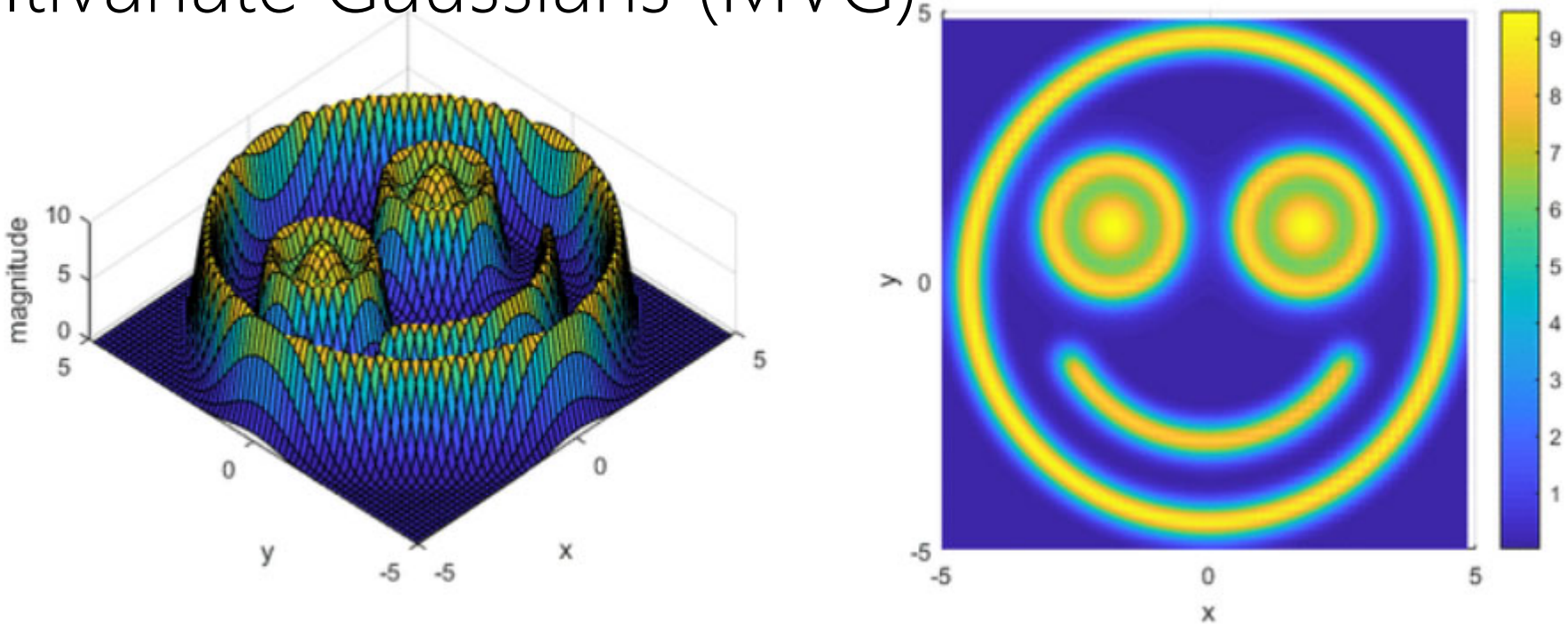
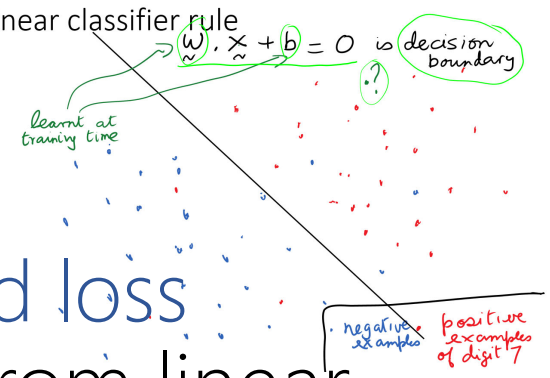
CS 189/289

Today's lecture:

1. Finish MLE from last class.
2. Multivariate Gaussians (MVG)



Maximum Likelihood Estimation (MLE)



This principle gives a useful, principled and widely-used loss function to estimate parameters of statistical models (from linear regression, to neural networks, and beyond).

- Training data set: $D = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad x_i \in R^D$, $y_i \in R$ or $y_i \in \{-1,1\}$
- Model class: aka hypothesis class $\quad f(x|w,b) = w^T x + b$ **Linear Models**
- Loss Function: $\quad L(a,b) = (a-b)^2$ **Squared Loss**
- Learning Objective: $\quad \underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f(x_i \mid w,b))$

Optimization Problem

RVs!

The basic set-up of MLE

- Given data $D = \{x_i\}_{i=1}^N$ for $x_i \in R^d$
- Assume a set (family) of distributions on R^d, $\{p_\theta(x) | \theta \in \Theta\}$.
- Assume D contains samples from one of these distributions:

$$x_i \sim p_{\hat{\theta}}(x)$$

- This assumes that each element of D is *identically and independently distributed* (iid).

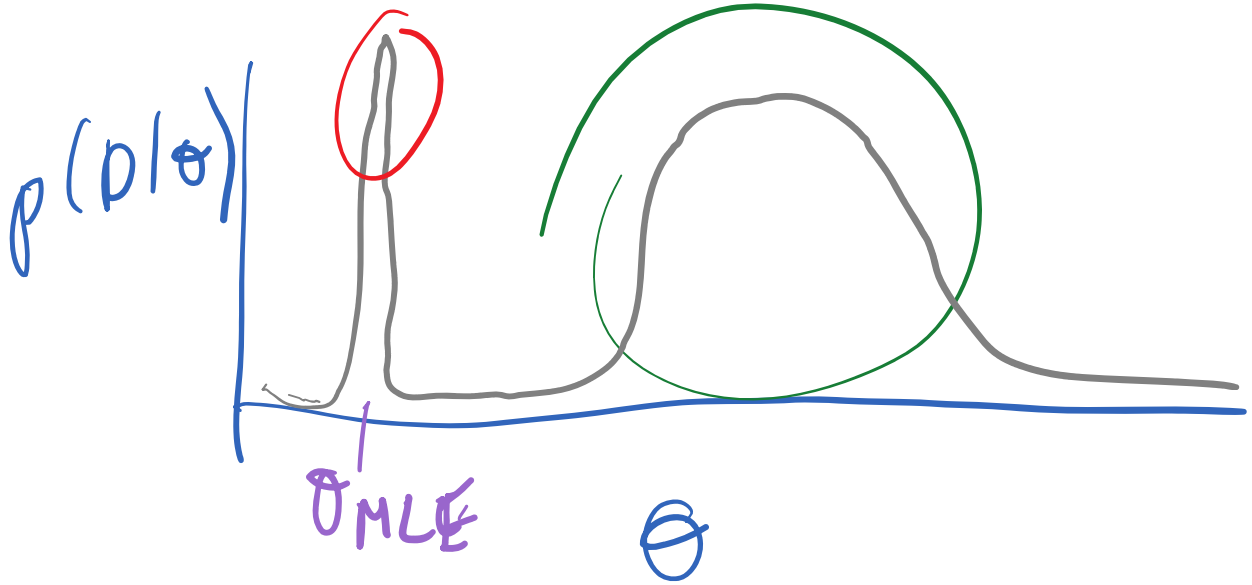
Goal of MLE: "learn"/estimate the value of θ that "pins down" the distribution from which the data came.

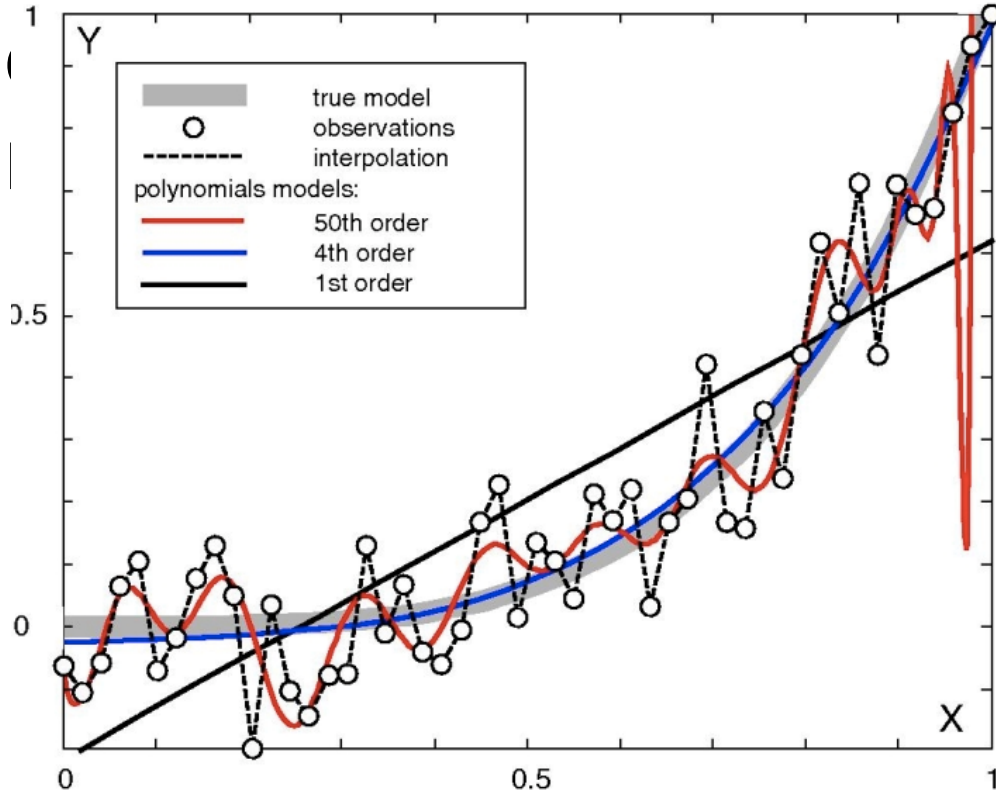
Definition: θ_{MLE} is a MLE for θ with respect to the data and family of distributions, if $\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} p(D|\theta)$.

"likelihood function" of θ.

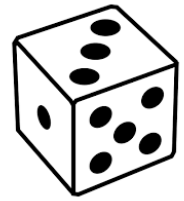
MLE yields a "point estimate" of our parameter

- When we perform MLE, we get just one single estimate of the parameter, $\boldsymbol{\theta}$, rather than a distribution over it which captures uncertainty.
- In Bayesian statistics, we obtain a (p
over $\boldsymbol{\theta}$. We will touch more on this i





e.g. MLE for the multinomial distribution



- Consider a six-sided die that we will roll: we want to know the probability of each side of the die turning up ($\theta = \theta_1 \dots \theta_6$).
- Assume we have observed N rolls, with RV, $X \sim p_\theta(X)$.
- We write that $P(X = k|\theta) = \theta_k$ (when k^{th} side faced up).
- Lets use MLE to estimate these parameters.
- First, since one side must always face up, we know that $1 = \sum_k \theta_k$.
- Second, let us denote $P(X = x|\theta) \equiv \theta_x$ (pick off the right parameter).
- Now we write the likelihood:

$$P(D|\theta) = p(x_1, \dots x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N}\prod_{k=1}^{6} \theta_k^{I[x_i=k]} = \prod_{k=1}^{6} \theta_k^{\sum_i^N I[x_i=k]} = \prod_{k=1}^{6} \theta_k^{n_k}$$

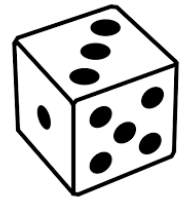
$n_k \equiv |\{i | x_i = k\}|$

Now our MLE problem becomes:

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\theta \in \{\Theta | 1 = \sum_k \theta_k\}}{\operatorname{argmax}} \sum_{k=1}^{6} \log \theta_k^{n_k}$$

constrained optimization

e.g. MLE for the multinomial distribution



Have a constrained optimization problem:

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta) = \underset{\theta \in \{\Theta | 1 = \sum_k \theta_k\}}{\operatorname{argmax}} \sum_{k=1}^{6} \log \theta_k^{n_k}$$

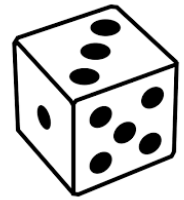
constrained optimization

What is one technique you should have learned in first year calculus to solve this?

The technique of Lagrange multipliers (Appendix C of textbook):

$J(\theta, \lambda) = \log p(D|\theta) + \lambda(1 - \sum_k \theta_k)$ (look for stationary points wrt θ, λ)

e.g. MLE for the multinomial distribution



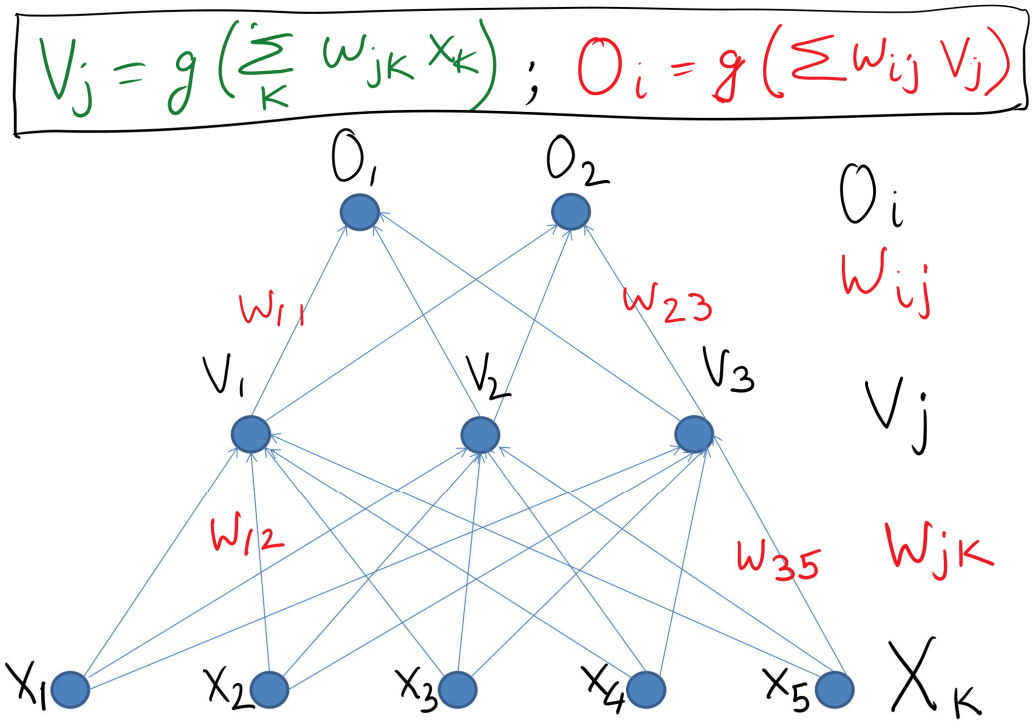
$$J(\theta,\lambda) = log\ p(D|\theta) + \lambda(1 - \sum_k \theta_k) = \sum_{k=1}^{6} \log \theta_k^{n_k} + \lambda(1 - \sum_k \theta_k)$$

1. $\frac{\partial J}{\partial \lambda} = 0 \Rightarrow 1 = \sum_k \theta_k$ (we just get the constraint back)
2. $\frac{\partial J}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \sum_{k=1}^{6} \log \theta_k^{n_k} - \frac{\partial}{\partial \theta_k} \lambda\theta_k = \frac{n_k}{\theta_k} - \lambda = 0 \Rightarrow \theta_k = \frac{n_k}{\lambda}$.
3. Lets plug this into 1), $1 = \sum_k \theta_k = \sum_k \frac{n_k}{\lambda} \Rightarrow \lambda = \sum_k n_k = N$.
4. All together then, $\theta_k = \frac{n_k}{N}$.

Doing MLE requires optimization

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(D|\theta)$$

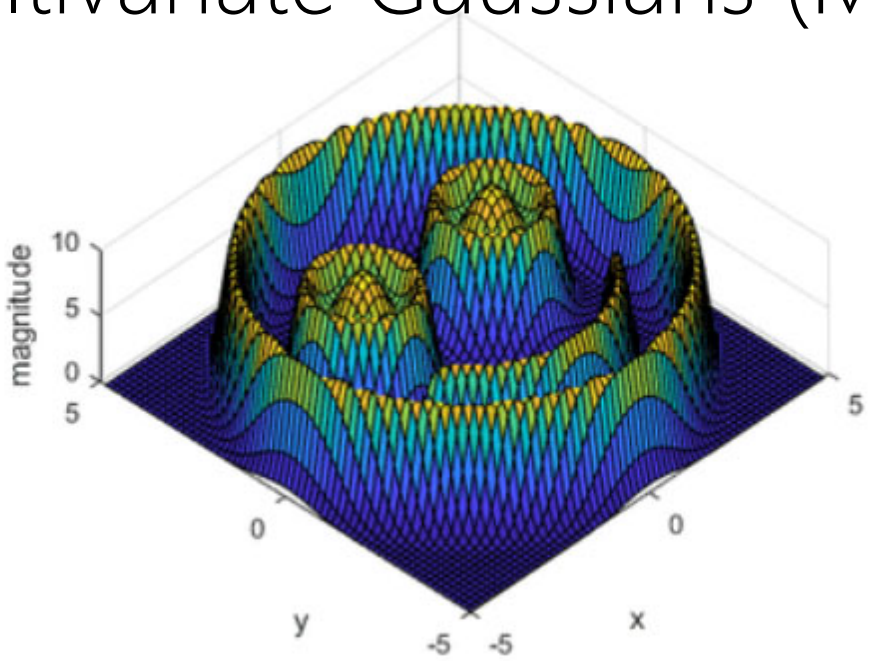
- For Gaussian, multinomial, and more, the MLE can be obtained in closed form by setting the derivative to zero.
- What if we had a neural network model such as mentioned in the first lecture?
- Here, we need *iterative optimization* (can take entire classes on special cases of this (e.g. Convex Optimization). More later.

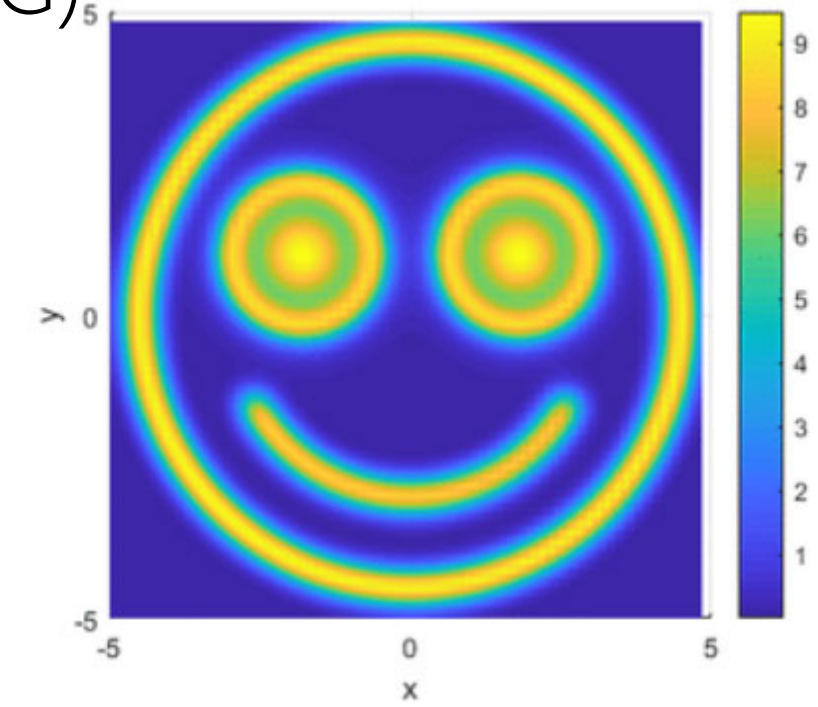


CS 189/289

Today's lecture:

1. Finish MLE from last class.
2. Multivariate Gaussians (MVG)





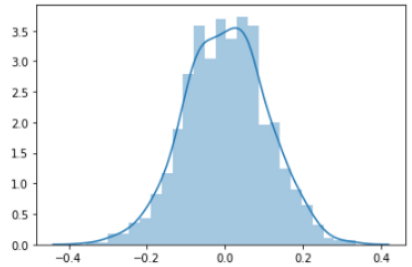
Assigned readings

- 3-3.2.3 (Multivariate Gaussians: Geometry, Moments, Covariance forms)

Multivariate Gaussian (MVG) distributions

Recall that the pdf of a univariate Gaussian (normal) distribution is:

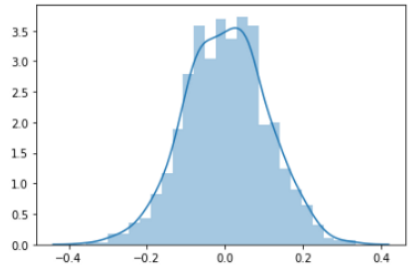
$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



Multivariate Gaussian (MVG) distributions

Recall that the pdf of a univariate Gaussian (normal) distribution is:

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

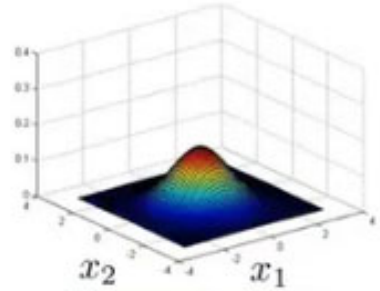


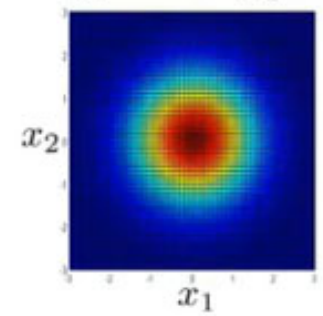
positive semi definate

The multivariate extension of this is for $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ and PSD.

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$\sqrt{|2\pi\Sigma|}$





https://regenerativetoday.com/univariate-and-bivariate-gaussian-distribution-clear-explanation-with-visuals/

Carl Friedrich Gauss

Portrait of Gauss by Christian Albrecht Jensen (1840)

Born	Johann Carl Friedrich Gauss 30 April 1777 Brunswick, Principality of Brunswick-Wolfenbüttel
Died	23 February 1855 (aged 77) Göttingen, Kingdom of Hanover, German Confederation

Why a lecture on MVGs?

MVGs permeate much of classical and modern day ML:

- Classification: generative vs. discriminative (later).
- Unsupervised models: Principal Components Analysis & autoencoders (later).
- Advanced topics: Gaussian Process Regression (and deep versions thereof), *etc.*

Why a lecture on MVGs?

- *All models are wrong but some are useful!*— George Box, *JASA* 1976.
- Ubiquitous in natural phenomena because of CLT.
- CLT: sum large # of independent RVs, their sum tends towards a Gaussian distribution.
  - *e.g.*, complex genetic traits such as height, blood pressure, *etc.*
- Convenient to work with (analytically tractable).

PCA Teaser: MVGs will let us reduce dimensions!

PCA “basis” images, $x \in \mathbb{R}^{1024}$

https://mikedusenberry.com/on-eigenfaces

Goals of this lecture:

1. Give you intuitive interpretation and manipulation of MVG (with technical underpinnings).
2. Teach you some of the properties of MVG that will come in handy for ML.

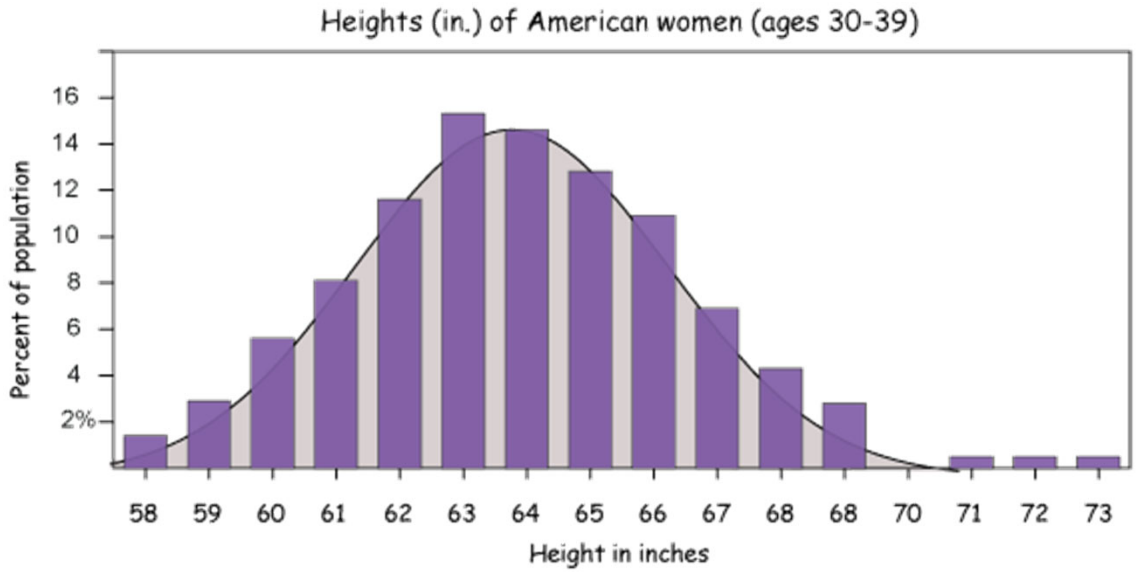
[may see *MVN* for "Multivariate Normal Distribution"]

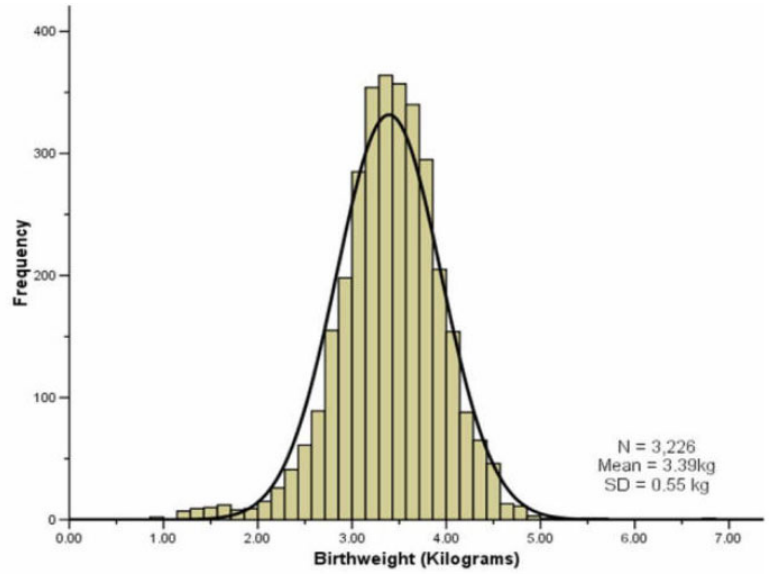
Multivariate Gaussian (MVG) distributions

- Consider two quantities, height and weight (of humans).
- Given the arguments of CLT with genetics, it's plausible that each of these is Gaussian distributed, so lets assume:

$$height = X_h \sim N(\mu_h, \sigma_h^2)$$

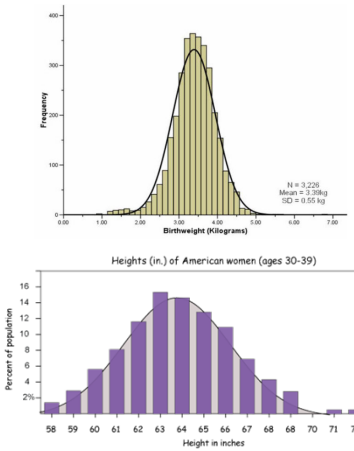
$$weight = X_w \sim N(\mu_w, \sigma_w^2)$$



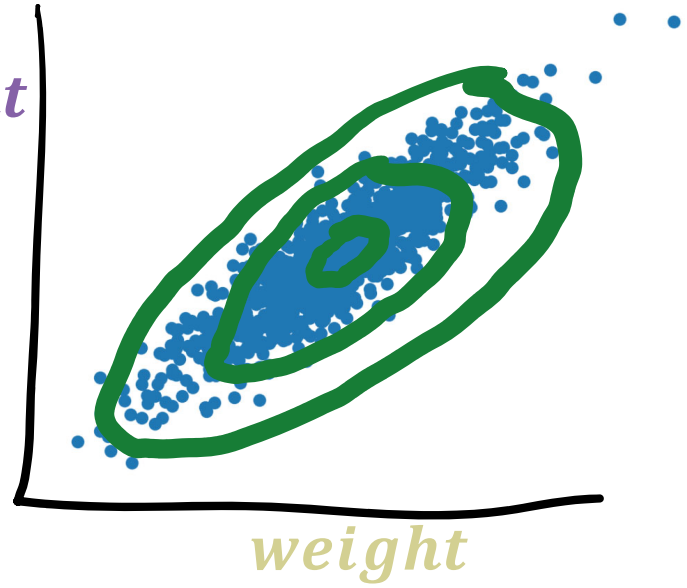


Suppose I want the *joint distribution*, $p([X_h = x_h, X_w = x_w])$, how would we write it down? (Shorthand: $p([x_h, x_w])$).

Multivariate Gaussian (MVG) distributions



$$p([x_h, x_w]) = N(?)$$

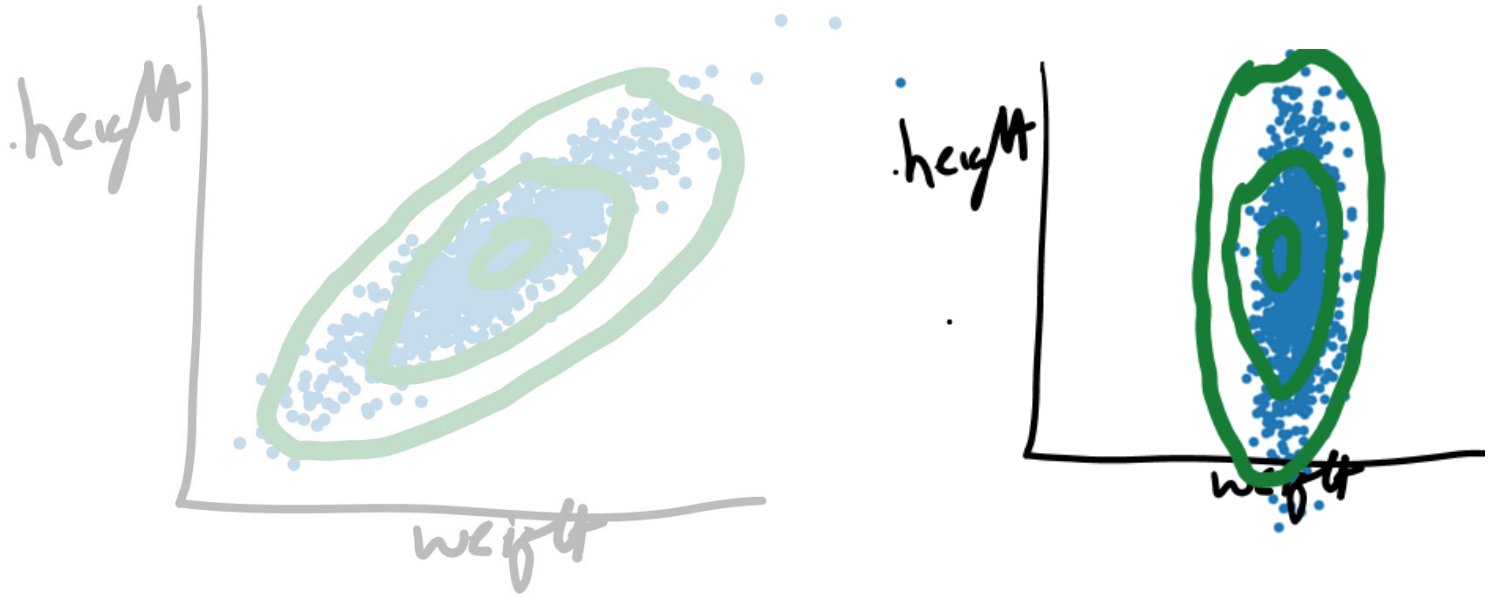


$$height = X_h \sim N(\mu_h, \sigma_h^2)$$
$$weight = X_w \sim N(\mu_w, \sigma_w^2)$$

- Each point is a sample from some 2D pdf, $p([x_h, x_w])$.
- If we computed the mean of this distribution, $u = [u_1, \mu_2]$, it would be..?.
- $u = [\mu_h, \mu_w]$
- How do we describe the "spread" of the points? What dimensionality would it even be?
- Can we use $p([x_h, x_w]) = N(x_h;\ \mu_h, \sigma_h^2)^* N(x_w;\ \mu_w, \sigma_w^2)$?

Multivariate Gaussian (MVG) distributions

$p([x_h, x_w]) = ?$

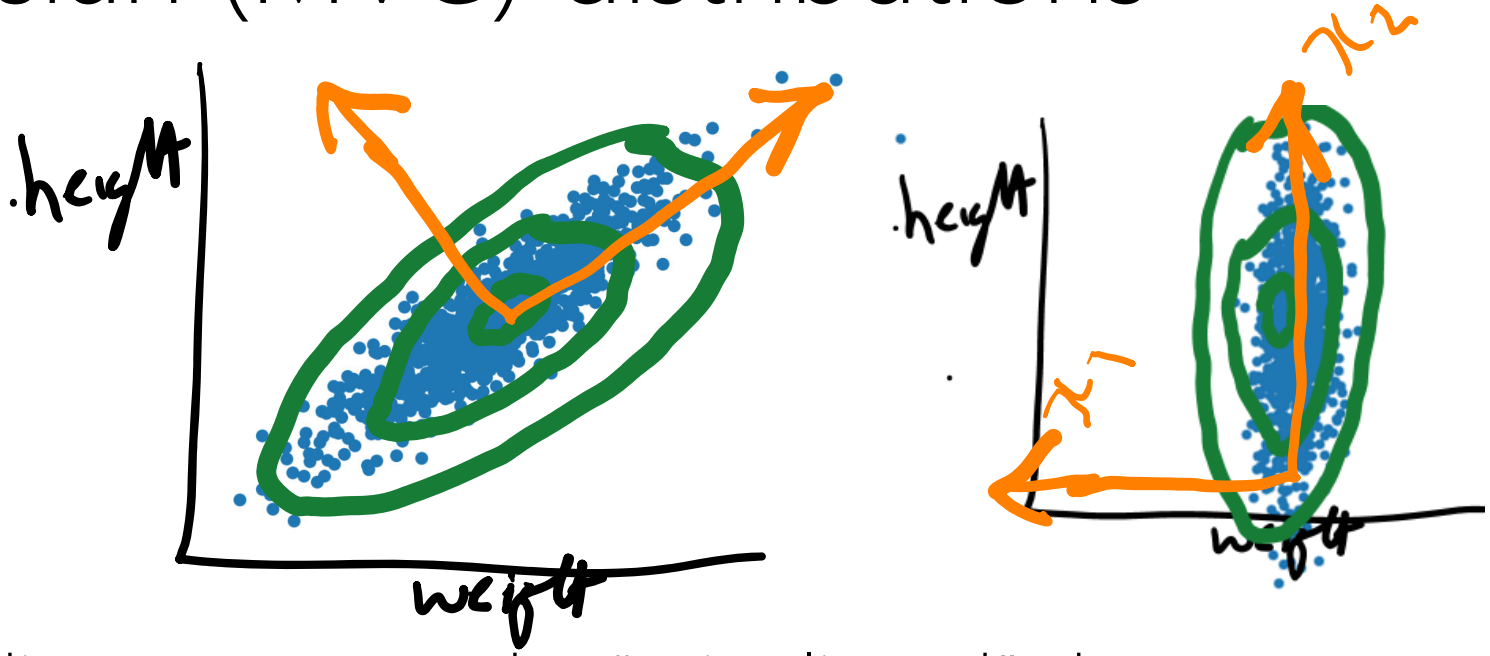


If independent RVs, $p([x_h, x_w]) = N(\mu_h, \sigma_h^2)^* N(\mu_w, \sigma_w^2)$.

$height = X_h \sim N(\mu_h, \sigma_h^2)$ $\quad$ $weight = X_w \sim N(\mu_w, \sigma_w^2)$

Multivariate Gaussian (MVG) distributions

$$p([x_h, x_w]) = ?$$



- If we could rotate the coordinate system to be "axis aligned", then

$$p([x_1, x_2]) = N(x_1; \mu_1, \sigma_1^2) * N(x_2; \mu_2, \sigma_2^2).$$

- How do we do a rotation?
- Multiply by an appropriate orthonormal matrix, Q:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = Q \begin{bmatrix} x_h \\ x_w \end{bmatrix}$$

$$height = X_h \sim N(\mu_h, \sigma_h^2) \qquad weight = X_w \sim N(\mu_w, \sigma_w^2)$$

Understanding the covariance matrix

"Baby" case: variables are independent, and each is 1D:

- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2)$
- Then $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$

Understanding the covariance matrix

"Baby" case: variables are independent, and each is 1D:

- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2)$
- Then $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$

$$= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$$

$$= P(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left\{[x-\mu_1, y-\mu_2]\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\right)$$

Understanding the covariance matrix

"Baby" case: variables are independent, and each is 1D:

- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2)$
- Then $p([x,y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$

$$= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$$

$$= P(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left\{[x-\mu_1, y-\mu_2]\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\right)$$

Math inverse: $\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_1^2\end{bmatrix}^{-1} = \begin{bmatrix}1/\sigma_1^2 & 0\\ 0 & 1/\sigma_2^2\end{bmatrix}$

$$[x,y]\,\Sigma^{-1}\begin{bmatrix}x\\ y\end{bmatrix}$$

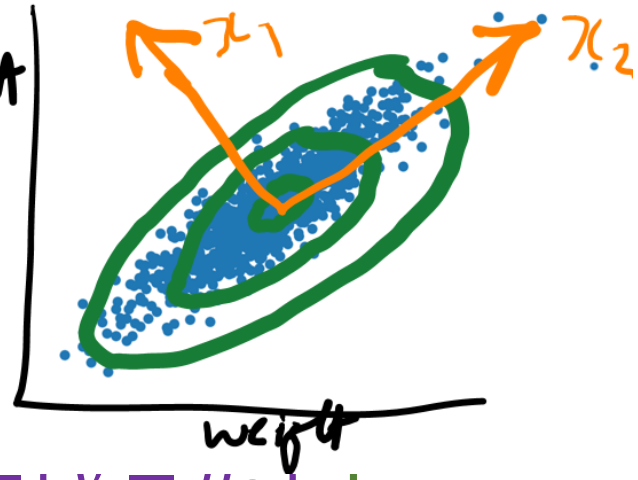
Σ is called the covariance matrix in the MVG (Σ^{-1} the precision matrix)

Understanding the covariance matrix

"Baby" case: variables are independent, and each is

- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2)$
- Then $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$

$$= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$$



$$= P(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left\{[x-\mu_1, y-\mu_2]\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\right)$$

math

inverse: $\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_1^2\end{bmatrix}^{-1} = \begin{bmatrix}1/\sigma_1^2 & 0\\ 0 & 1/\sigma_2^2\end{bmatrix}$

$$[x, y]\,\Sigma^{-1}\begin{bmatrix}x\\ y\end{bmatrix}$$

Σ is called the covariance matrix in the MVG (Σ^{-1} the precision matrix)

Review of expectations, variance, covariance

EXPECTATION

$E[X]$ ← often denoted by μ for mean.

for discrete $\sum x P(x)$

for continuous $\int x p(x) dx$

$$E_x[h(x,y)] = \int h(x,y) p(x) dx$$

Properties of expectation

- Linearity $E(\sum_i \alpha_i X_i) = \sum_i \alpha_i E(X_i)$ ← doesn't require independence
- Let $X_1 \ldots X_n$ be independent random variables

  $E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i)$

- $E(X+C) = E(X) + C$ ← constant

Review of expectations, variance, covariance

VARIANCE

Let X be a r.v. with Mean (expectation) $\mu = E[X]$

Then variance is defined as $E(X-\mu)^2$

Properties:

- $\text{Var}(X) = E(X^2) - \mu^2$
- $\text{Var}(aX+b) = a^2 \text{Var}(X)$
- if $X_1, \ldots X_n$ are independent and $\alpha_1, \ldots \alpha_n$ constants

$$\text{Var}\left(\sum \alpha_i X_i\right) = \sum \alpha_i^2 \text{Var}(X_i)$$

Review of expectations, variance, covariance

COVARIANCE of two R.V.

$$Cov(X,Y) = E\big((X-E(X))(Y-E(Y))\big) = E\big((X-\mu_X)(Y-\mu_Y)\big)$$

R.V with mean 0

$$= E[XY] - E[X]E[Y]$$

$\Rightarrow Cov(X,X) = Var(X)$. independent $X, Y \Rightarrow Cov(X,Y) = 0$

CORRELATION

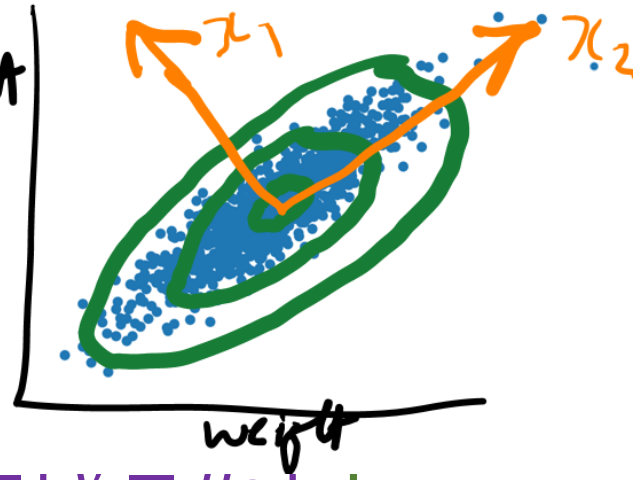
$$CORR(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \Rightarrow \rho \in [-1,1]$$

$\rho \rightarrow r$

Back to this example

Lets work out the baby case, variables are indepenc

- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2$
- Then $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2}}$



$$= \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{}{2\sigma_2^2}(y-\mu_2)\right]$$

$$= P(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left\{[x-\mu_1, y-\mu_2]\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\right)$$

lets revisit

$$[x\ y]\,\Sigma^{-1}\begin{bmatrix}x\\ y\end{bmatrix}$$

Σ is called the covariance matrix in the MVG (Σ^{-1} the precision matrix)

The covariance matrix contains covariances!

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2) & \mathrm{Cov}(X_1,X_3) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2) & \mathrm{Cov}(X_2,X_3) \\ \mathrm{Cov}(X_3,X_1) & \mathrm{Cov}(X_2,X_3) & \mathrm{Cov}(X_3,X_3) \end{bmatrix}$$

σ_1^2 → top-left entry; σ_3^2 ← bottom-right entry

Symmetric. Since $\mathrm{Cov}(X_1,X_2) = \mathrm{Cov}(X_2,X_1)$

also positive semi definite.

In our example $\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ is the Cov. Matrix. $\mathrm{Cov}(X_1,X_2) = 0$ since We started from independant variables X, Y

Multivariate Gaussian (MVG) distributions

Fact: If $X \in \mathbb{R}^d$ is distributed as a MVG, then $\forall i, j \in \{1, \dots d\}$

$cov(X_i, X_j) = 0$ iff X_i, X_j are independent.

Generally (beyond MVG), weaker statement: if X_i, X_j are independent then $cov(X_i, X_j) = 0$.

.

Intuition?

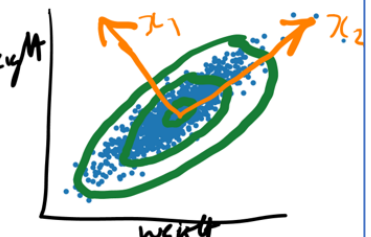
Lets work out the baby case, variables are indepenc

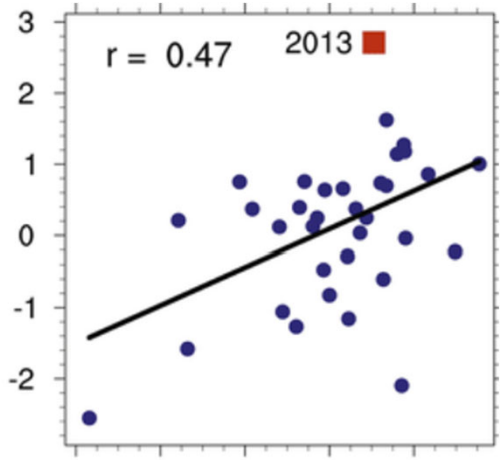
- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2$
- Then $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right]$

$$= \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$$

$$= P(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left\{[x-\mu_1, y-\mu_2]\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\right)$$

lets

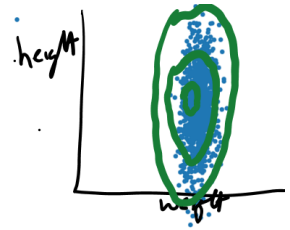




From the baby case to the general case

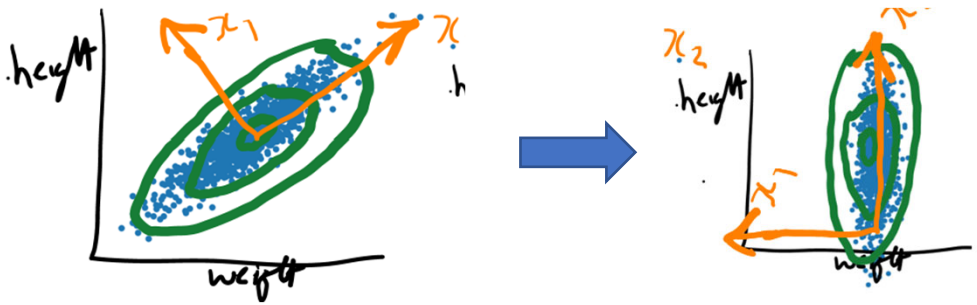
We worked the baby case, variables are independent, and each is 1D:
$X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2)$, so that $p([x, y]) = p(x)p(y)$

$$P(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left\{[x-\mu_1, y-\mu_2]\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\right)$$



How can we better understand the general case, with $\mathbf{X} \in \mathbb{R}^d$ and non-independence between the components?

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$



The MVG is at its core a quadratic form

MVG has 2 main terms:

1. Quadratic term, where most of the "*action happens*".

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

The MVG is at its core a quadratic form

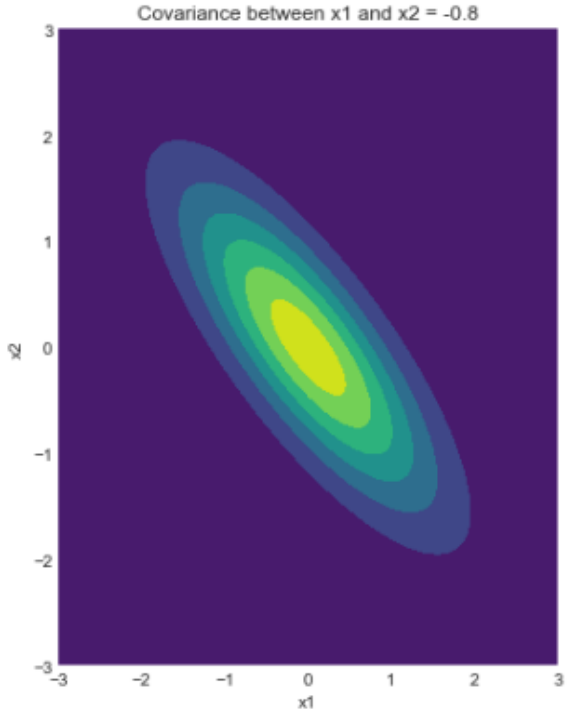
MVG has 2 main terms:

1. Quadratic term, where most of the "*action happens*".
2. Normalizing constant, which ensures that the distribution integrates to 1.

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx_1 dx_2 \cdots dx_n = 1.$$

Quadratic term⇒level sets of MVG pdf are ellipses



One level set (contour line) of MVG pdf comprises the values x for which $p(x)$ is a constant:

$$x^T \Sigma^{-1} x \simeq \text{constant}$$

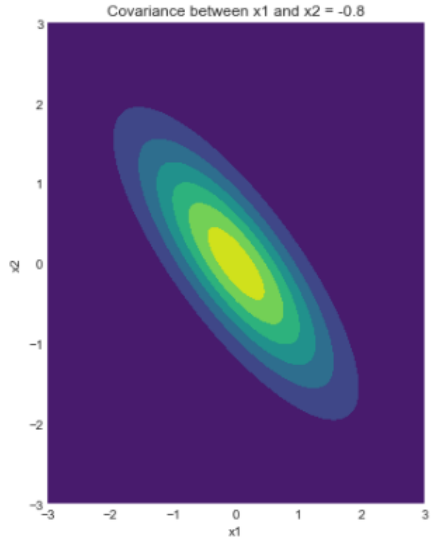
$1 \times d \quad d \times d \quad d \times 1$

e.g. $x \in \mathbb{R}^{d=2}$ then $0 = ax_1^2 + bx_1x_2 + cx_2^2 + d$

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

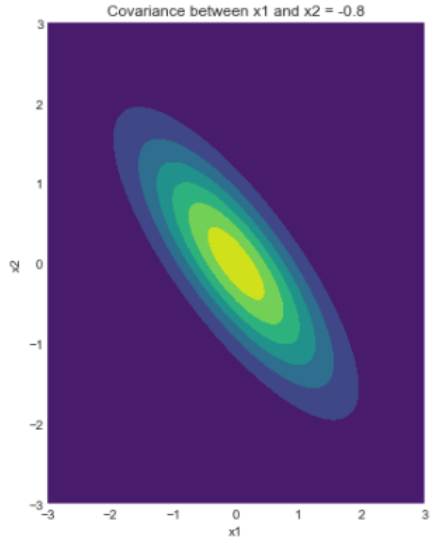
What do these look like in high dimensions?

To deal with a 14-dimensional space, visualize a 3D space…



—Geoff Hinton, "grandfather" of deep neural networks (U. Toronto).

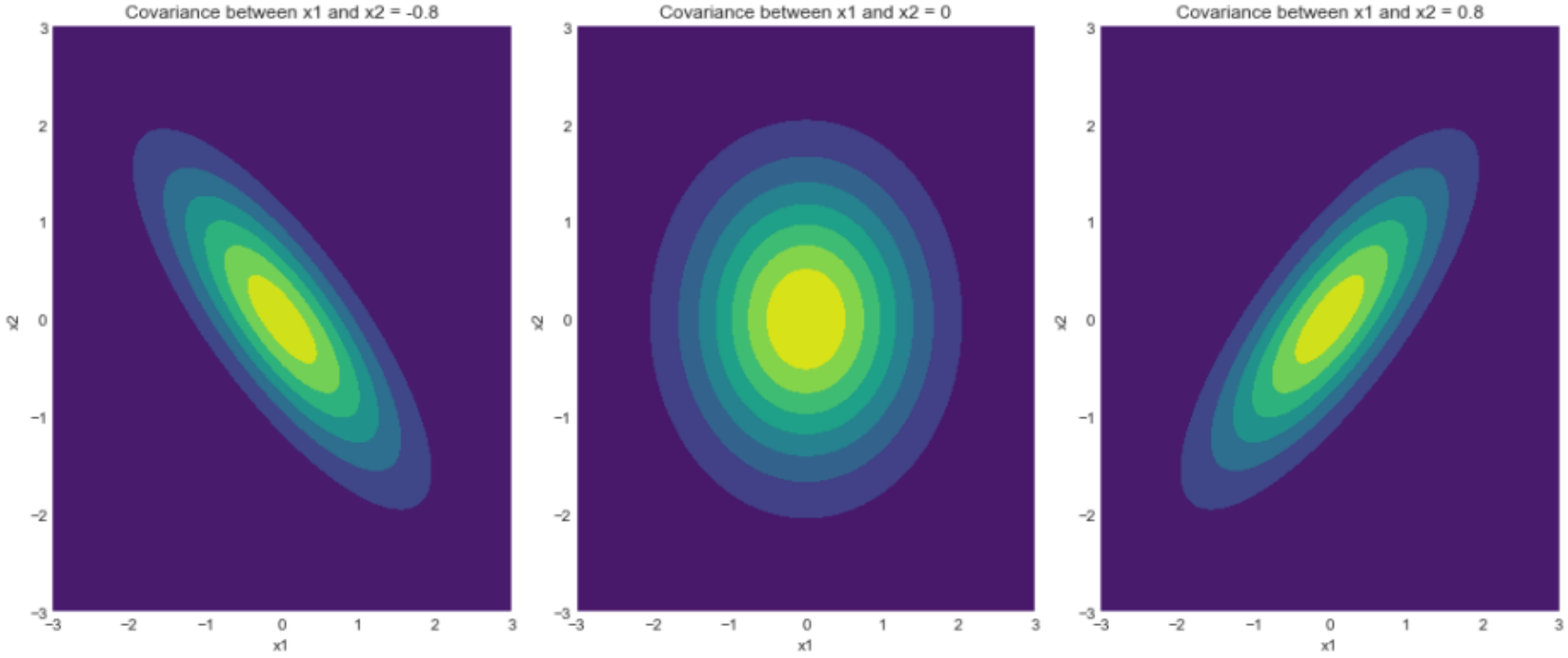
What do these look like in high dimensions?



To deal with a 14-dimensional space, visualize a 3D space and say "fourteen" to yourself very loudly. Everyone does it.

—Geoff Hinton, "grandfather" of deep neural networks (U. Toronto).

Still, lets try to get an intuition.



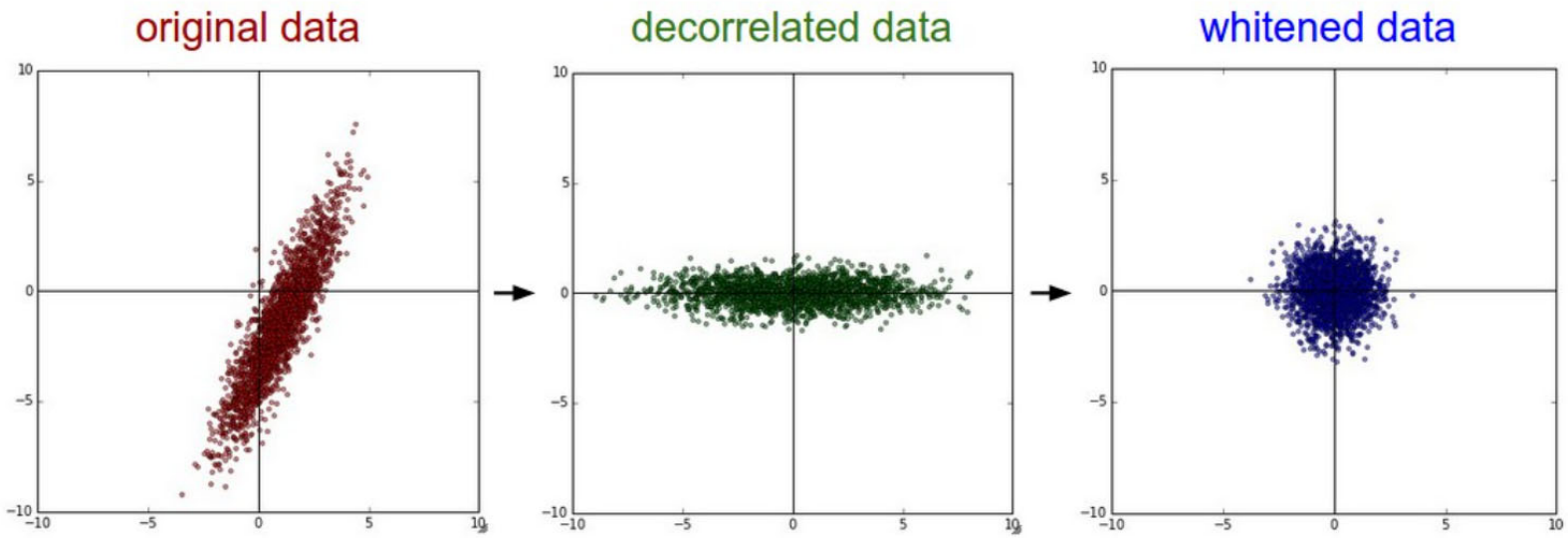
could visualize each pair of variables x_i and x_j

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Sphering a MVG

"diagonalized"

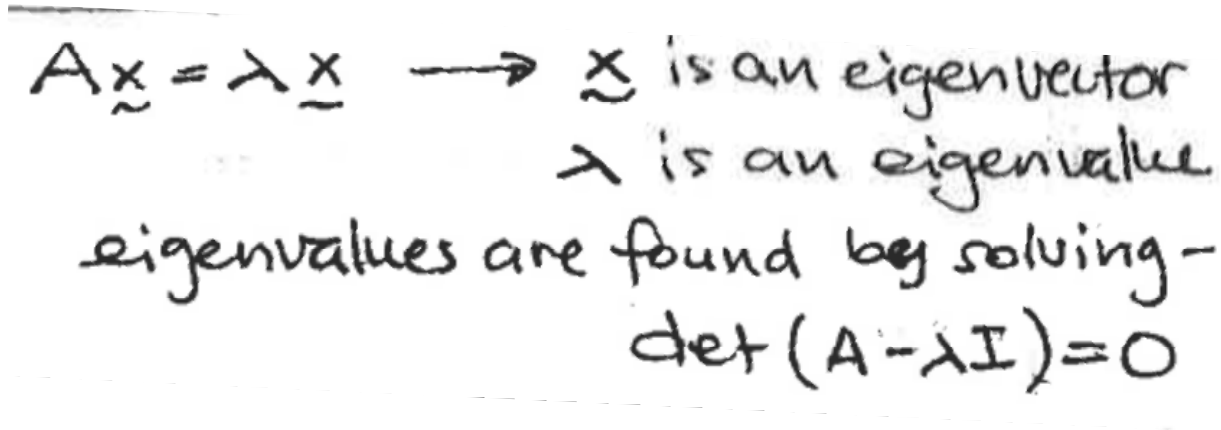
- To "sphere" a MVG is to transform it so as make all its contour lines be spheres (also called "whitening"):
- (useful for manipulation of MVGs related to PCA, advanced linear regressions, etc.)



https://pantelis.github.io/cs677/docs/common/lectures/optimization/whitening/

Linear Algebra: Diagonalizing a matrix

- For the MVG, the covariance matrix (and its inverse) is *symmetric* and *positive semi-definite* (PSD).
- Symmetric because covariance is symmetric $cov(x, y) = cov(y, x)$.
- Recall a symmetric matrix $C \in \mathbb{R}^{d \times d}$ is PSD iff $u^T C u \geq 0$ for every $u \in \mathbb{R}^d$. (PD if strictly > 0). It follows that all eigenvalues are ≥ 0.
- Recall eigenvalues:



Linear Algebra: Diagonalizing a matrix

Spectral theorem:

When A is symmetric $A = A^T$

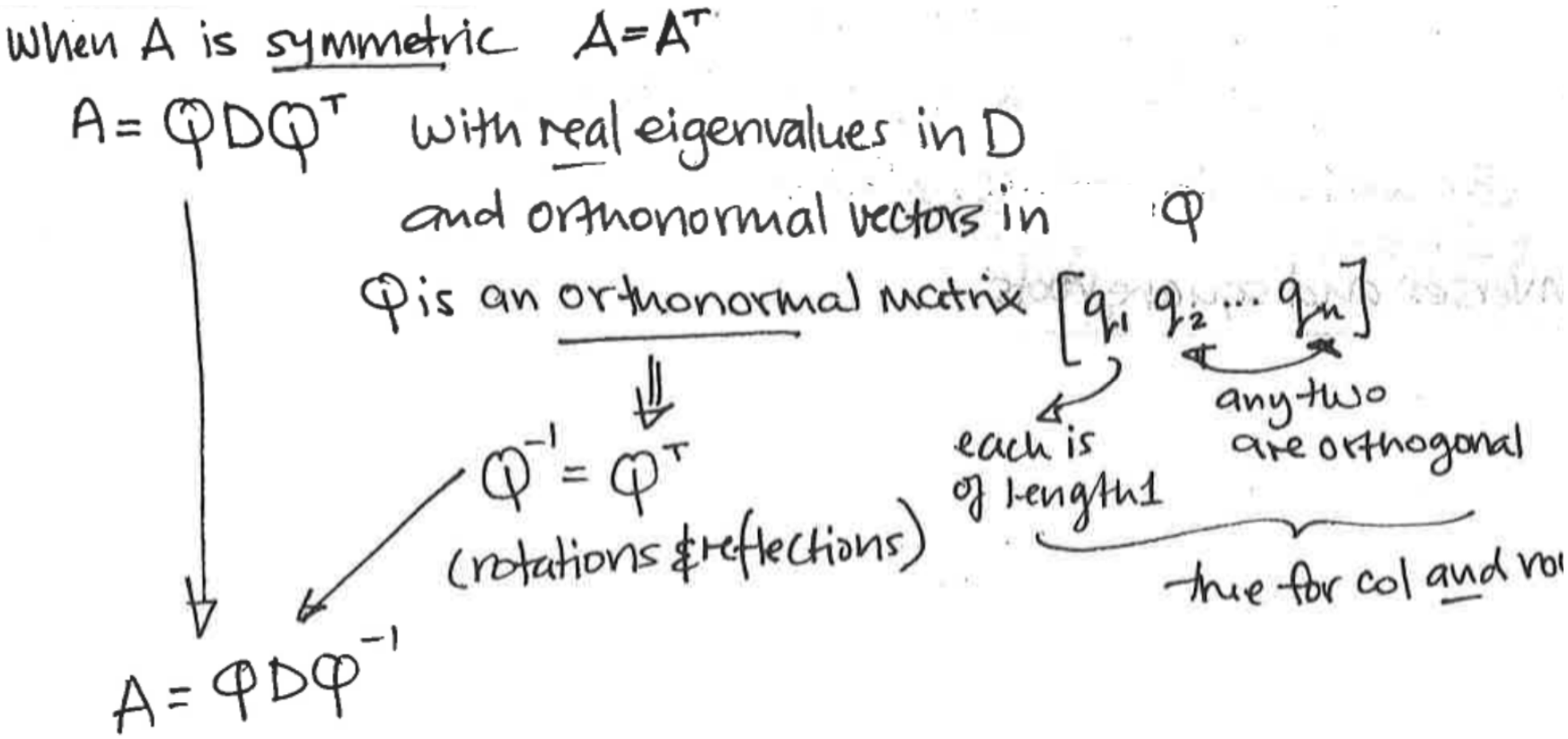
$A = \Phi D \Phi^T$ with real eigenvalues in D

and orthonormal vectors in Φ

Next we will use this theorem to "de-rotate" (to sphere) an ellipse.

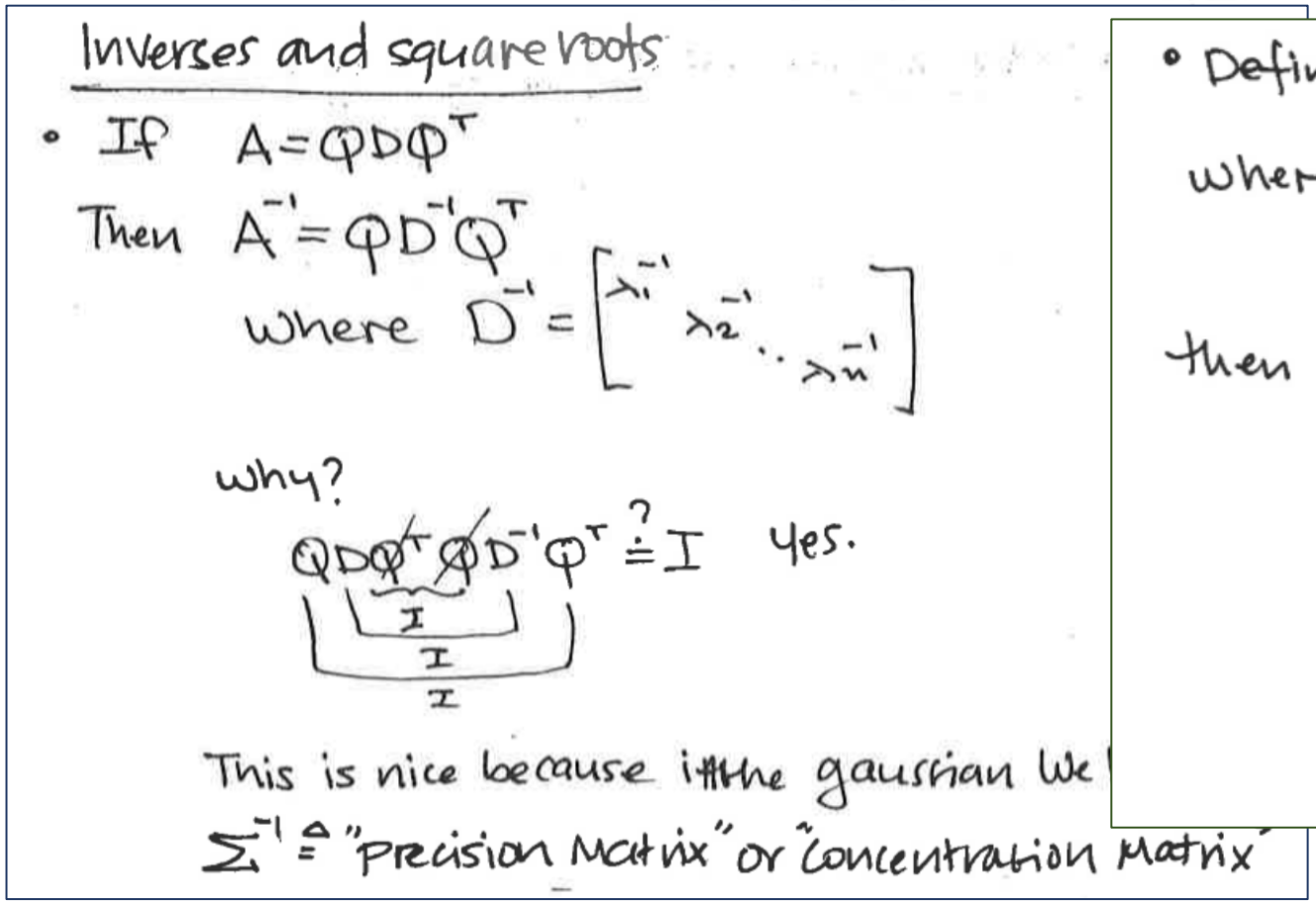
Linear Algebra: Diagonalizing a matrix

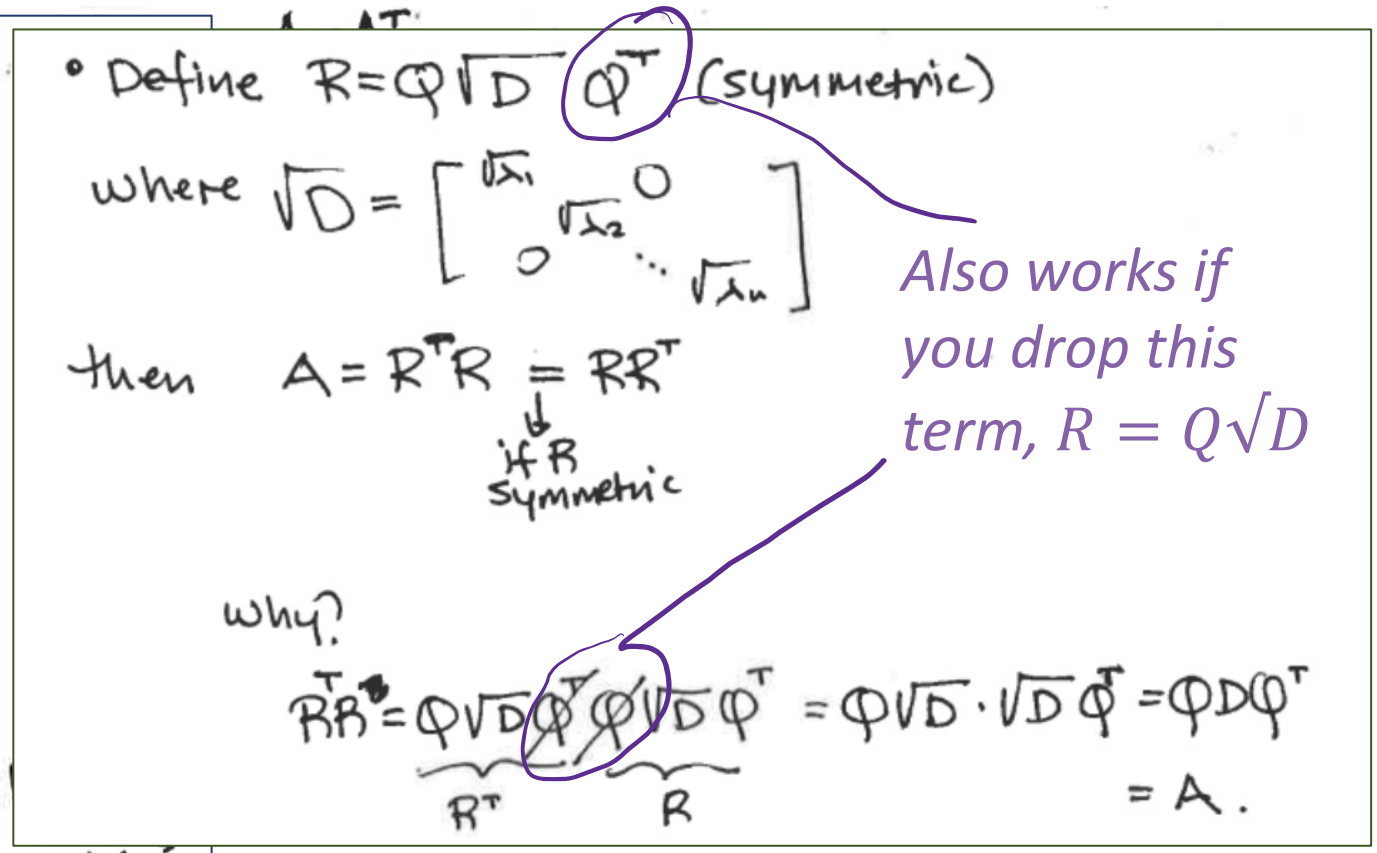
Spectral theorem:



Next we will use this theorem to "de-rotate" (to sphere) an ellipse.

Linear Algebra: Diagonalizing a matrix





Also works if you drop this term, $R = Q\sqrt{D}$

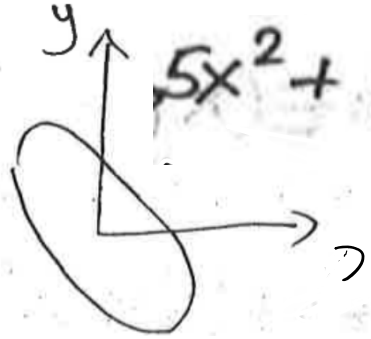
can be a useful factorization of A

Diagonalizing an ellipse

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 1$$

$A = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$ is positive definite

$$A = \Phi D \Phi^T$$



$5x^2 + 8xy + 5y^2 = 1$

1. Eigenvalues of A –

$Ax = \lambda x$

$$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 9 \\ 9 \end{bmatrix} \rightarrow v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \lambda_1 = 9$$

$$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \rightarrow v_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \lambda_2 = 1$$

2. Want to make this into an orthonormal Φ

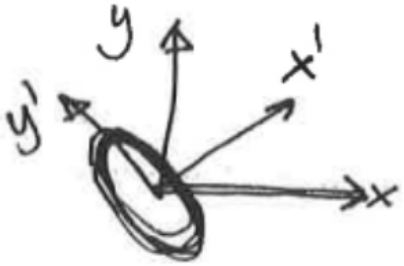
→ make eigenvectors normalized by dividing by $\sqrt{2}$

$$\Phi = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\rightarrow A = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

3. Change coordinate system along the eigenvectors –

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{\Phi} \begin{bmatrix} x \\ y \end{bmatrix} \rightarrow \begin{aligned} x' &= \frac{x+y}{\sqrt{2}} \\ y' &= \frac{x-y}{\sqrt{2}} \end{aligned}$$



⇒ in the new system: $9x'^2 + 1y'^2 = 1$

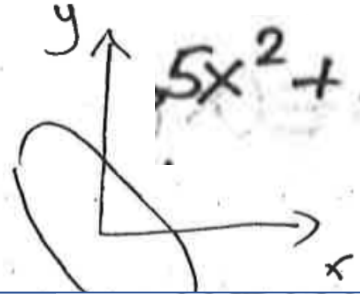
⇒ axis aligned!

(no cross terms)

Sphering an ellipse

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 1$$

$A = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$ is positive definite



$5x^2 + 8xy + 5y^2 = 1$

The axes point along the eigenvectors

the minor and major axes lengths are $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$.

⟹ So finding a coordinate system that makes the ellipse axis aligned is the same as diagonilizing A.

1. Eigenvalues & eigenvectors

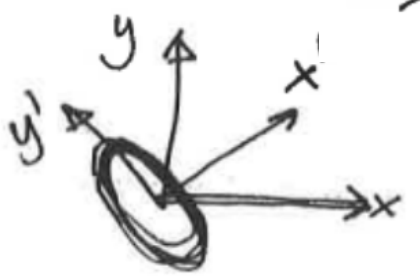
$Ax = \lambda x$

$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

3. Change coordinate system along the eigenvectors—

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{\varphi} \begin{bmatrix} x \\ y \end{bmatrix} \longrightarrow \begin{aligned} x' &= \frac{x+y}{\sqrt{2}} \\ y' &= \frac{x-y}{\sqrt{2}} \end{aligned}$$



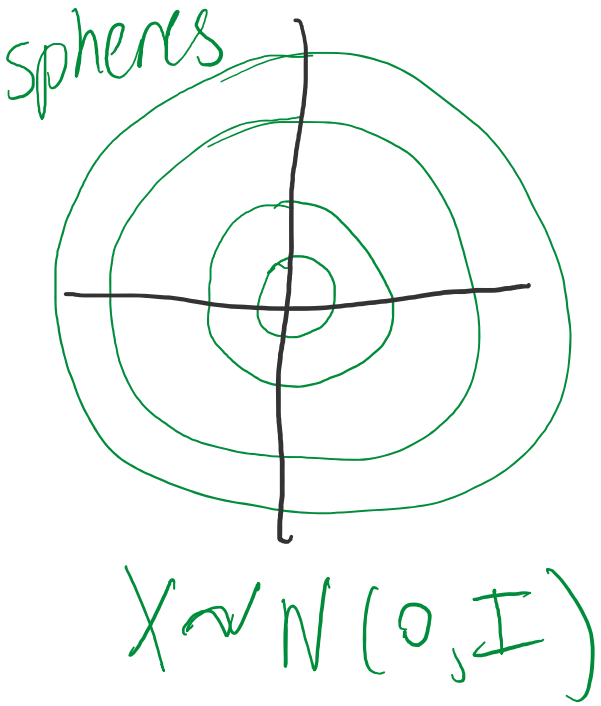
⟹ in the new system: $9x'^2 + 1y'^2 = 1$

⟹ axis aligned!

(no cross terms)

Geometric intuition: MVG: $\boldsymbol{\Sigma = I \rightarrow}$ general $\boldsymbol{\Sigma}$

- Let $X \sim N(0, I)$.
- Let $\Sigma = QDQ^T$ be a covariance matrix factored into its eigenvectors and diagonal matrix. Can also write it as $\Sigma = (QD^{\frac{1}{2}})(D^{\frac{1}{2}}Q^T) = AA^T$.
- Let $Y = AX + \mu$. Then by affine property $Y \sim N(\mu, \Sigma)$.



(1) $x_1, x_2 \ldots x_n \sim N(0,1)$ all independent.

$\underset{\sim}{X} \sim N(\underset{\sim}{0}, I)$.

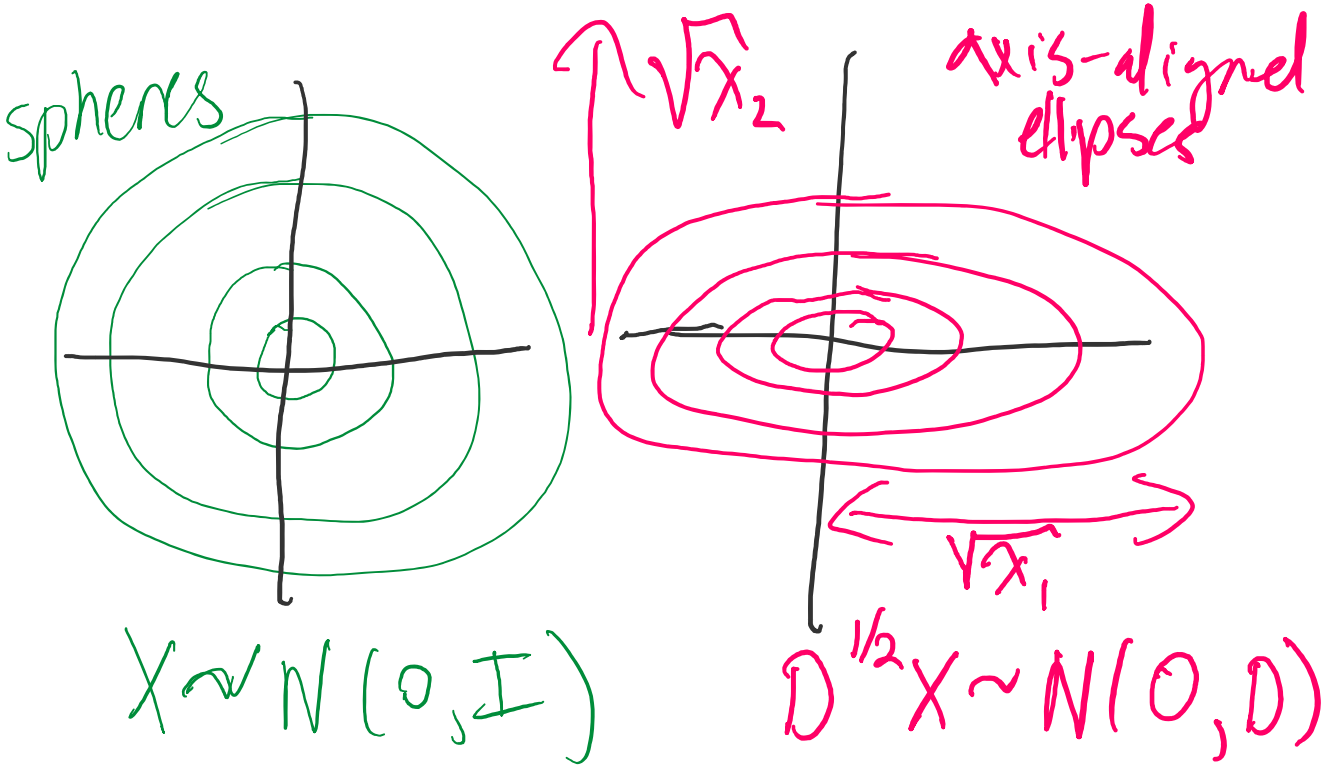
$\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{\mu} \sim N(\underset{\sim}{\mu}, AA^T)$

(2) If Σ is positive definite then if $Y \sim N(\mu, \Sigma)$ then -

$A^{-1}(Y - \mu) \sim N(0,1)$.

Geometric intuition: MVG: $\boldsymbol{\Sigma = I \rightarrow}$ general $\boldsymbol{\Sigma}$

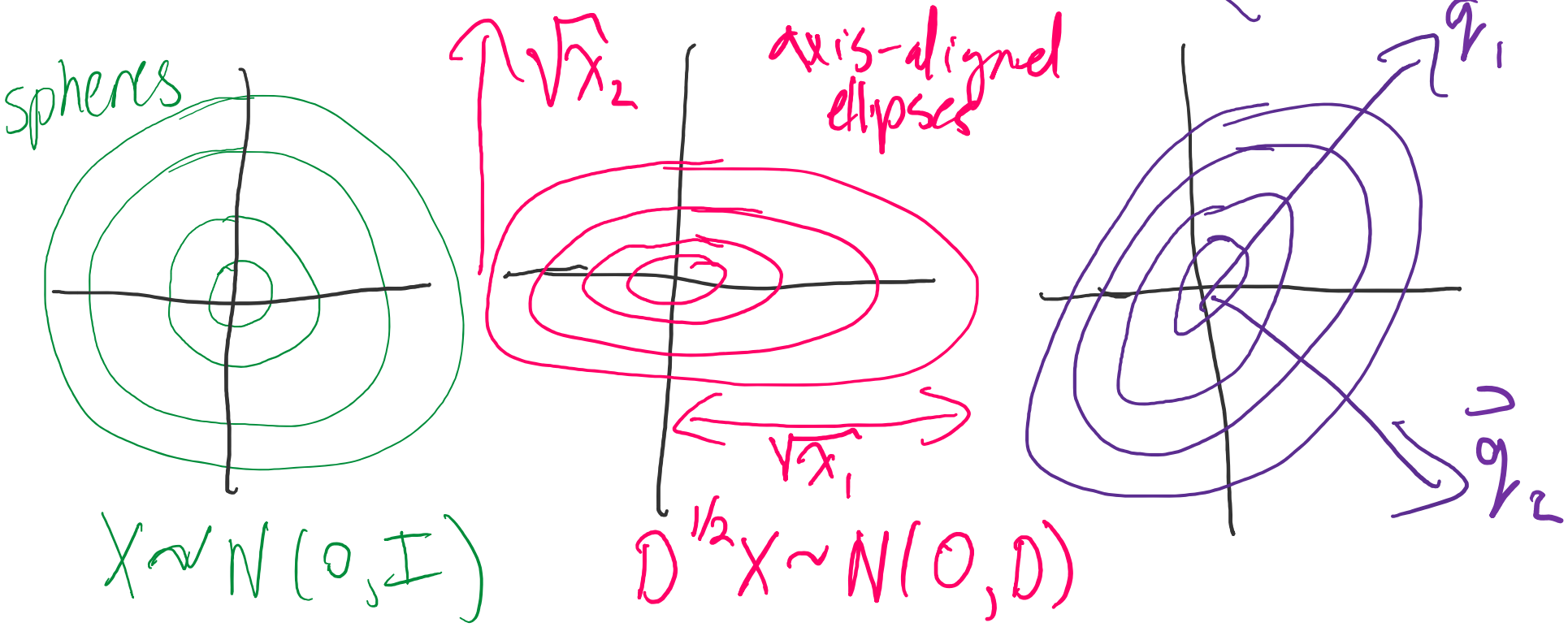
- Let $X \sim N(0, I)$.
- Let $\Sigma = QDQ^T$ be a covariance matrix factored into its eigenvectors and diagonal matrix. Can also write it as $\Sigma = (QD^{\frac{1}{2}})(D^{\frac{1}{2}}Q^T) = AA^T$.
- Let $Y = AX + \mu$. Then by affine property $Y \sim N(\mu, \Sigma)$



Geometric intuition

$QD^{1/2}X = AX \sim N(0, \Sigma)$

- Let $X \sim N(0, I)$.
- Let $\Sigma = QDQ^T$ be a covariance matrix factored into its eigenvectors and diagonal matrix. Can also write it as $\Sigma = (QD^{\frac{1}{2}})(D^{\frac{1}{2}}Q^T) = AA^T$.
- Let $Y = AX + \mu$. Then by affine property $Y \sim N(\mu, \Sigma)$

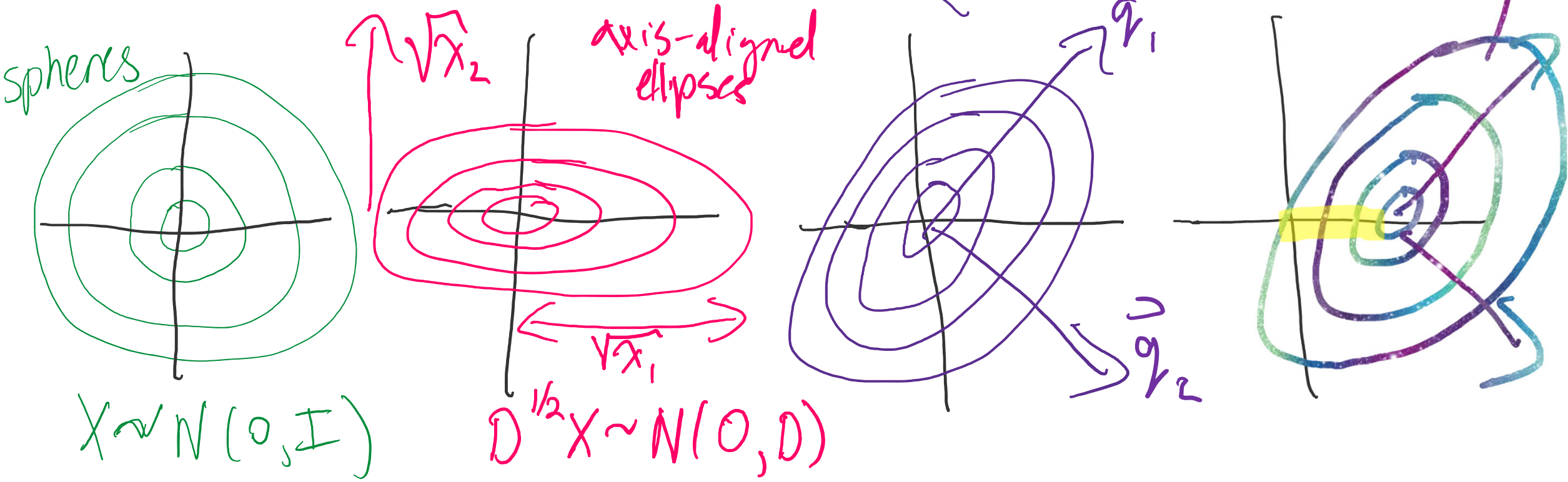


Geometric intuition

$QD^{1/2}X = AX \sim N(0, \Sigma)$

$AX + \mu \sim N(\mu, \Sigma)$

- Let $X \sim N(0, I)$.
- Let $\Sigma = QDQ^T$ be a covariance matrix factored into its eigenvectors and diagonal matrix. Can also write it as $\Sigma = (QD^{\frac{1}{2}})(D^{\frac{1}{2}}Q^T) = AA^T$.
- Let $Y = AX + \mu$. Then by affine property $Y \sim N(\mu, \Sigma)$



Geometric intuition

- Can decompose any MVG in terms of a "scaling", "rotation" and "shift" operator with respect to the standard $N(0, I)$ form.
- Went from sphere to general, but can also go in the other direction ("sphering").
  - Start with general $X \sim N(\mu, \Sigma)$.
  - Use spectral decomposition, $\Sigma = QDQ^T$
  - Then $\Sigma^{-1} = QD^{-1}Q^T$, and then $\Sigma^{-1/2} = QD^{-1/2}$
  - Thus $QD^{-1/2}(X - \mu) \sim N(0, I)$ (affine property)

(1) $x_1, x_2 \ldots x_n \sim N(0,1)$ all independent.

$\underset{\sim}{X} \sim N(\underset{\sim}{0}, I)$.

$\underset{\sim}{Y} = A\underset{\sim}{X} + \underset{\sim}{\mu} \sim N(\underset{\sim}{\mu}, AA^T)$

(2) If Σ is positive definite then if $Y \sim N(\mu, \Sigma)$ then $A^{-1}(Y - \mu) \sim N(0,1)$.