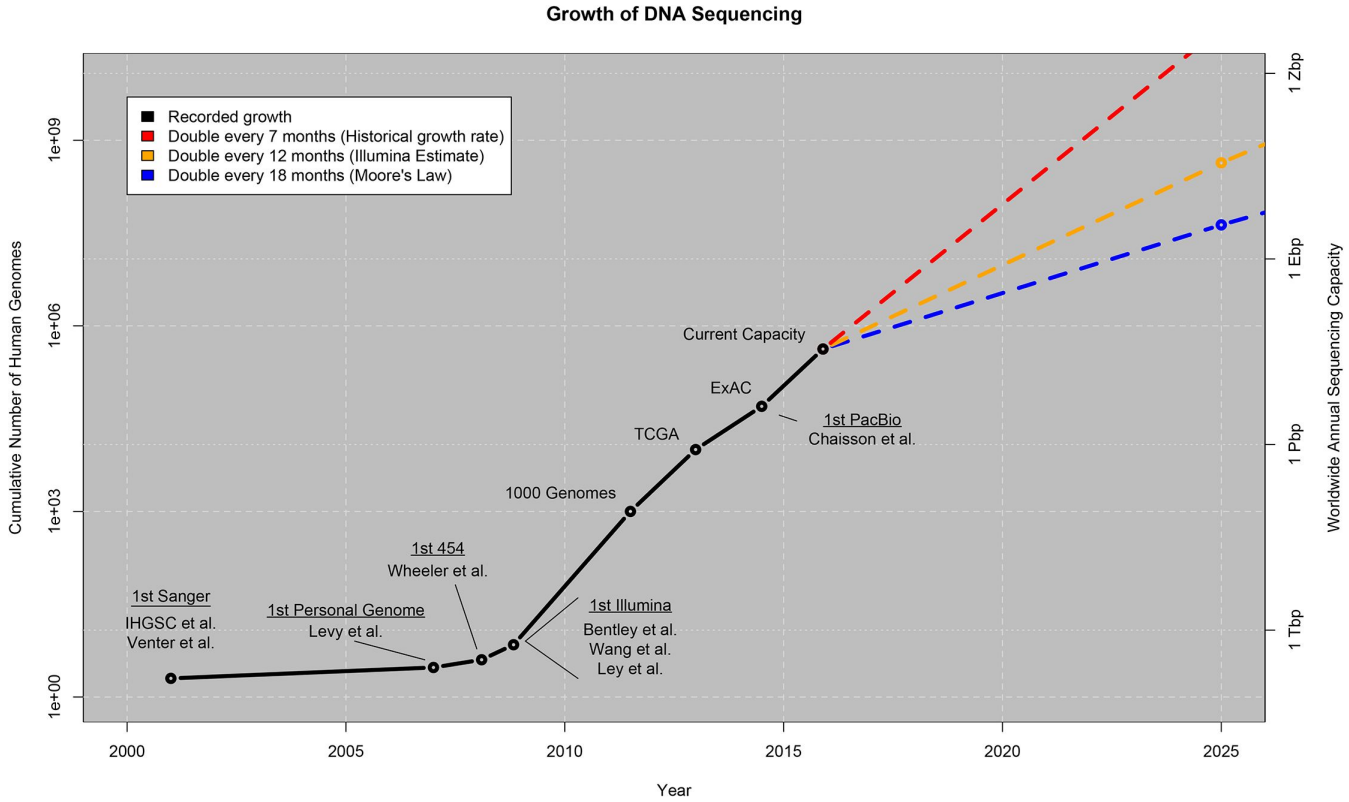
Latent variable models and dimension reduction

Neo Christopher Chung

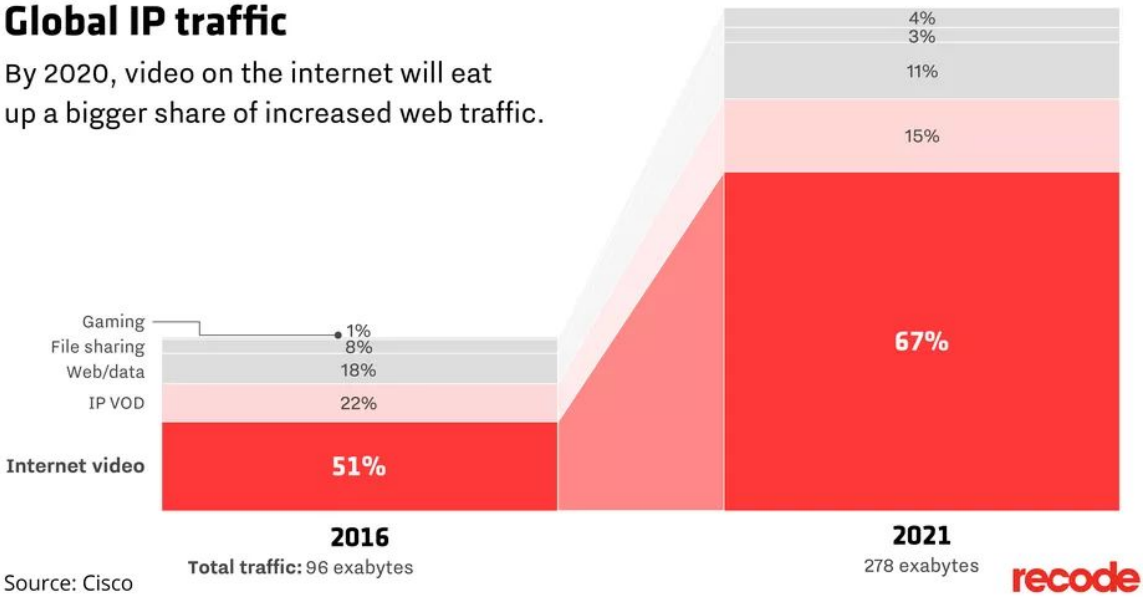
Lecture 3, 1000-719bMSB

Explosive growth of genomic data



Stephens et al. (2015) PLoS Biology

Data growth, everywhere



300 hours uploaded to youtube per minute

5 billion videos watched on youtube every day

700 million photos shared on Snapchat per day

4.7 trillion photos stored

1 exabyte is ~1 billion movies

Exploratory vs. Confirmatory data analysis

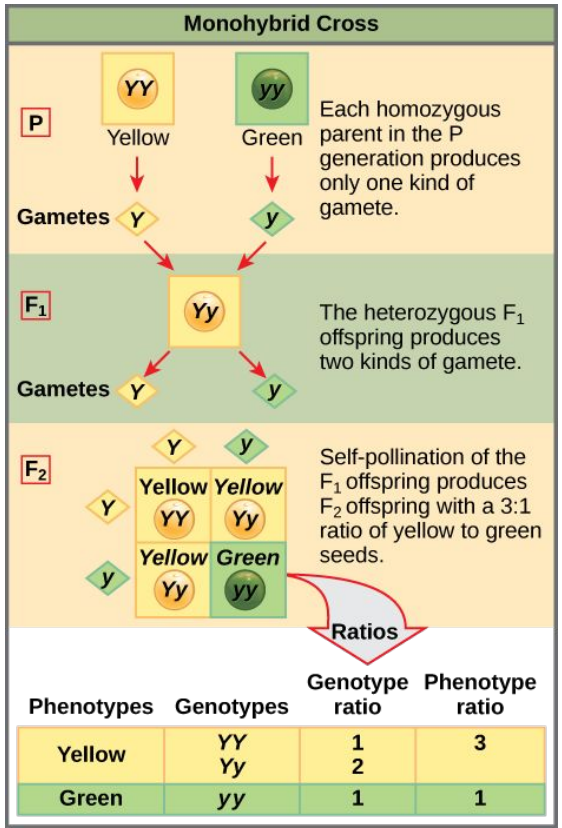
Exploratory data analysis (EDA): summarize the data

Always look min/max, median, quantiles, empirical distribution, and so on.
Lean on robust statistics and nonparametric methods
Not necessarily, but could employ statistical models

In contrast to statistical tests, EDA doesn’t rely on a hypothesis.
EDA may help generate hypotheses to test
EDA may help you go beyond CDA
EDA becomes more relevant in high dimensional data

Gregor Mendel's principles of heredity

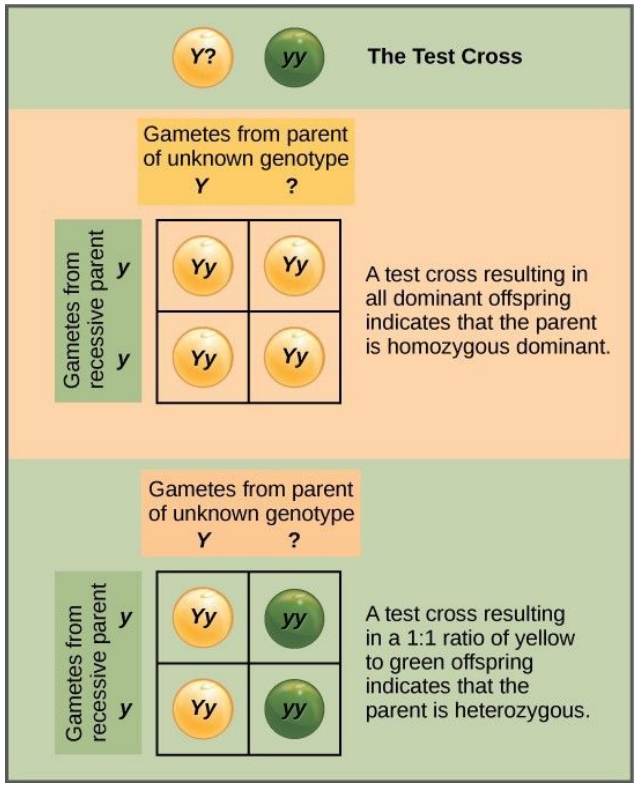
First statistical/math results in biology (1860s)



In the P generation, pea plants that are true-breeding for the dominant yellow phenotype are crossed with plants with the recessive green phenotype. This cross produces F1 heterozygotes with a yellow phenotype. Punnett square analysis can be used to predict the genotypes of the F2 generation.

MCB: https://rwu.pressbooks.pub/bio103/chapter/mendelian-genetics/

Gregor Mendel's principles of heredity



A test cross can be performed to determine whether an organism expression a dominant trait is a homozygote or a heterozygote.

MCB: https://rwu.pressbooks.pub/bio103/chapter/mendelian-genetics/

Three principles of heredity

Law of Dominance
*Hybrid offspring will only inherit the **dominant trait** in the phenotype. The alleles that are suppressed are called the **recessive traits** while the alleles that determine the trait are known as the **dominant traits.***

Law of Segregation
The law of segregation states that during the production of gametes, two copies of each hereditary factor segregate so that offspring acquire one factor from each parent. In other words, allele (alternative form of the gene) pairs segregate during the formation of gamete and re-unite randomly during fertilization.

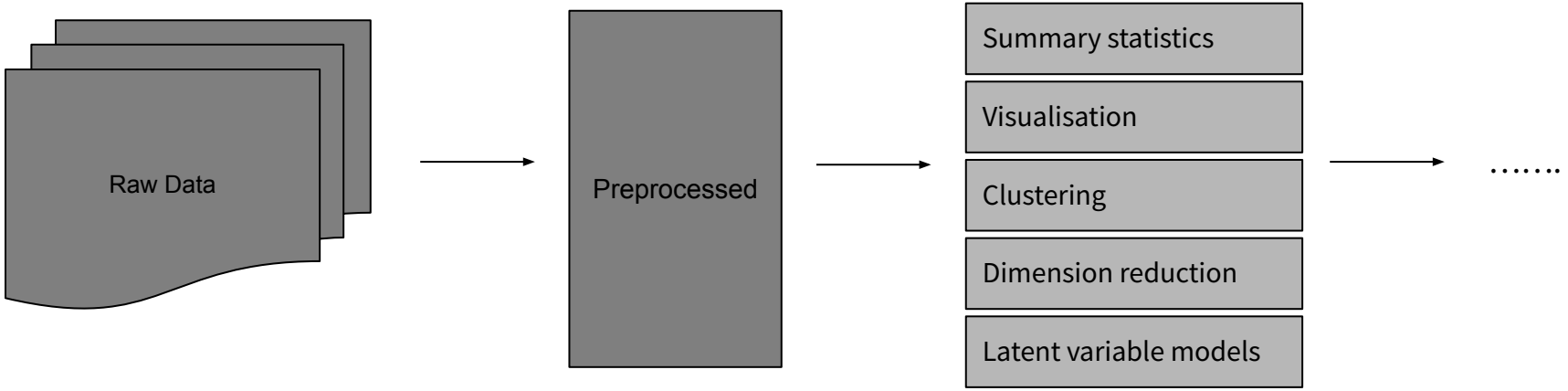
Law of Independent Assortment
A pair of traits segregates independently of another pair during gamete formation. As the individual heredity factors assort independently, different traits get equal opportunity to occur together.

MCB: https://rwu.pressbooks.pub/bio103/chapter/mendelian-genetics/

J. W. Tukey's Exploratory Data Analysis (1997)

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

"Numerical quantities focus on expected values, graphical summaries on unexpected values."



Previously looked at visualization techniques, from a boxplot to a density plot

Now, we look at how to summarize high-dimensional data in a low-dimensional space

Particularly, focus on what does it mean to reduce the dimensions

Clustering

Much of unsupervised learning are rooted in classic clustering algorithms.

Group similar observations together – identify similarity and dissimilarity in the data

Typically hard clustering, we aim to group *n* variables (samples) into *k* clusters

If samples come from naturally occuring latent groups, our goal elevates to identifying how many clusters exist and which observations belong to which cluster.

K-means Clustering

m variables (e.g., genes as rows in the data matrix) are represented as $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$

Each $\mathbf{x}_i$ has n observations (e.g., samples as columns)

Euclidean distance: $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\sum[\mathbf{x}_i - \mathbf{x}_j])^2}$

We consider/identify that there are K clusters, $C_1, C_2, \ldots, C_K$

μ_k is the mean (centroid) of all $\mathbf{x}_i$ that belongs that kth cluster

Hartigan-Wong algorithm (1979)

Minimizes the within-cluster sum of squared distances (WCSS)

For a kth cluster C_k, between $\mathbf{x_i}$ and the corresponding centroid:

$$W(C_k) = \sum[d(\mathbf{x_i}, \mu_\mathbf{k})], \text{ where } x_i \in C_k$$

Total within-cluster variation is then,

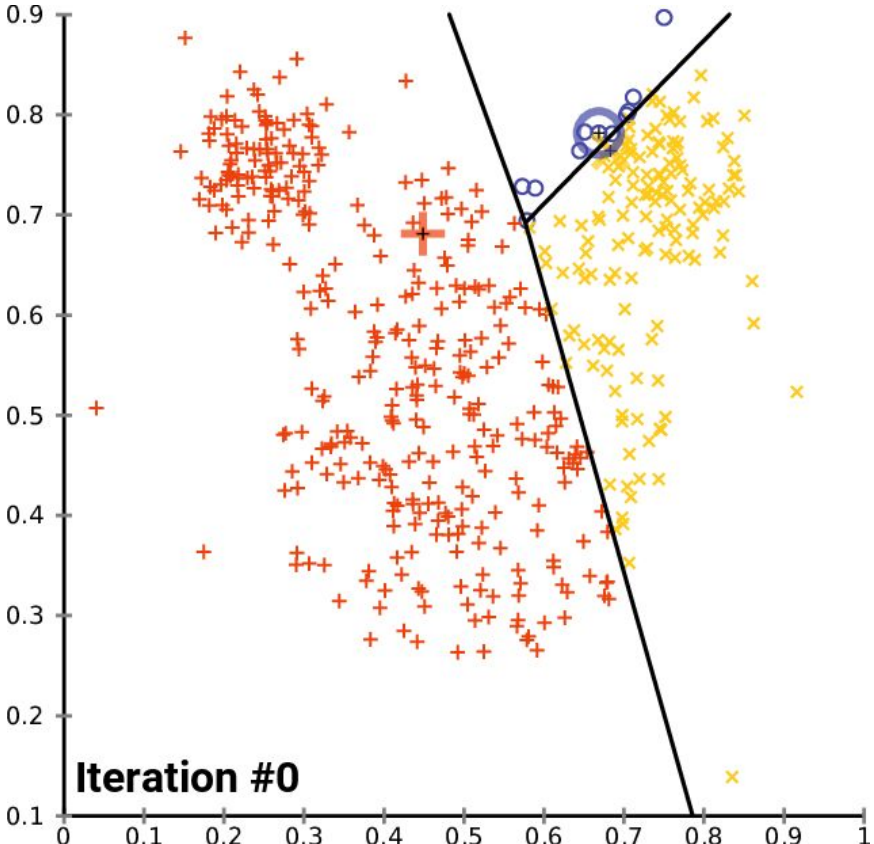
$$WCSS = \sum W(C_k), \text{ for } k = 1, \ldots, K$$

Hartigan-Wong algorithm (1979)

1. Specify the number of clusters (K) to be created (by the analyst)
2. Select randomly k objects from the data set as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the k clusters, update the cluster centroid by calculating the new mean values of all the data points in the cluster.
5. Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached.

https://uc-r.github.io/kmeans_clustering

K-means clustering



K-means Clustering

Small changes to this algorithm yield many other “advanced” clustering algorithms:

K-means clustering, minimizing the 2-norm distance metric

k-medoids clustering, aka PAM (Partitioning Around Medoids)

Mini-batch k-means, a scalable clustering algorithm based on k-means

Many variants of fuzzy (soft) clustering algorithms

k-nearest neighbor classifier is closely related to k-means clustering

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

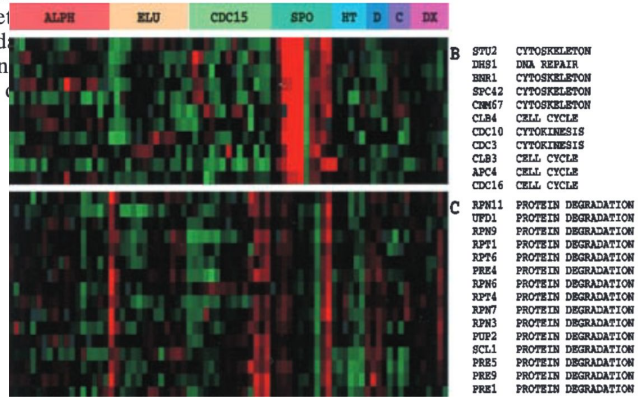
*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Contributed by David Botstein, October 13, 1998

ABSTRACT **A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, conveying the clustering and the underlying expression data simultaneously in a form intuitive for biologists. We have found in the budding yeast *Saccharomyces cerevisiae* that clustering gene expression data groups together efficiently genes of known similar function, and we find a similar tendency in human data. Thus patterns seen in genome-wide expression experiments can be interpreted as indications of the status of cellular processes. Also, coexpression of genes of known function with poorly characterized or novel genes may provide a simple means of gaining leads to the functions of many genes for which information is not available currently.**

be used, such as the Euclidean distance, angle, or dot products of the two *n*-dimensional vectors representing a series of *n* measurements. We have found that the standard correlation coefficient (i.e., the dot product of two normalized vectors) conforms well to the intuitive biological notion of what it means for two genes to be "coexpressed;" this may be because this statistic captures similarity in "shape" but places no emphasis on the magnitude of the two series of measurements.

It is not the purpose of this paper to survey the various methods available to cluster genes on the basis of their expression patterns, but rather to illustrate how such methods can be useful to biologists in the analysis of gene expression data. We aim to use these met
tables containing primary da
that can be reduced, in the en
Clustering methods can be



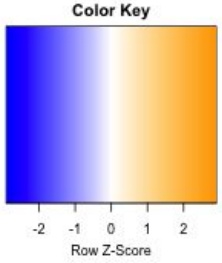
Cluster analysis and display of genome-wide expression patterns

MB **Eisen**, PT Spellman, PO Brown… - Proceedings of the …, **1998** - National Acad Sciences

… Therefore, we always combine **clustering** methods with a … **cluster** analysis (5) to gene expression data collected in our laboratories. This method is a form of **hierarchical clustering**, …

☆ Save 99 Cite Cited by 20443 Related articles All 80 versions

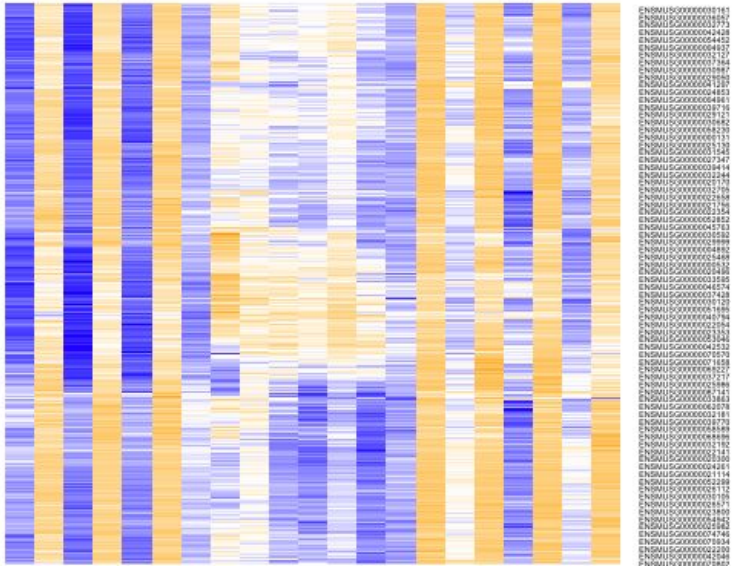
Bottomly et al. 2011 data on mouse gene exp.



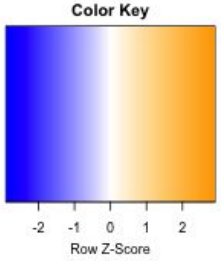
Bottomly et al. Raw

Data from [Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays.]

Each row's scaled and centered.



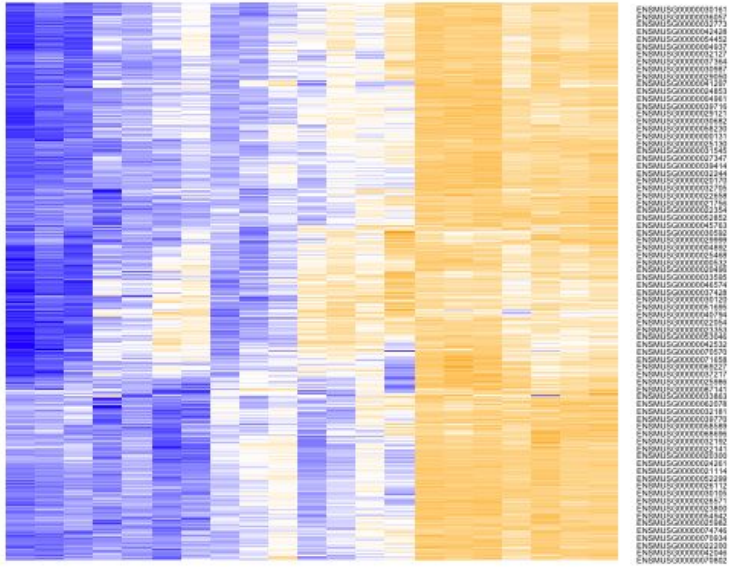
Bottomly et al. 2011 data on mouse gene exp.



Bottomly et al. Clustered

Data from [Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays.]

Each row's scaled and centered.
Columns are clustered revealing systematic patterns



Systematic variation

How do we evaluate, extract, and/or model the systematic variation in data?

- Computing variances and means of rows and/or columns
- Group rows and/or columns according to their characteristics (e.g., clustering)

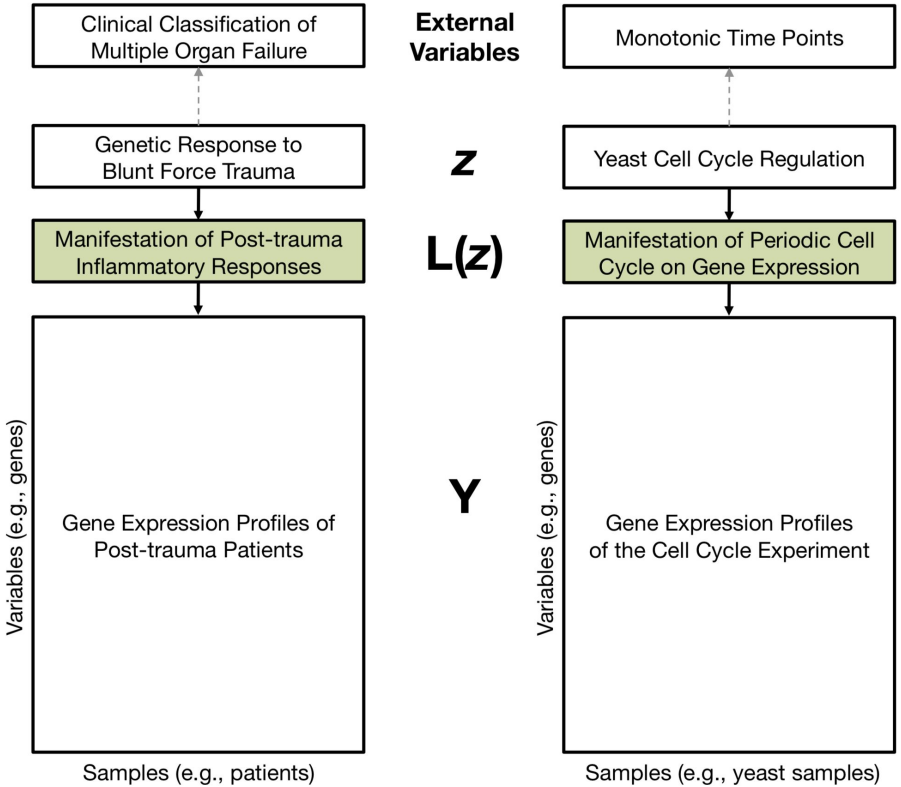
More advanced approaches would consider

- how the data are generated
- how the variables are related

Dimension reduction and latent variables

- Compress the high-dimensional data using fewer variables while minimizing the information loss
- Get a new basis (or multivariate variables) that could explain maximal variance across variables
- Find a low-dimensional space in which the relationships among original variables are preserved
- Identify hidden and unobserved (latent) variables that may underly the original variables

Manifestation of latent variables



Types of models

Table 1.1: Classifications of Latent Variable Models

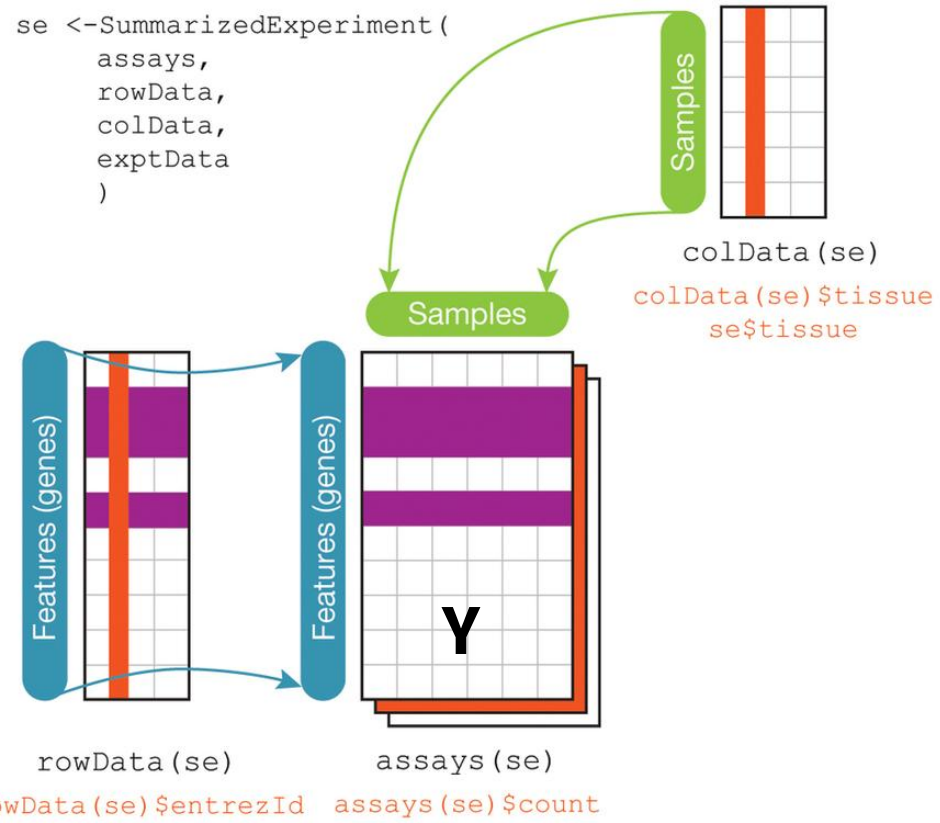
	Manifest Observed Variables	
Latent Unobserved Variables	Continuous	Categorical
Continuous	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

	Manifest Observed Variables	
Latent Unobserved Variables	Continuous	Categorical
Continuous	Factor analysis	Latent trait analysis
Categorical	Latent profile analysis	Latent class analysis

EXAMPLES

- Abundances of mRNAs may be considered continuous observed variables
- MS/MS data on protein concentrations may be considered continuous
- Genotypes (SNPs) are categorical
- Batch effects may be categorical latent variables
- Population structures may be modeled as continuous or categorical
- Etc.

Genomic data



Huber et al. 2015

General latent variable models

m variables $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}$, measured over n observations

Organize into a matrix $\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m})^T$

Expected influence of the latent variables on **Y** by $E[\mathbf{Y}|\mathbf{z}]$,

$$\mathbf{Y} = E[\mathbf{Y}|\mathbf{z}] + \mathbf{E}$$

Estimate **L(z)**, that is a row basis for $E[\mathbf{Y}|\mathbf{z}]$

This low dimensional matrix **L(z)** can be thought of as the manifestation of the latent variables in the observed data.

Graphical representation of LVM

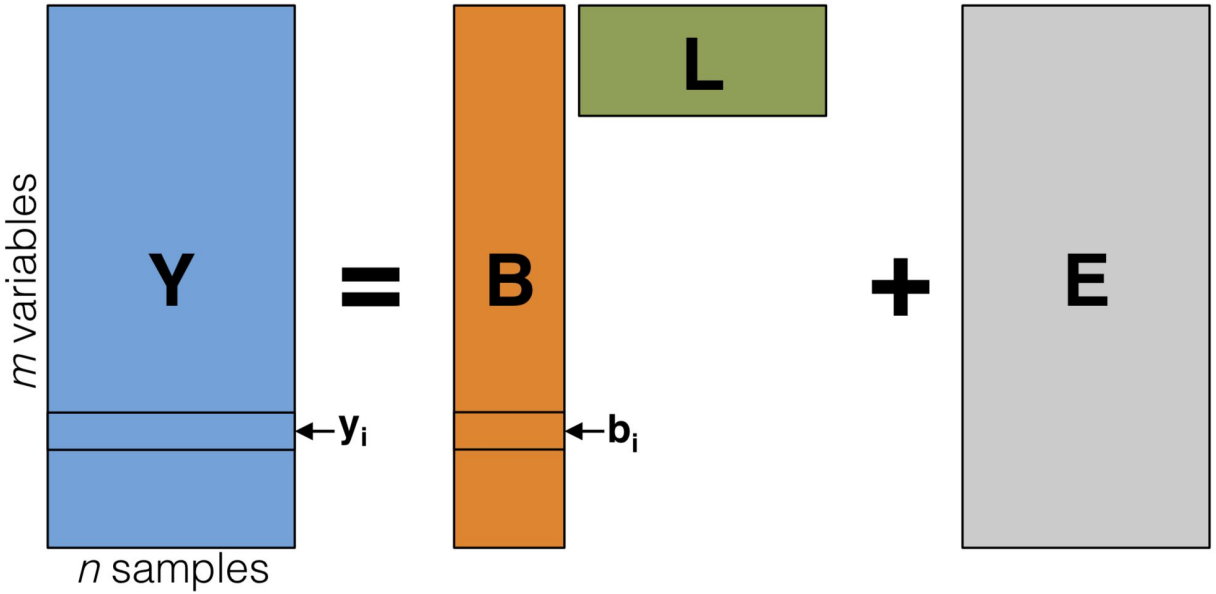


Figure 1.2: Diagram of the latent variable model (1.1). The latent variable basis $\mathbf{L}$ is not observable, but may be estimated from $\mathbf{Y}$ using the top r right singular vectors $\mathbf{V}_{(r)}^T$. The noise term $\mathbf{E}$ is independent random variation. $\mathbf{B}$ is a $m \times r$ matrix of unknown parameters of interest.

Estimating latent variables

Factor analysis (FA), often based on eigendecomposition, was originally developed in psychology where a number of variables aren't that high

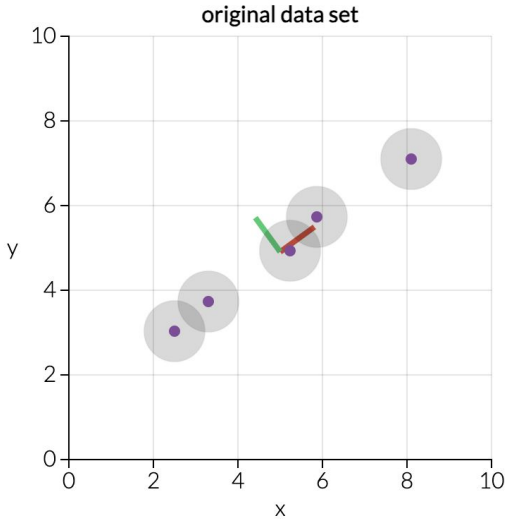
Leek 2011 "Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data" proves that SVD/PCA with a rank r estimates the latent variables in high-dimensional data where m >> n

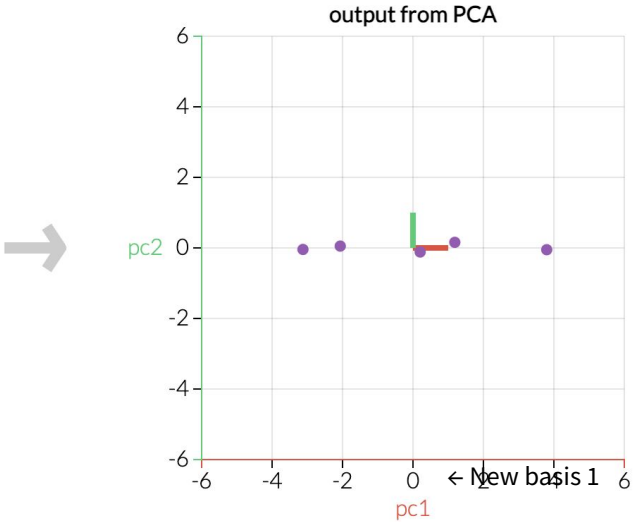
Recent approaches using variational autoencoders (VAE) and related ML methods may be also seen as estimating the latent variables

For more, see Bartholomew's textbook

Latent Variable Models and Factor Analysis: A Unified Approach

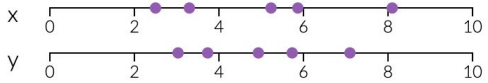
Principal Component Analysis

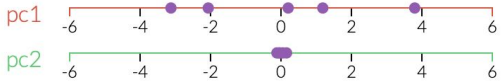




PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.

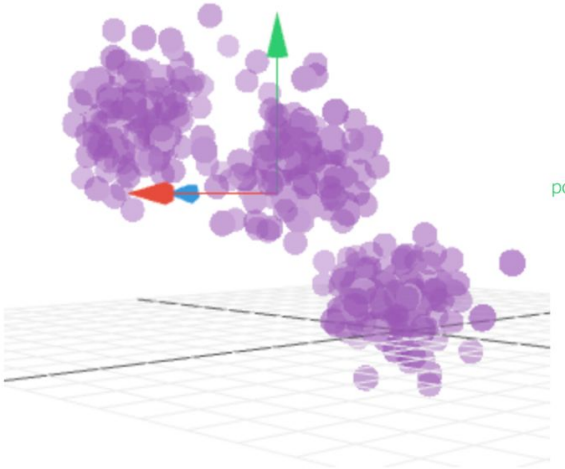
If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.

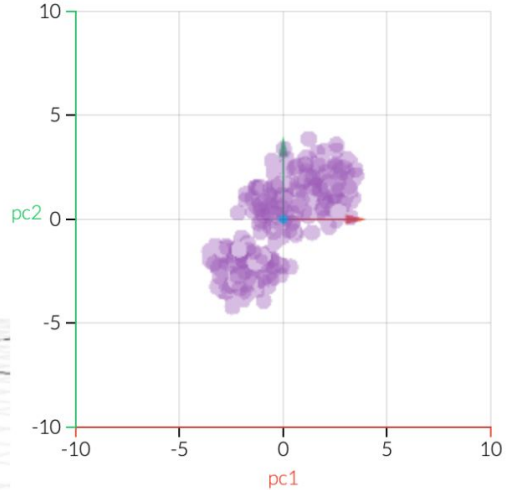


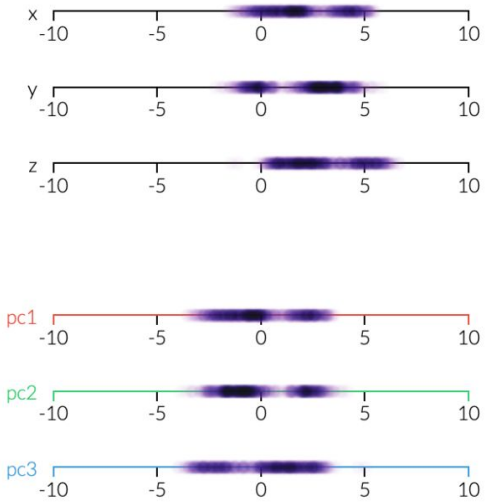


https://setosa.io/ev/principal-component-analysis/

Principal Component Analysis







show PCA reset

https://setosa.io/ev/principal-component-analysis/

Notations

1. **y** is a vector of m random variables
2. $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_n$ are combined to form a matrix **Y**
3. **u** is a vector of m constants
4. $\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_r$ are combined to form a matrix **U**
5. $$\mathbf{x}_1 = \mathbf{u}_1^T \mathbf{Y} = \sum_{i=1}^{m} u_{i1} \mathbf{y}_i$$

Sequential algebraic derivation Hotelling, 1933

1. The 1st PC can be found by searching for a weighted sum of m variables with maximum variance, where a set of m loadings is constrained to be an unit vector

$$\mathbf{x}_1 = \mathbf{u}_1^T \mathbf{Y} = \sum_{i=1}^{m} u_{i1} \mathbf{y}_i$$

2. The maximization of var(x_1) leads to $\mathbf{u}_1$ that is the loadings for the 1st PCs, $\mathbf{x}_1$.
3. The 2nd PC is then a linear function $\mathbf{x}_2 = \mathbf{u}^T_2 \mathbf{Y}$ with maximum variance that is subject to $\mathbf{x}^T_1 \mathbf{x}_2 = 0$ (orthogonality) and $\mathbf{u}^T_2 \mathbf{u}_2 = 1$ (unit length).
4. Subsequently, we can derive $r < \min(m, n)$ PCs, which are mutually orthogonal.

Minimizing the sum of squared residuals

We can estimate $\mathbf{Y}$ by superimposing the top r PCs and the corresponding loadings. This matrix is often called an eigenmatrix:

$$\widehat{\mathbf{Y}}^{(r)} = \sum_{k=1}^{r} \mathbf{u}_k \mathbf{x}_k$$

Then, the sum of squared residuals (SSR) is,

$$\mathrm{SSR} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \|\mathbf{y}_\mathrm{i} - \widehat{\mathbf{y}}_\mathrm{i}^{(\mathrm{r})}\|^2$$, where $\|\cdot\|$ is the L_1 norm.

When estimating $\mathbf{Y}$ with any set of r arbitrary vectors, using the top r PCs always leads to the minimal SSR.

Singular value decomposition

PCA is the most efficiently computed by SVD in practice:

$$\mathbf{Y}_{(m\times n)} = \mathbf{U}_{(m\times n)}\mathbf{S}_{(n\times n)}\mathbf{V}^{T}_{(n\times n)}$$

U is a m×n orthonormal matrix, the left singular vectors
D is a n×n diagonal matrix, where the diagonal elements are the singular values
V is a n×n orthonormal matrix, the right singular vectors

PCs are the rows of $\mathbf{DV}^T$, where the i^{th} PC is found in the i^{th} row of $\mathbf{DV}^T$.
The right singular vectors of **Y** are equivalent to the eigenvectors of $m^{-1}\mathbf{Y}^T\mathbf{Y}$.

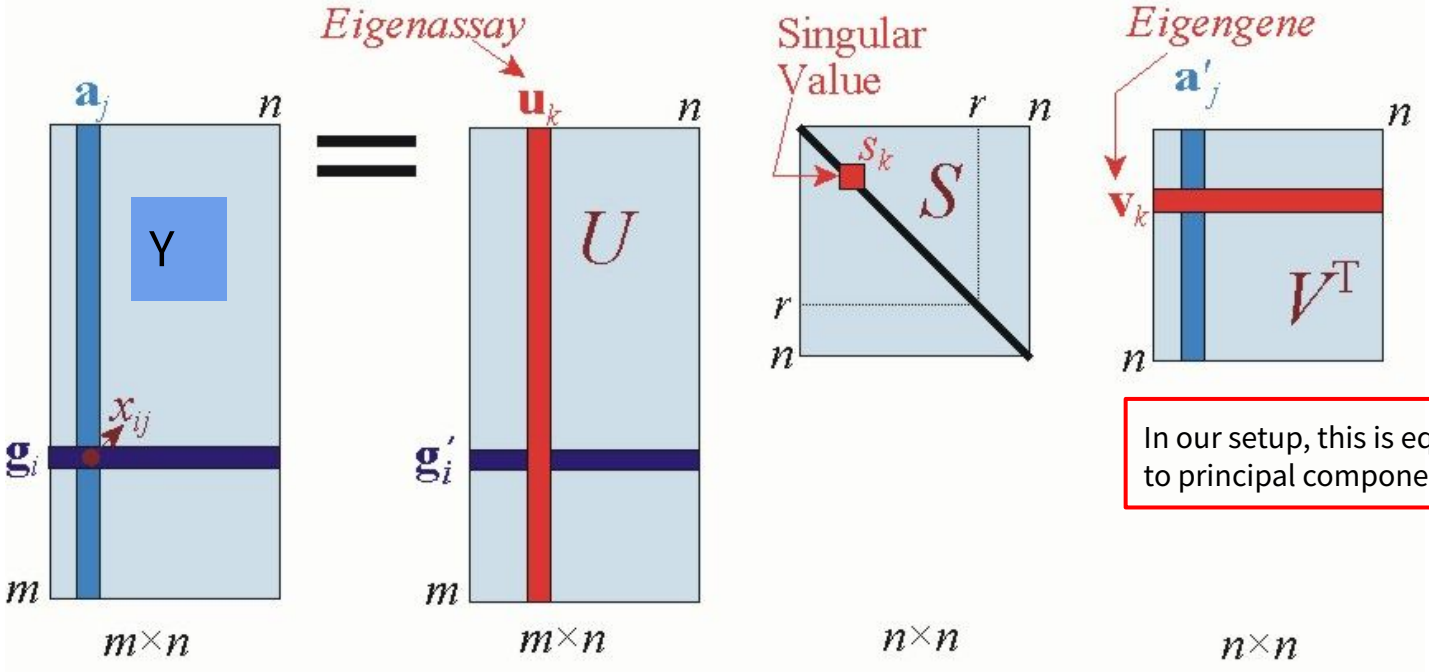
Asymptotic Conditional SVD

Leek 2010 Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data. *Biometrics* proves that in large-scale genomic data, SVD (therefore PCA) can accurately capture the latent variables.

> As $m \to \infty$, the top r right singular vectors of **Y** converge with probability 1 to a matrix whose row space is equivalent to that of **L** (Leek, 2010)

Singular value decomposition

$$X = USV^{\mathrm{T}}$$



In our setup, this is equivalent to principal components

Wall et al. 2003

Using eigenmatrices for imputation

Missing data imputation (**SVDimpute** from Troyanskaya et al. 2001 *Bioinformatics*)

1. Consider data Y with m rows and n columns
2. For missing values, use the row means as the first approximations
3. Compute SVD
4. Take the eigenmatrix of k rank
5. Impute missing values with corresponding values from this eigenmatrix

SVD/PCA for netflix recommendation

1 Million $ Prize.

We have a very large data of reviews (5 stars)

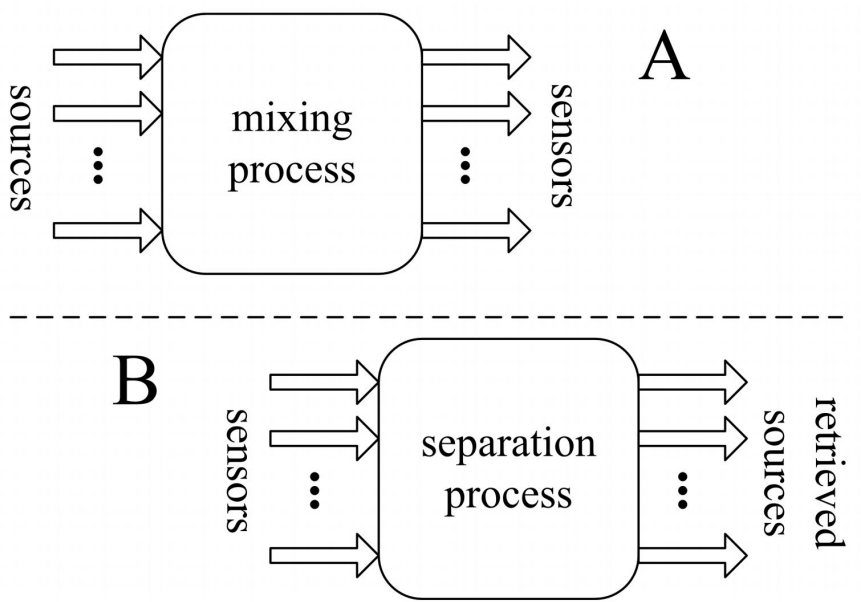
How do we recommend the best movies to users.

Aka, how do we predict what a user’s rating

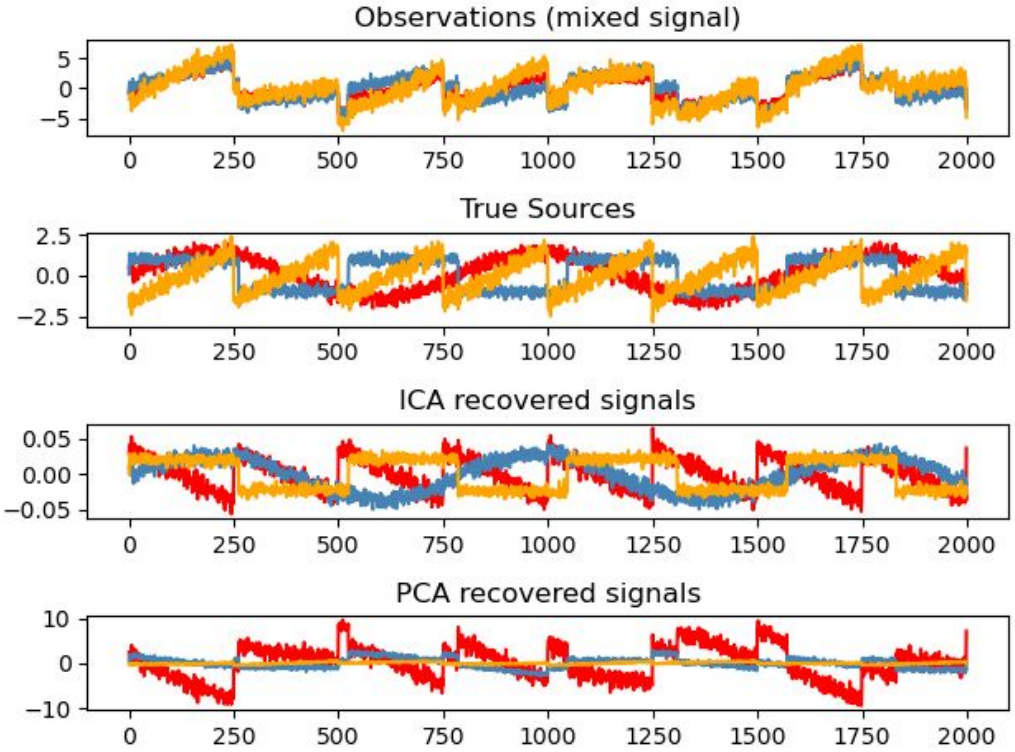
For any (unseen/unrated) movie

Independent Component Analysis

Closely related to PCA
Originated in signal processing for “cocktail party problem”



Recovering “mixed” signals



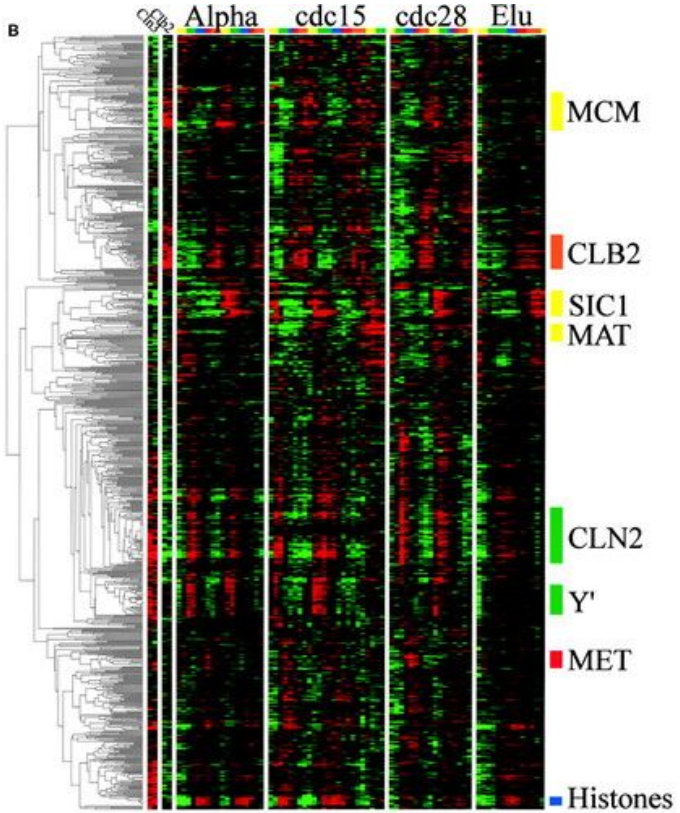

t = 14160000 step
kurt = −0.68
kurt = −1.56
kurt = −1.26
kurt = −1.20

Alter, Brown, and Botstein (2000) Singular value decomposition for genome-wide expression data processing and modeling. PNAS

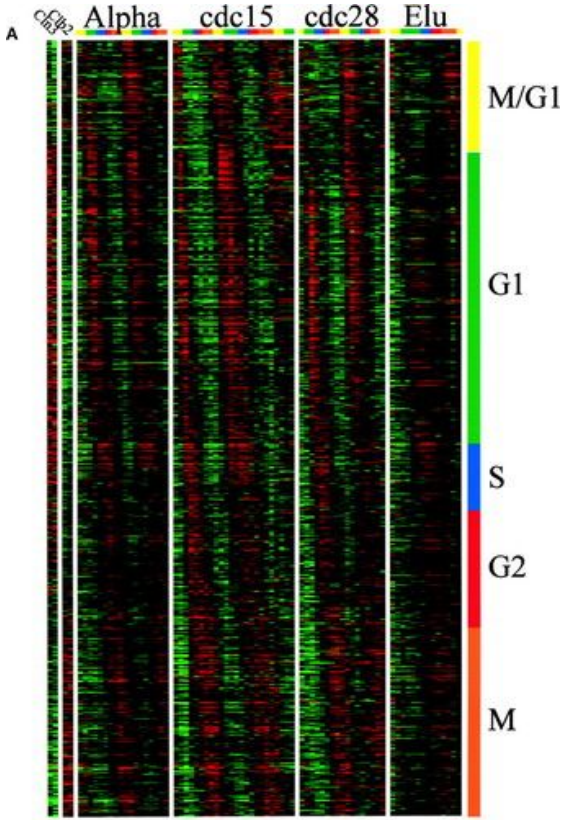
Spellman et al. (3) monitored genome-wide mRNA levels, for 6,108 ORFs of the budding yeast Saccharomyces cerevisiae simultaneously, over approximately one cell cycle period, T ≈ 390 min, in a yeast culture synchronized by elutriation, relative to a reference mRNA from an asynchronous yeast culture, at 30-min intervals.

→ How do we capture the cell cycle regulation?

See the original data in Spellman et al. (1998) MBoC

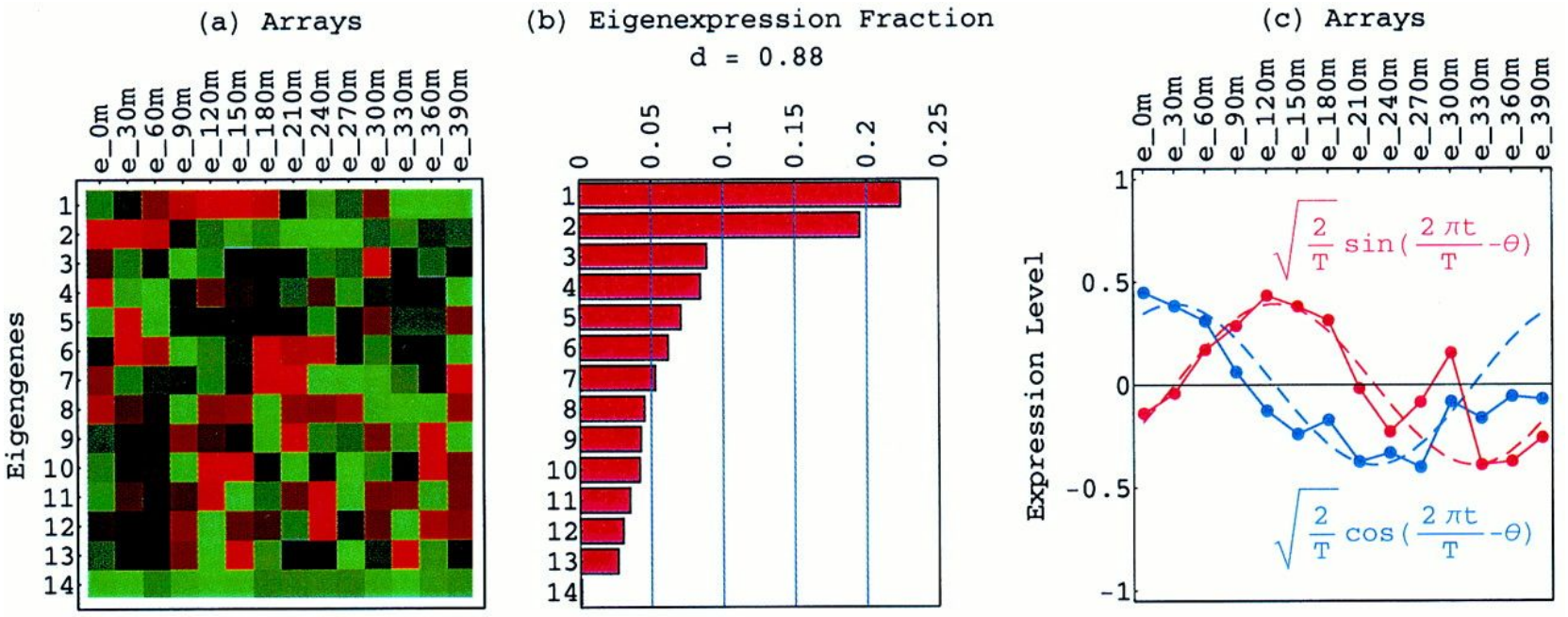


(A) Gene expression patterns for cell cycle–regulated genes. The 800 genes are ordered by the times at which they reach peak expression.



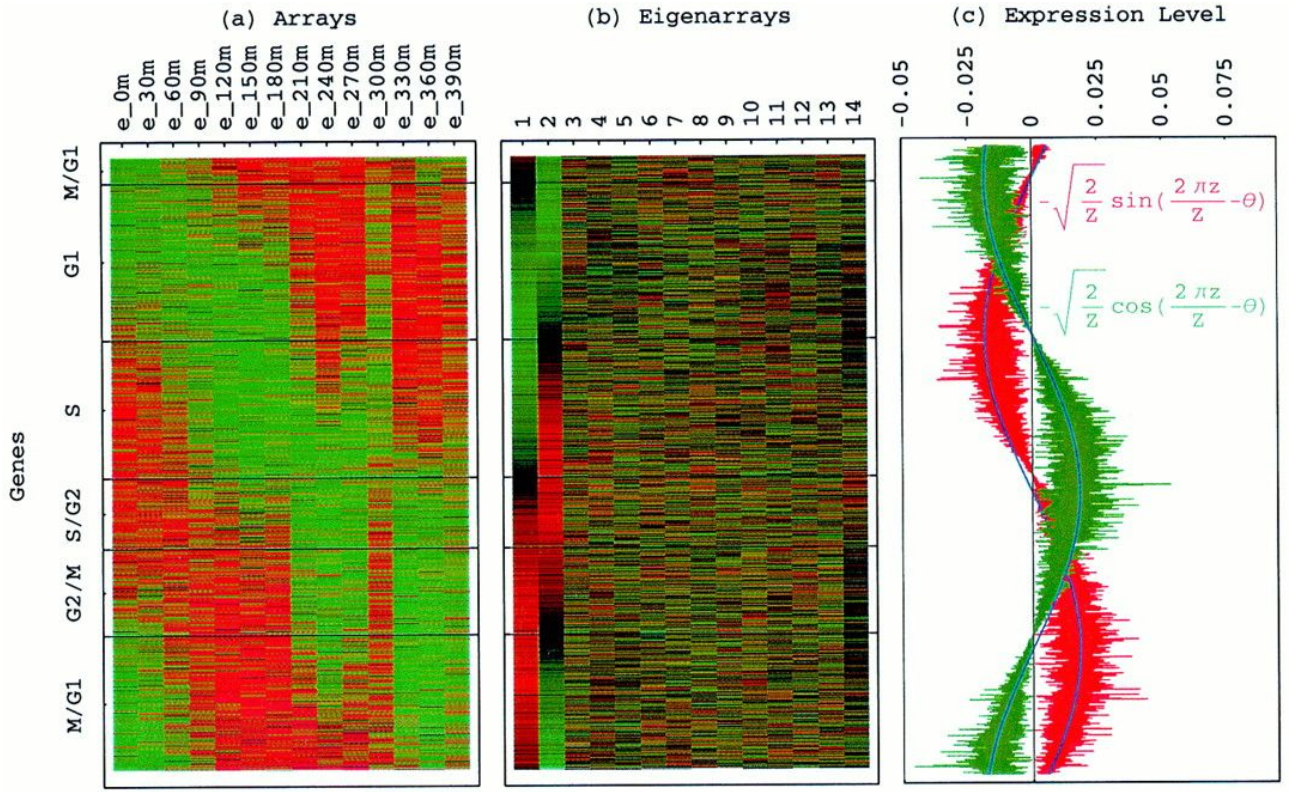
(B) Genes that share similar expression profiles are grouped by a (hierarchical) clustering algorithm

SVD/PCA in gene expression data



Alter et al 2000

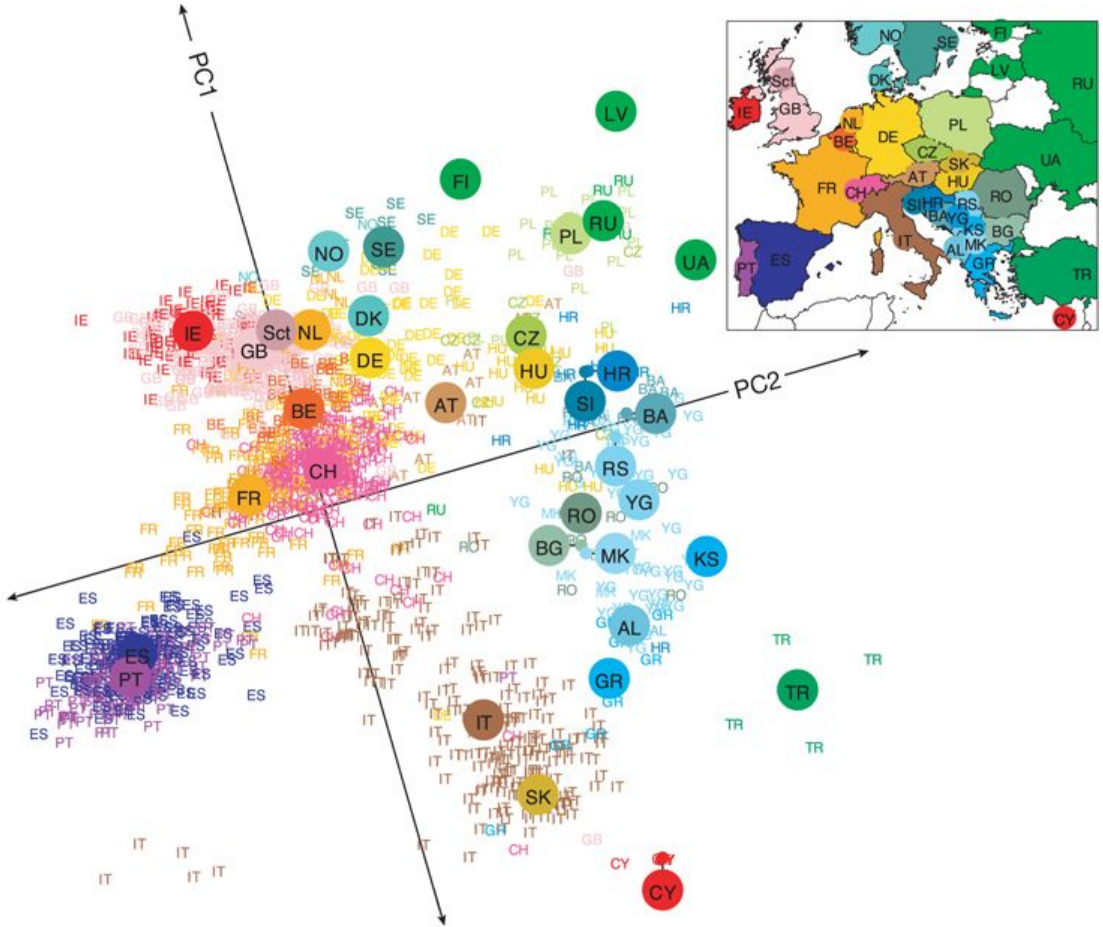
Eigenarrays



Alter et al 2000

Novembre et al. (2008) Genes mirror geography within Europe. Nature

[...] we characterize genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans.



A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe.

Accounting for population structure

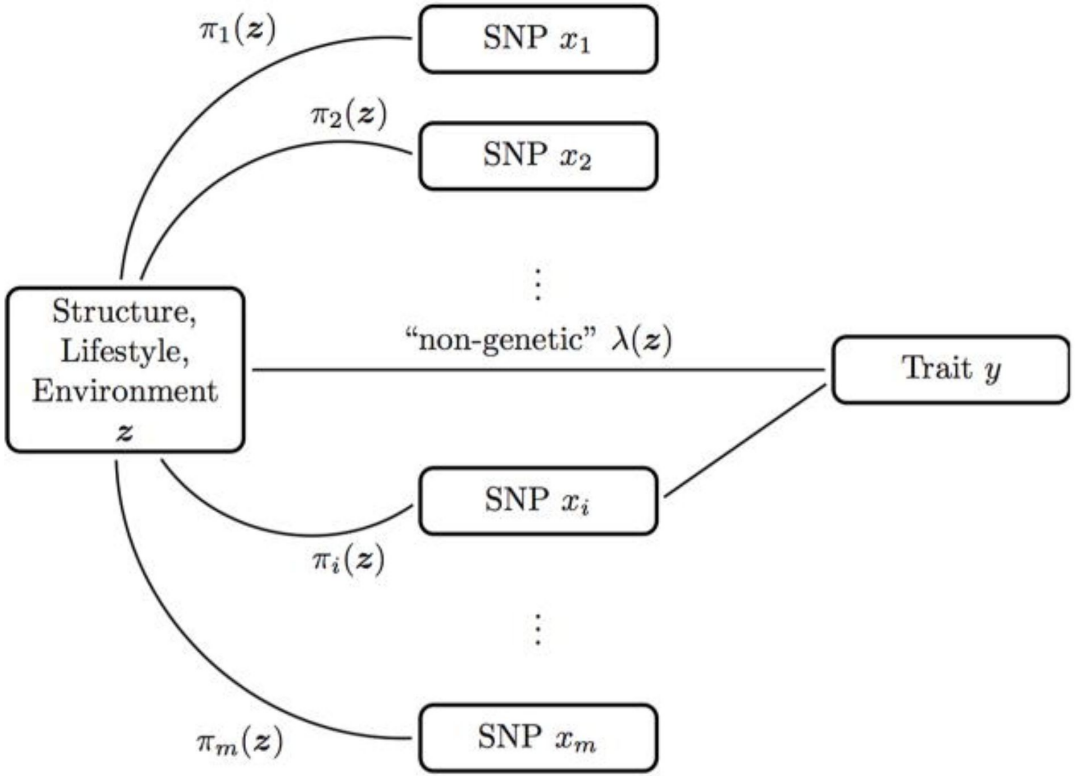
Importantly, “population structure” is needed in assessing association between genetics and diseases. Without this type of methods, we may not be able to distinguish or identify genes (or loci) that are contributing to susceptibility to a disease

1. Model the SNP data using latent variable models
2. Estimate population structure by PCA, LMM, LFA, or related methods
3. Include the top r latent variables in an association test - GWAS: disease ~ gene

Price et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies

Kang et al. (2010) Variance component model to account for sample structure in genome-wide association studies

Song et al. (2015) Testing for genetic associations in arbitrarily structured populations



A graphical model describing population structure and its effects on a trait of interest. Population structure is captured by a common latent variable $\boldsymbol{z}$ among a set of loci x_i (i =1,2, …, m), via the allele frequencies $\pi_i(\boldsymbol{z})$. When one locus has a causal effect on the trait, this induces spurious associations with other loci affected by population structure. At the same time, population structure may be correlated with lifestyle and environment as these are all possibly related to ancestry and geography.

LVM for population structure

There are n individuals, each with m measured SNP genotypes.

The genotype for SNP i in individual j is denoted by $x_{ij} \in \{0,1,2\}$, $i = 1,2, \ldots, m$, $j = 1,2, \ldots, n$. We collected these SNP genotypes into an $m \times n$ matrix $\boldsymbol{X}$, where the (i, j) entry is $\boldsymbol{x}_{ij}$. We denote the genotypes for individual j by $\boldsymbol{x}^j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$.

Introduce $\boldsymbol{Z}$ as an unobserved variable capturing an individual's structure

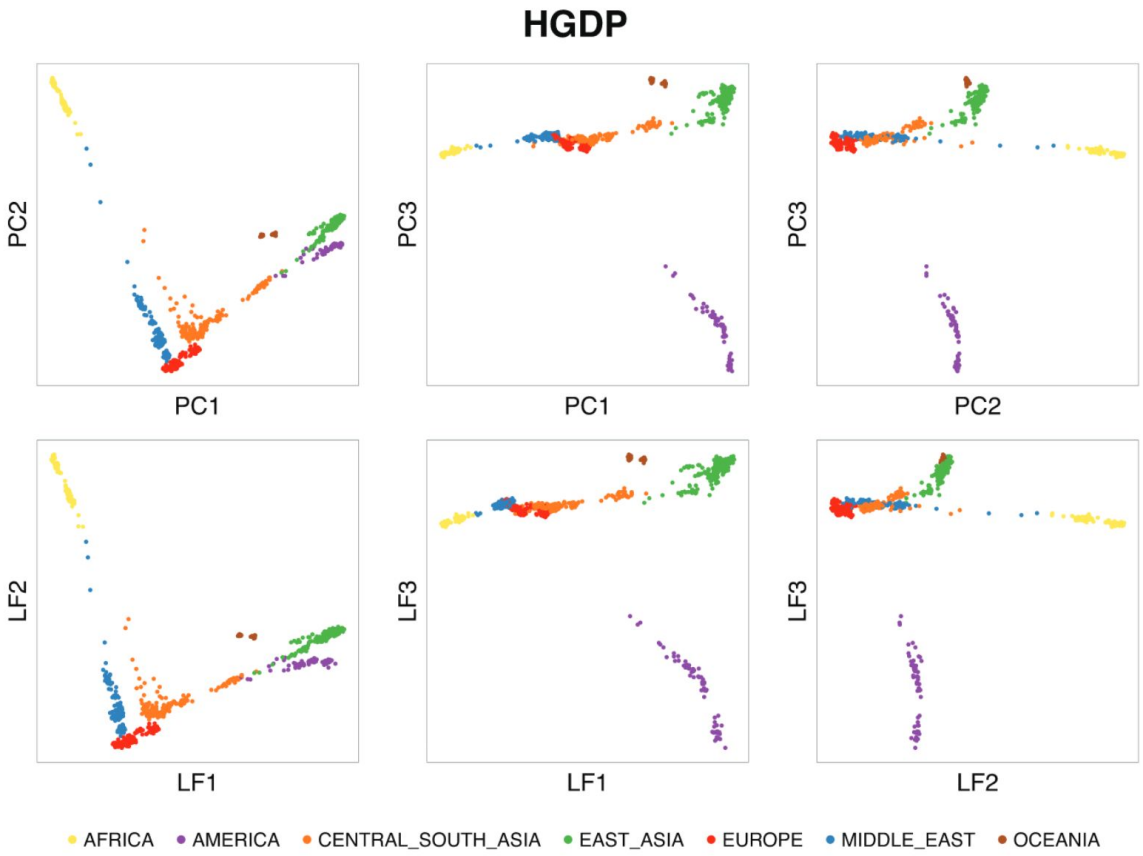
For SNP i, the allele frequency can be viewed as a function of $\boldsymbol{Z}$, i.e. $\pi_i(\boldsymbol{Z})$.

For a sampled individual j from an overall population, we have 'individual-specific allele frequencies' defined as $\pi_{ij} \equiv \pi_i(\boldsymbol{z}_j)$ at SNP i.

Each value of π_{ij} informs us as to the expectation of that particular SNP/individual pair under the scenario we observed a new individual at that locus with the same structure, specifically as $E[x_{ij}]/2 = \pi_{ij}$.

If an observed SNP genotype x_{ij} is treated as a random variable, then we assume that π_{ij} serves to model x_{ij} as a Binomial parameter: $x_{ij}|\boldsymbol{Z} = \boldsymbol{z}_j \sim \text{Binomial}(2, \pi_i(\boldsymbol{z}_j))$.

Logistic Factor Analysis (LFA),
Extending PCA for binomial data



Principal component and logistic factor biplots for the Human Genome Diversity Project dataset.