Lecture 7
Principal Component Analysis (PCA)

CREIGS 2020

Lewis E Tomalin, PhD
Assistant Professor of Biostatistics

Icahn School of Medicine at Mount Sinai
Department of Population Health Science & Policy

Mount Sinai

Lecture Overview

1. **Part 1: Introduction to PCA**
    1. What is PCA used for?
    2. What is a principle component?
    3. How to interpret PCA results
    4. Mathematics underlying PCA
2. **Part 2: Performing PCA in R**
    1. Installing packages
    2. Formatting the data
    3. Running PCA
    4. Making plots

What is PCA used for?

When working with ‘high-throughput’ data such as DNA/RNA-seq, each sample can have measurements of 100’s or even 10,000’s of genes.

This high-number of ‘features/variables/dimension’ makes the data hard to interpret.

PCA is an un-supervised modelling technique, that decreases the number of dimensions in the data and thus helps us visualize characteristics of the data.

In RNA-seq we can use PCA to answer two important questions:

1. **Do samples with similar/different phenotypes have similar/different gene-expression profiles?**
    1. This is an important QC check, eg: do samples taken pre-treatment have similar expression profiles?
    2. Do post-treatment samples look different to pre-treatment?

2. **Which genes are most responsible for these similarities/differences?**
    1. PCA can provide a rough indication of which genes are different, however, there are ‘better’ methods for properly answering this question.

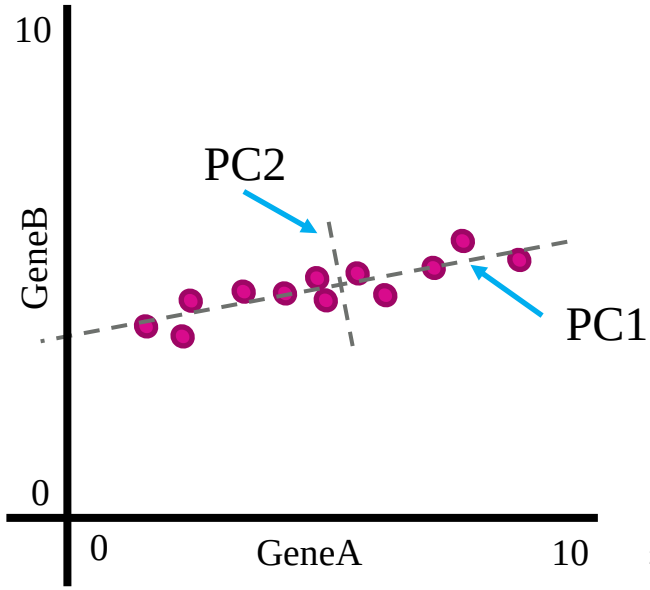
What is PCA?

Terminology

The First Principle Component (PC1): The First Principle Component is a line/plane in the data that explains most of the variation in that data. This plane will have fewer dimensions that the the original data.

The Second Principle Component (PC2): The First Principle Component is a line/plane in the data, perpendicular to PC1 that explains the 2nd most of the variation in that data.

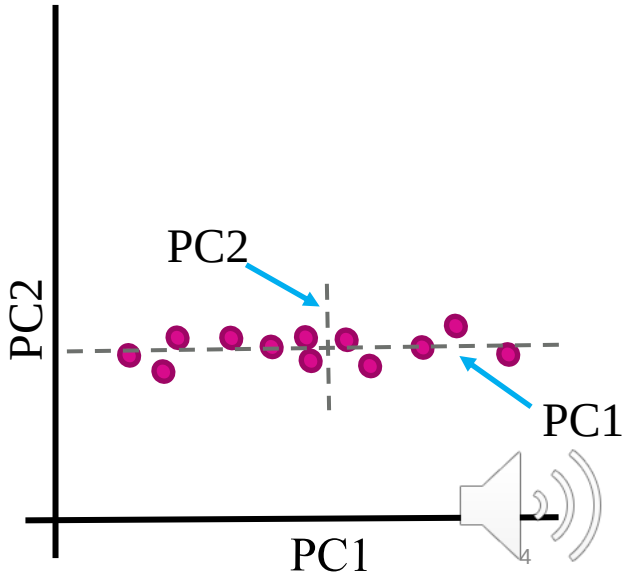
Dimension Reduction: PCA is sometimes referred to as a 'dimension reduction' technique, since it can summarize large dimensional data into smaller dimensions. Ie: summarize 1,000 genes/dimensions into just 2 components/dimensions.



Two genes, 2 dimensions

PC1 is a 1D line.

Can use each PC as an axis and plot data relative to each PC

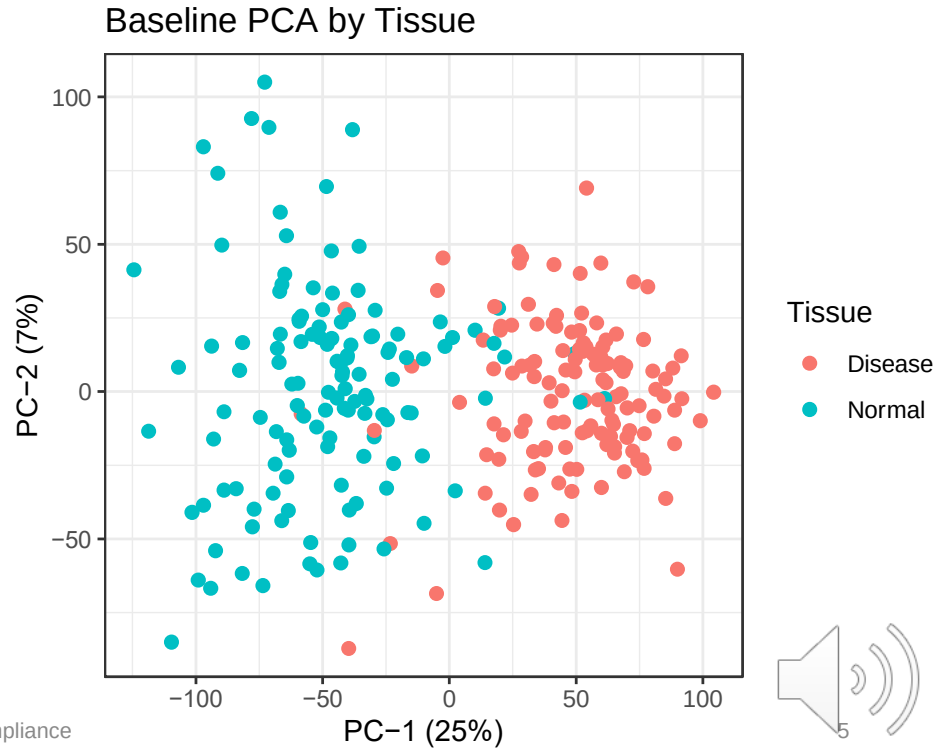


Interpreting a PCA plot

Example#1: Skin samples were taken from Psoriasis patients before treatment, samples of diseased skin and normal skin were taken, gene-expression profiles were measured and PCA was performed.

Interpretation

- Samples that are close together have similar gene-expression profiles.
- Disease skin expression profiles are different to Normal skin.
- PC1 by definition represents most of the variation.
- Since skin type varies across PC1 we can say that Skin type accounts for most of the variation in the data.



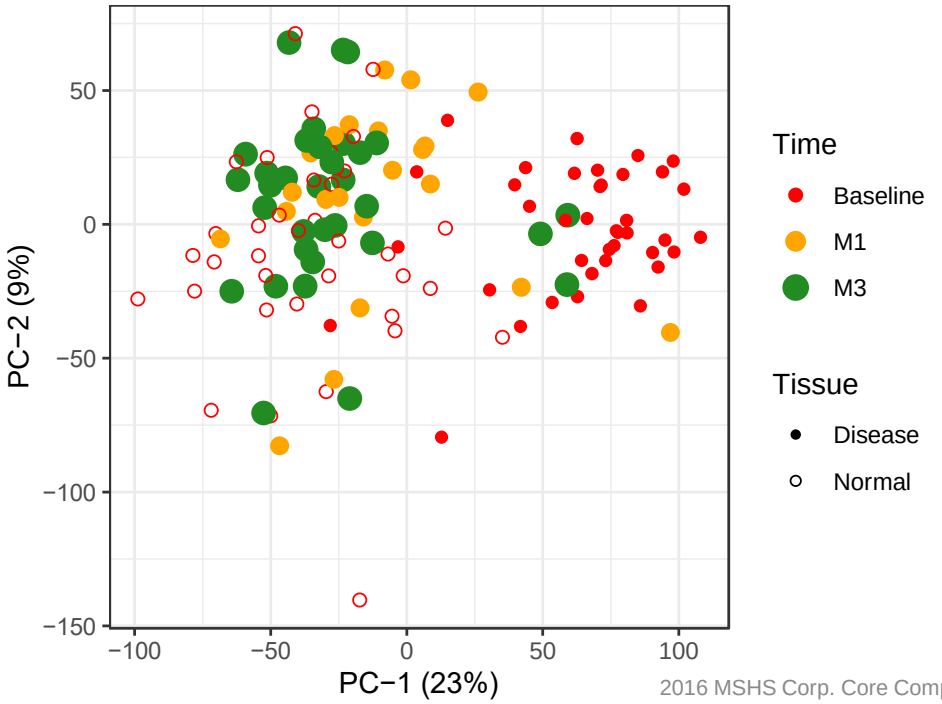
Interpreting a PCA plot

Example#2: Skin samples were taken from Psoriasis patients before and after treatment (1 month and 3 months). Samples of diseased skin and normal skin were taken, gene-expression profiles were measured and PCA was performed.

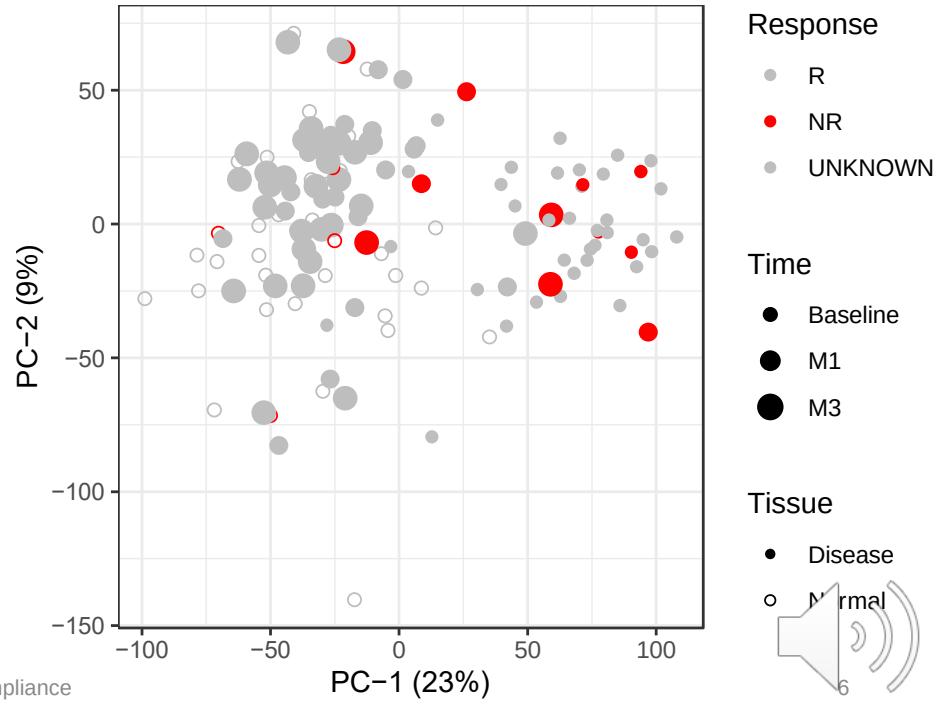
Interpretation

- Post treatment skin has similar profile to Normal skin, suggesting that treatment worked in these patients.
- Some samples still look diseased, perhaps these patients did not respond.

Treatment over time



Color by Response

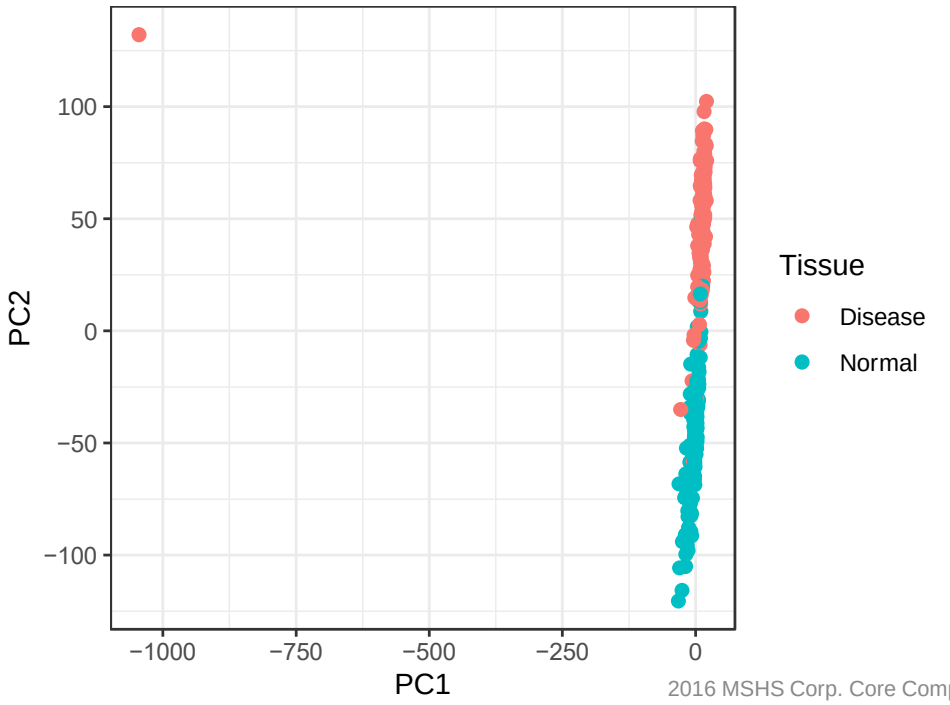


Interpreting a PCA plot

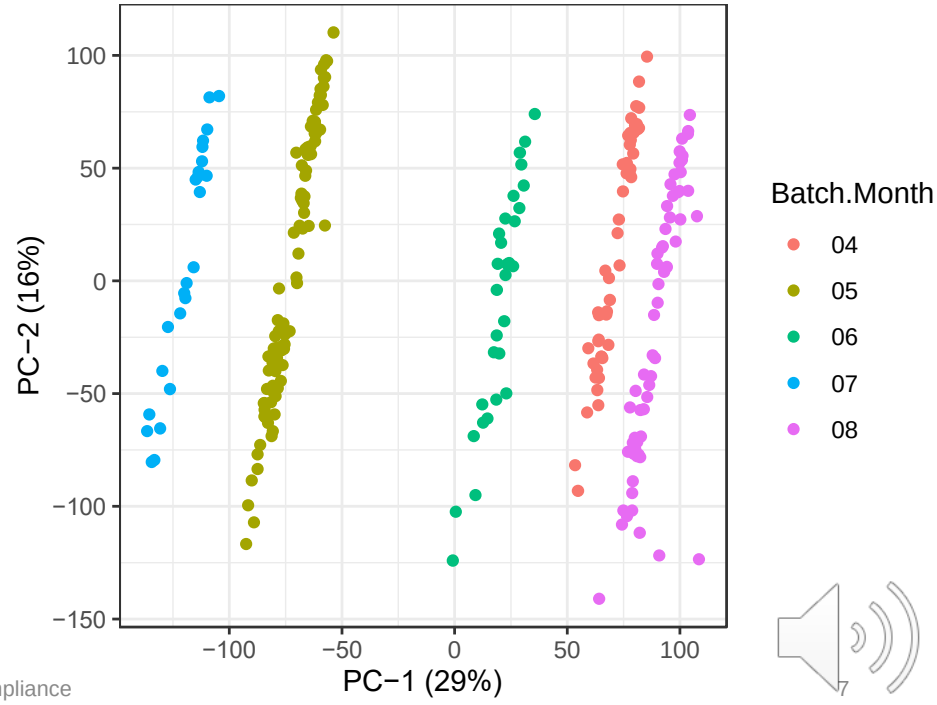
Bad Examples

- One sample is completely different to the rest, check this sample, probably just delete it.
- Samples analyzed on same date are grouped together, suggests a batch effect, consider batch adjustment.

Plot with outlier



Batch Effect



Mathematics of PCA (how are PCs calculated)

- **You do not need to fully understanding how to calculate a PC in order to use it in your research** (you don't need to know how an engine works to drive a car)

- However, understanding the mathematics will help you understand and understand PCA at a deeper level, and will also help you understand other/similar techniques.

- I will give a brief introduction here, but I recommend watching the Chapter 10 videos on the following site to get the details (https://www.r-bloggers.com/2014/09/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/). (goldilocks zone)

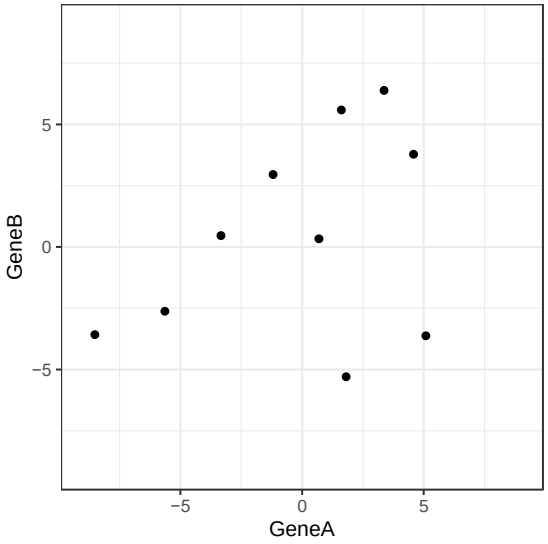
- PC1 (Z_1) is calculated using the formula: $\boldsymbol{Z_1 = \phi_1 X_1 + \phi_2 X_2 + \cdots + \phi_p X_p}$ **,** where X is the expression of each gene *p,* and values for each $\boldsymbol{\phi}$ are optimized to maximize the variation whilst constraining the sum as all $\boldsymbol{\phi^2}$ to be equal to 1.

- Thus genes that contribute most to the variation will have higher $\boldsymbol{\phi}$ values, which are often referred to a weights or loadings.

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose some weights

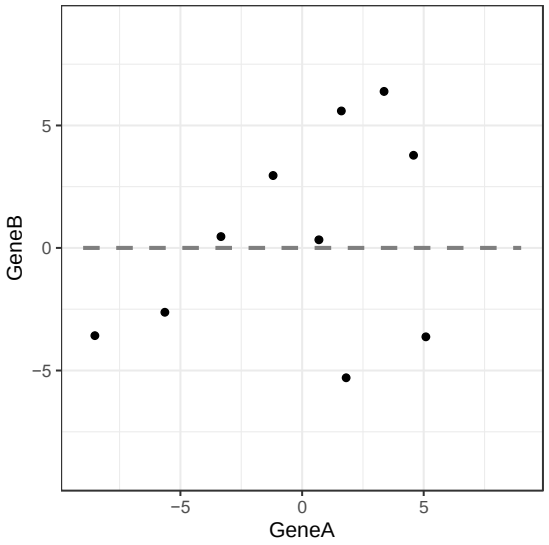
ϕ_A= 1, ϕ_B= 0:

	Wt_1	Wt_2	Wt_3
ϕ_A	1		
ϕ_B	0		
$(\phi_A^2+\phi_B^2)$	1		
Var			

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose some weights

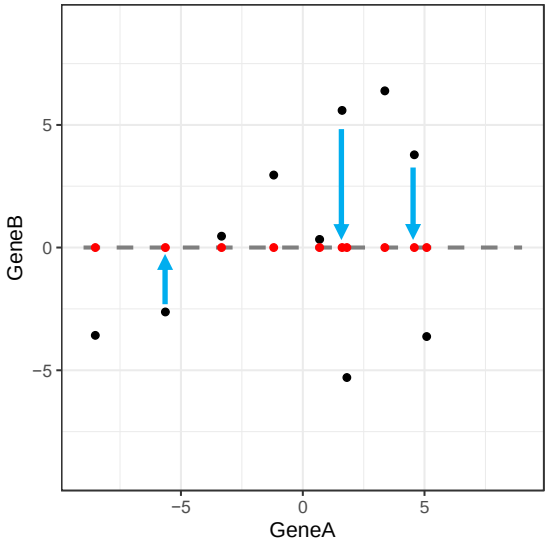
ϕ_A= 1, ϕ_B= 0:

	Wt_1	Wt_2	Wt_3
ϕ_A	1		
ϕ_B	0		
$(\phi_A^2+\phi_B^2)$	1		
Var			

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose some weights

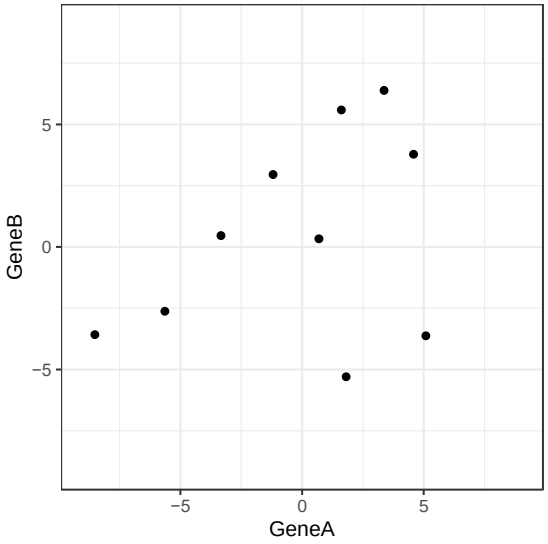
ϕ_A= 1, ϕ_B= 0: Just uses GeneA values for variance calculation.
Var=20.1

	Wt_1	Wt_2	Wt_3
ϕ_A	1		
ϕ_B	0		
$(\phi_A^2+\phi_B^2)$	1		
Var	20.1		

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose some more weights

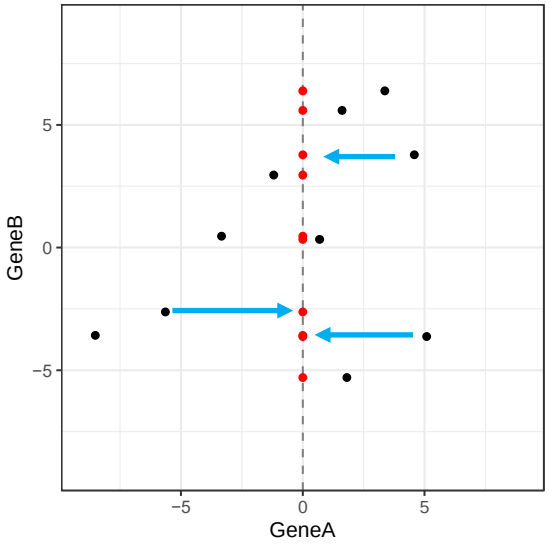
$\phi_A = 0$, $\phi_B = 1$:

	Wt_1	Wt_2	Wt_3
ϕ_A	1	0	
ϕ_B	0	1	
$(\phi_A^2+\phi_B^2)$	1	1	
Var	20.1		

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose some more weights

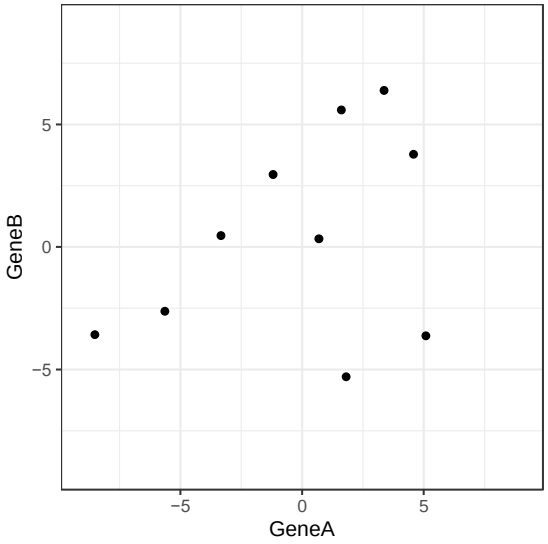
ϕ_A= 0, ϕ_B= 1: Essentially the variance of GeneB, var=17.2

	Wt_1	Wt_2	Wt_3
ϕ_A	1	0	
ϕ_B	0	1	
$(\phi_A^2+\phi_B^2)$	1	1	
Var	20.1	17.2	

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose a third set of weights

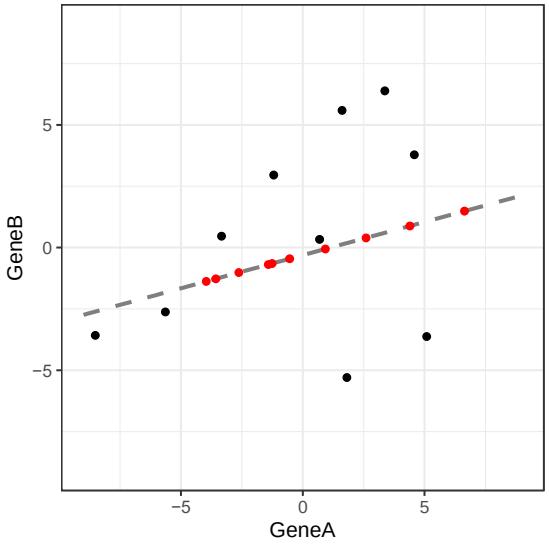
ϕ_A= -0.8, ϕ_B= -0.6:

	Wt_1	Wt_2	Wt_3
ϕ_A	1	0	-0.8
ϕ_B	0	1	-0.6
$(\phi_A^2+\phi_B^2)$	1	1	1
Var	20.1	17.2	

Mathematics of PCA (deeper dive)

- Let's use a toy example to really breakdown how the loadings for PC1 are estimated.
- Imagine we have **10 samples**, with measurements for 2 genes, **GeneA** and **GeneB**.

SampleID	GeneA	GeneB
S1	5.1	-3.6
S2	-5.6	-2.6
S3	-3.3	0.5
S4	3.4	6.4
S5	1.6	5.6
S6	-8.5	-3.6
S7	1.8	-5.3
S8	4.6	3.8
S9	-1.2	3.0
S10	0.7	0.3
Var	20.1	17.2



$$Z_1 = \phi_A X_A + \phi_B X_B$$

Let's choose a third set of weights

ϕ_A= -0.8, ϕ_B= -0.6

Use these weights calculate new data points for each sample.

The variance of these new points is **25.6.** Higher than the other weights.

	Wt_1	Wt_2	Wt_3
ϕ_A	1	0	-0.8
ϕ_B	0	1	-0.6
$(\phi_A^2+\phi_B^2)$	1	1	1
Var	20.1	17.2	

SampleID	(GA*-0.8)	(GB*-0.6)	SUM
S1	-4.0	2.3	-1.7
S2	4.4	1.6	6.0
S3	2.6	-0.3	2.3
S4	-2.6	-4.0	-6.6
S5	-1.3	-3.5	-4.8
S6	6.6	2.2	8.9
S7	-1.4	3.3	1.9
S8	-3.6	-2.4	-5.9
S9	0.9	-1.9	-0.9
S10	-0.5	-0.2	-0.7
Var			25.6

Thank You