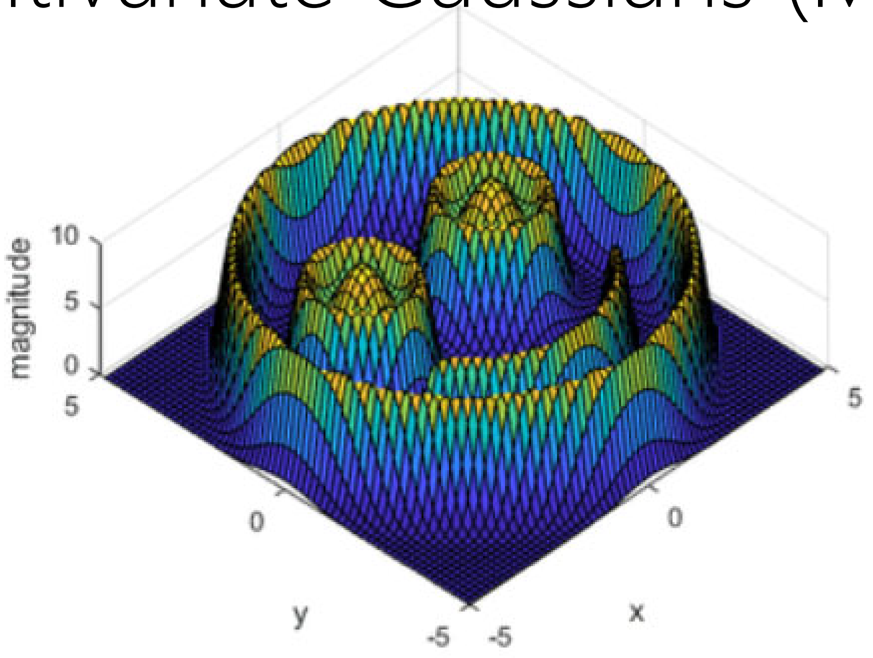
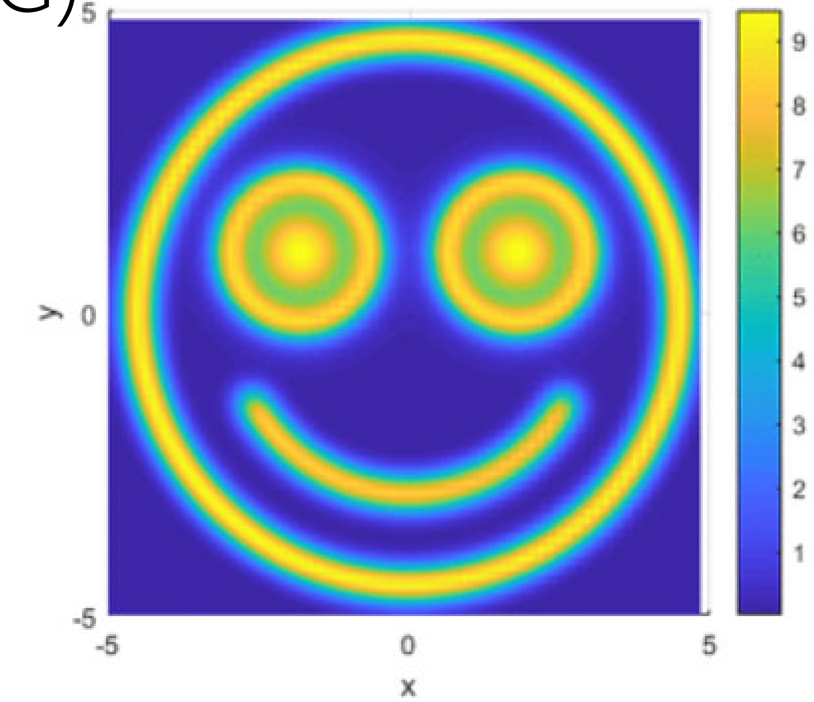
CS 189/289

Today's lecture:

1. Administrative FAQ
2. Multivariate Gaussians (MVG)





FAQs 1

Website is up and running finally! https://eecs189.org

I'm on the waitlist or a CE student. Do I have to keep up with assignments? Yes, you have to keep up with assignments like an enrolled student.

Will I get off the waitlist/have CE enrollment approved?

The EECS department handles waitlist/enrollment, so we don't have control over that. I don't know any more. Don't reach out to course staff/profs about CE approval: these applications are being processed.

FAQs 2

Can I get added to the bCourses?

- We are not using bCourses for the semester, as everything is self-contained within Ed, Gradescope, and the course website (eecs189.org). (Slides also posted to google folder)

Are discussions recorded/available by zoom?

- No. However, discussion sheets and solutions will be posted periodically.

What are the Ed and Gradescope codes?

- Posted on eecs189.org as well but also included info here:
- Gradescope Code: E73744
- Ed: **https://edstem.org/us/join/fCBF32**
- **Still having difficulty, contact** arvind.rajaraman@berkeley.edu (Head GSI).

FAQs 3

- Do we have alternate exam times?
  - No. You must take the midterm and final in-person with the rest of the class (if you're DSP, your exam time accommodation will be honored).
- Can lecture slides be posted before lecture?
  - Yes, there is now a Google Drive folder (check homepage of eecs189.org) which contains these slides for faster access.

Please check eecs189.org's syllabus first, as most students' questions so far have been answered there. If you have any other questions, please initiate contact in this order of priority.

- Public Ed post
- Private Ed post
- Email Head GSI (arvind.rajaraman@berkeley.edu)

FAQs 4

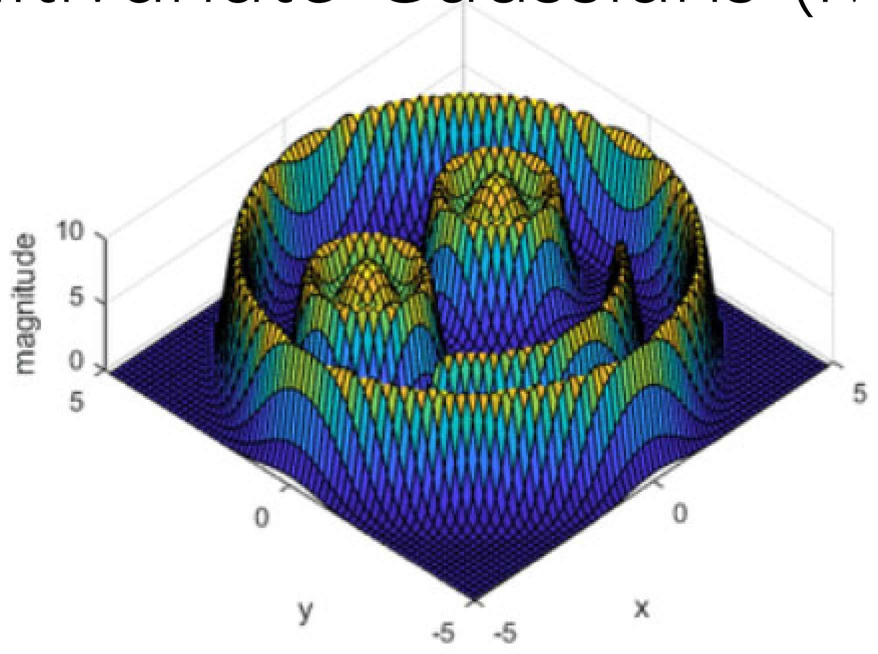
When and where are professor office hours.

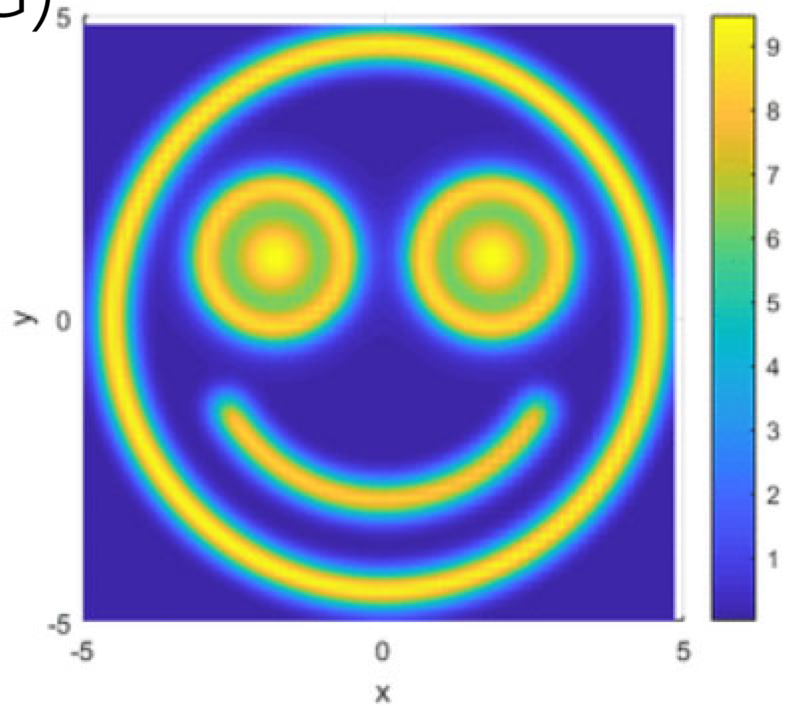
- These will be held by the lecturing professor directly after the lecture for one hour.
- We are struggling to find an appropriate room, so for now, follow the professor out of the lecture hall and we can meet outside somewhere near Dwinelle.

CS 189/289

Today's lecture:

1. Administrative FAQ
2. Multivariate Gaussians (MVG)





Recall last class:

e.g. MLE for univariate Gaussian

- In simple settings, we can find the setting of the parameters that set the partial derivatives to zero in closed-form:

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\mu,\sigma^2)$$

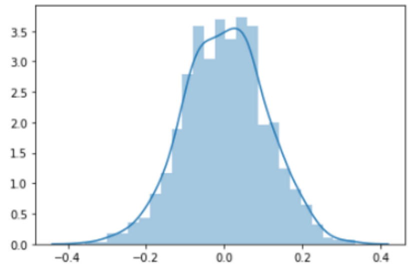
- Lets expand out so we can take the derivative:

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log p(x_i | \mu, \sigma^2) = \sum_i \frac{\partial}{\partial \mu} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right]$$

Today: Multivariate Gaussian (MVG) distributions

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

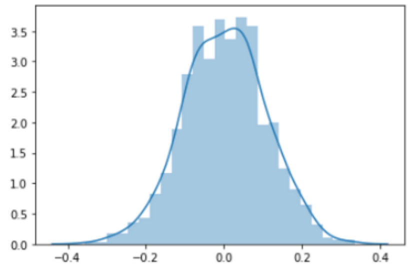
$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



Today: Multivariate Gaussian (MVG) distributions

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

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

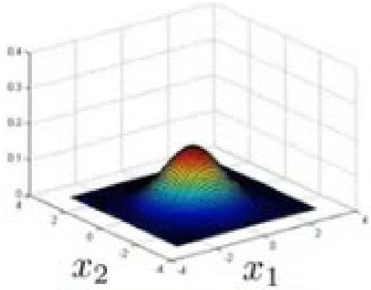


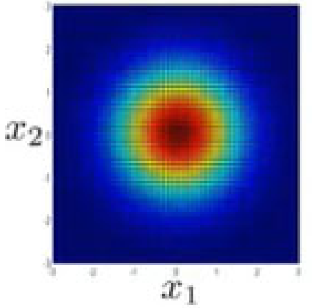
positive
semi
definate

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 (this class).
- Unsupervised models: Principle Components Analysis & autoencoders (this class).
- 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).

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.

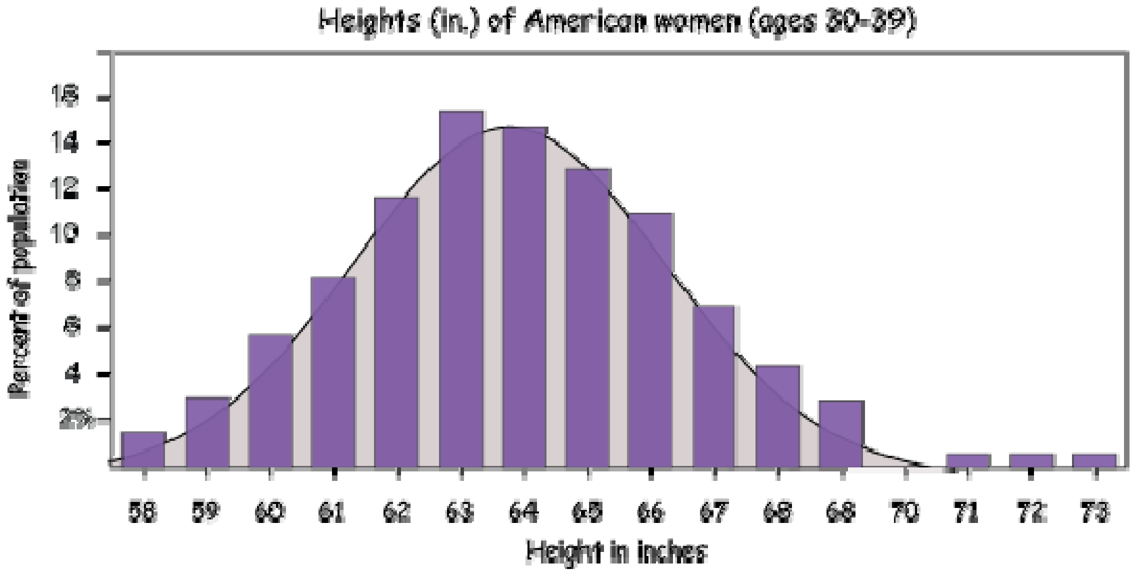
[may see *MVN* for "Multivariate Normal Distribution"]

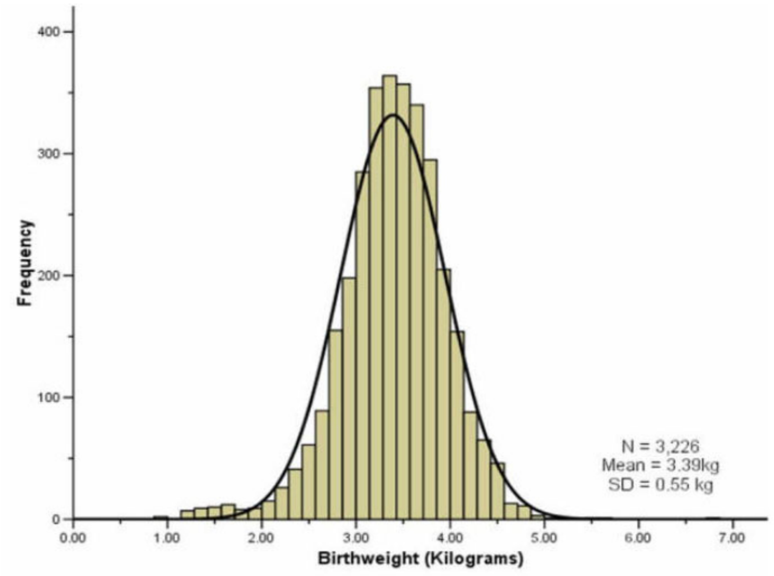
Multivariate Gaussian (MVG) distributions

- 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:

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

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

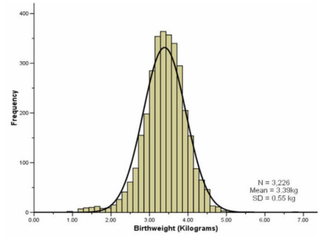


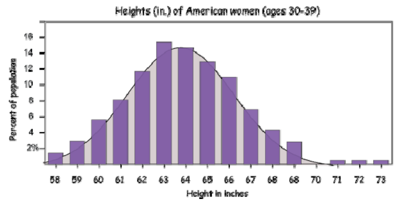


<u>joint</u> distribution *vs* <u>marginal</u> distributions

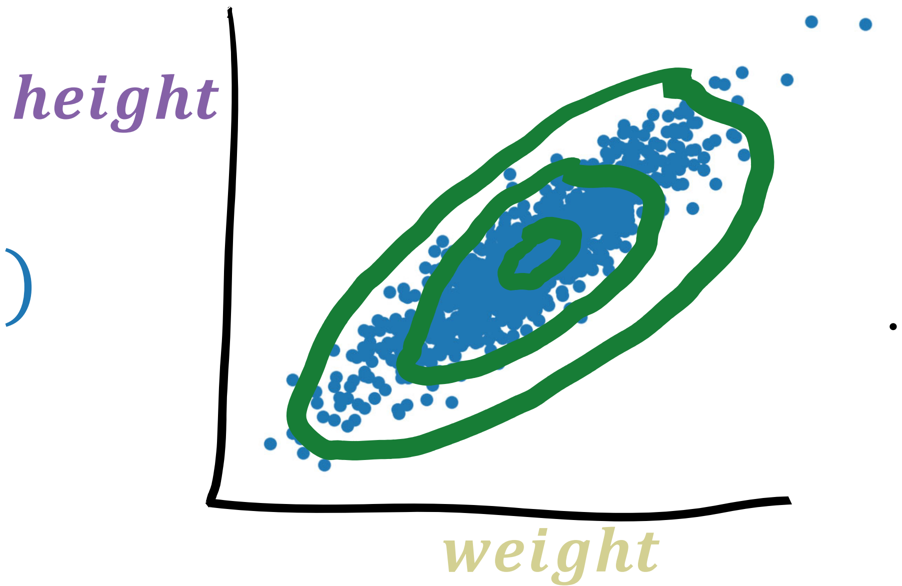
Suppose I want the *joint distribution*, $p([X_h = x_h, X_w = x_w])$, how would we write it down? (Shorthand: $p([x_h, x_w])$).

Multivariate Gaussian (MVG) distributions





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

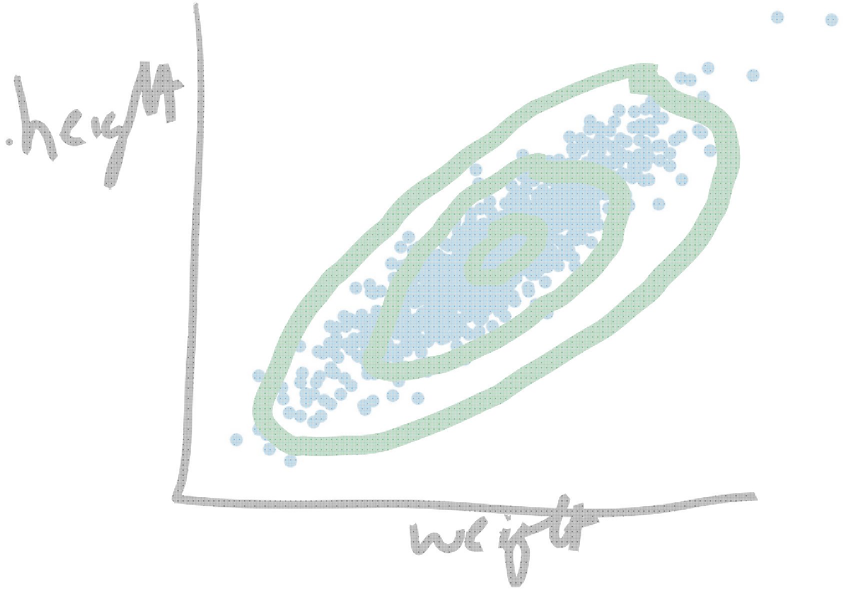


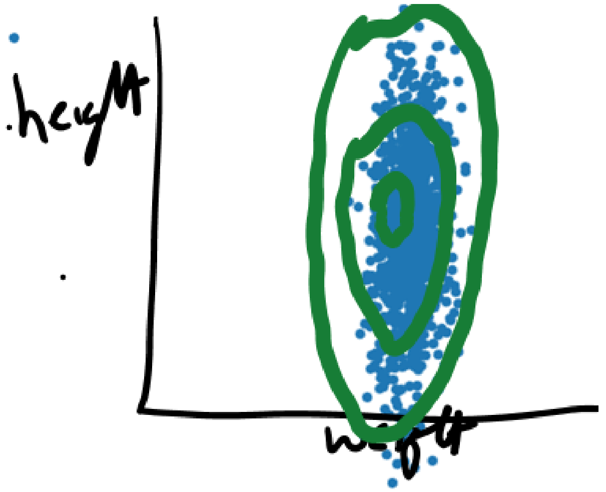
- Each point is a sample from some 2D pdf, $p([x_h, x_w])$.
- If we computed the mean of this distribution, $u = [u_1, \mu_2]$, it would be..?.
- $u = [\mu_h, \mu_w]$
- How do we compute/write the "spread" of the points?
- Can we use $p([x_h, x_w]) = N(x_h;\ \mu_h, \sigma_h^2)*N(x_w;\ \mu_w, \sigma_w^2)$?

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

Multivariate Gaussian (MVG) distributions

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



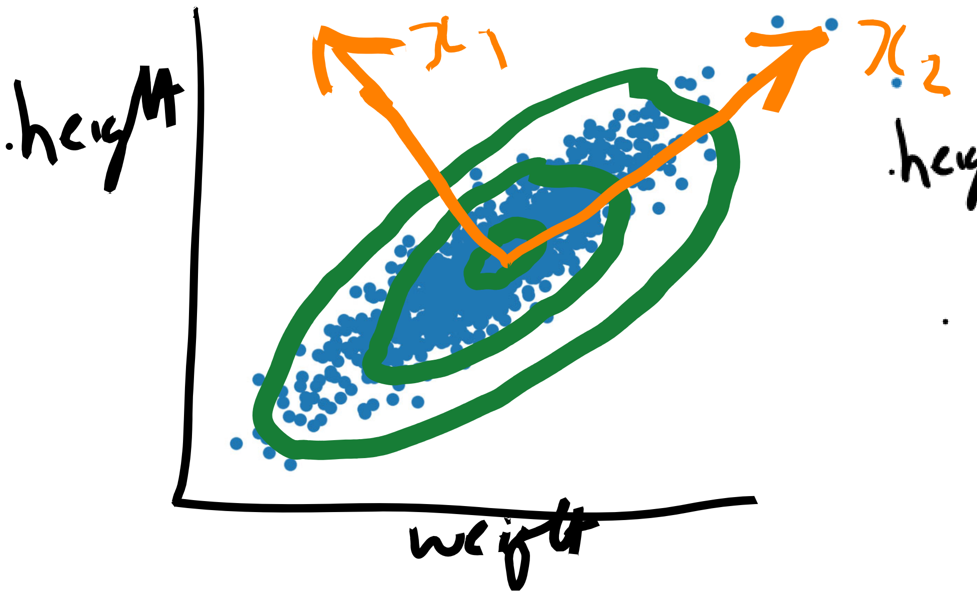


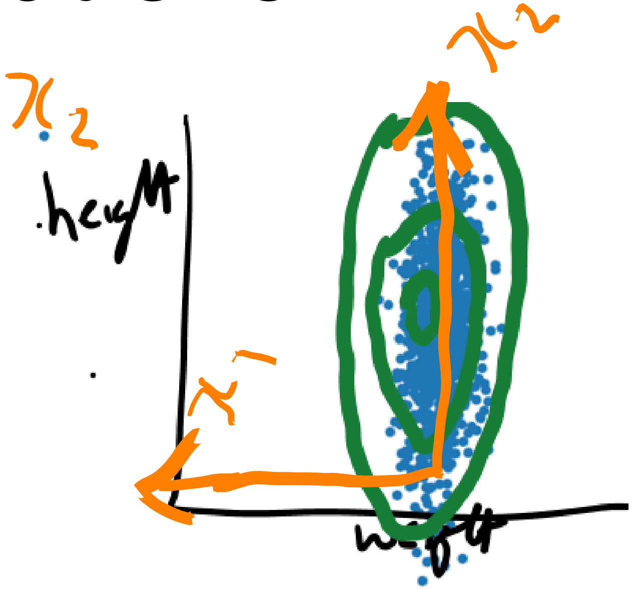
If independent RVs, $p([x_h, x_w]) = N(\mu_h, \sigma_h^2)^* N(\mu_w, \sigma_w^2)$.

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

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:

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

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

MVGs: Finding the right rotation 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 $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x - \mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y - \mu_2)^2\right]$

MVGs: Finding the right rotation 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 $p([x, y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$

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

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

MVGs: Finding the right rotation 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 $p([x,y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right]$

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

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

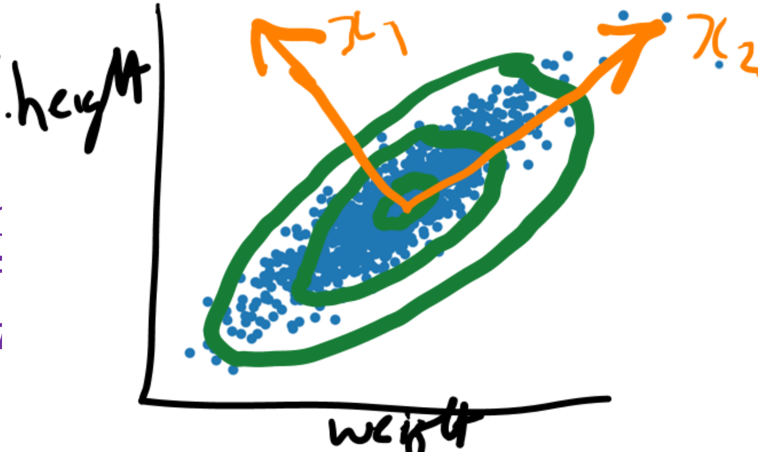
$$[x,y]\Sigma^{-1}\begin{bmatrix}x\\ y\end{bmatrix}$$

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

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

MVGs: Finding the right rotation 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 $p([x,y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2\pi}}$

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

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

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

$E[X]$ for discrete $\sum x P(x)$

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

often denoted by μ for mean.

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

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

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

- $E(X+c) = E(X) + c$ 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:

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

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

Review of expectations, variance, covariance

COVARIANCE of two R.V.

$$\mathrm{Cov}(X,Y) = E\big(\underbrace{(X-E(X))}_{\text{R.V with mean 0}}\underbrace{(Y-E(Y))}\big) = E\big((X-\mu_x)(Y-\mu_y)\big)$$

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

$\Rightarrow \mathrm{cov}(X,X) = \mathrm{Var}(X)$. independent $X,Y \Rightarrow \mathrm{cov}(X,Y)=0$

CORRELATION

$$\mathrm{CORR}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} \rightarrow \rho \in [-1,1]$$

$\rho \rightarrow r$

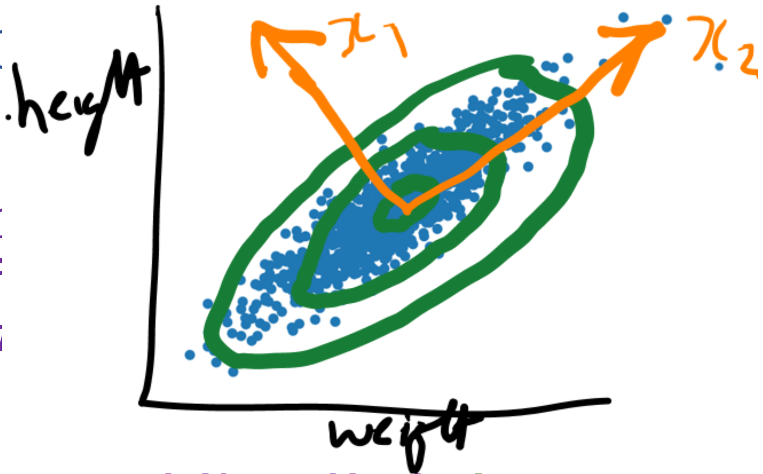
Back to this example

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

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

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



lets 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!

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \overset{\sigma_1^2}{\mathrm{cov}(X_1, X_1)} & \mathrm{cov}(X_1, X_2) & \mathrm{cov}(X_1, X_3) \\ \mathrm{cov}(X_2, X_1) & \mathrm{cov}(X_2, X_2) & \mathrm{cov}(X_2, X_3) \\ \mathrm{cov}(X_3, X_1) & \mathrm{cov}(X_2, X_3) & \underset{\sigma_3^2}{\mathrm{cov}(X_3, X_3)} \end{bmatrix}$$

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?

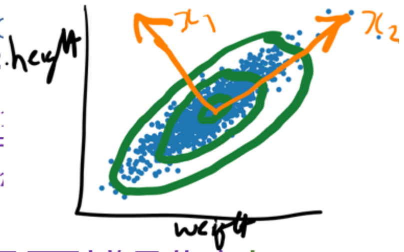
Lets work out the baby case, variables are indepen

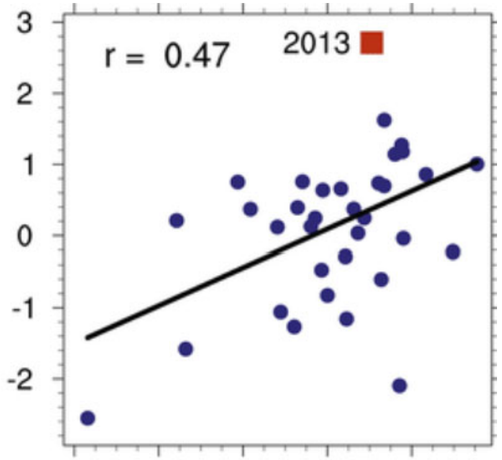
- $X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2$
- Then $p([x,y]) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right] \frac{1}{\sqrt{2}}$

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

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

lets



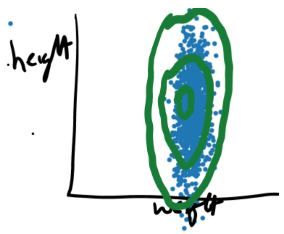


From the baby case to the general case

We worked the baby case, variables are independent, and each is 1D:

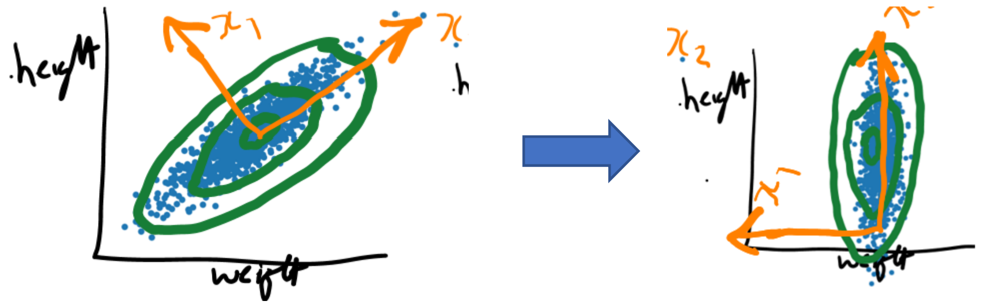
$X \sim p(x) = N(\mu_1, \sigma_1^2)$ and $Y \sim p(y) = N(\mu_2, \sigma_2^2)$, so that $p([x,y]) = p(x)p(y)$

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



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

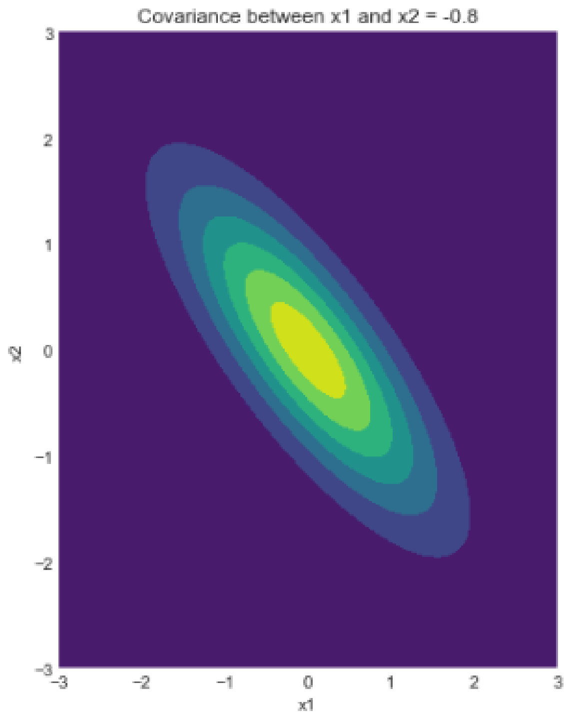
MVG has 2 main terms:

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



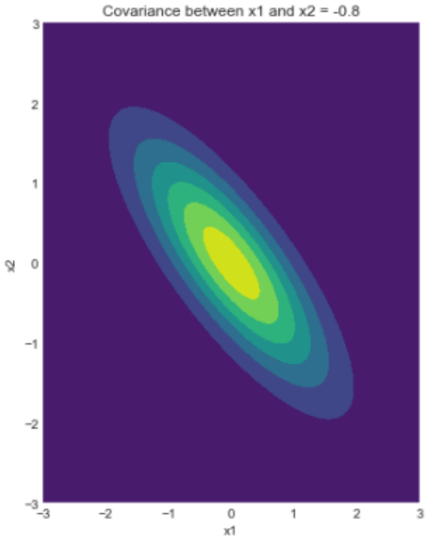
One level set (contour line) of MVG pdf comprises the values $\boldsymbol{x}$ for which $\boldsymbol{p(x)}$ is a constant:

$$\underset{1\times d}{x^T}\ \underset{d\times d}{\Sigma^{-1}}\ \underset{d\times 1}{x} = \text{constant}$$

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

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

What do these look like in high dimensions?



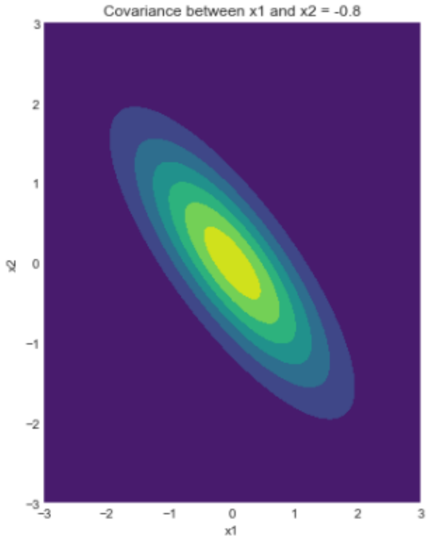
To deal with a 14-dimensional space, visualize a 3D space

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

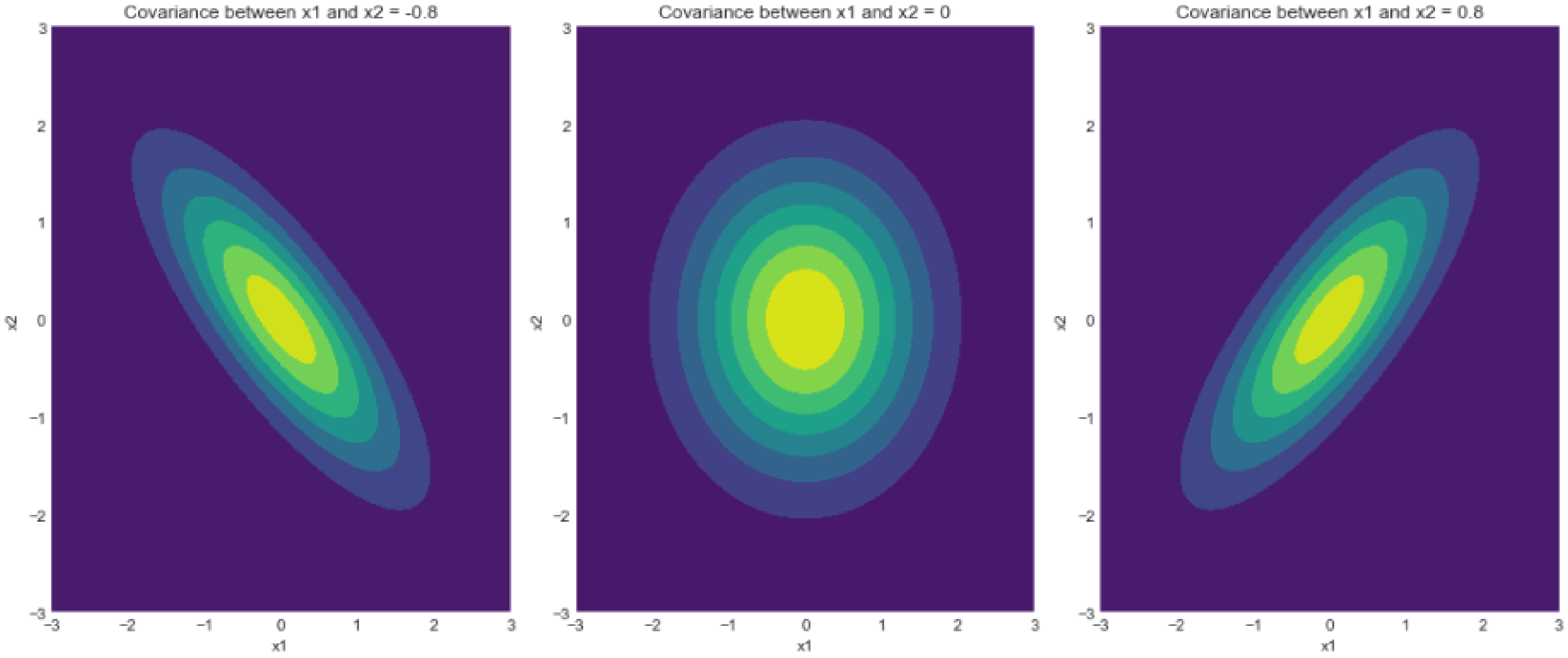
What do these look like in high dimensions?

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.



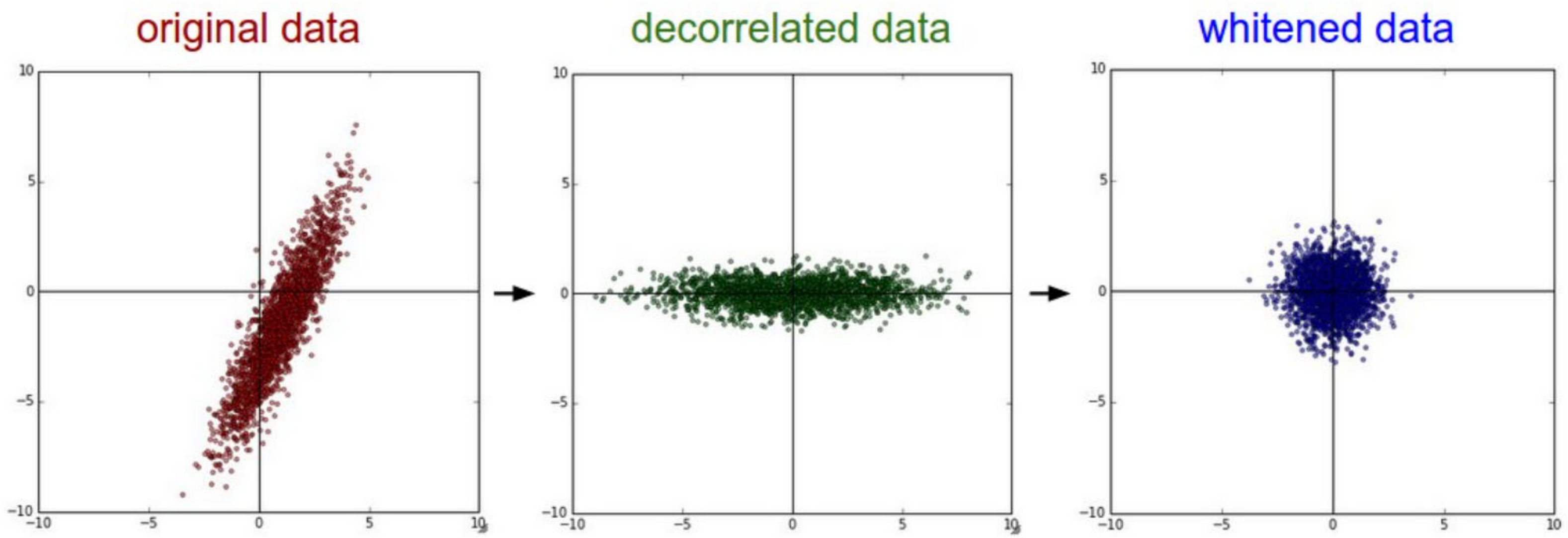
could look at each pair of variables x_i and x_j

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

Sphering a MVG

"diagonalized"

- To "sphere" a MVG is to alter 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 $\boldsymbol{cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{cov}(\boldsymbol{y}, \boldsymbol{x})$.
- Recall a symmetric matrix $\boldsymbol{C} \in \mathbb{R}^{d \times d}$ is PSD iff $\boldsymbol{u}^T \boldsymbol{C} \boldsymbol{u} \geq \boldsymbol{0}$ for every $\boldsymbol{u} \in \mathbb{R}^d$. (PD if strictly $> \boldsymbol{0}$). It follows that all eigenvalues are $\geq \boldsymbol{0}$.
- Recall eigenvalues:

$A\underset{\sim}{x} = \lambda \underset{\sim}{x} \longrightarrow \underset{\sim}{x}$ is an eigenvector

λ is an eigenvalue

eigenvalues are found by solving -

$\det(A - \lambda I) = 0$

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

nal

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

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

Φ is an orthonormal matrix $[q_1 \; q_2 \cdots q_n]$

each is of length 1

any two are orthogonal

true for col and ro

$\Downarrow$

$\Phi^{-1} = \Phi^T$

(rotations & reflections)

$A = \Phi D \Phi^{-1}$

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

Linear Algebra: Diagonalizing a matrix

Inverses and square roots

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

Then $A^{-1} = \Phi D^{-1} \Phi^T$

where $D^{-1} = \begin{bmatrix} \lambda_1^{-1} & & & \\ & \lambda_2^{-1} & & \\ & & \ddots & \\ & & & \lambda_n^{-1} \end{bmatrix}$

why?

$\Phi D \cancel{\Phi^T} \cancel{\Phi} D^{-1} \Phi^T \stackrel{?}{=} I$ yes.

I

I

I

This is nice because in the gaussian we

$\Sigma^{-1} \triangleq$ "precision Matrix" or "concentration matrix"

- Define $R = \Phi \sqrt{D}\, \Phi^T$ (symmetric)

where $\sqrt{D} = \begin{bmatrix} \sqrt{\lambda_1} & & & 0 \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{\lambda_n} \end{bmatrix}$

then $A = R^T R = RR^T$

if R symmetric

why?

$R^T R = \underbrace{\Phi \sqrt{D} \cancel{\Phi^T}}_{R^T} \underbrace{\cancel{\Phi} \sqrt{D}\, \Phi^T}_{R} = \Phi \sqrt{D} \cdot \sqrt{D}\, \Phi^T = \Phi D \Phi^T$

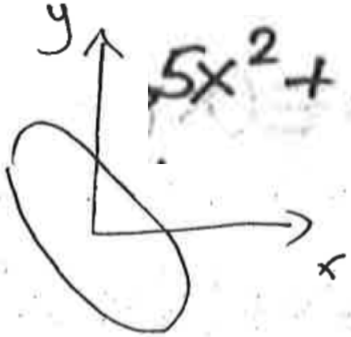
$= A.$

can be a useful factorization of A

Diagonalizing an ellipse

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

$$\boxed{A = \Phi D \Phi^T}$$



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

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

1. Eigenvalues of A –

$$Ax=\lambda x$$

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

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

2. Want to make this into an orthonormal Φ

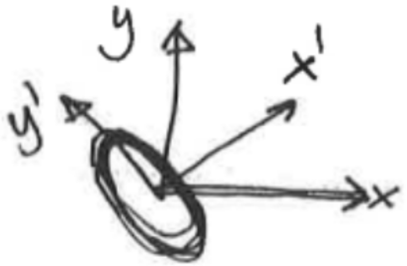
→ make eigenvectors normalized by dividing by $\sqrt{2}$

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

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

3. change coordinate system along the eigenvectors –

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

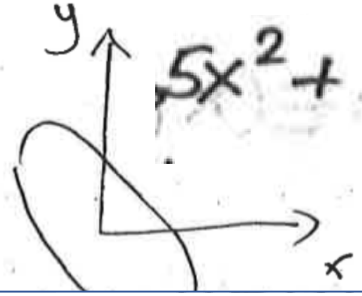


$\Longrightarrow$ in the new system: $9x'^2+1y'^2=1$

$\Longrightarrow$ axis aligned!

(no cross terms)

Sphering an ellipse



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

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

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

1. Eigenvalues

$Ax = \lambda x$

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

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

$\Rightarrow$ The axes point along the eigenvectors

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

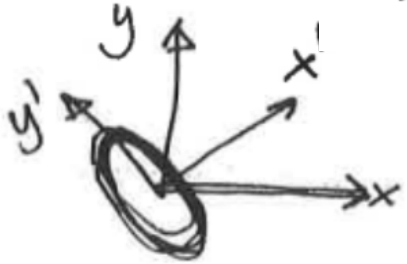
So finding a coordinate system that makes the ellipse axis aligned is the same as diagonilizing A.

φ

y dividing

3. change coordinate system along the eigenvectors —

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



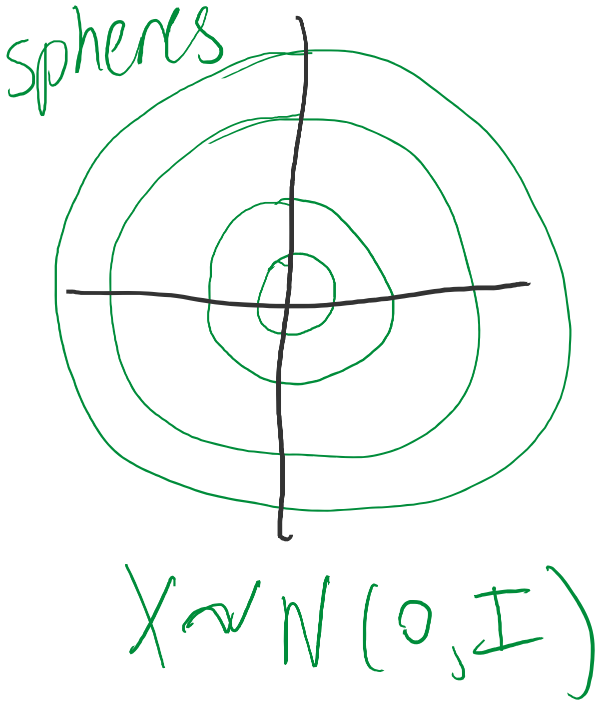
$\Rightarrow$ in the new system: $9x'^2 + 1y'^2 = 1$

$\Rightarrow$ axis aligned!

(no cross terms)

Geometric intuition: "de-sphering" a MVG

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

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

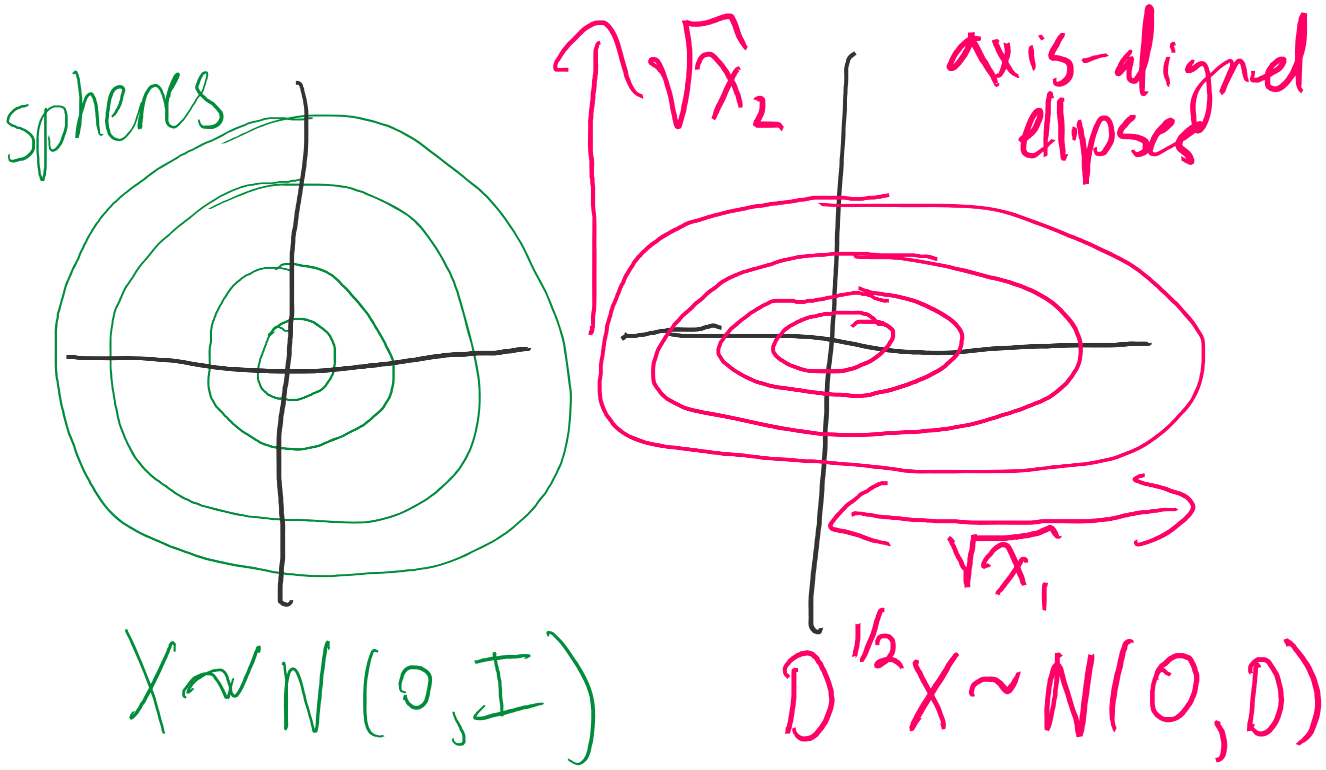
$Y = AX + \underset{\sim}{\mu} \sim N(\underset{\sim}{\mu}, AA^T)$

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

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

Geometric intuition: “de-sphering” a MVG

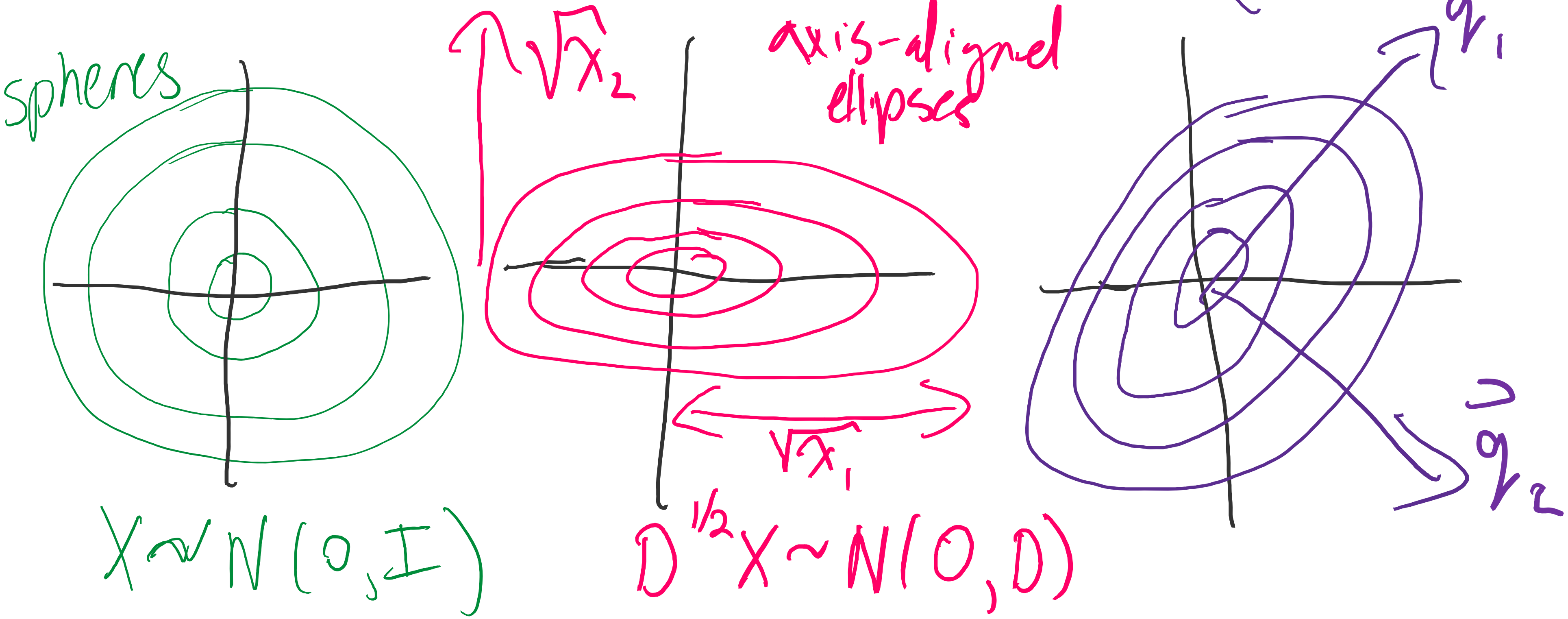
- Let $X \sim N(0, I)$.
- Let $\Sigma = QDQ^T$ be a covariance matrix factored into its eigenvectors and diagonal matrix. Can also write it as $\Sigma = (QD^{\frac{1}{2}})(D^{\frac{1}{2}}Q^T) = AA^T$.
- Let $Y = AX + \mu$. Then by affine property $Y \sim N(\mu, \Sigma)$



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 $Y \sim N(\mu, \Sigma)$

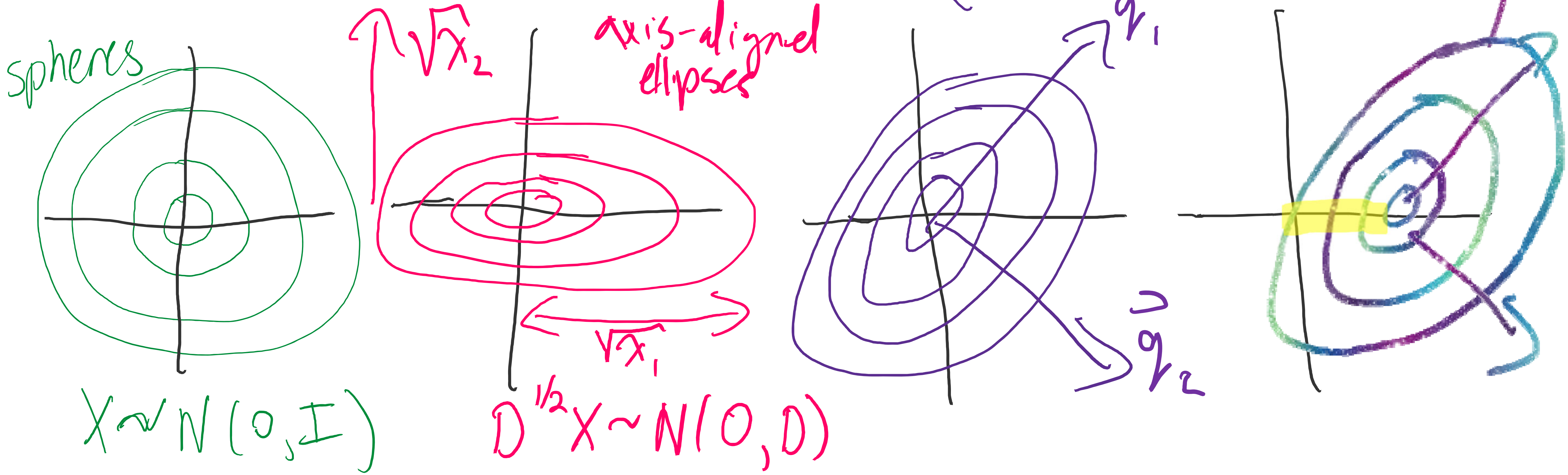


Geometric intuition

- 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 $Y \sim N(\mu, \Sigma)$

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

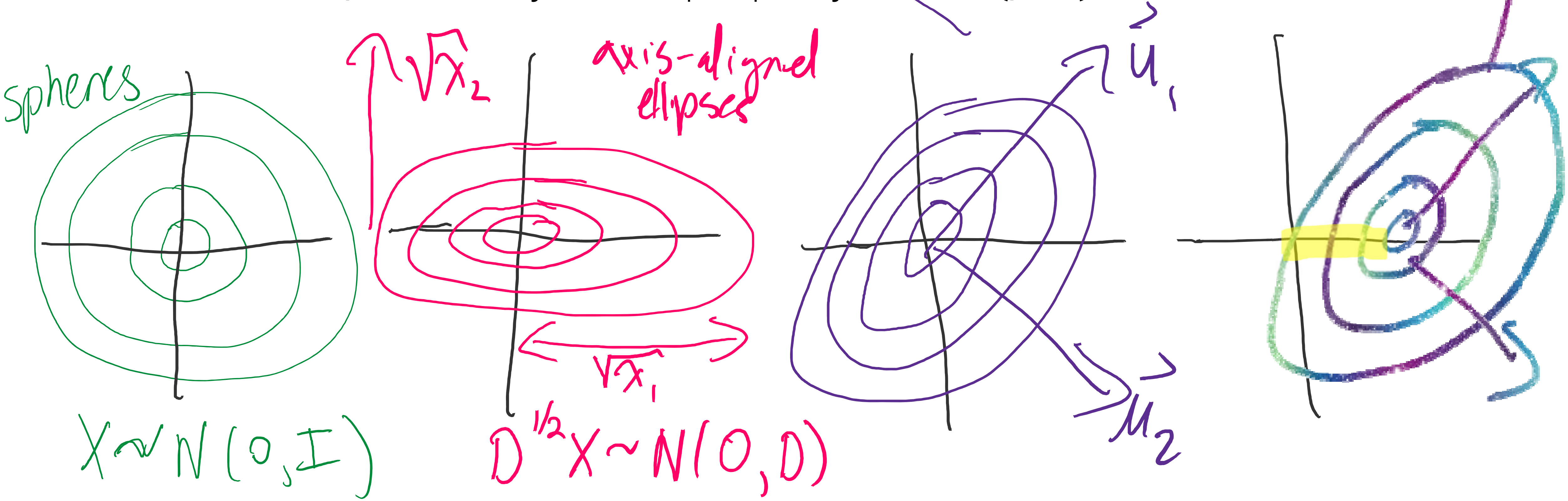
$AX + \mu \sim N(\mu, \Sigma)$



Geometric intuition

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

- Can decompose any MVG in terms of a "scaling", "rotation" and "shift" operator with respect to the standard $\boldsymbol{N(0,I)}$ form.
- Let $\boldsymbol{\Sigma = QDQ^T}$ be a covariance matrix factored into its eigenvectors and diagonal matrix. Can also write it as $\boldsymbol{\Sigma} = (QD^{\frac{1}{2}})(D^{\frac{1}{2}}Q^T) = AA^T$.
- Let $Y = AX + \mu$. Then by affine property $Y \sim N(\mu, \Sigma)$



Extra slides (not responsible for)

Real application: Genome-Wide Association Studies (GWAS)

Input:

- A set of people with/without a disease
- Measure a large set of genetic markers for each person (*e.g.*, SNPs: s*ingle* n*ucleotide* p*olymorphism*).

Desired output:

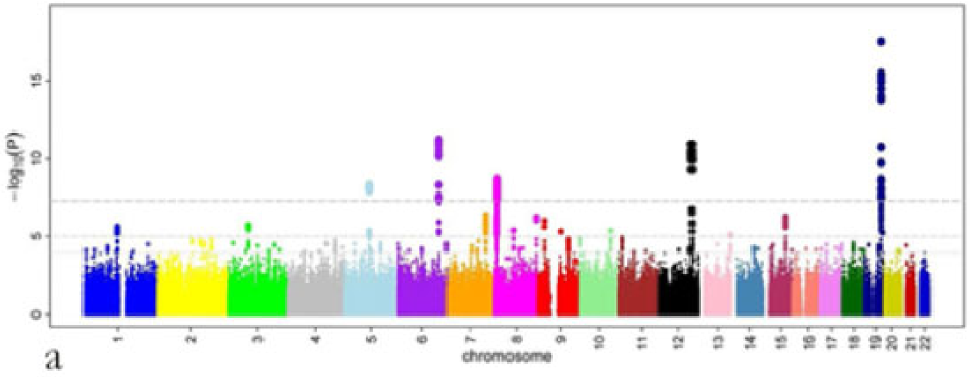
- A list of genetic markers underlying the disease.

+

ATGTTGAATCTG
AAAGTGAAATGT
TATTATACGAAG
AAGTATTTGCTA
GACCTCAAAACC
CTTCATCATAAC

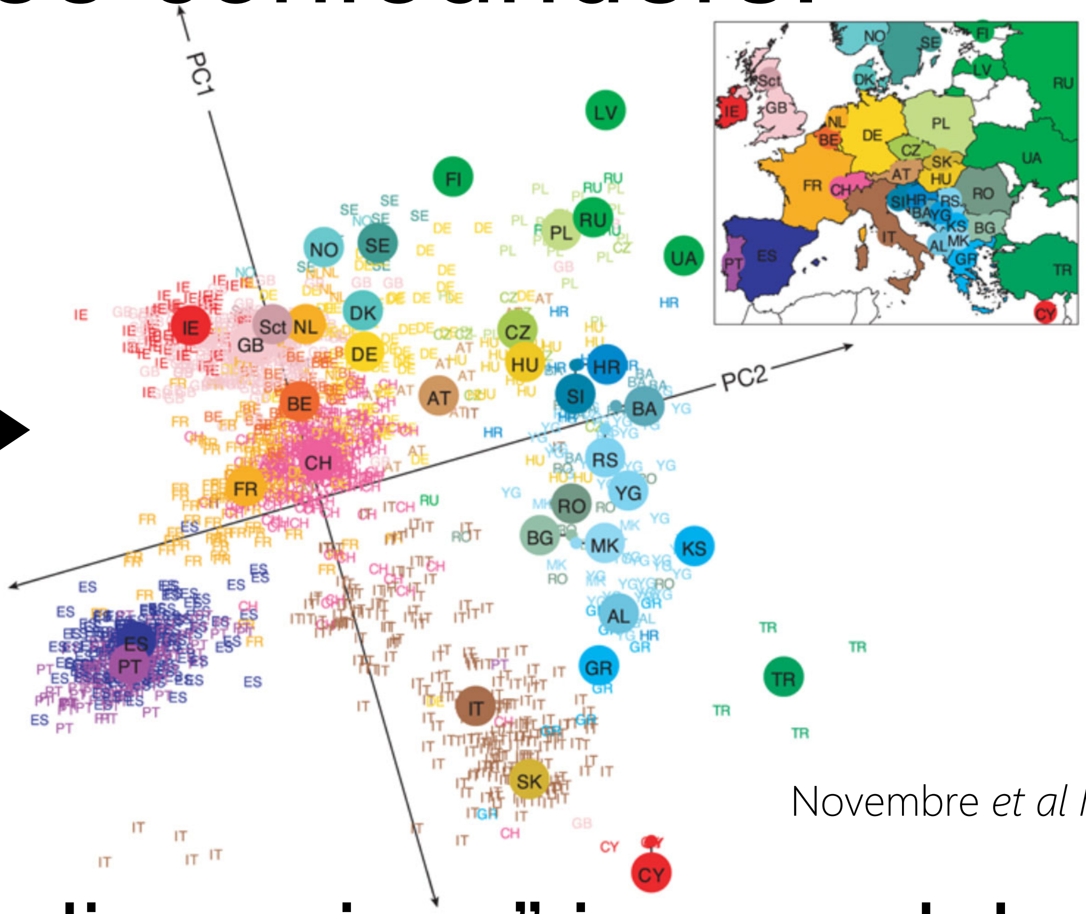
DNA from set of individuals

list of genetic markers



Spurious "signal" in the genetic markers owing to racial/pedigree confounders.

$N \times N$ covariance matrix



Novembre *et al Nature* 2008

By including these "hidden dimensions" in a model, we can correct the problem.

http://guardianlv.com/2014/06/dna-sequencing-on-the-spot-saved-life/

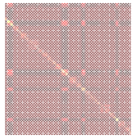
Nat. Genetics 2013, *Nat. Methods* 2012, 2011, 2014 *etc.*

Mixed Model Approach (GPR)

Generative model:

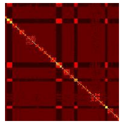
- sample a "latent ancestry vector"
- add as covariate

"amount of confounding genetic signal"

$p(\vec{u}) = N(\vec{0},$  $)$

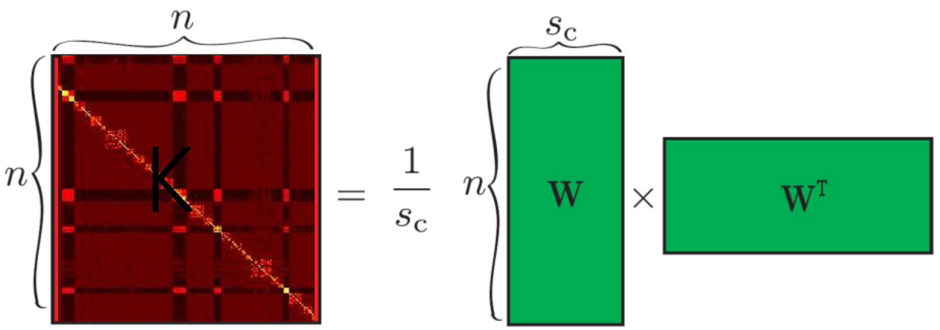
$$p(\vec{y}|\vec{u}) = N(X\beta + \eta\vec{u}, I\sigma^2)$$

LIKELIHOOD $p(\vec{y}) = \int p(\vec{y}|\vec{u})p(\vec{u})d\vec{u}$

$p(\vec{y}) = N(X\beta, \eta^2$  $+ I\sigma^2)$

2. Exploit rank when few SNPs are used

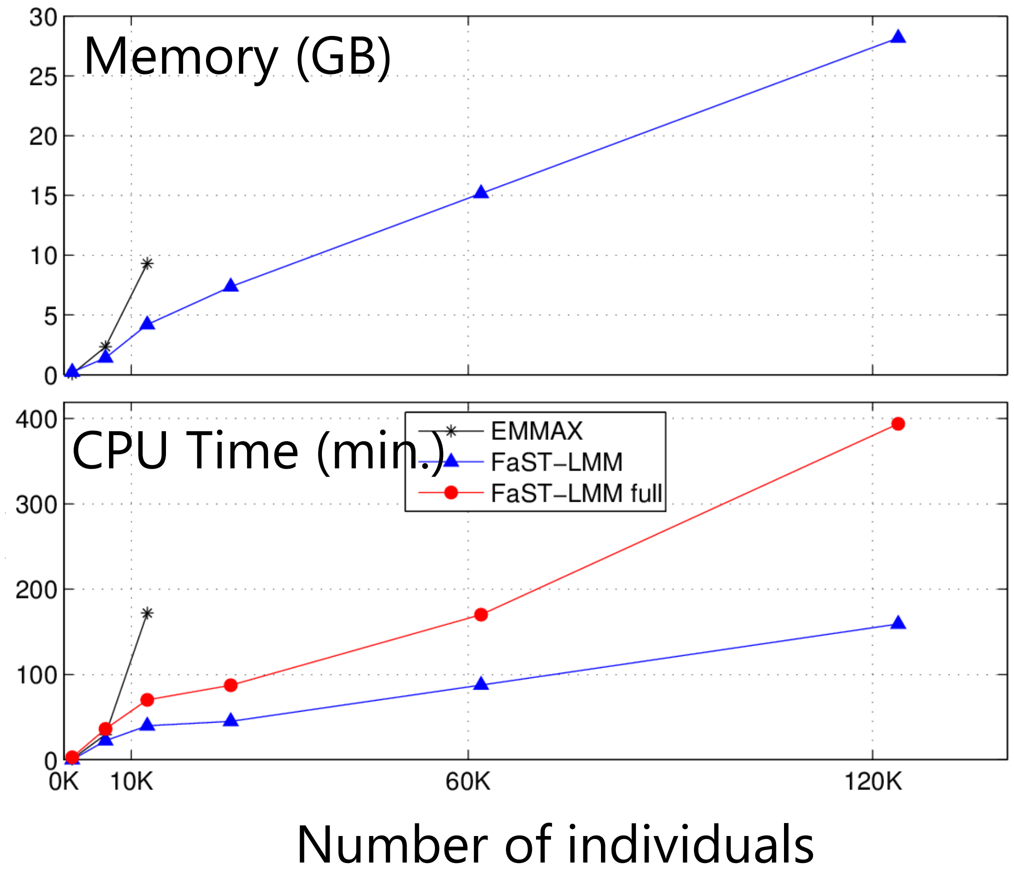
If $\mathbf{K}$ is computed in a particular way, and $n > s_c$ then computations are linear in n.



e.g., the realized relationship matrix (RRM) [Visscher et al]

n is # people
s_c is # of SNPs used to compute K

Experimental running time and memory



- **EMMAX** out of memory after 13K individuals
- **FaST-LMM** fixed variance components
- **FaST-LMM full** re-estimate variance components