

ECE-271A
Statistical Learning I:
Dimensionality and
dimensionality reduction

Nuno Vasconcelos
ECE Department, UCSD

Motivation

► recall, in Bayesian decision theory we

- world, states Y in $\{1, \dots, M\}$
- observations X
- class conditional densities $P_{X|Y}(x|y)$
- class probabilities $P_Y(i)$,
- Bayes decision rule (BDR)

$$i^* = \arg \max_i P_{X|Y}(x|i) P_Y(i)$$

► we have hinted that the dimension of observation space can play a significant role in the quality of the BDR

Example

► cheetah Gaussian classifier, DCT space

8 best features



Prob. of error: 4%

all 64 features



8%

► more features = higher error!

Plan for today

- ▶ high dimensional spaces are STRANGE!!!
- ▶ introduction to dimensionality reduction
- ▶ principal component analysis (PCA)

High dimensional spaces

- ▶ are strange!

- ▶ first thing to know:

“never trust your intuition in high dimensions!”

- ▶ more often than not you will be wrong!

- ▶ there are many **examples** of this

- ▶ we will do a couple

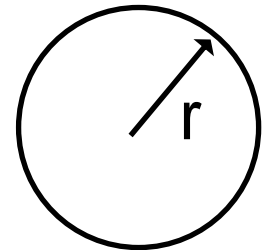
The hyper-sphere

- ▶ Consider the sphere of radius r on a space of dimension d

$$\mathcal{S} = \left\{ \mathbf{x} \mid \sum_{i=1}^d x_i^2 \leq r^2 \right\}$$

- ▶ **Homework:** show that its volume is

$$V_d(r) = \frac{r^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right)}$$

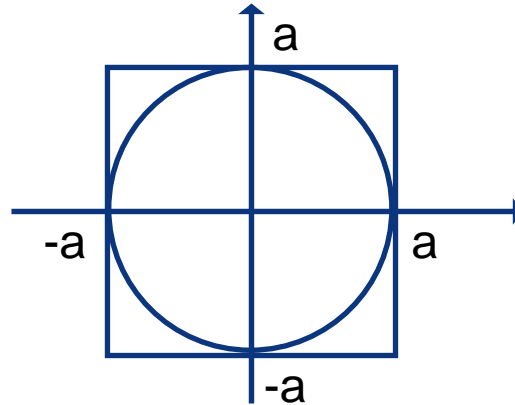


- ▶ where $\Gamma(n)$ is the Gamma function

$$\Gamma(n) = \int_0^{\infty} e^{-x} x^{n-1} dx$$

Hyper-cube vs hyper-sphere

- ▶ next consider the hyper-cube $[-a, a]^d$ and the inscribed hyper-sphere, i.e.



- ▶ Q: what does your intuition tell you about the relative sizes of these two objects?
 1. volume of sphere \approx volume of cube
 2. volume of sphere \gg volume of cube
 3. volume of sphere \ll volume of cube

Answer

- ▶ we can just compute this

$$f_d = \frac{\text{Vol}(sphere)}{\text{Vol}(cube)} = \frac{a^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right) (2a)^d} = \frac{\pi^{\frac{d}{2}}}{2^d \Gamma\left(\frac{d}{2} + 1\right)}$$

- ▶ sequence that does not depend on a , just on the dimension d !

d	1	2	3	4	5	6	7
f_d	1	.785	.524	.308	.164	.08	.037

- ▶ it goes to zero, and goes to zero fast!

Hyper-cube vs hyper-shpere

► this means that:

“as the dimension of the space increases, the volume of the sphere is much smaller (infinitesimal) than that of the cube!”

► how is this going against intuition?

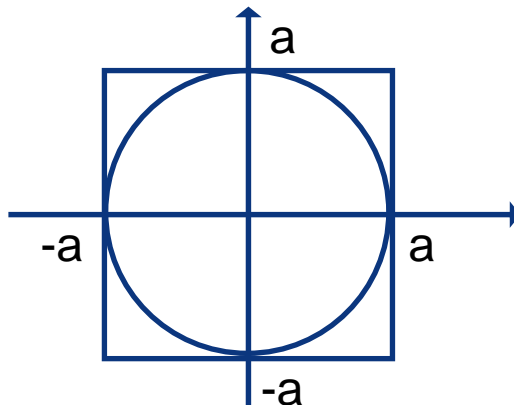
► it is actually not very surprising. we can see it even in low dimensions

1. $d = 1$



volume is the same

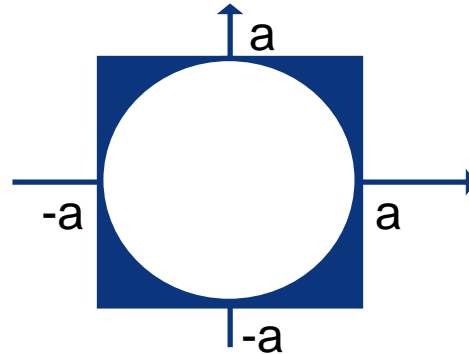
2. $d = 2$



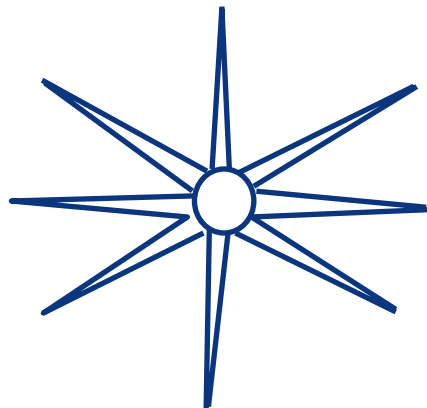
volume of sphere is already smaller

Hyper-cube vs hyper-sphere

- ▶ as the dimension increases the volume of the shaded corners becomes larger



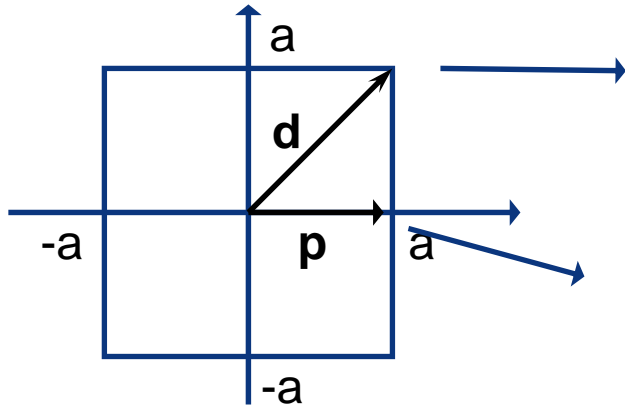
- ▶ in high dimensions the picture you should have in mind is



all the volume of the cube
is in these spikes!

Believe it or not

- ▶ we can check mathematically: consider \mathbf{d} and \mathbf{p}



$$\mathbf{d} = (a, a, \dots, a) \in \mathcal{R}^d$$

$$\mathbf{p} = (a, 0, \dots, 0) \in \mathcal{R}^d$$

- ▶ note that

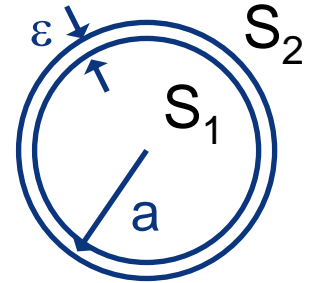
$$\frac{\|\mathbf{d}\|^2}{\|\mathbf{p}\|^2} = \frac{da^2}{a^2} = d \rightarrow \infty$$

$$\begin{aligned} \cos\theta &= \frac{\mathbf{d}^T \mathbf{p}}{\sqrt{\|\mathbf{d}\|^2 \|\mathbf{p}\|^2}} \\ &= \frac{a^2}{\sqrt{da^2 a^2}} = \frac{1}{\sqrt{d}} \rightarrow 0 \end{aligned}$$

- ▶ \mathbf{d} orthogonal to \mathbf{p} as d increases and infinitely larger!!!

But there is more

- ▶ consider the crust of unit sphere of thickness ϵ
- ▶ we can compute its volume



$$Vol(crust) = \left[1 - \frac{Vol(S_1)}{Vol(S_2)} \right] Vol(S_2)$$

$$\frac{Vol(S_1)}{Vol(S_2)} = \frac{\frac{(a-\epsilon)^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}}{\frac{a^d \pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}} = \frac{a^d \left(1 - \frac{\epsilon}{a}\right)^d}{a^d} = \left(1 - \frac{\epsilon}{a}\right)^d$$

- ▶ no matter how small ϵ is, ratio goes to zero as d increases
- ▶ i.e. “all the volume is in the crust!”

High dimensional Gaussian

- ▶ Homework: show that if

$$\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}), \quad \mathbf{x} \in \mathcal{R}^n$$

and one considers the hyper-sphere where the probability density drops to 1% of peak value

$$S_{0.01}(\mathbf{x}) = \left\{ \mathbf{x} \left| \frac{G(\mathbf{x}, \mathbf{0}, \mathbf{I})}{G(\mathbf{0}, \mathbf{0}, \mathbf{I})} \leq 0.01 \right. \right\}$$

- ▶ the probability mass outside this sphere is

$$P_n = P[\chi^2(n) \geq 9.21]$$

- ▶ where $\chi^2(n)$ is a chi-squared random variable with n degrees of freedom

High-dimensional Gaussian

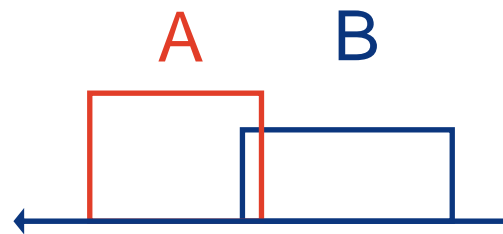
- ▶ if you evaluate this, you'll find out that

n	1	2	3	4	5	6	10	15	20
$1-P_n$.998	.99	.97	.94	.89	.83	.48	.134	.02

- ▶ as the dimension increases, all probability mass is on the tails
- ▶ the point of maximum density is still the mean
- ▶ really strange: in high-dimensions the Gaussian is a very heavy-tailed distribution
- ▶ take-home message:
 - “in high dimensions never trust your intuition!”
- ▶ Q: how does all this affect decision rules?

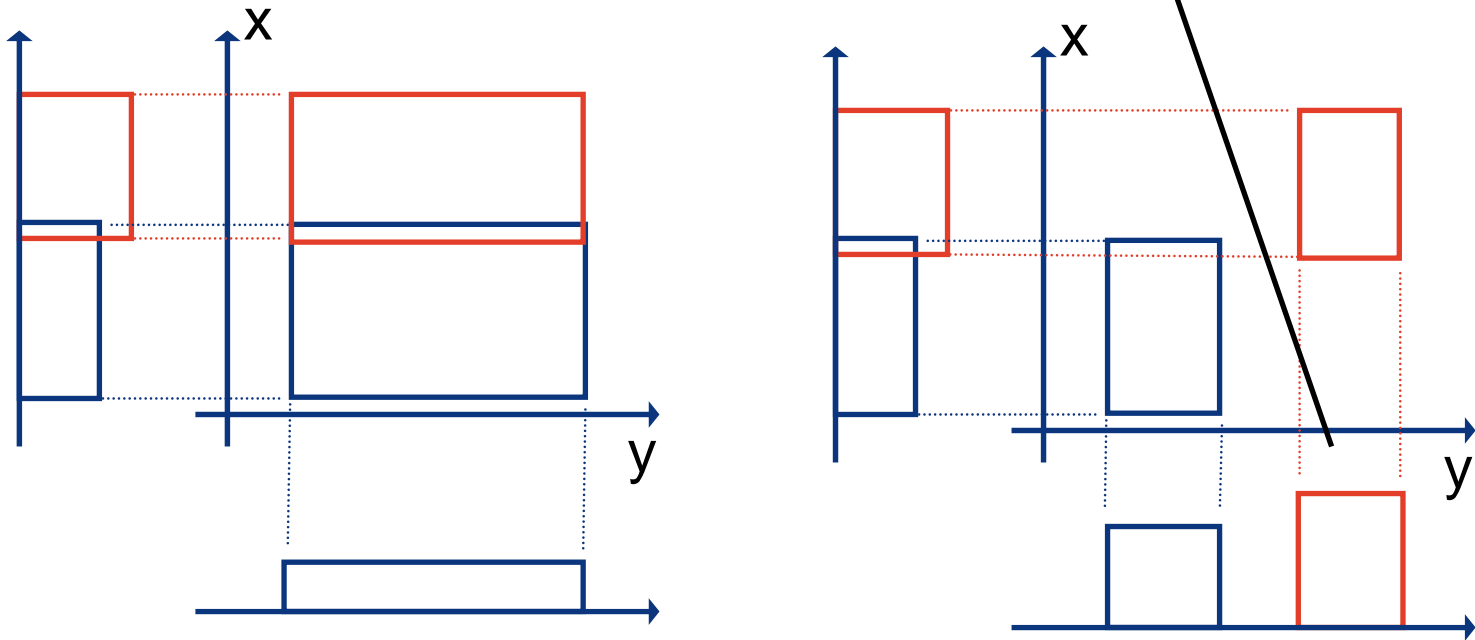
The curse of dimensionality

- ▶ typical observation in Bayes decision theory:
 - error increases when number of features is large
- ▶ highly unintuitive since, theoretically:
 - if I have a problem in n -D I can always generate a problem in $(n+1)$ -D with smaller probability of error
- ▶ e.g. two uniform classes in 1D



- ▶ can be transformed into a 2D problem with same error
- ▶ just add a non-informative variable y .

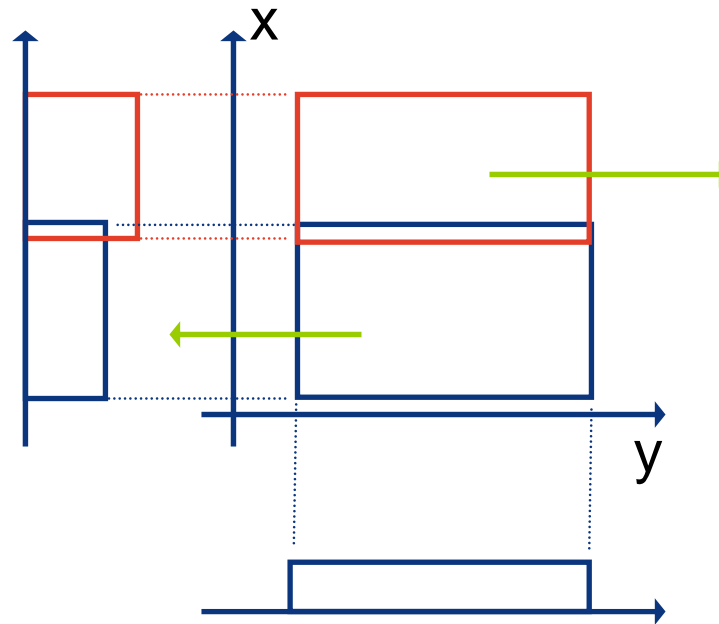
Curse of dimensionality



- ▶ but it is also possible to reduce the error by adding a second variable which is informative
- ▶ on the left there is no decision boundary that will achieve zero error
- ▶ on the right, the decision boundary shown has zero error

Curse of dimensionality

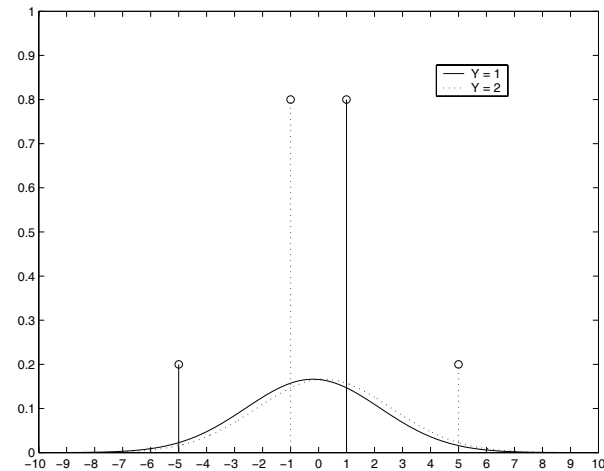
- ▶ in fact, it is impossible to do worse in 2D than 1D



- ▶ if we move the classes along the lines shown in green the error can only go down, since there will be less overlap

Curse of dimensionality

- ▶ so why do we observe this curse of dimensionality?
- ▶ the problem is the quality of the density estimates
- ▶ all we have seen so far, assumes perfect estimation of the BDR
- ▶ as we have seen in PS2, when the estimates are different from the true densities the BDR can be quite poor
- ▶ DHS give a perfect example of this where, by tweaking the location of the delta functions (and assuming a Gaussian model) you can make the error go to 100%!



Curse of dimensionality

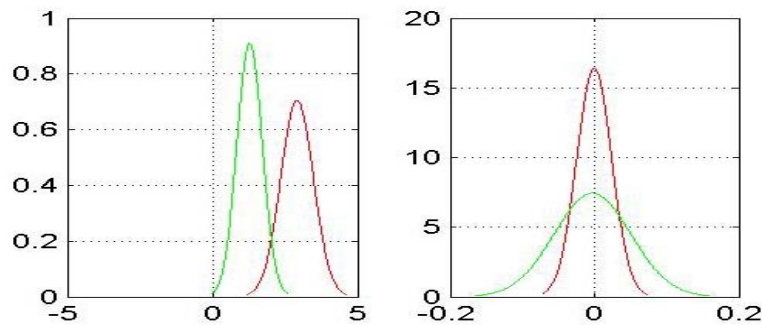
- ▶ we have seen that the variance of an estimator tends to be inversely proportional to the number of points n
 - e.g. ML estimate of the mean of a Gaussian has variance σ^2/n
- ▶ hence, we need a large n to have good estimates
- ▶ Q: what does “large” mean? This depends on the dimension of the space
- ▶ the best way to see this is to think of an histogram
 - suppose you have 100 points and you need at least 10 bins per axis in order to get a reasonable quantization
- ▶ for uniform data you get, on average,
- ▶ decent in 1D, bad in 2D, terrible in 3D (9 out of each 10 bins empty)

dimension	1	2	3
points/bin	10	1	0.1

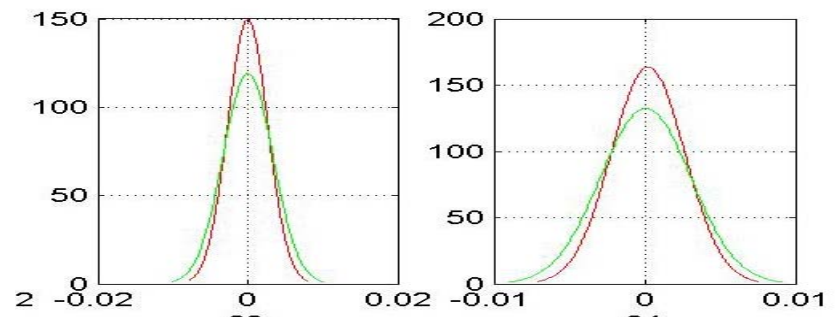
Dimensionality reduction

- ▶ what do we do about this? we avoid unnecessary dimensions
- ▶ unnecessary can be measured in two ways:
 1. features are not discriminant
 2. features are not independent
- ▶ non-discriminant means that they do not separate the classes well

discriminant



non-discriminant

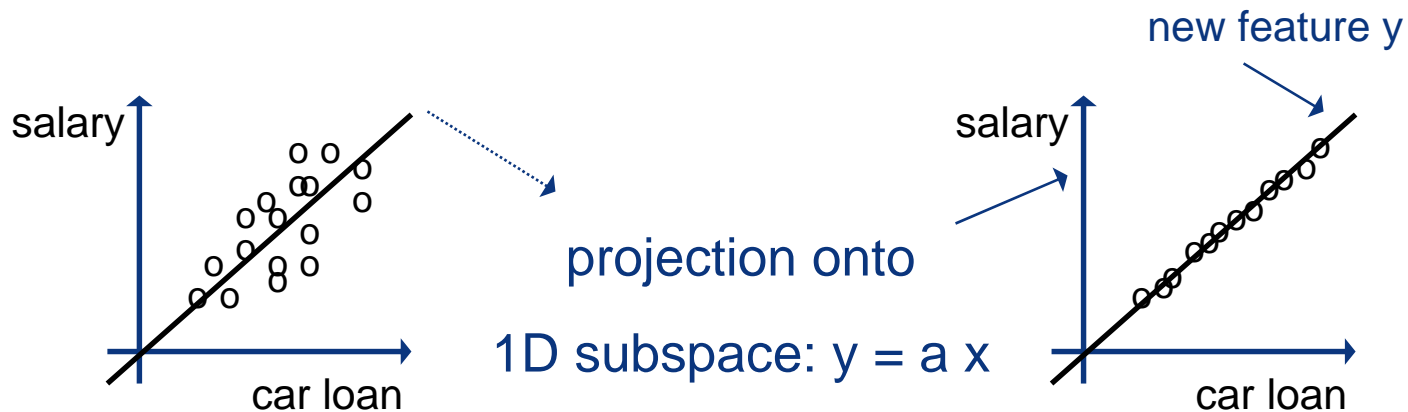


Dimensionality reduction

- ▶ dependent features, even if very discriminant, are not needed - one is enough!
- ▶ e.g. data-mining company studying consumer credit card ratings
- ▶ $X = \{\text{salary, mortgage, car loan, \# of kids, profession, ...}\}$
- ▶ the first three features tend to be highly correlated:
 - “the more you make, the higher the mortgage, the more expensive the car you drive”
 - from one of these variables I can predict the others very well
- ▶ including features 2 and 3 does not increase the discrimination, but increases the dimension and leads to poor density estimates

Dimensionality reduction

- ▶ Q: how do we detect the presence of these correlations?
- ▶ A: the data “lives” in a low dimensional subspace (up to some amounts of noise). E.g.



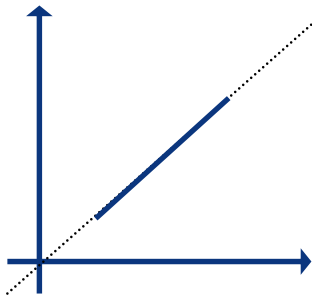
- ▶ in the example above we have a 3D hyper-plane in 5D
- ▶ if we can find this hyper-plane we can
 - project the data onto it
 - get rid of half of the dimensions without introducing significant error

Principal component analysis

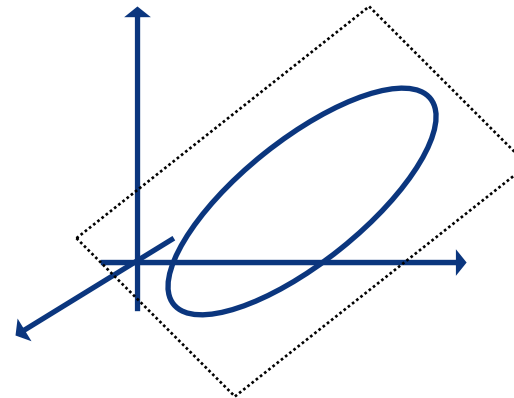
► basic idea:

- if the data lives in a subspace, it is going to look very flat when viewed from the full space, e.g.

1D subspace in 2D



2D subspace in 3D



- this means that if we fit a Gaussian to the data the equiprobability contours are going to be highly skewed ellipsoids

Gaussian review

- ▶ the equiprobability contours of a Gaussian are the points such that

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = K$$

- ▶ let's consider the **change of variable** $z = x - \mu$, which only moves the origin by μ . The equation

$$z^T \Sigma^{-1} z = K$$

is the equation of an **ellipse**.

- ▶ this is **easy to see when Σ is diagonal**:

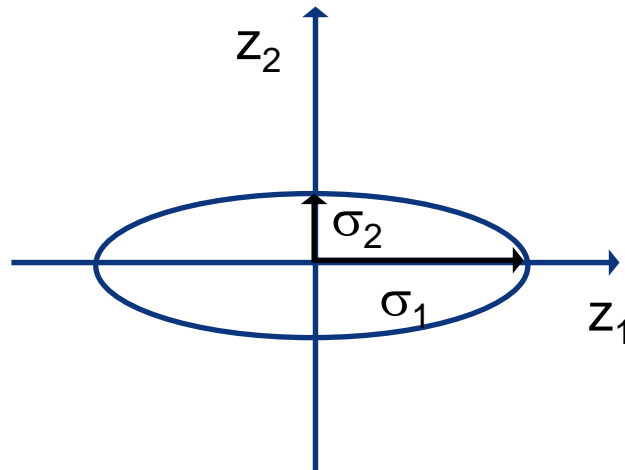
$$\Sigma = \Lambda = \text{diag}(\sigma_1^2, \dots, \sigma_d^2) \Rightarrow z^T \Sigma^{-1} z = \sum_i \frac{z_i^2}{\sigma_i^2} = K$$

Gaussian review

- ▶ this is the equation of an ellipse with principal lengths σ_i
 - e.g. when $d = 2$

$$\frac{z_1^2}{\sigma_1^2} + \frac{z_2^2}{\sigma_2^2} = 1$$

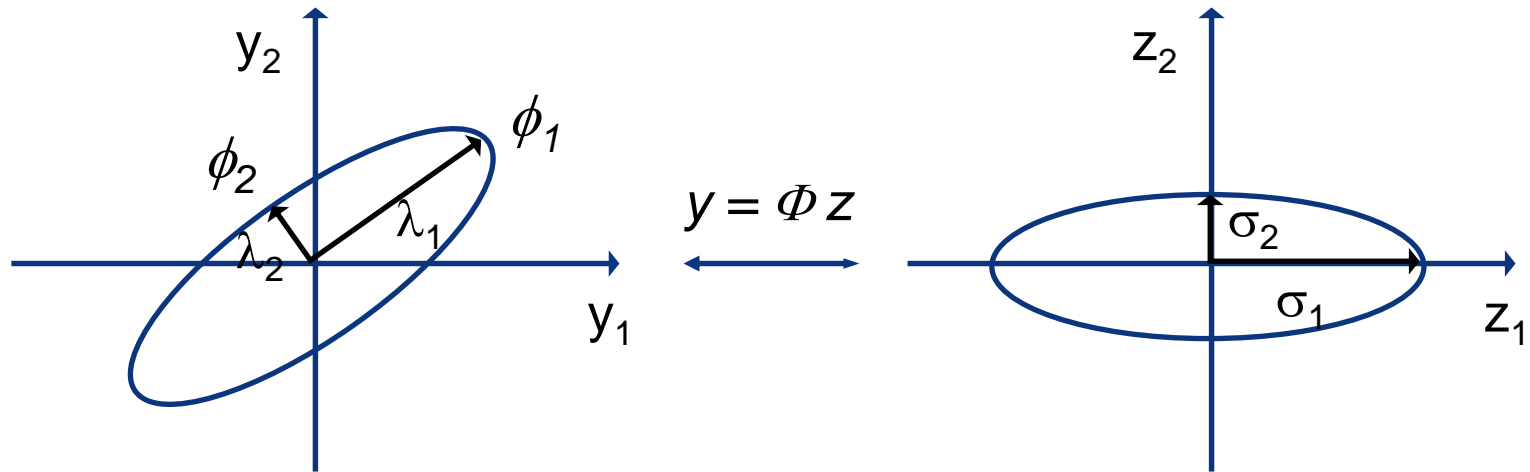
- ▶ is the ellipse



- ▶ introduce the transformation $y = \Phi z$

Gaussian review

- ▶ introduce the transformation $y = \Phi z$
- ▶ then y has covariance $\Sigma_y = \Phi \Sigma_z \Phi^T = \Phi \Lambda \Phi^T$
- ▶ if Φ is orthonormal this is just a rotation and we have

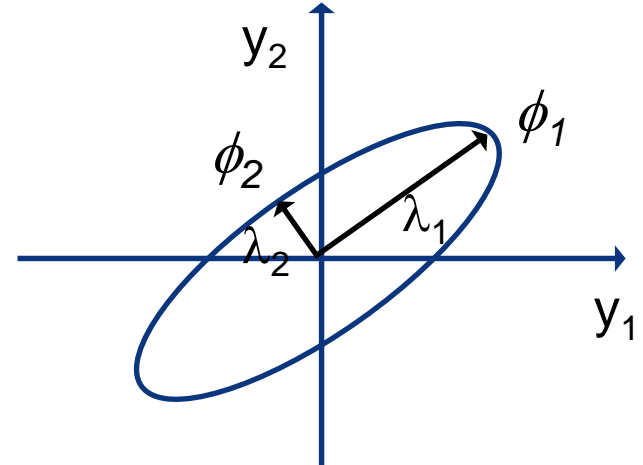


- ▶ we obtain a rotated ellipse with principal components ϕ_1 and ϕ_2 which are the columns of Φ
- ▶ note that $\Sigma_y = \Phi \Lambda \Phi^T$ is the eigen-decomposition of Σ_y

Principal component analysis

► If y is Gaussian with covariance Σ , the equiprobability contours are the ellipses whose

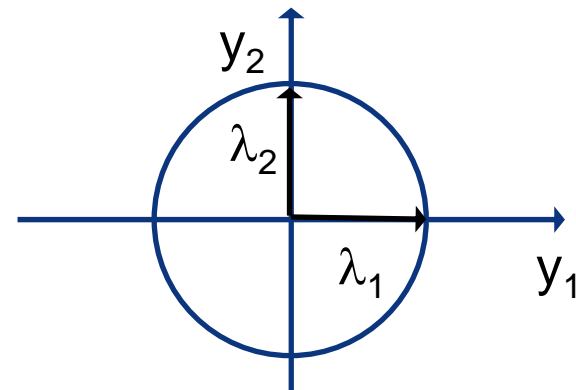
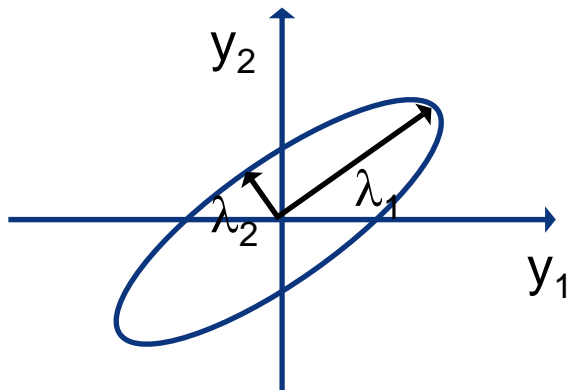
- principal components ϕ_i are the eigenvectors of Σ
- principal lengths λ_i are the eigenvalues of Σ



► by computing the eigenvalues we know if the data is flat

$\lambda_1 \gg \lambda_2$: flat

$\lambda_1 = \lambda_2$: not flat



Principal component analysis (learning)

► Given sample $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $x_i \in \mathcal{R}^d$

- compute sample mean: $\hat{\mu} = \frac{1}{n} \sum_i (\mathbf{x}_i)$

- compute sample covariance: $\hat{\Sigma} = \frac{1}{n} \sum_i (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T$

- compute eigenvalues and eigenvectors of $\hat{\Sigma}$

$$\hat{\Sigma} = \Phi \Lambda \Phi^T, \quad \Lambda = \text{diag}(\sigma_1^2, \dots, \sigma_n^2) \quad \Phi^T \Phi = I$$

- order eigenvalues $\sigma_1^2 > \dots > \sigma_n^2$

- if, for a certain k , $\sigma_k \ll \sigma_1$ eliminate the eigenvalues and eigenvectors above k .

Principal component analysis

► Given principal components $\phi_i, i \in 1, \dots, k$ and a test sample $\mathcal{T} = \{\mathbf{t}_1, \dots, \mathbf{t}_n\}, t_i \in \mathcal{R}^d$

- subtract mean to each point $\mathbf{t}'_i = \mathbf{t}_i - \hat{\mu}$
- project onto eigenvector space $\mathbf{y}_i = \mathbf{A}\mathbf{t}'_i$ where

$$\mathbf{A} = \begin{bmatrix} \phi_1^T \\ \vdots \\ \phi_k^T \end{bmatrix}$$

- use $\mathcal{T}' = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ to estimate class conditional densities and do all further processing on \mathbf{y} .

Principal components

- ▶ what are they? in some cases it is possible to see
- ▶ example: eigenfaces
 - face recognition problem: can you identify who is the person in this picture
 - training:
 - assemble examples from people's faces
 - compute the PCA basis
 - project each image into PCA space
 - recognition:
 - project image to classify into PCA space
 - find the closest vector in that space
 - label the image with the identity of this nearest neighbor

Principal components

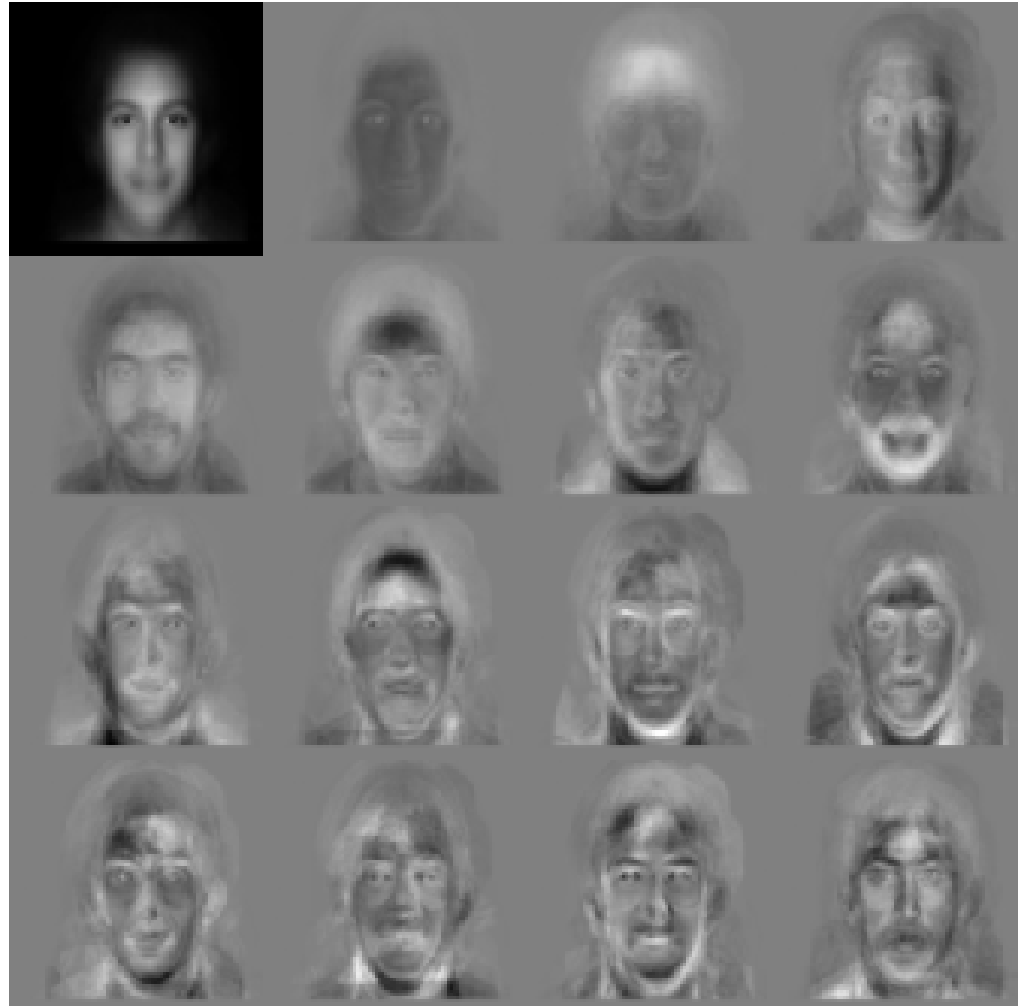
▶ face examples



Principal components

► Principal components (eigenfaces)

- high-energy ones tend to have low-frequency
- capture average face, illumination, etc.
- at the intermediate-level we have face detail
- low-energy tends to be high-frequency noise



Any questions?