

Introduction to Principal Component Analysis and Dimensionality Reduction

Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

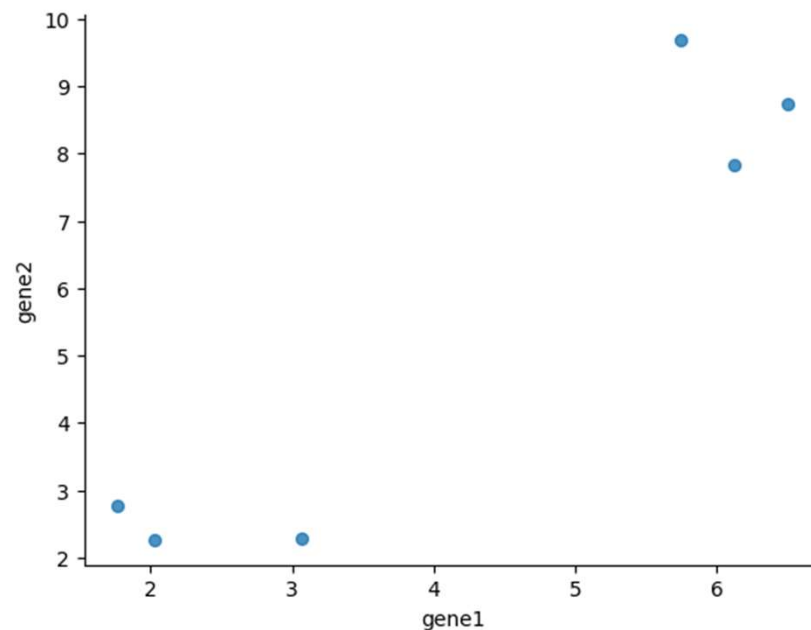
	gene1	gene2
mouse1	1.768969	2.762043
mouse2	2.027924	2.265112
mouse3	3.070331	2.296477
mouse4	5.749098	9.692229
mouse5	6.503984	8.731366
mouse6	6.118130	7.838842

This is a synthetic dataset representing **gene expression** levels for two genes—**gene1** and **gene2**—across six samples labeled **mouse1** to **mouse6**.

Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

	gene1	gene2
mouse1	1.768969	2.762043
mouse2	2.027924	2.265112
mouse3	3.070331	2.296477
mouse4	5.749098	9.692229
mouse5	6.503984	8.731366
mouse6	6.118130	7.838842



Why we need PCA and Dimensionality Reduction

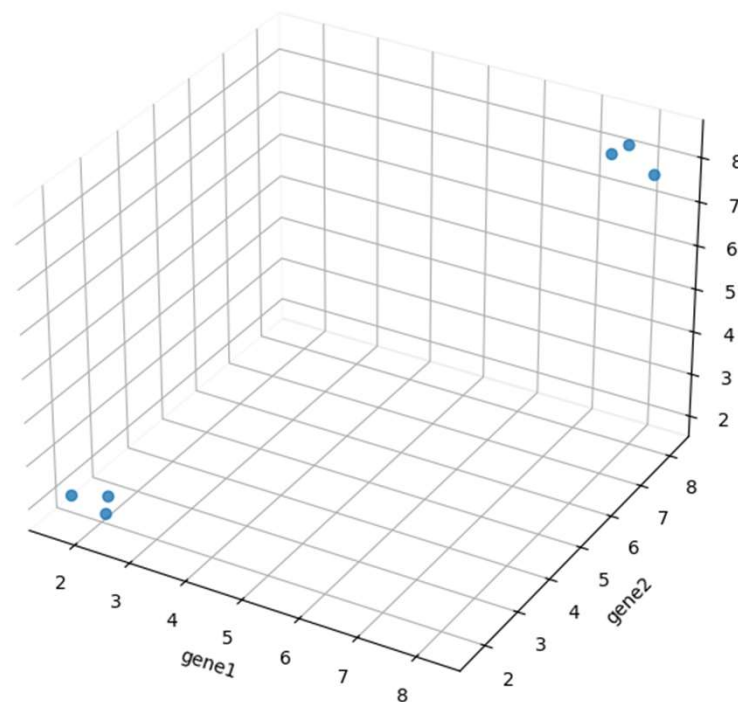
The Curse of Dimensionality

	gene1	gene2	gene3
mouse1	1.802367	1.912414	1.761788
mouse2	1.501424	2.061744	2.173459
mouse3	2.269733	2.007214	2.107653
mouse4	7.567415	8.466957	8.168858
mouse5	8.160152	7.836521	8.332693
mouse6	8.035728	7.441409	7.680521

Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

	gene1	gene2	gene3
mouse1	1.802367	1.912414	1.761788
mouse2	1.501424	2.061744	2.173459
mouse3	2.269733	2.007214	2.107653
mouse4	7.567415	8.466957	8.168858
mouse5	8.160152	7.836521	8.332693
mouse6	8.035728	7.441409	7.680521



Why we need PCA and Dimensionality Reduction

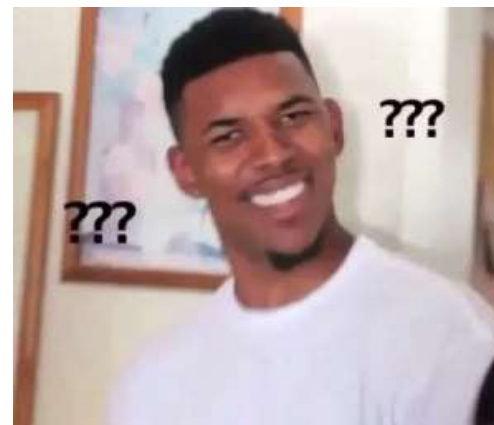
The Curse of Dimensionality

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
mouse1	2.039921	2.423343	1.670159	1.843498	2.013310	2.458927	1.797103	1.761939	1.621228	1.737895
mouse2	2.678077	1.827688	2.497142	2.178316	2.063992	2.114370	2.467147	2.312575	2.567257	1.587723
mouse3	1.586499	2.037714	1.584581	2.023876	1.954108	1.889529	1.918150	1.679878	2.092005	1.555760
mouse4	8.368589	8.027876	8.440553	8.180589	8.187888	8.451959	8.094824	8.039292	8.065158	8.040934
mouse5	8.199671	7.430767	8.119223	7.728076	7.794994	7.575481	8.348658	7.711904	8.584504	7.819865
mouse6	8.231404	8.339778	8.030988	8.315850	8.385829	8.317694	8.044136	7.855494	7.824927	7.628482

Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

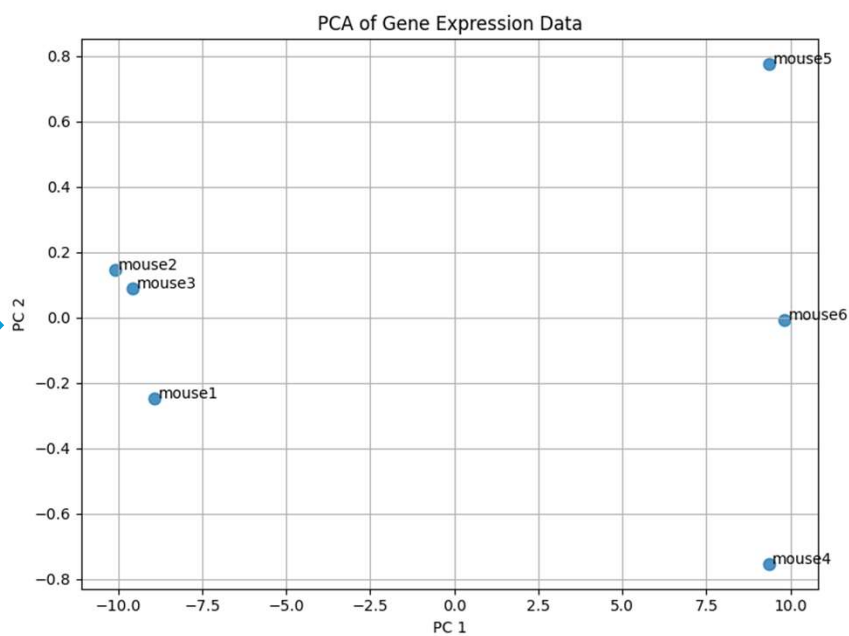
	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
mouse1	2.039921	2.423343	1.670159	1.843498	2.013310	2.458927	1.797103	1.761939	1.621228	1.737895
mouse2	2.678077	1.827688	2.497142	2.178316	2.063992	2.114370	2.467147	2.312575	2.567257	1.587723
mouse3	1.586499	2.037714	1.584581	2.023876	1.954108	1.889529	1.918150	1.679878	2.092005	1.555760
mouse4	8.368589	8.027876	8.440553	8.180589	8.187888	8.451959	8.094824	8.039292	8.065158	8.040934
mouse5	8.199671	7.430767	8.119223	7.728076	7.794994	7.575481	8.348658	7.711904	8.584504	7.819865
mouse6	8.231404	8.339778	8.030988	8.315850	8.385829	8.317694	8.044136	7.855494	7.824927	7.628482



Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
mouse1	2.039921	2.423343	1.670159	1.843498	2.013310	2.458927	1.797103	1.761939	1.621228	1.737895
mouse2	2.678077	1.827688	2.497142	2.178316	2.063992	2.114370	2.467147	2.312575	2.567257	1.587723
mouse3	1.586499	2.037714	1.584581	2.023876	1.954108	1.889529	1.918150	1.679878	2.092005	1.555760
mouse4	8.368589	8.027876	8.440553	8.180589	8.187888	8.451959	8.094824	8.039292	8.065158	8.040934
mouse5	8.199671	7.430767	8.119223	7.728076	7.794994	7.575481	8.348658	7.711904	8.584504	7.819865
mouse6	8.231404	8.339778	8.030988	8.315850	8.385829	8.317694	8.044136	7.855494	7.824927	7.628482



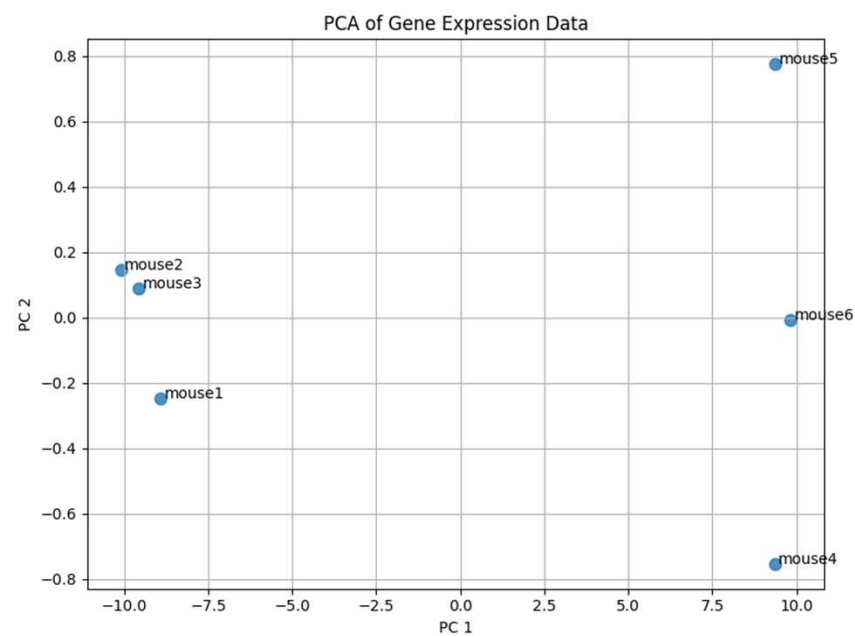
Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
mouse1	2.039921	2.423343	1.670159	1.843498	2.013310	2.458927	1.797103	1.761939	1.621228	1.737895
mouse2	2.678077	1.827688	2.497142	2.178316	2.063992	2.114370	2.467147	2.312575	2.567257	1.587723
mouse3	1.586499	2.037714	1.584581	2.023876	1.954108	1.889529	1.918150	1.679878	2.092005	1.555760
mouse4	8.368589	8.027876	8.440553	8.180589	8.187888	8.451959	8.094824	8.039292	8.065158	8.040934
mouse5	8.199671	7.430767	8.119223	7.728076	7.794994	7.575481	8.348658	7.711904	8.584504	7.819865
mouse6	8.231404	8.339778	8.030988	8.315850	8.385829	8.317694	8.044136	7.855494	7.824927	7.628482

PCA will

- Help us plot multidimensional data on 2D



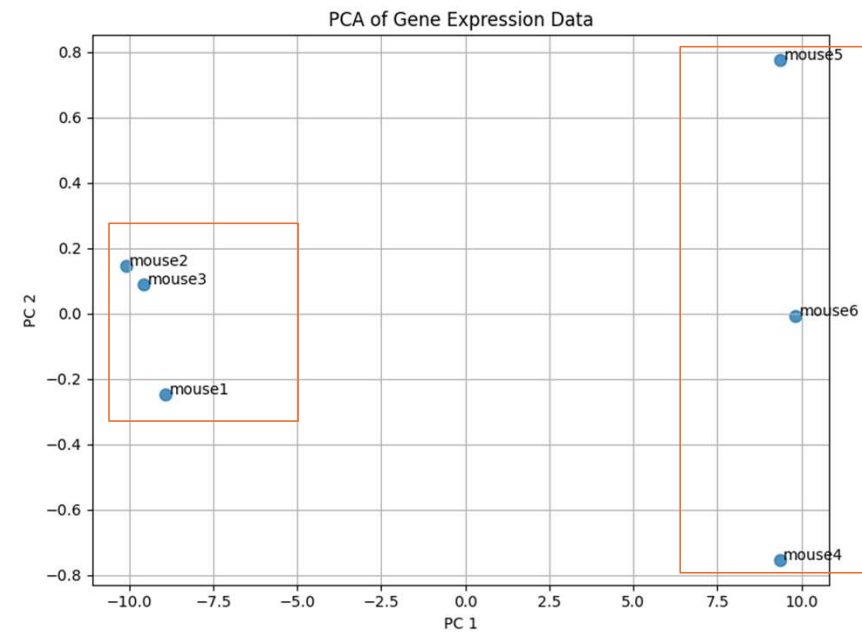
Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
mouse1	2.039921	2.423343	1.670159	1.843498	2.013310	2.458927	1.797103	1.761939	1.621228	1.737895
mouse2	2.678077	1.827688	2.497142	2.178316	2.063992	2.114370	2.467147	2.312575	2.567257	1.587723
mouse3	1.586499	2.037714	1.584581	2.023876	1.954108	1.889529	1.918150	1.679878	2.092005	1.555760
mouse4	8.368589	8.027876	8.440553	8.180589	8.187888	8.451959	8.094824	8.039292	8.065158	8.040934
mouse5	8.199671	7.430767	8.119223	7.728076	7.794994	7.575481	8.348658	7.711904	8.584504	7.819865
mouse6	8.231404	8.339778	8.030988	8.315850	8.385829	8.317694	8.044136	7.855494	7.824927	7.628482

PCA will

- Help us plot multidimensional data on 2D
- The similar samples will be grouped (clustered) together



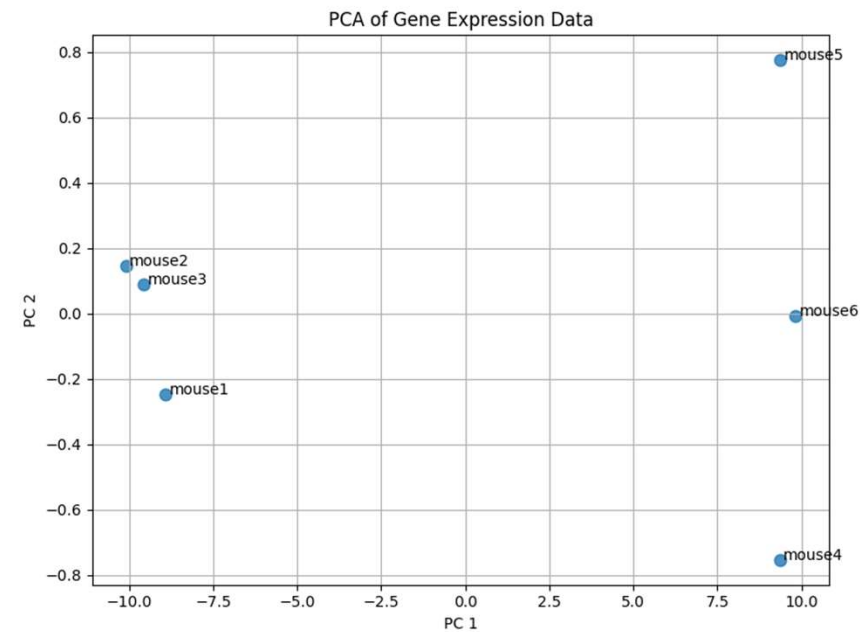
Why we need PCA and Dimensionality Reduction

The Curse of Dimensionality

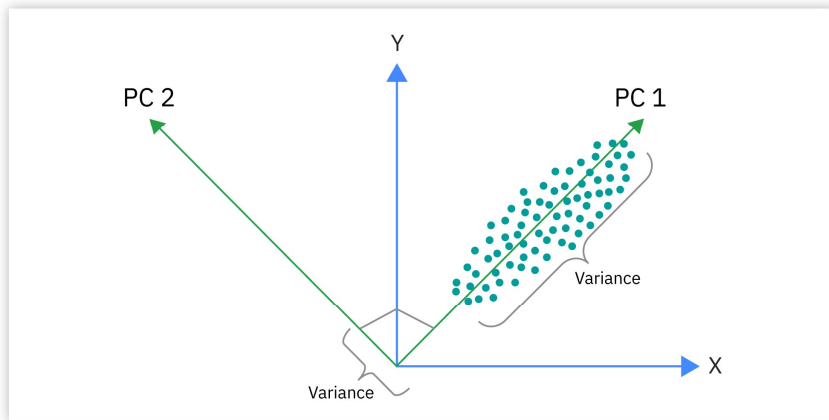
	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10
mouse1	2.039921	2.423343	1.670159	1.843498	2.013310	2.458927	1.797103	1.761939	1.621228	1.737895
mouse2	2.678077	1.827688	2.497142	2.178316	2.063992	2.114370	2.467147	2.312575	2.567257	1.587723
mouse3	1.586499	2.037714	1.584581	2.023876	1.954108	1.889529	1.918150	1.679878	2.092005	1.555760
mouse4	8.368589	8.027876	8.440553	8.180589	8.187888	8.451959	8.094824	8.039292	8.065158	8.040934
mouse5	8.199671	7.430767	8.119223	7.728076	7.794994	7.575481	8.348658	7.711904	8.584504	7.819865
mouse6	8.231404	8.339778	8.030988	8.315850	8.385829	8.317694	8.044136	7.855494	7.824927	7.628482








PCA will

- Help us plot multidimensional data on 2D
- The similar samples will be grouped (clustered) together
- Tell us about the most important variables(genes)

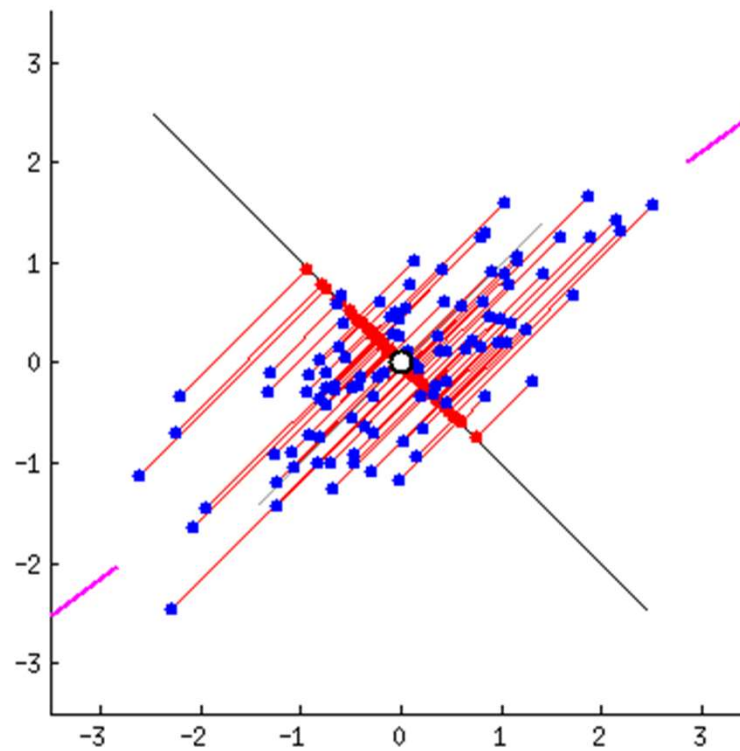
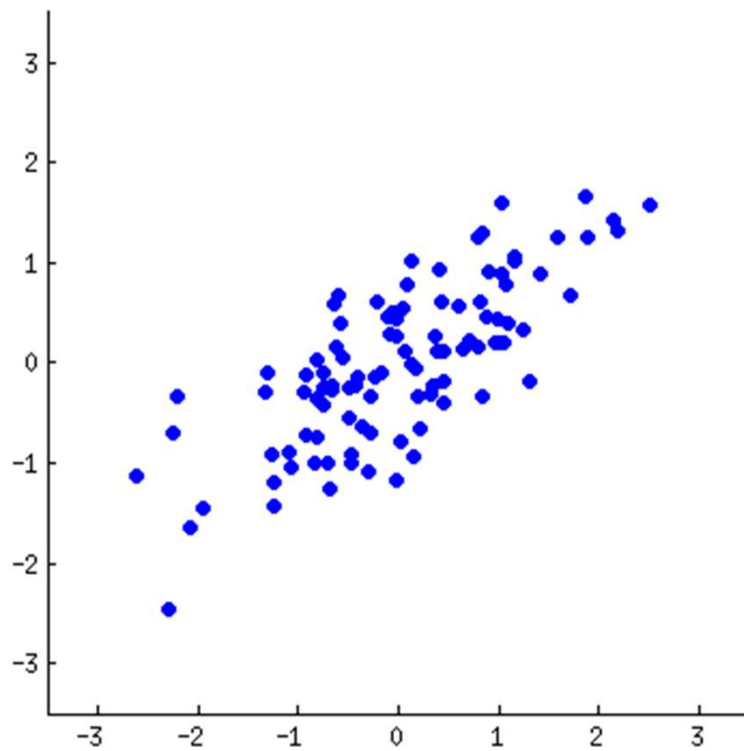


How PCA works



-  **Standardize the features**
 - Make sure each feature has the same scale (mean = 0, similar variance).
-  **Understand the variance of the data**
 - Understand how features vary together.
-  **Find the principal directions**
 - These are the directions where the data spreads out the most.
-  **Rank components by importance**
 - The first few components capture the most variation in the data.
-  **Select top components**
 - Choose 2 or 3 components to reduce dimensions.
-  **Transform the data**
 - Project the original data onto the new, lower-dimensional space.
-  **Use the result**
 - Visualize clusters, simplify analysis, or feed into models.

How PCA works



How PCA works

Each principal component (PC) captures a percentage of the total variation in the data:

- **PC1 = 40%** → This captures the largest pattern in the data.
- **PC2 = 23%** → The second-most important direction, independent of PC1.
- **PC3 = 15%** → Adds more detail, but less than PC1 or PC2.
- ...

We usually **choose the top components** that together explain a **large portion of the total variance** (e.g., 80–90%).

This helps us **simplify the data** while still keeping the most important information.