix is no longer non−invertible

$\gamma$ "controls" the eigenvalues of $\mathbf{S}_W + \gamma\mathbf{I}$

# The Rayleigh Quotient

- in summary,

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$\mathbf{S}_B, \mathbf{S}_W$ symmetric
positive−semidefinite

- 1) $\mathbf{S}_W$ invertible
  - $\mathbf{w}^*$ is the eigenvector of **largest eigenvalue** of $\mathbf{S}_W^{-1}\mathbf{S}_B$
  - the **max value** is $\lambda K$, where $\lambda$ is the largest eigenvalue

- 2) $\mathbf{S}_W$ not−invertible
  - regularize: $\mathbf{S}_W \to \mathbf{S}_W + \gamma \mathbf{I}$
  - $\mathbf{w}^*$ is the eigenvector of **largest eigenvalue** of $(\mathbf{S}_W + \gamma \mathbf{I})^{-1}\mathbf{S}_B$
  - the **max value** is $\lambda K$, where $\lambda$ is the largest eigenvalue

# Regularized Discriminant Analysis

- back to LDA:
  - when the within scatter matrix is non−invertible, instead of

between class scatter

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T$$
$$\mathbf{S}_W = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$$

within class scatter

we use

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T$$
$$\mathbf{S}_W = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 + \gamma \mathbf{I}$$

regularized within class scatter

  - this is called Regularized Discriminant Analysis (RDA)

# Regularized Discriminant Analysis

- noting that

$$\boxed{\begin{aligned}\mathbf{S}_W &= \mathbf{\Sigma}_0 + \mathbf{\Sigma}_1 + \gamma\mathbf{I}\\ &= \mathbf{\Sigma}_0 + \gamma_0\mathbf{I} + \mathbf{\Sigma}_1 + \gamma_1\mathbf{I}\end{aligned}}$$

$$\boxed{\gamma_0 + \gamma_1 = \gamma}$$

- this can also be seen as regularizing **each** covariance matrix individually
- the regularization parameters $\gamma_i$ are determined by **cross–validation**
  - more on this later
  - basically means that we try several possibilities and keep the best

# Principal Component Analysis

- Back to PCA: given $\mathbf{X}$ with one example per column
  - create the centered–data matrix

$$\boxed{\mathbf{X}_c^T = \left(\mathbf{I} - \frac{1}{n}\,\mathbf{1}\mathbf{1}^T\right)\mathbf{X}^T}$$

$$\boxed{\mathbf{X}_c^T = \begin{bmatrix} \mathbf{x}_1^T - \boldsymbol{\mu}^T \\ \vdots \\ \mathbf{x}_n^T - \boldsymbol{\mu}^T \end{bmatrix}}$$

  - this has **one** example per row
  - note that the projection of all examples $\mathbf{x}_i$ on principal component $\boldsymbol{\phi}$ is

$$\boxed{\mathbf{z} = \mathbf{X}_c^T\,\boldsymbol{\phi}}$$

$$\boxed{\begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu})^T\boldsymbol{\phi} \\ \vdots \\ (\mathbf{x}_n - \boldsymbol{\mu})^T\boldsymbol{\phi} \end{bmatrix}}$$

# Principal Component Analysis

$$\mathbf{z} = \mathbf{X}_c^T\, \boldsymbol{\phi}$$

$$\begin{bmatrix} \mathrm{z}_1 \\ \vdots \\ \mathrm{z}_n \end{bmatrix} = \begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu})^T \boldsymbol{\phi} \\ \vdots \\ (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\phi} \end{bmatrix}$$

- since

$$\mu_{\mathrm{Z}} = \frac{1}{n}\sum_i z_i = \frac{1}{n}\sum_i (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\phi} = \left(\frac{1}{n}\sum_i \mathbf{x}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\phi} = 0$$

- the **sample variance** of the random variable Z is given by the **norm** squared of vector $\mathbf{z}$

$$\boxed{\mathrm{var}(\mathrm{Z}) = \frac{1}{n}\sum_i (\mathrm{z}_i - \mu_{\mathrm{Z}})^2 = \frac{1}{n}\sum_i \mathrm{z}_i^2 = \frac{1}{n}\|\mathbf{z}\|^2}$$

▶ recall that PCA looks for the **largest** variance component

$$\max_{\boldsymbol{\phi}} \|\mathbf{z}\|^2 = \max_{\boldsymbol{\phi}} \|\mathbf{X}_c^T\, \boldsymbol{\phi}\|^2 = \max_{\boldsymbol{\phi}} (\mathbf{X}_c^T\, \boldsymbol{\phi})^T \mathbf{X}_c^T\, \boldsymbol{\phi} = \max_{\boldsymbol{\phi}} \boldsymbol{\phi}^T \mathbf{X}_c \mathbf{X}_c^T\, \boldsymbol{\phi}$$

$$\mathbf{z} = \mathbf{X}_c^T\, \boldsymbol{\phi}$$

# Principal Component Analysis

- recall that the sample covariance is

$$\mathbf{\Sigma} = \frac{1}{n}\sum_i (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T = \frac{1}{n}\sum_i \mathbf{x}_i^c(\mathbf{x}_i^c)^T$$

  where $\mathbf{x}_i^c$ is the $i^{\text{th}}$ column of $\mathbf{X}_c$

- this can be written as

$$\mathbf{\Sigma} = \frac{1}{n}\begin{bmatrix} | & & | \\ \mathbf{x}_1^c & \cdots & \mathbf{x}_n^c \\ | & & | \end{bmatrix}\begin{bmatrix} - & \mathbf{x}_1^c & - \\ & \vdots & \\ - & \mathbf{x}_n^c & - \end{bmatrix} = \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^T$$

# Principal Component Analysis

- hence, the PCA problem is

$$\max_{\boldsymbol{\phi}} \ \boldsymbol{\phi}^T \mathbf{X}_c \mathbf{X}_c^T \ \boldsymbol{\phi} = \max_{\boldsymbol{\phi}} \ \boldsymbol{\phi}^T \boldsymbol{\Sigma} \ \boldsymbol{\phi}$$

$$\max_{\boldsymbol{\phi}} \ ||\mathbf{z}||^2 = \max_{\boldsymbol{\phi}} \ \boldsymbol{\phi}^T \mathbf{X}_c \mathbf{X}_c^T \ \boldsymbol{\phi}$$

$$\boldsymbol{\Sigma} = \frac{1}{n} \mathbf{X}_c \mathbf{X}_c^T$$

- as in LDA, this can be made arbitrarily large by simply scaling $\boldsymbol{\phi}$
- to **normalize**, we constraint $\boldsymbol{\phi}$ to have **unit norm**

$$\max_{\boldsymbol{\phi}} \ \boldsymbol{\phi}^T \boldsymbol{\Sigma} \ \boldsymbol{\phi} \quad \text{subject to} \quad ||\boldsymbol{\phi}||^2 = 1$$

which is equivalent to

$$\max_{\boldsymbol{\phi}} \frac{\boldsymbol{\phi}^T \boldsymbol{\Sigma} \ \boldsymbol{\phi}}{\boldsymbol{\phi}^T \boldsymbol{\phi}}$$

and shows that PCA = maximization of a Rayleigh quotient

$$\max_{\phi} \frac{\phi^T \Sigma \phi}{\phi^T \phi}$$

# Principal Component Analysis

- in this case,

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$\mathbf{S}_B, \mathbf{S}_W$ symmetric
positive−semidefinite

with $\mathbf{S}_B = \boldsymbol{\Sigma}$ $\mathbf{S}_W = \mathbf{I}$

- $S_W$ is clearly invertible
  - no regularization problems
  - $\mathbf{w}^*$ is the eigenvector of largest eigenvalue of $\mathbf{S}_W^{-1}\mathbf{S}_B$ and this is just the largest eigenvalue of the covariance $\boldsymbol{\Sigma}$
  - the max value is $\lambda K$, where $\lambda$ is the largest eigenvalue

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

# The Rayleigh Quotient Dual

- let's **assume**, for a moment, that the solution is of the form

$$\boxed{\mathbf{w} = \mathbf{X}_c \boldsymbol{\alpha}}$$

i.e. a linear combination of the centered–datapoints

$$\mathbf{X}_c = \begin{bmatrix} | & & | \\ \mathbf{x}_1^c & \cdots & \mathbf{x}_n^c \\ | & & | \end{bmatrix}$$

$$\boldsymbol{\alpha} \in \mathbb{R}^n$$

- hence, the problem is equivalent to

$$\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T \mathbf{X}_c^T \mathbf{S}_B \mathbf{X}_c \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{X}_c^T \mathbf{S}_W \mathbf{X}_c \boldsymbol{\alpha}}$$

- this does **not** change its form, the solution is
  - $\boldsymbol{\alpha}^*$ is the eigenvector of **largest eigenvalue** of $\left(\boldsymbol{X}_c^T \mathbf{S}_W \boldsymbol{X}_c\right)^{-1} \boldsymbol{X}_c^T \mathbf{S}_B \boldsymbol{X}_c$
  - the **max value** is $\lambda K$, where $\lambda$ is the largest eigenvalue

# The Rayleigh Quotient Dual

- for PCA
  - $\mathbf{S}_W = \mathbf{I}$ and $\mathbf{S}_B = \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^T$
  - the solution satisfies

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w} \quad \Leftrightarrow \quad \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^T\mathbf{w} = \lambda\mathbf{w} \quad \Leftrightarrow \quad \boxed{\mathbf{w} = \mathbf{X}_c \underbrace{\frac{1}{n\lambda}\mathbf{X}_c^T\mathbf{w}}_{\boldsymbol{\alpha}}}$$

  - and, therefore, we have

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} \qquad \max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T\mathbf{X}_c^T\mathbf{S}_B\mathbf{X}_c\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{X}_c^T\mathbf{S}_W\mathbf{X}_c\boldsymbol{\alpha}}$$

    - $\mathbf{w}^*$ is the eigenvector of $\mathbf{S}_W^{-1}\mathbf{S}_B = \mathbf{X}_c\mathbf{X}_c^T$
    - $\boldsymbol{\alpha}^*$ is the eigenvector of $(\mathbf{X}_c^T\mathbf{S}_W\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{S}_B\mathbf{X}_c = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{X}_c\mathbf{X}_c^T\mathbf{X}_c = \mathbf{X}_c^T\mathbf{X}_c$
  - i.e. we have two alternative manners to compute PCA

# Principal Component Analysis

primal

- assemble matrix

$$\mathbf{\Sigma} = \mathbf{X}_c\mathbf{X}_c^T$$

- compute eigenvectors $\boldsymbol{\phi}_i$
- these are the principal components

dual

- assemble matrix

$$\mathbf{K} = \mathbf{X}_c^T\mathbf{X}_c$$

- compute eigenvectors $\boldsymbol{\alpha}_i$
- the principal components are $\boldsymbol{\phi}_i = \mathbf{X}_c\boldsymbol{\alpha}_i$

- in both cases, we have an **eigenvalue problem**
  - primal on the sum of the outer−products $\mathbf{\Sigma} = \sum_i \mathbf{x}_i^c\left(\mathbf{x}_j^c\right)^T$
  - dual on the matrix of the inner−products $\mathrm{K}_{ij} = (\mathbf{x}_i^c)^T\mathbf{x}_j^c$

# The Rayleigh Quotient

- this is a property that holds for <u>**many**</u> Rayleigh quotient problems
  - the **primal solution** is a **linear combination of datapoints**
  - the **dual solution** only depends on **dot−products of the datapoints**

- whenever **both** of these hold
  - the <u>**problem can be kernelized**</u>
  - this has various interesting properties
  - we will talk about them

- many examples
  - kernel PCA, kernel LDA, manifold learning, etc.