Current topics in computational biology

VII: Principal component analysis

Christoph Lippert

Material

- Jon Shlens, 2003:

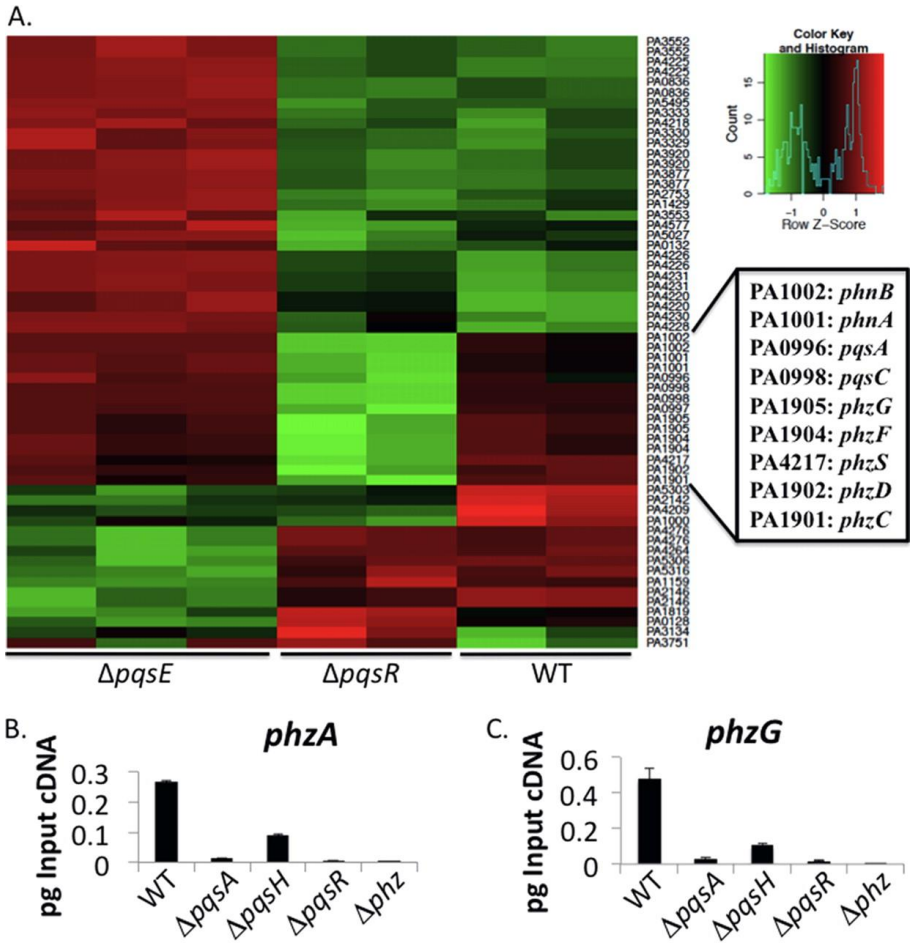
  A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS - Derivation, Discussion and Singular Value Decomposition

- Chris Bishop, 2006

  Pattern Recognition and Machine Learning, Chapter 12

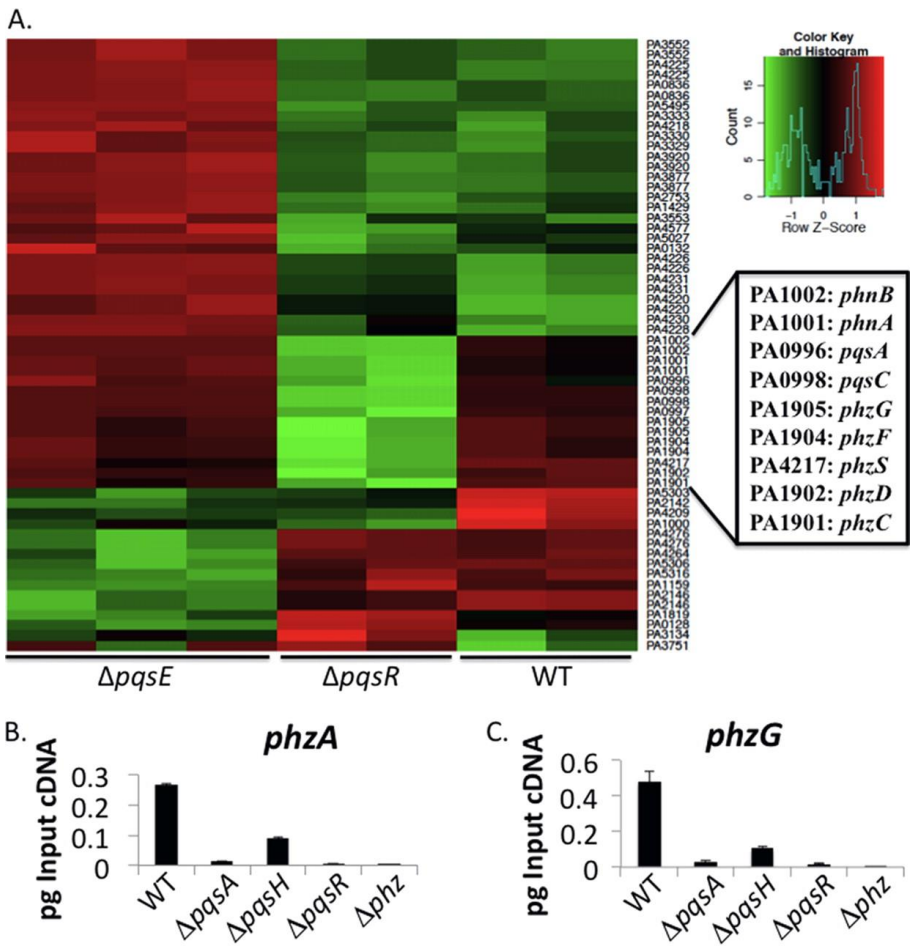
The curse of dimensionality

- Many dimensions measured
- E.g. in linear regression:
  - Variance in β_{ML} increases drastically
- Hard to interpret
- Hard to visualize



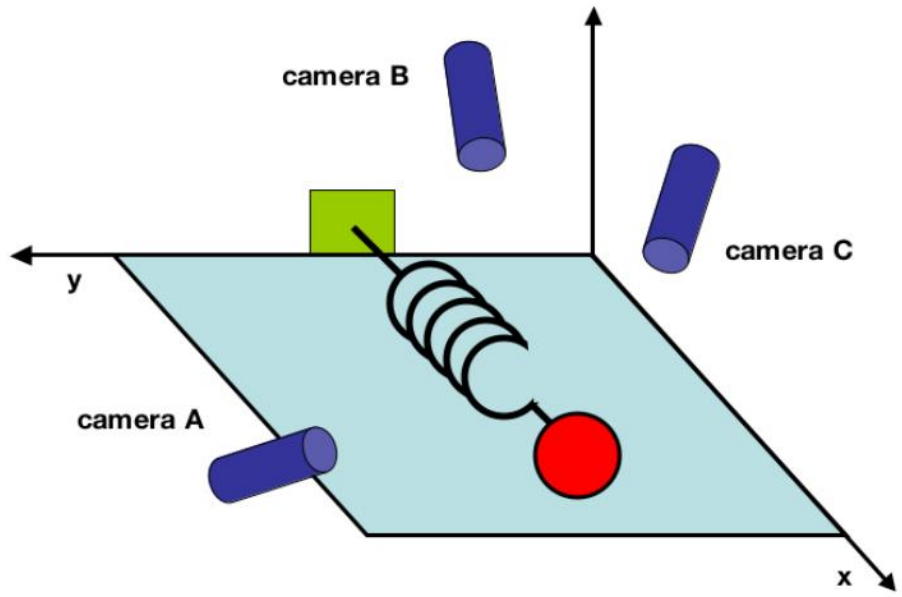
The blessing of dimensionality

- Typically there are only a small number of phenomena underlying the data
- Observed data are redundant representations
- Concentration of measure
  - If a function is smooth across dimensions, then it is almost constant in a high dimensional space => easy to describe



Example: The naïve physicist

- We record a ball on a string over time
- Original signal is 1-dimensional
- Three cameras placed arbitrarily in 3D space
- 2D measurements of each camera are distorted by noise
- Can we recover the original phenomenon?



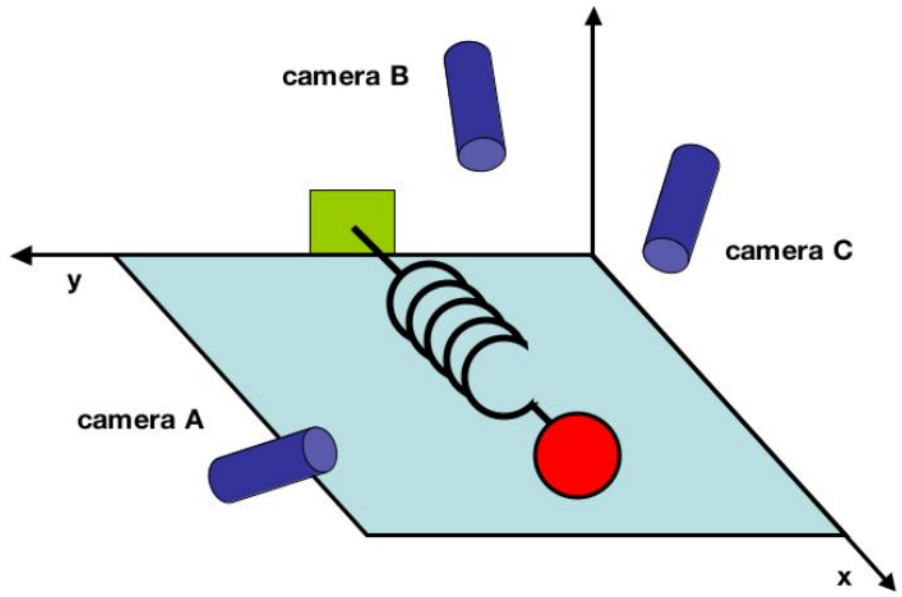
[Shlens 2003]

Example: The naïve physicist

- Sample location over time
  - 2-dimensional projection per camera for each time point
  - Each sample is 6 dimensional

$$X = [x_A, y_A, x_B, y_B, x_C, y_C\ ]$$

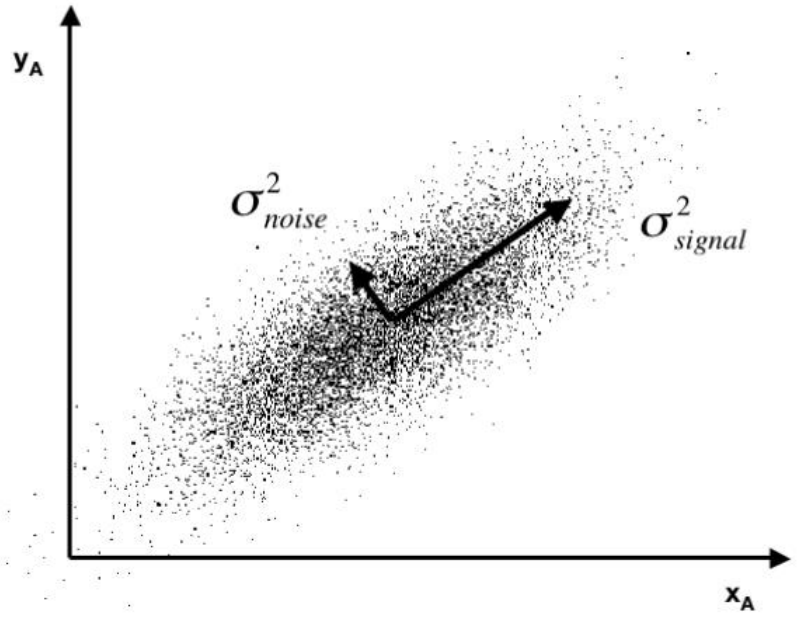
- Goal: compute the most meaningful basis for the data
  - In the example: recover the x-axis



[Shlens 2003]

Signal-to-noise ratio

- In theory each camera in the example should record a straight line
- Deviation from straight line due to noise
- High signal to noise ratio
  - High precision data
- Low signal to noise ratio
  - Noise contaminated data



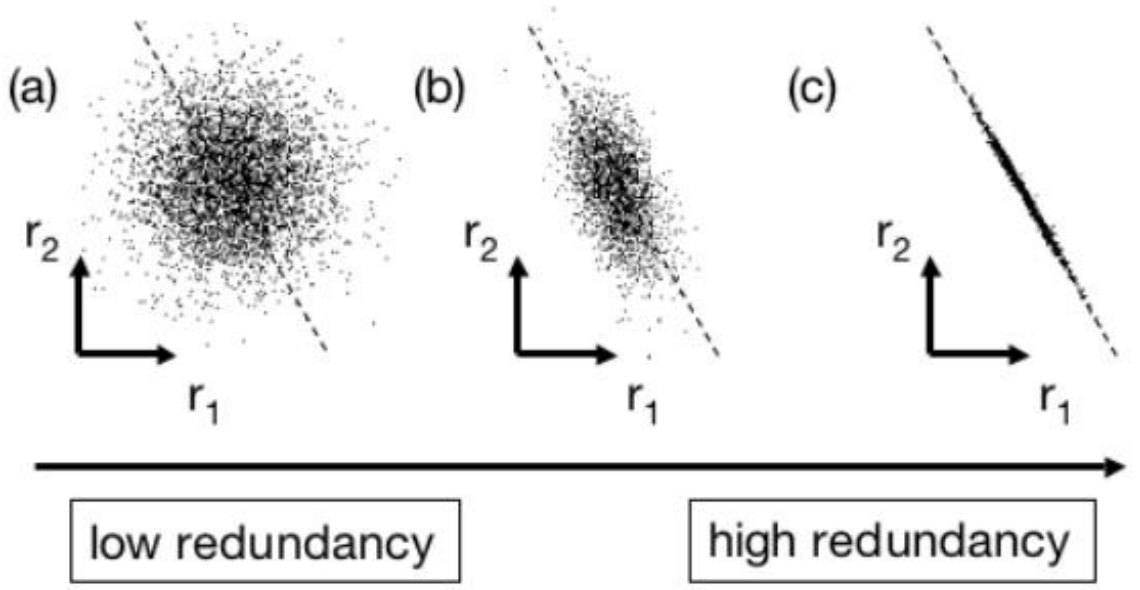
Measurements from camera A

[Shlens 2003]

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

Redundancy

- Low redundancy corresponds to low correlation
  - (a): $(x_A, humidity)$
- High redundancy implies high correlation
  - (c): $(x_A, \tilde{x}_A)$ x_A sensor in meters, $\tilde{x}_A$ sebnsor in inches
  - Recording only one if the two would help reduce the number of recordings
  - Ideal recording:
    - $r_2 - r_1\hat{\beta}$
  - => dimensionality reduction



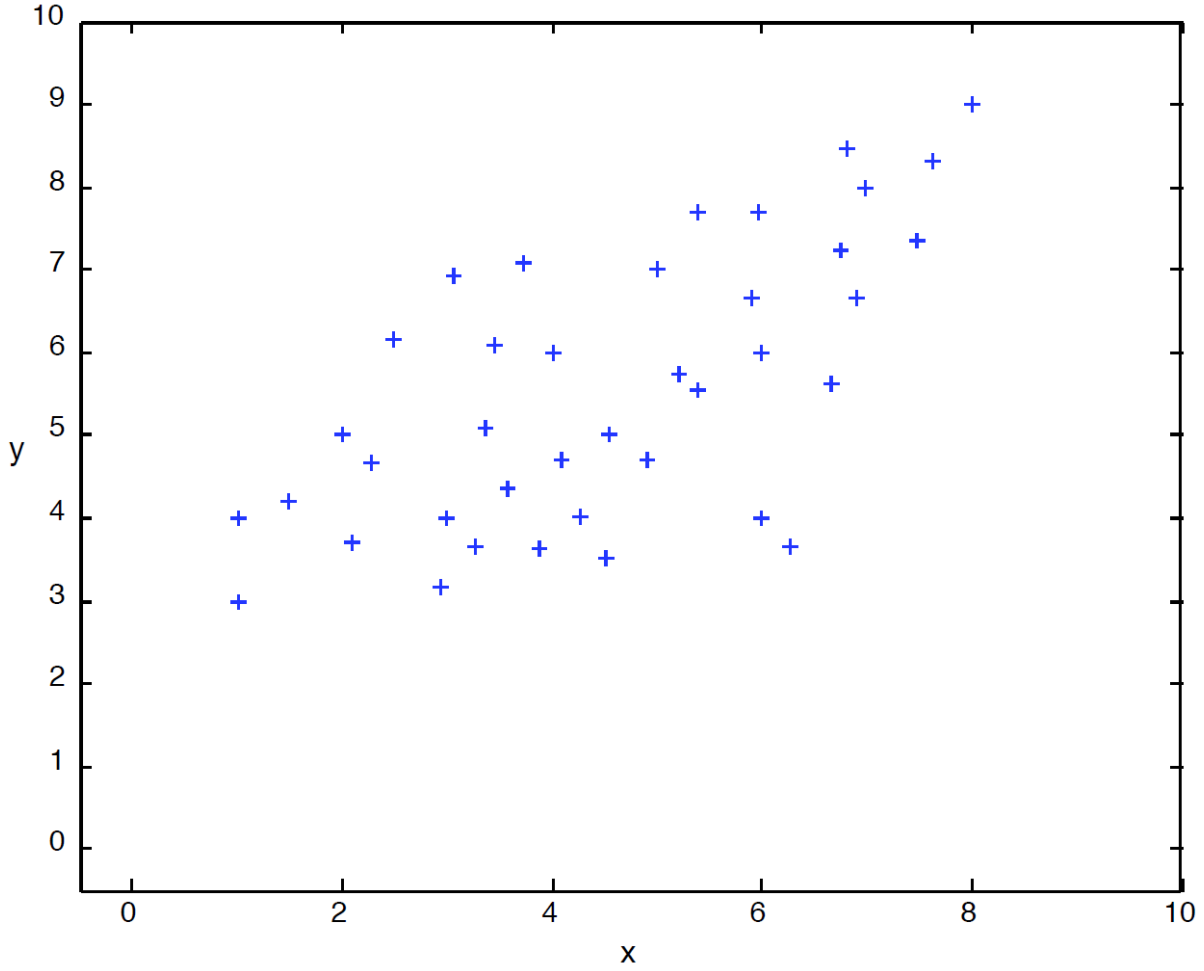
Dashed line: least squares fit: $r_1\hat{\beta}$

with

$r_2 = r_1\beta + \epsilon$

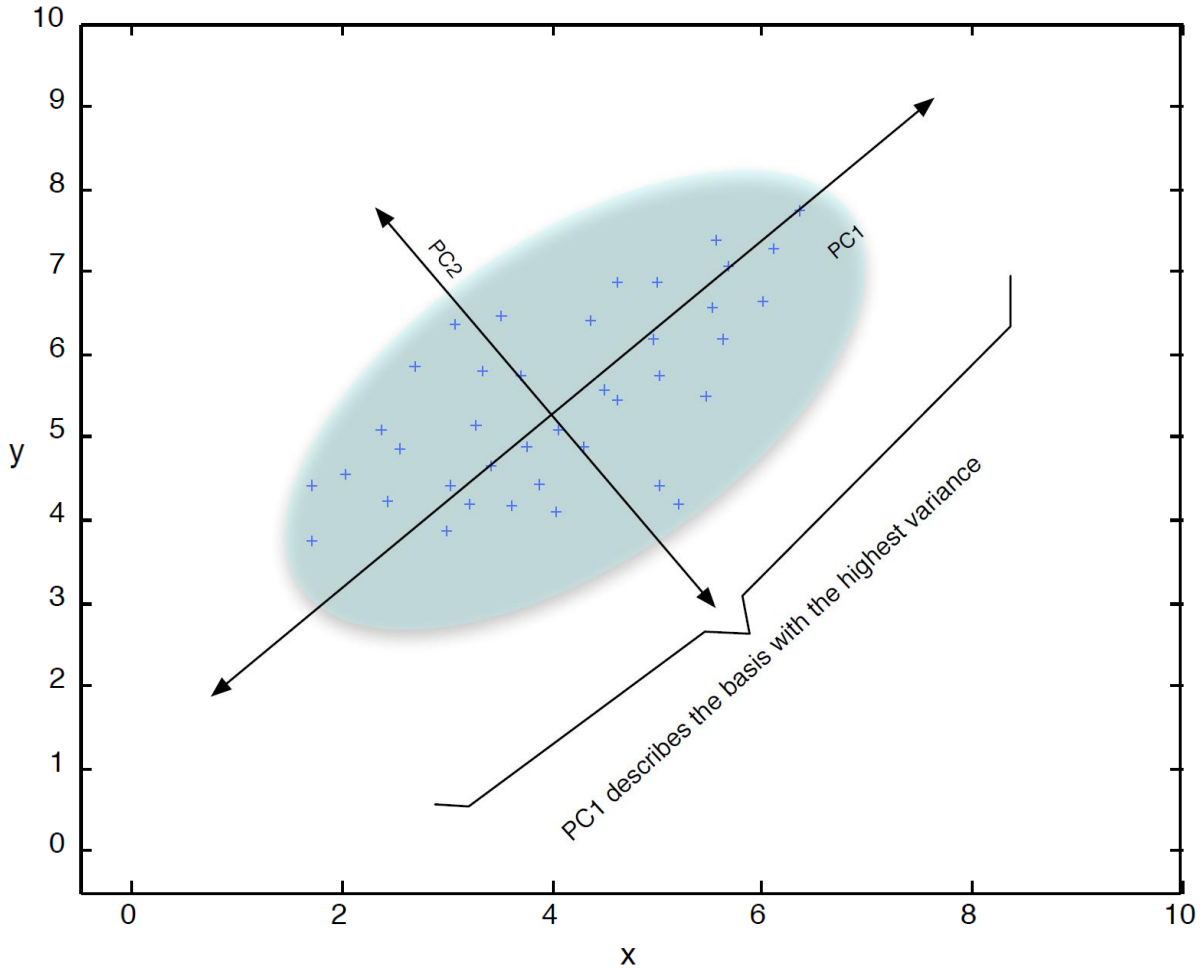
[Shlens 2003]

Principal components analysis



- High dimensional data

Principal components analysis



- High dimensional data
- Find most important axes of variation (e.g. PC1)
  - Maximize signal to noise ratio
    - => principal components
  - Minimize redundancy
    - => orthogonal components

Principal components analysis

- *U* forms a new basis for the data in *X* as a linear combination of the original basis
- *Y* is the projection of the of *X* onto the basis $\{u_1, \ldots, u_M\}$
  - What is the best way to re-express *X*?
  - What is a good choice for *P*?
    - Maximize signal to noise ratio
      - => principal components
    - Minimize redundancy
      - => orthogonal components

$$U^T X = Y$$

$$U^T X = \begin{bmatrix} u_1 \\ \cdots \\ u_M \end{bmatrix} [x_1 \ldots x_M]$$

$$Y = \begin{bmatrix} u_1^T \cdot x_1, \ldots, u_M^T \cdot x_N \\ \vdots \quad \ddots \quad \vdots \\ u_1^T \cdot x_1, \ldots, u_M^T \cdot x_N \end{bmatrix}$$

Covariance and variance

- Empirical covariance of the *M* dimensions of *X*
  - Diagonal entries: variances
    - Measure amount of signal in that dimension
  - Off-diagonal entries: co-variances
    - Measure redundancy between dimensions
- How to find a good *u*?
- Signal to noise ratio is maximized
  - => maximize variances λ^2

$$S_X = \frac{1}{N-1}\sum_{n=1}^{N}[x_n - \bar{x}]\,[x_n - \bar{x}]^T$$

$$S_{Y_1} = \frac{1}{N-1}\sum_{n=1}^{N} u_1^T[x_n - \bar{x}]\,[x_n - \bar{x}]^T u_1$$

$$S_{Y_1} = u_1^T S_X u_1$$

Covariance and variance

- Empirical covariance of the *M* dimensions of *X*
  - Diagonal entries: variances
    - Measure amount of signal in that dimension
  - Off-diagonal entries: co-variances
    - Measure redundancy between dimensions
- How to find a good *u*?
- Signal to noise ratio is maximized
    - => maximize variance

$$S_X = \frac{1}{N-1}\sum_{n=1}^{N}[x_n - \bar{x}]\,[x_n - \bar{x}]^T$$

$$S_{Y_1} = \frac{1}{N-1}\sum_{n=1}^{N} u_1^T[x_n - \bar{x}]\,[x_n - \bar{x}]^T u_1$$

$$S_{Y_1} = u_1^T S_X u_1$$

Finding an optimal u_1

- Maximize variance $S_{Y_1} = u_1^T S_X u_1$
- Under the constraint $u_1^T u_1 = 1$
  - λ_1*: Lagrange multiplier* enforcing constraint
- Set to zero
- It follows:
  - u_1 is an eigenvector of S_X
  - Variance of Y_1 is equal to the eigenvalue λ_1
- Variance of Y_1 is maximized if we chose the eigenvector with largest eigenvalue!

$$u_1^T S_X u_1 + \lambda_1 (1 - u_1^T u_1)$$

$$\frac{\nabla}{\nabla u_1} u_1^T S_X u_1 + \lambda_1 (1 - u_1^T u_1) = S_X u_1 - \lambda_1 u_1$$

$$S_X u_1 - \lambda_1 u_1 = 0$$

$$S_X u_1 = \lambda_1 u_1$$

$$u_1^T S_X u_1 = \lambda_1$$

Finding an optimal u_1 to u_M

- Empirical covariance of the *M* dimensions of *X*
  - Diagonal entries: variances
    - Measure amount of signal in that dimension
  - Off-diagonal entries: co-variances
    - Measure redundancy between dimensions
- How to find a good *u*?
- Signal to noise ratio is maximized
    - => maximize variance
- **Redundancy is minimized**
  - **=>covariances = 0**
  - Eigenvectors are orthogonal $u_i^T u_j = \delta(i,j)$
    $\delta(i,j)$ =(1 if i=j, 0 otherwise)
  - => u_1 to u_M are eigenvectors corresponding to largest M eigenvalues $\lambda_1, \dots, \lambda_M$

$$S_X = \frac{1}{N-1}\sum_{n=1}^{N}[x_n - \bar{x}]\,[x_n - \bar{x}]^T$$

$$S_{Y_1} = \frac{1}{N-1}\sum_{n=1}^{N} u_1^T[x_n - \bar{x}]\,[x_n - \bar{x}]^T u_1$$

$$S_{Y_1} = u_1^T S_X u_1$$

$$S_Y = U^T S_X U = diag([\lambda_1, \dots, \lambda_M])$$

Equivalent formulation: Minimizing the squared reconstruction error

- If $M = D$ principal components are used, then $\{u_1, \dots, u_D\}$ form a complete orthogonal ($u_i^T u_i = \delta(i,j)$) basis of the D-dim space.
- For *M=D* x_n can exactly be represented by u_i.
- For *M<D* x_n can only be approximated by reconstruction $\tilde{x}$.

$$x_n = \sum_{i=1}^{D} \alpha_{ni} u_i = \sum_{i=1}^{D} (x_n u_i^T) u_i$$

$$\tilde{x}_n = \sum_{i=1}^{M} (x_n u_i^T) u_i + \bar{x}$$

- Minimizing the squared error

  (= Frobenius norm)

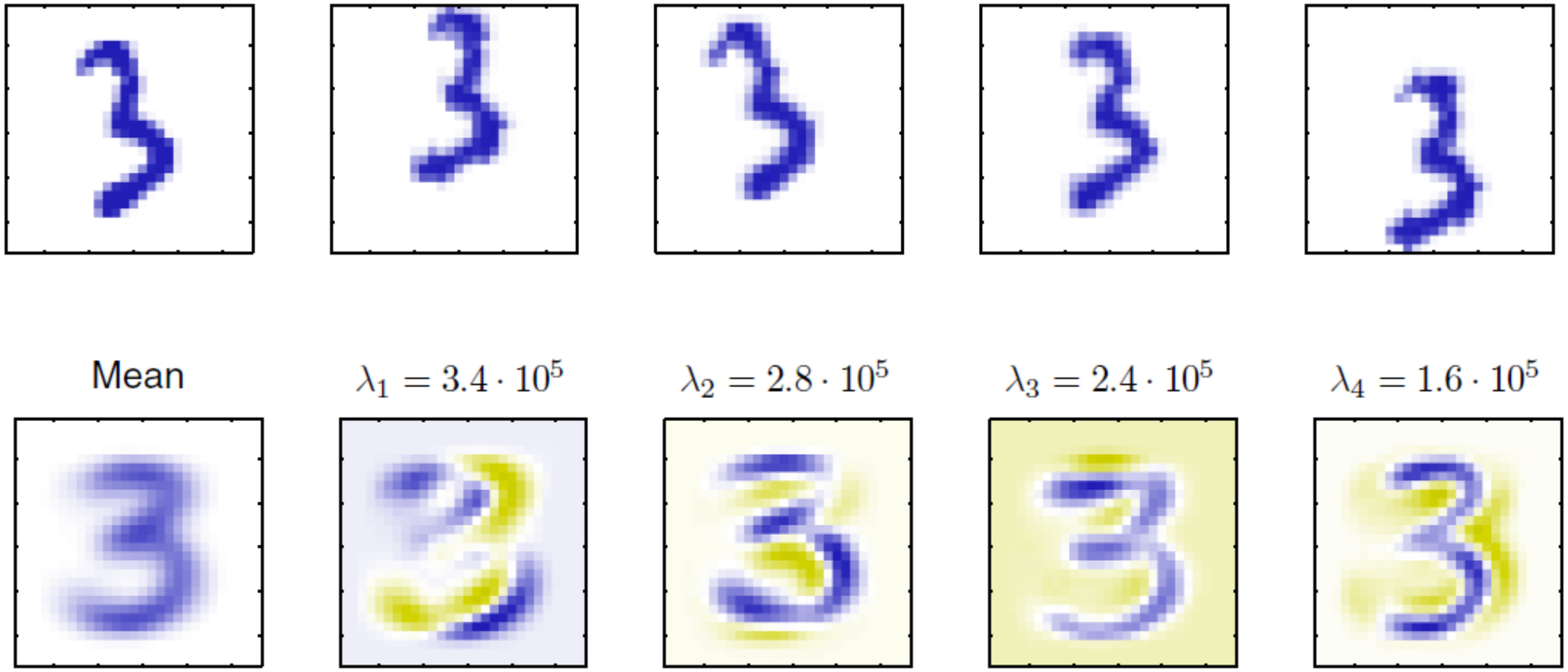
  - equivalent to minimizing the residual variances

    =>equivalent to PCA

$$J = \frac{1}{N-1} \sum_{n=1}^{N} |x_n - \tilde{x}_n|^2 = \sum_{i=M+1}^{D} u_i^T S_X u_i^T$$

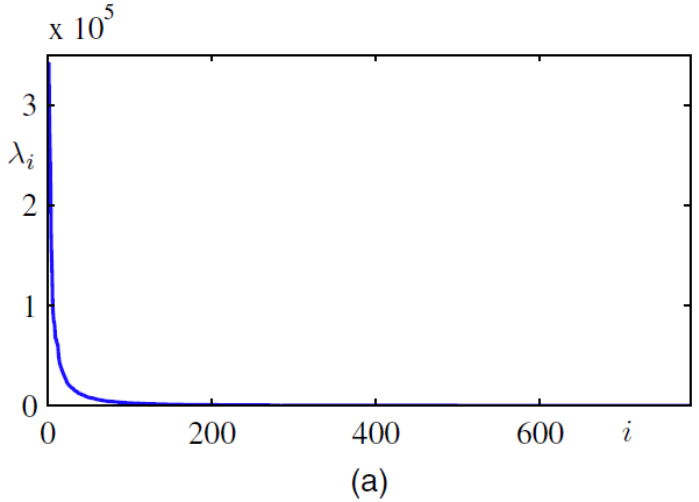
$$J = \frac{1}{N-1} \sum_{n=1}^{N} \sum_{d=1}^{D} (x_{nd} - \tilde{x}_{nd})^2 = \sum_{i=M+1}^{D} \lambda_i^2$$

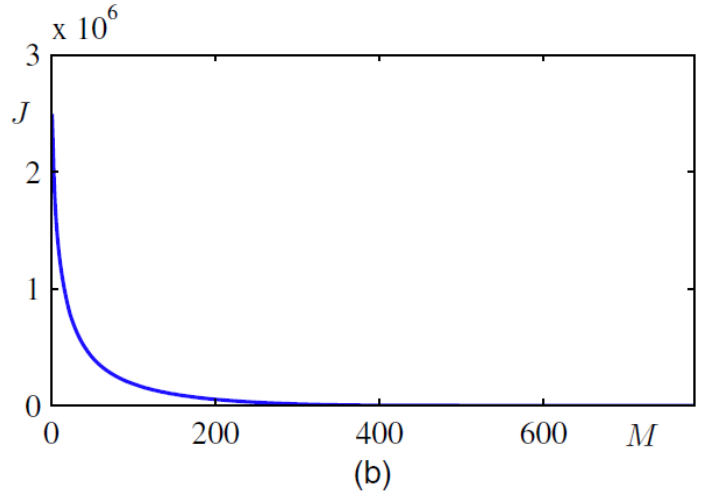
Digits example [Bishop 2006]

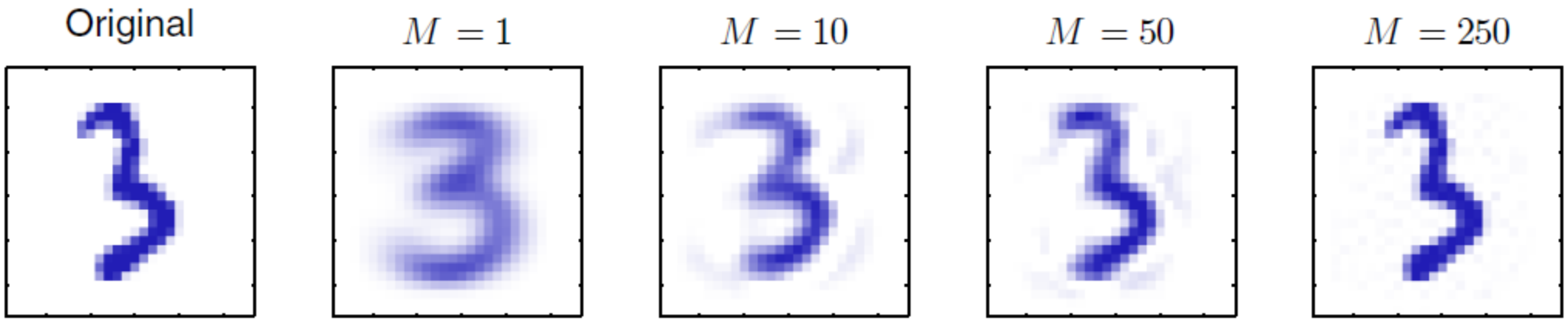


Digit reconstruction [Bishop 2006]

(a) Eigenvalues vs. rank
(b) Sum of discarded eigenvalues vs. rank

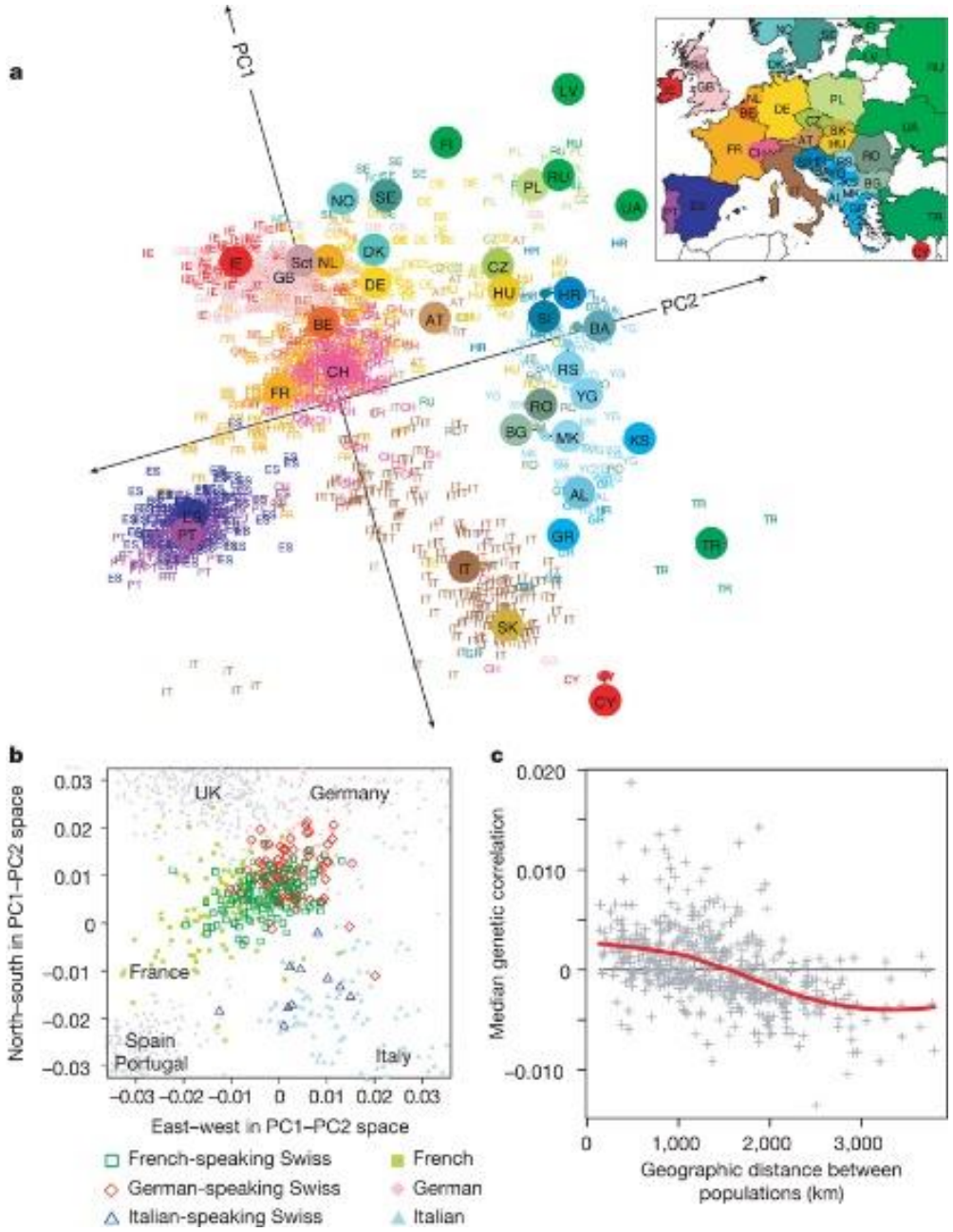






PCA in genetics

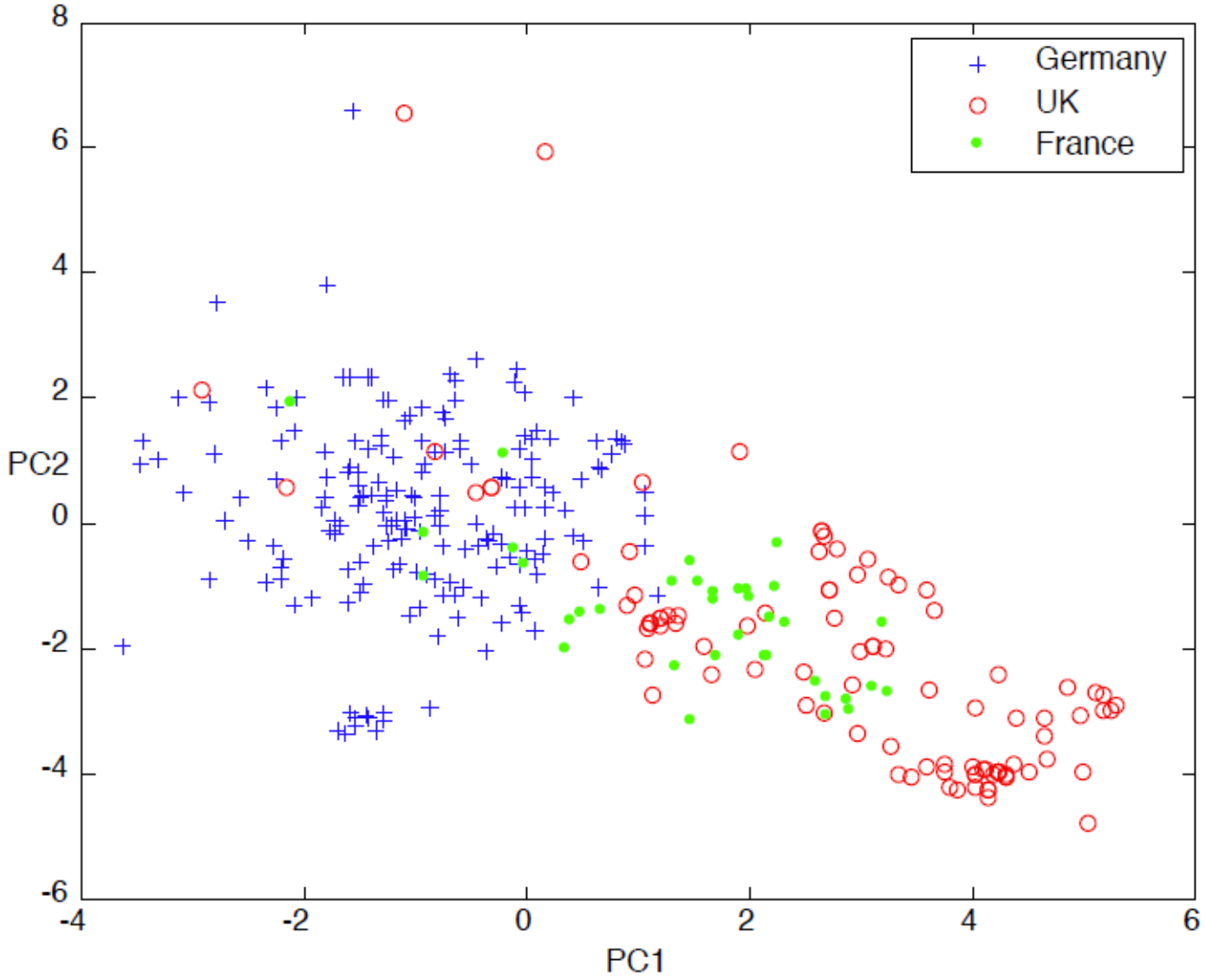
- Population structure causes genome-wide correlations between SNPs
- A large part of the total variation in the SNPs can be explained by population differences.
  - PCA represents population structure on a continuous scale (admixture)



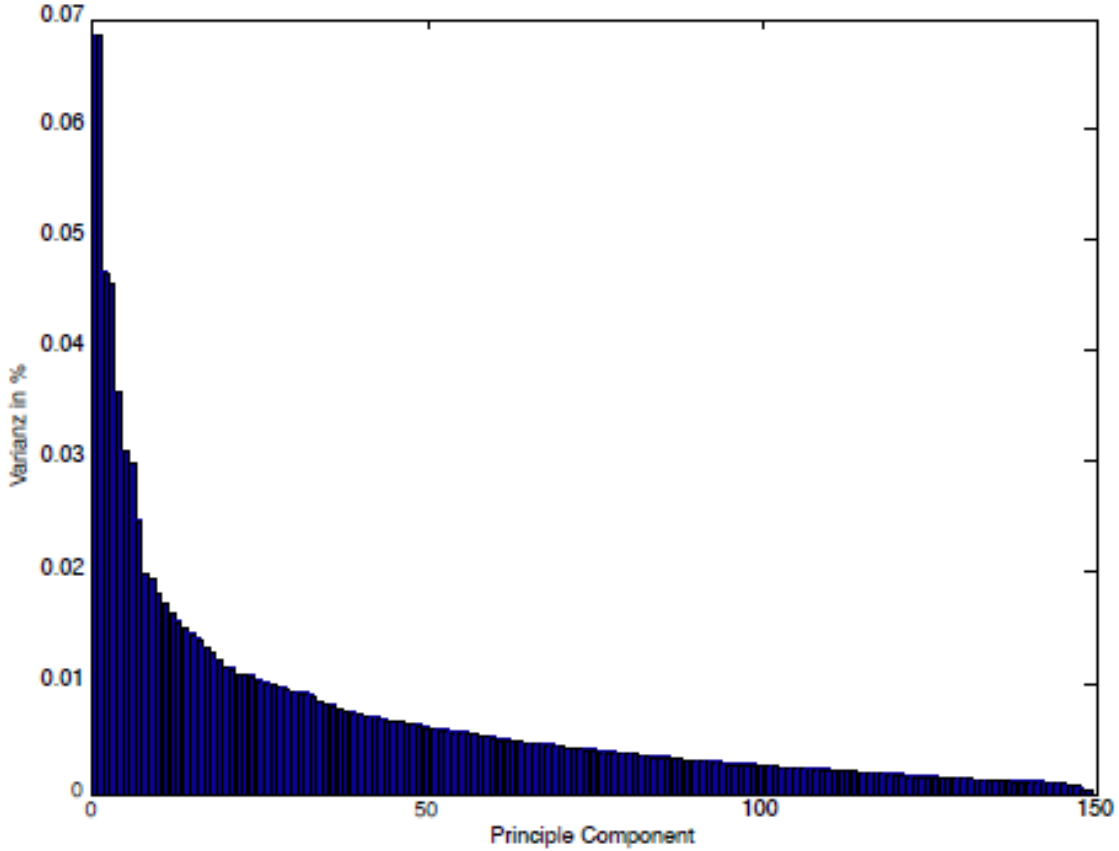
J. Novembre and T. Johnson et al. Genes mirror geography within Europe. Nature, 456(98-103), 2008.

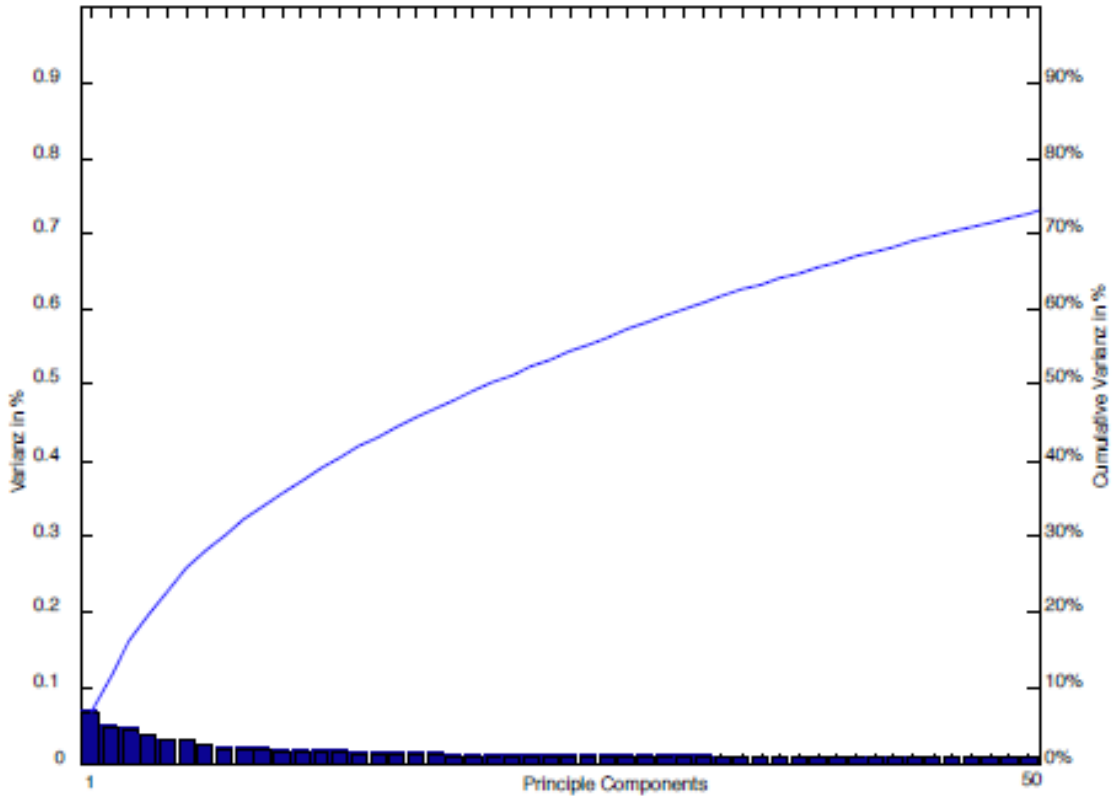
PCA on *A. thaliana*

- Stockori data
- 149 genotypes for 697 plants
- Country:
  - sampling origin



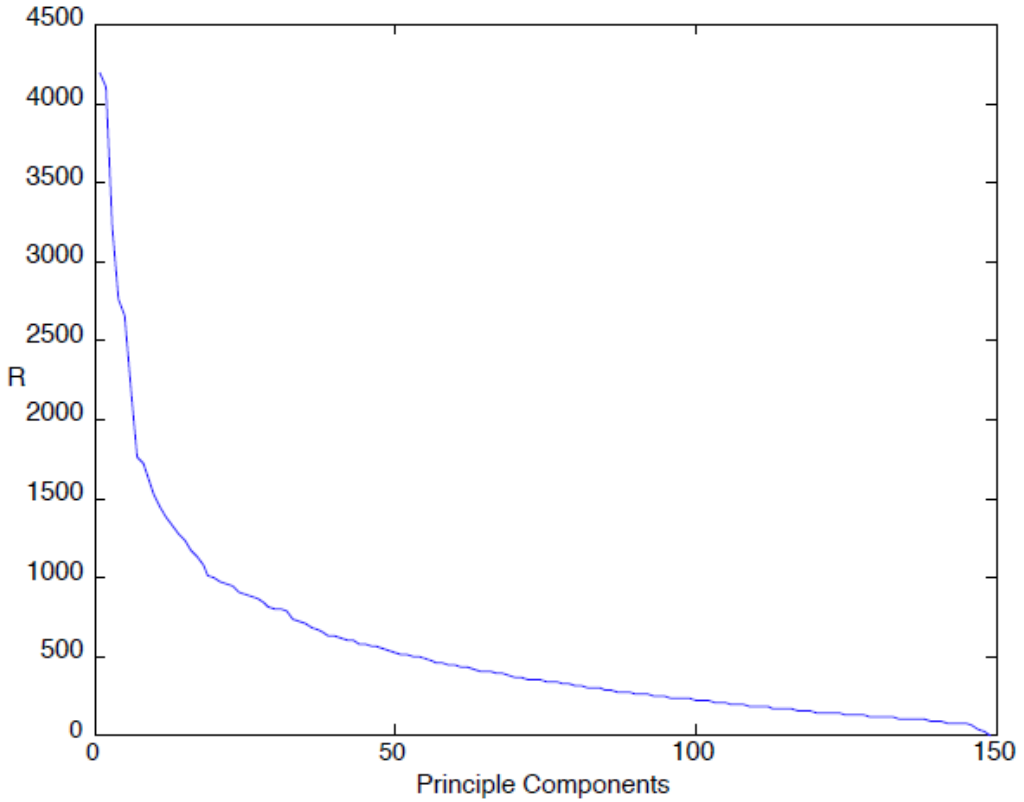
PCA on *A. thaliana* - Variances





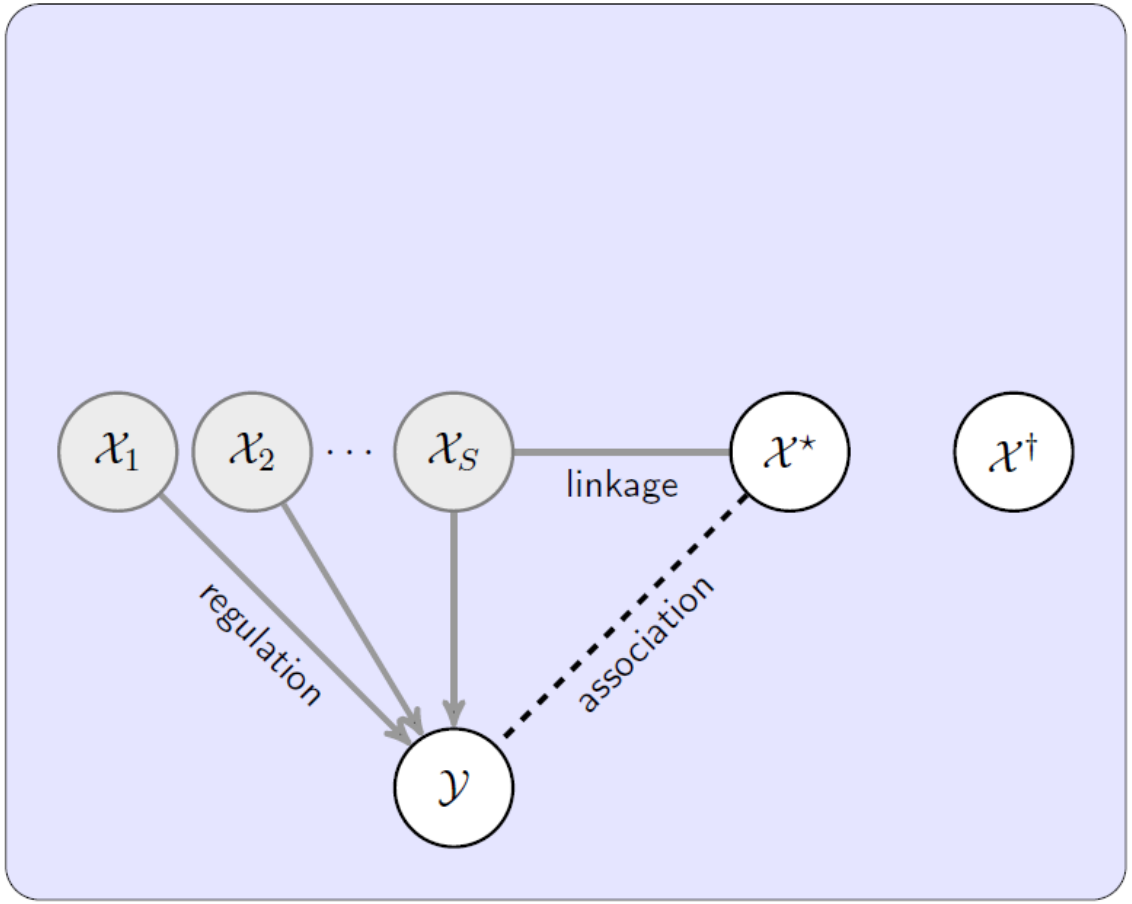
PCA on *A. thaliana* - reconstruction error

$R = |x - \hat{x}_d|^2$ where x is the original data and $\hat{x}$ is the reconstructed data using d principle components



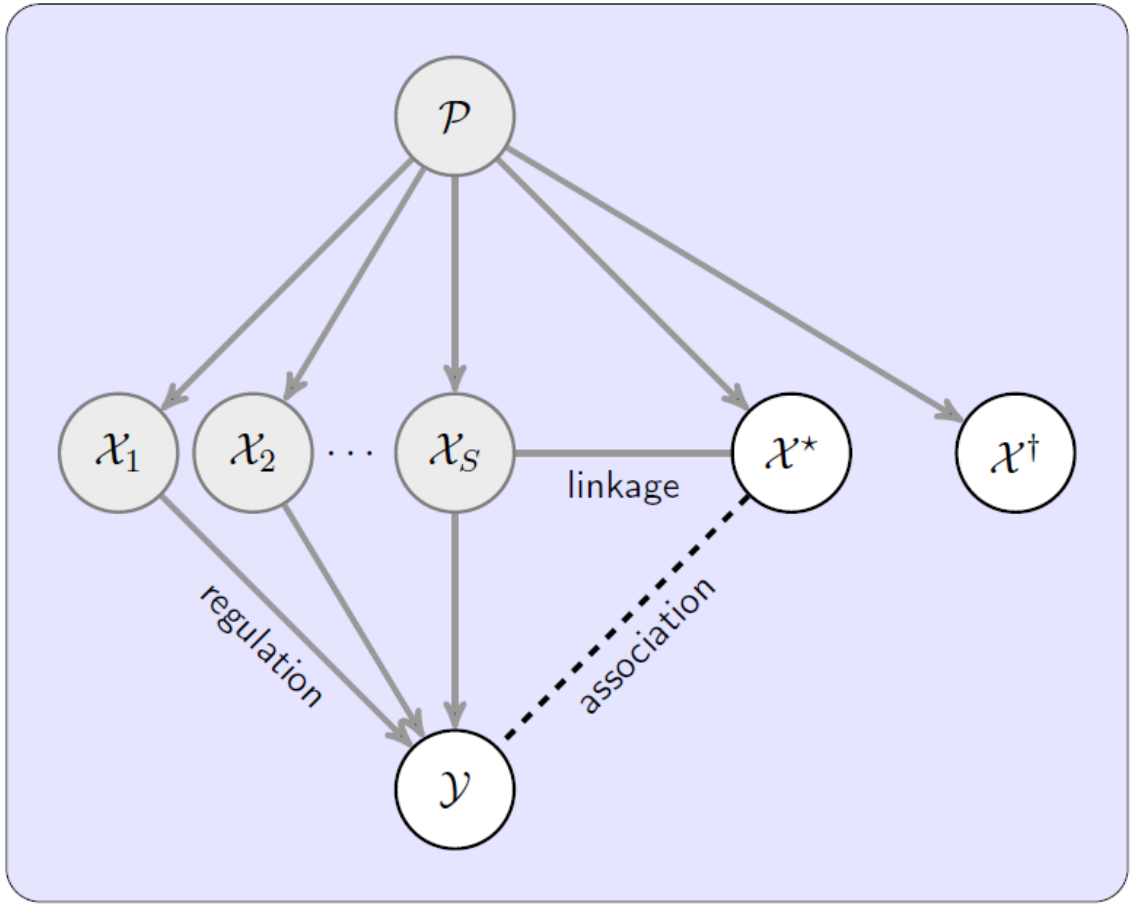
How about GWAS?

- Linkage allows to test for associations between phenotype and genetic markers



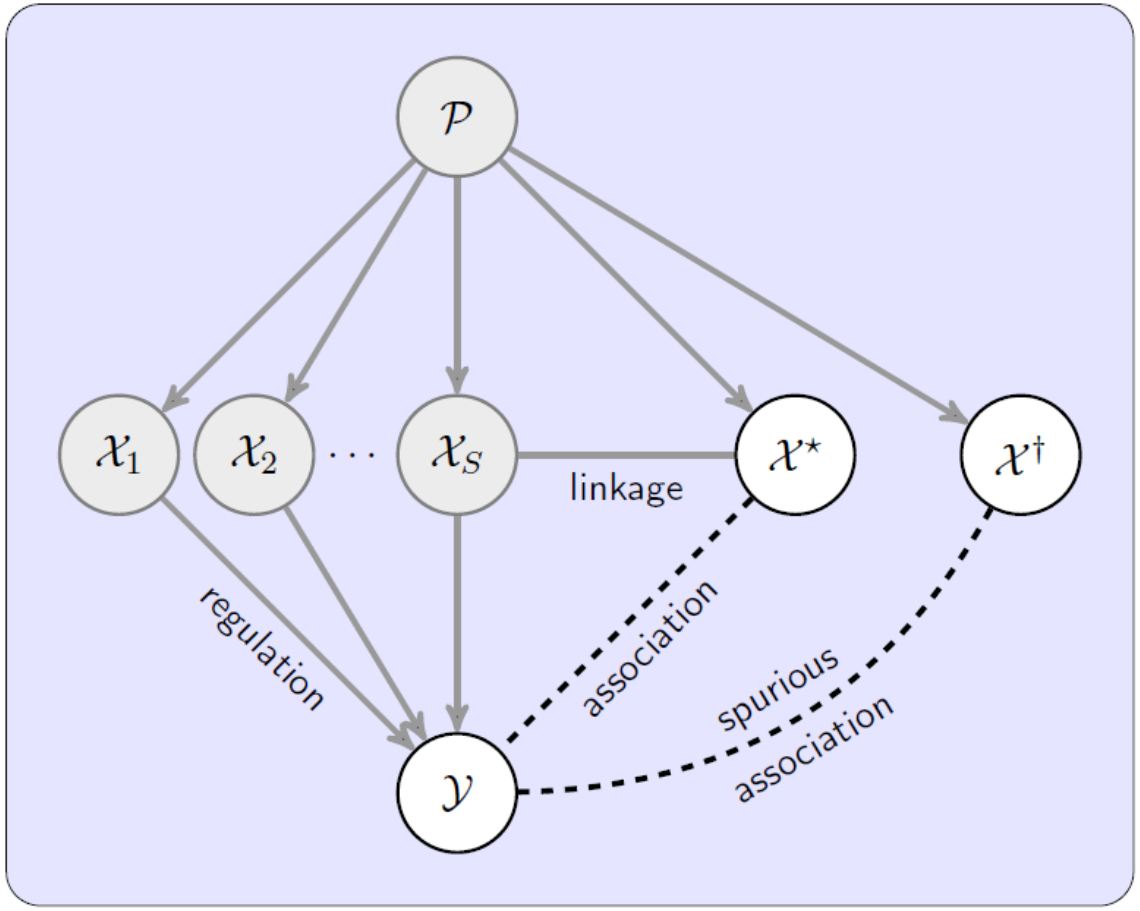
How about GWAS?

- Linkage allows to test for associations between phenotype and genetic markers
- Hidden population structure causes correlations between SNPs



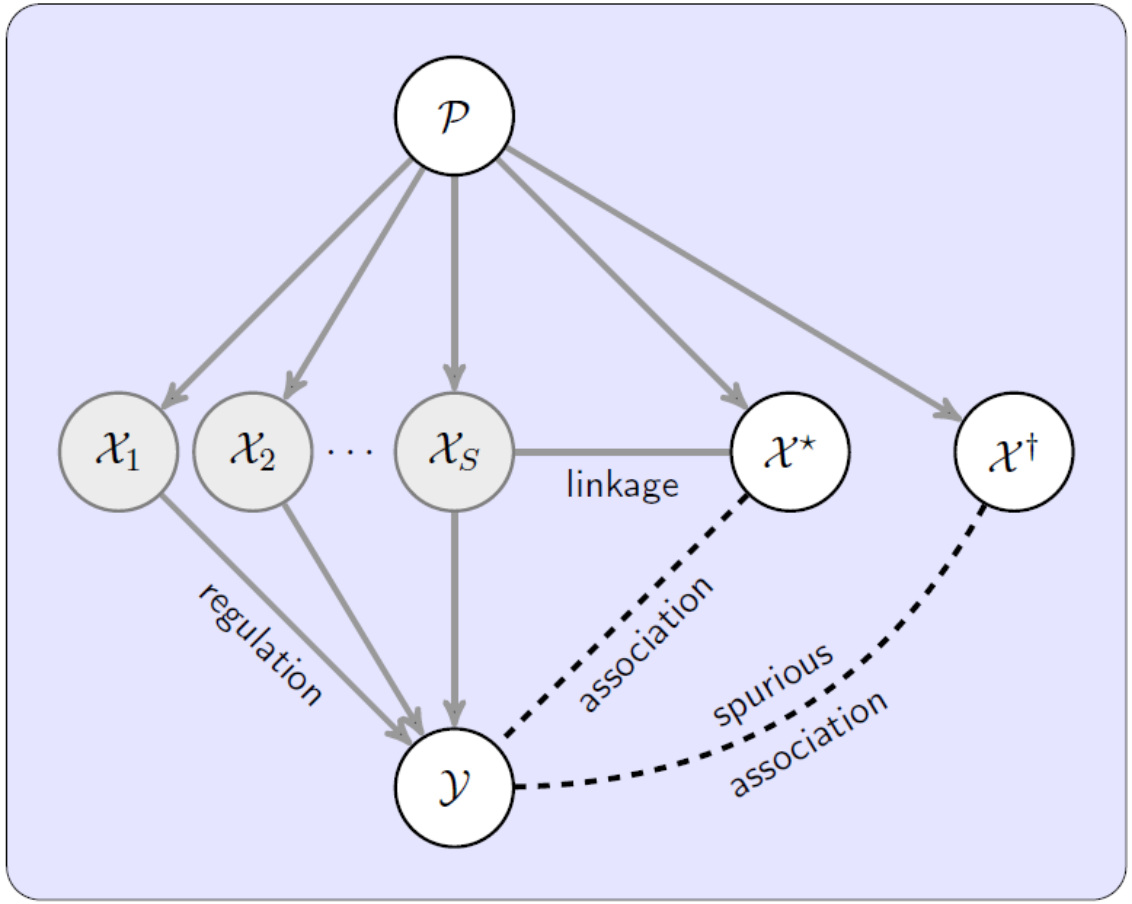
How about GWAS?

- Linkage allows to test for associations between phenotype and genetic markers
- Hidden population structure causes correlations between SNPs
- Causing associations to non-linked SNPs



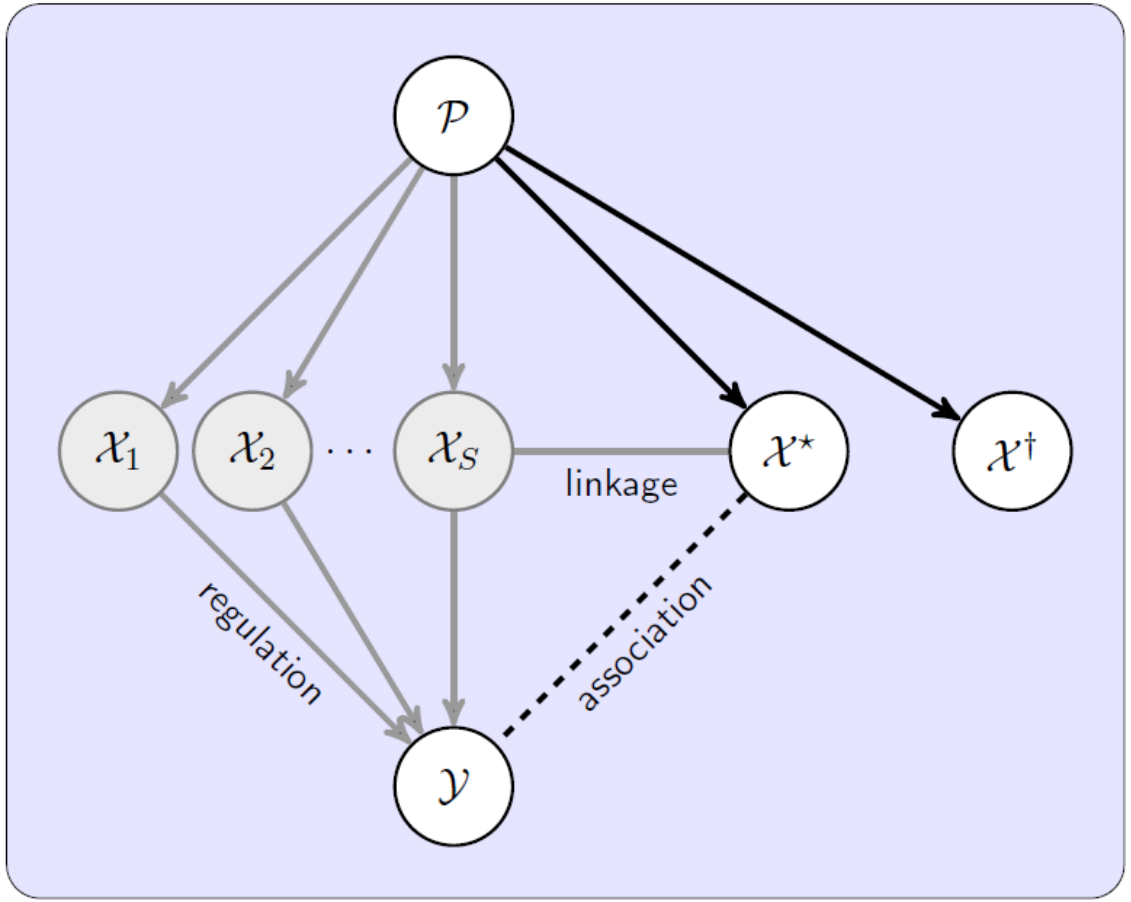
How about GWAS?

- Linkage allows to test for associations between phenotype and genetic markers
- Hidden population structure causes correlations between SNPs
- Causing associations to non-linked SNPs
- Take population structure into account



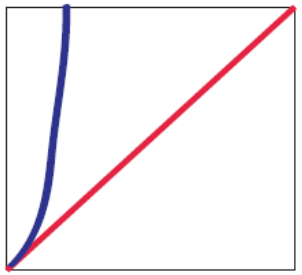
How about GWAS?

- Linkage allows to test for associations between phenotype and genetic markers
- Hidden population structure causes correlations between SNPs
- Causing associations to non-linked SNPs
- Take population structure into account
- Use PCA!



Eigenstrat [Price et al 2006]

$$\mathcal{N}\left(\mathbf{y} \,\middle|\, \underbrace{\begin{matrix}C\\C\\C\\T\\T\\T\end{matrix}}_{\mathbf{X}}\beta\,;\,\sigma^2\underbrace{\begin{matrix}\blacksquare&&&\\&\blacksquare&&\\&&\blacksquare&\\&&&\blacksquare\end{matrix}}_{\mathbf{I}}\right)$$



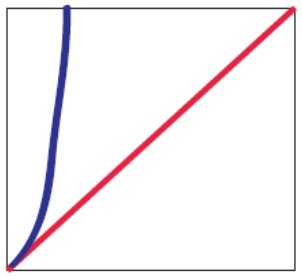
Eigenstrat [Price et al 2006]

- Compute covariance from SNPs

$$\frac{1}{S}\overbrace{\mathbf{X}-\mathbb{E}[\mathbf{X}]}^{S}\times\overbrace{\mathbf{X}^{\mathrm{T}}-\mathbb{E}[\mathbf{X}]^{\mathrm{T}}}^{N}$$

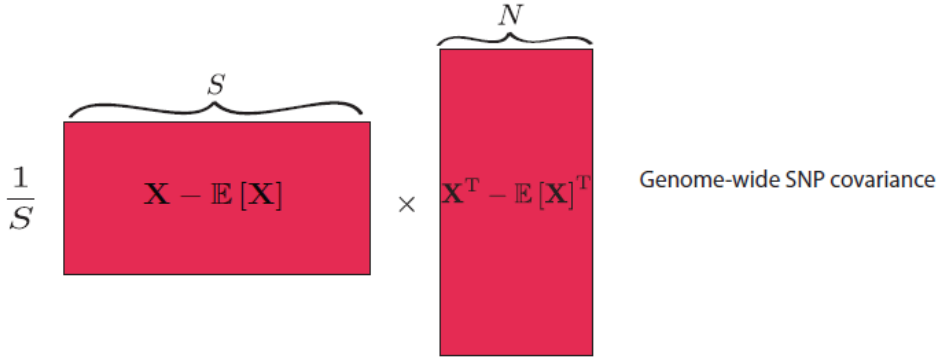
Genome-wide SNP covariance

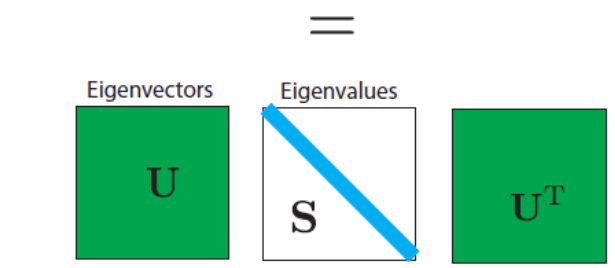
$$\mathcal{N}\left(\mathbf{y}\,\middle|\,\underbrace{\begin{matrix}C\\C\\C\\T\\T\\T\end{matrix}}_{\mathbf{X}}\beta\;;\;\sigma^2\underbrace{\begin{matrix}\blacksquare&&\\&\blacksquare&\\&&\blacksquare\end{matrix}}_{\mathbf{I}}\right)$$

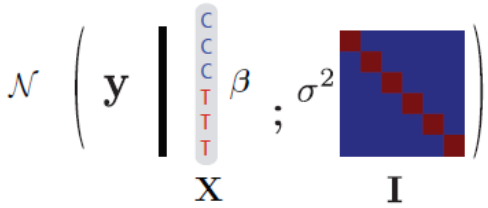


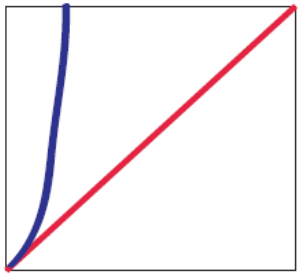
Eigenstrat [Price et al 2006]

- Compute covariance from SNPs
- Compute spectral decomposition



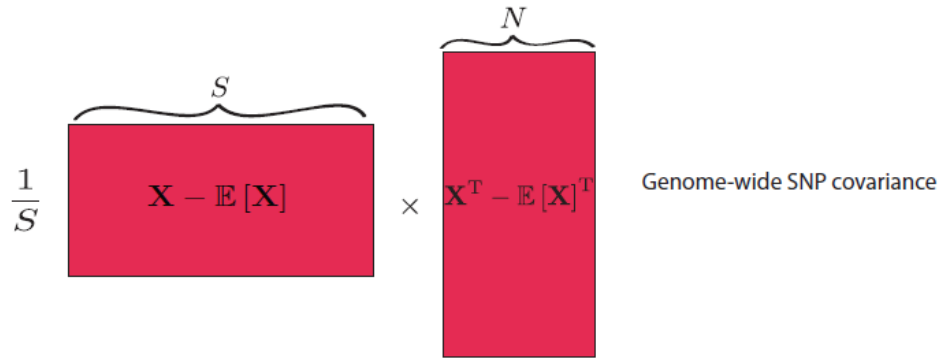


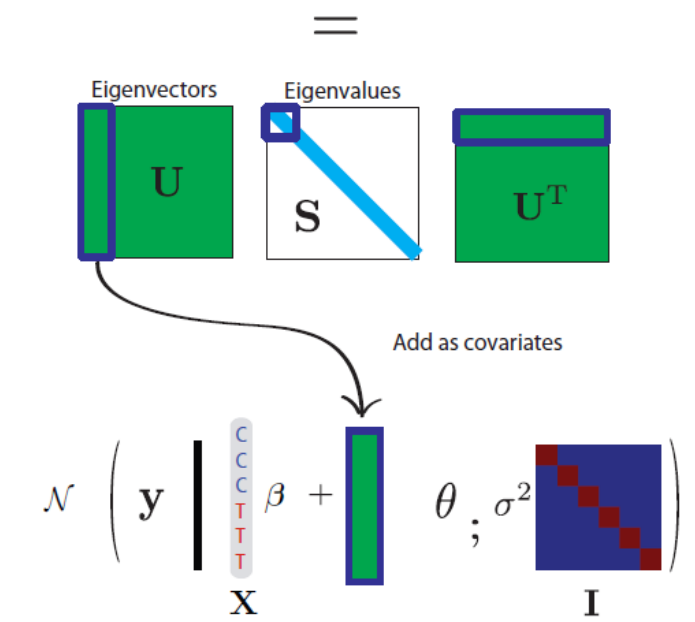


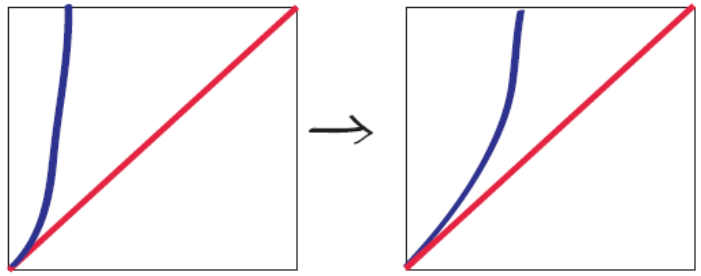


Eigenstrat [Price et al 2006]

- Compute covariance from SNPs
- Compute spectral decomposition
- Add PC with largest eigenvalue to model

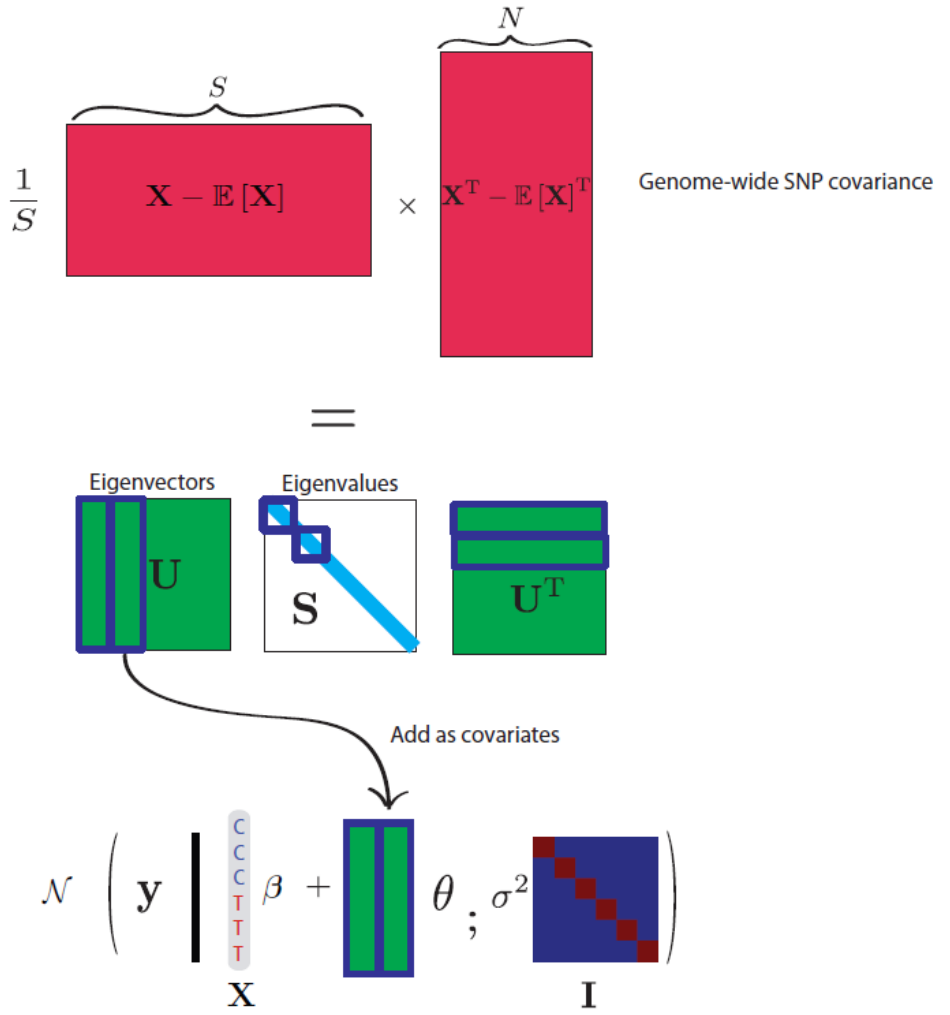


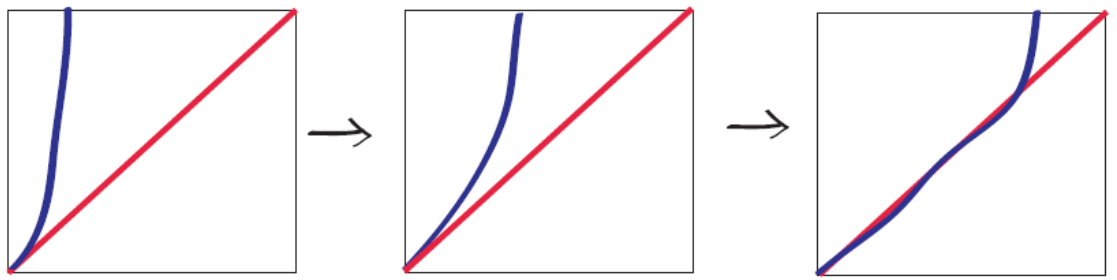




Eigenstrat [Price et al 2006]

- Compute covariance from SNPs
- Compute spectral decomposition
- Add PC with largest eigenvalue to model
- Iterate.
- **Note:**
  - PCA corrects well for population structure
  - But: cannot correct for relatedness/family structure
  - Can be combined with LMMs (sometimes useful!)





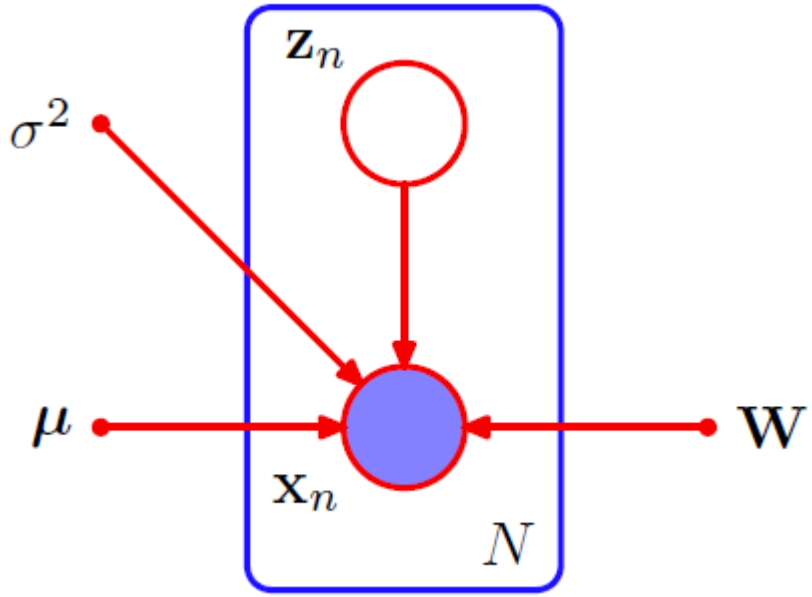
Probabilistic PCA [Tipping & Bishop 1999]

- Only mean $E[x]$ and co-variance matters
- Minimizing squared error

$$J = \frac{1}{N-1}\sum_{n=1}^{N}\sum_{d=1}^{D}(x_{nd} - \tilde{x}_{nd})^2 = \sum_{i=M+1}^{D}\lambda_i^2$$

=>Gaussian noise model

- bi-linear Gaussian model
  - $x_n \sim N(\mu + Wz_n, \sigma^2 I_M)$
  - z_n *hidden* variables (principal component)
  - μ, W, σ^2 parameters



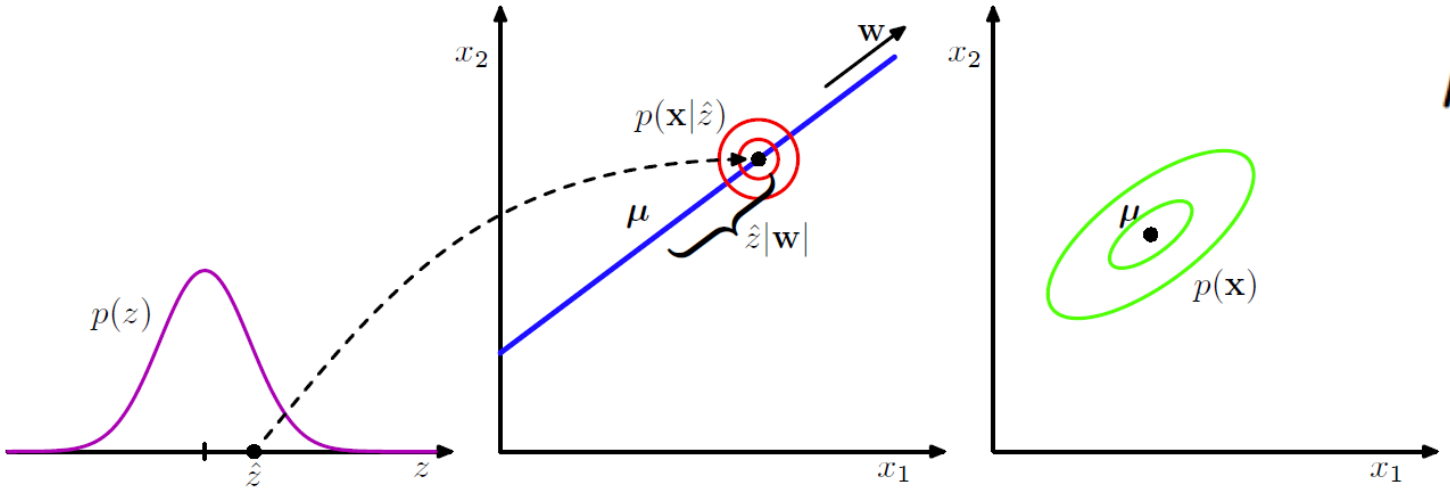
[Bishop 2006]

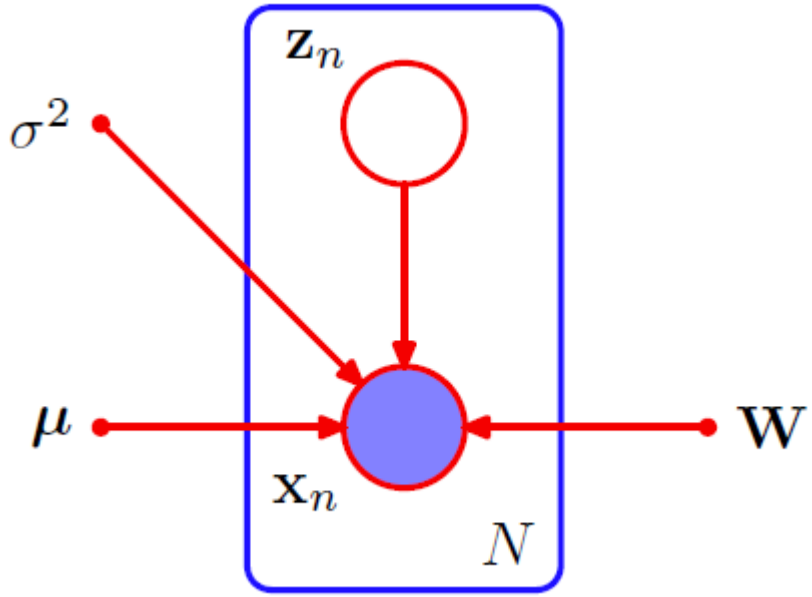
Generative process

Sample $z_n \sim p(z) = N(0, I)$

Sample $x_n \sim p(x|z) = N(\mu + Wz_n, \sigma^2 I)$

$$p(x) = \int p(x|z)p(z)\ \mathrm{d}z$$
$$= N(x|\mu, WW^T + \sigma^2 I)$$





[Bishop 2006]