Principal Components Analysis

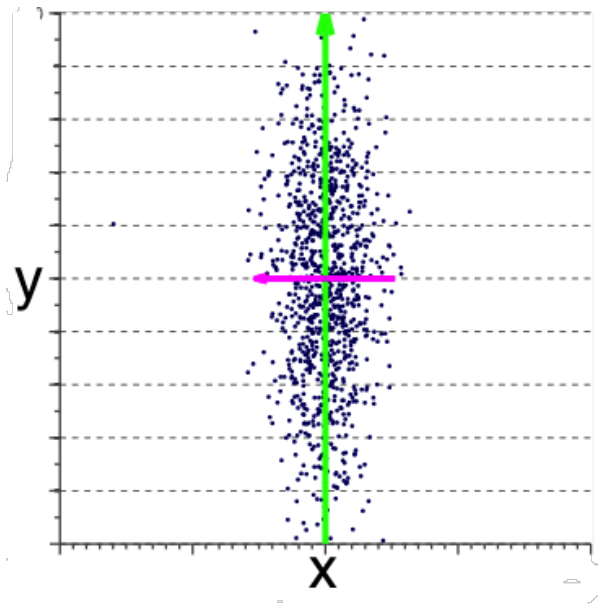
David Benjamin, Broad DSDE Methods

February 10, 2016

What is PCA?

PCA turns high-dimensional data into low-dimensional data by throwing out directions with low variance.

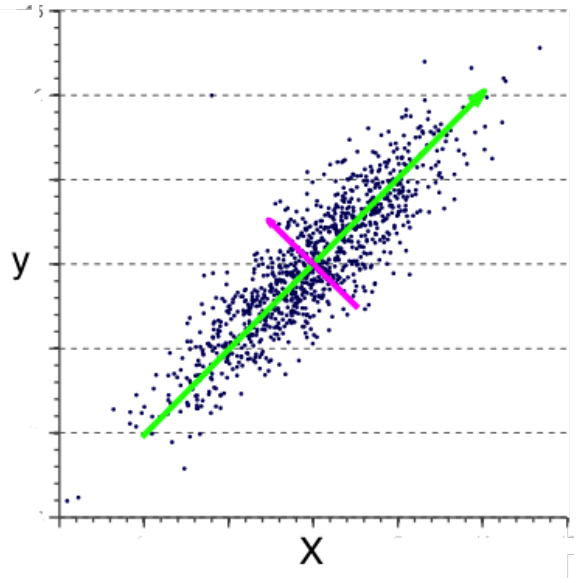
- Keep y, throw out x.
- Assumption: noise smaller than signal.



What about correlations?

PCA turns high-dimensional data into low-dimensional data by throwing out directions with low variance.

- Find the pink and green axes.
- Throw out the pink component.
- Resulting low-dimensional data is projection onto green axis.



Covariance matrix

$\Sigma_{ij} = \frac{1}{N}\sum_n (x_{ni} - \mu_i)(x_{nj} - \mu_j) \neq 0$ if x_i and x_j are correlated.

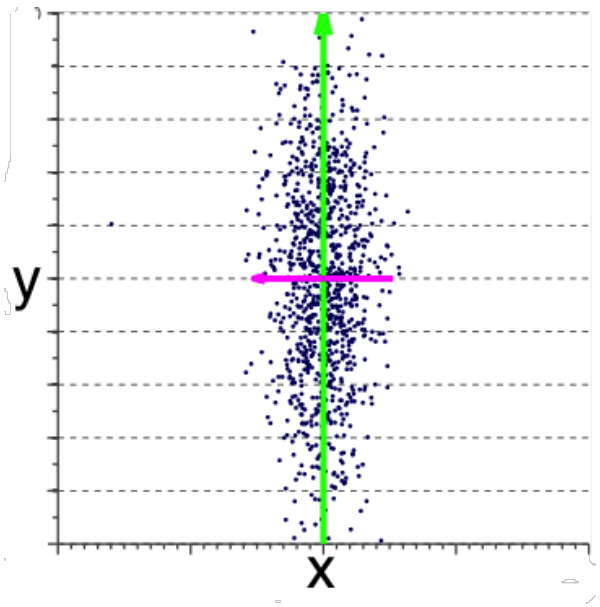


Figure: $\Sigma = \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix}$

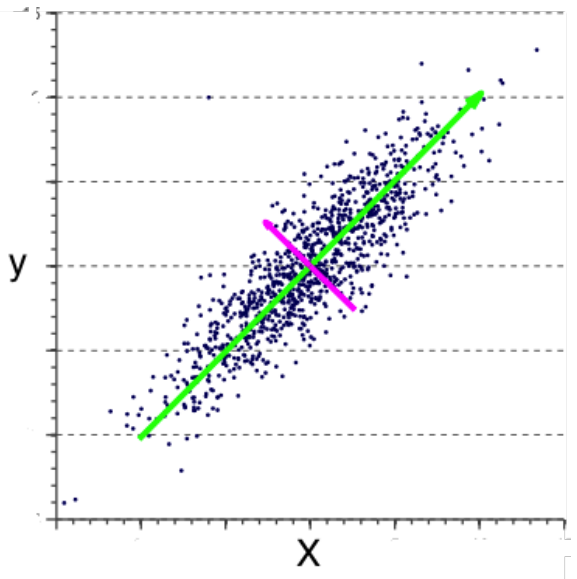


Figure:
$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} > 0 \\ \Sigma_{xy} > 0 & \Sigma_{yy} \end{pmatrix}$

We want coordinates that make Σ diagonal.

PCA recipe

Coordinates (principal components) that make Σ diagonal are the eigenvectors of Σ.

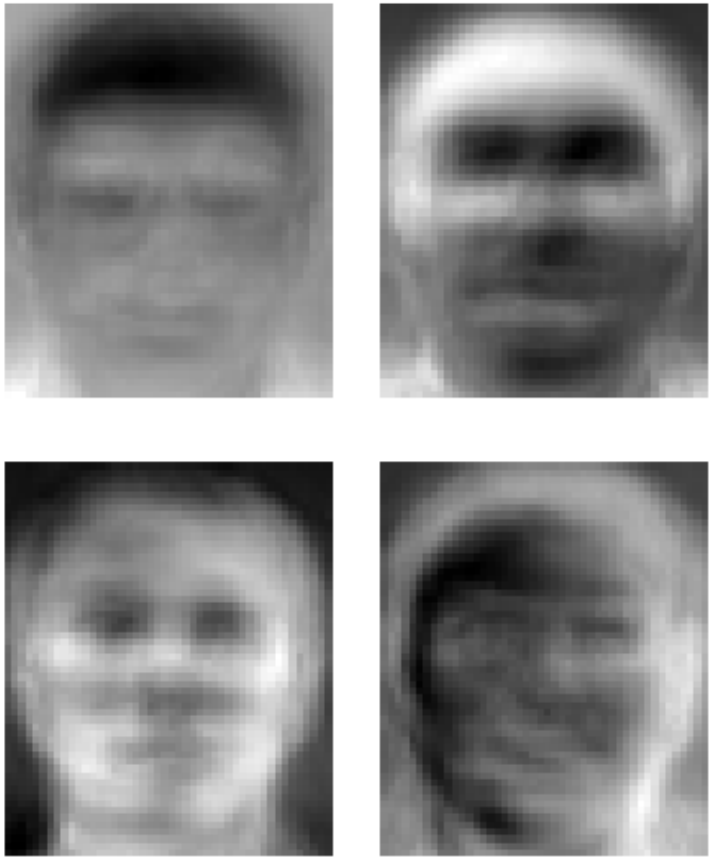
PCA recipe

- Calculate covariance matrix Σ.
- Find eigenvectors $\mathbf{v}$ and eigenvalues λ such that $\Sigma \mathbf{v}_k = \lambda_k \mathbf{v}_k$.
- λ_k is the variance in the $\mathbf{k}_k$ direction.
- Use heuristic to choose K eigenvectors to keep.
- Data is now K-dimensional: $\mathbf{x} \approx \mu + \sum_{k=1}^{K} c_k \mathbf{v}_k$,

  $c_k = (\mathbf{x} - \mu) \cdot \mathbf{v_k}$
- Generative model: $\mathbf{x} = \mu + \sum_{k=1}^{K} c_k \mathbf{v}_k + \text{noise}$

Eigenfaces

Pixel images are very high-dimensional vectors. Run PCA and look at the principal components. . .



Not strictly "eigenfaces," but eigen-variation in faces relative to average face.

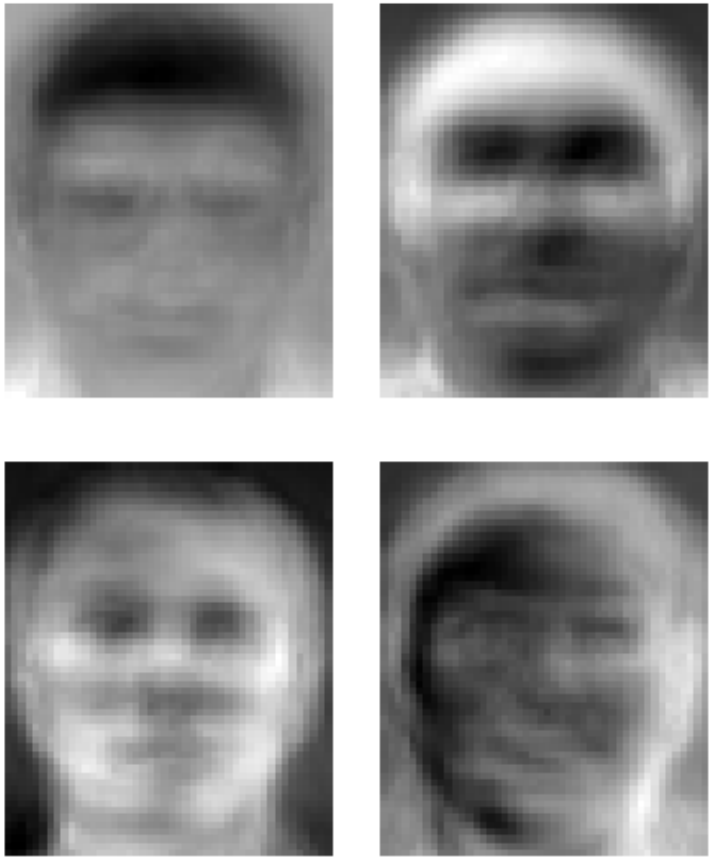
Eigenfaces

Pixel images are very high-dimensional vectors. Run PCA and look at the principal components. . .

Clockwise from top left

- full head of hair
- sunken eyes
- war paint
- your interpretation goes here

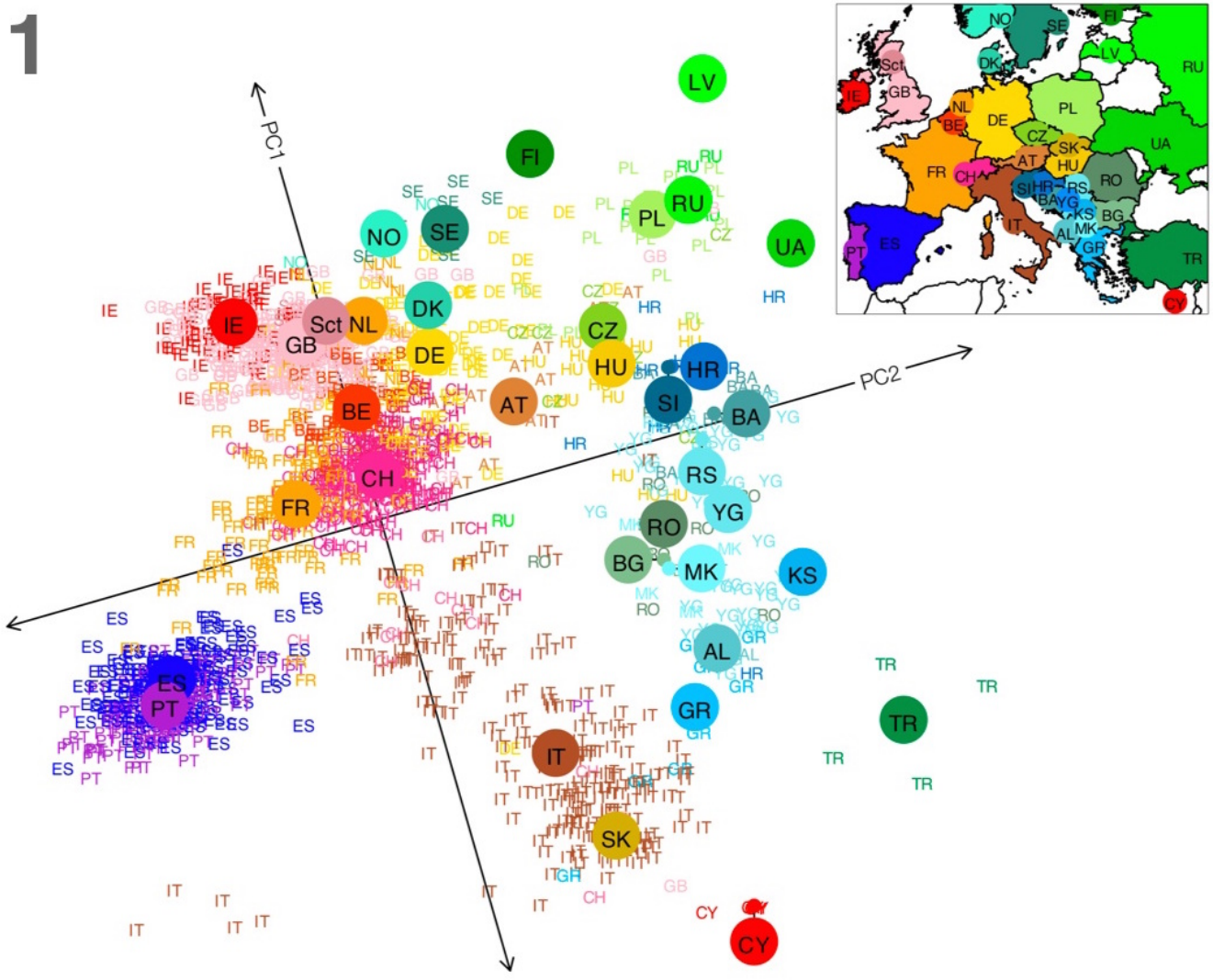
Not strictly "eigenfaces," but eigen-variation in faces relative to average face.



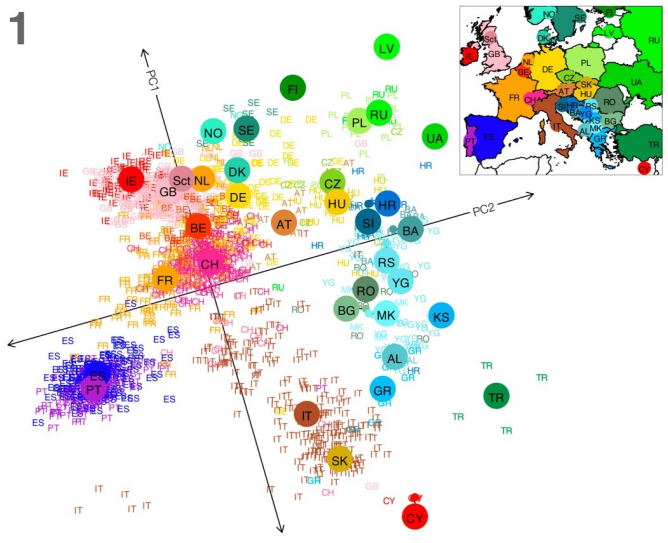
Eigenfaces

PCA map of Europe

Data: x_{ni} = genotype (0, 1, 2) of SNP i in person n.



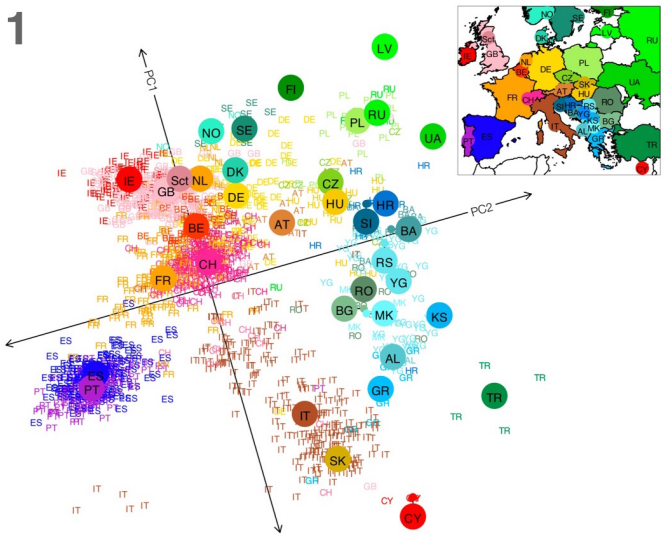
PCA map of Europe



Applications

- Classification / geneaology
- Population stratification in GWAS (regress against PCs)

PCA map of Europe



Applications

- Classification / geneaology
- Population stratification in GWAS (regress against PCs)

- Do the PCs correspond to the map *suspiciously* well?
- Why do the genes of a population migrating north keep going straight along the first PC?
- Why is Hungary - Austria parallel to Switzerland - France?

Copy number variation from exome capture

crash course in exome capture

- get DNA
- exon DNA hybridizes to baits, throw out remaining DNA
- sequence exon DNA

Copy number variation from exome capture

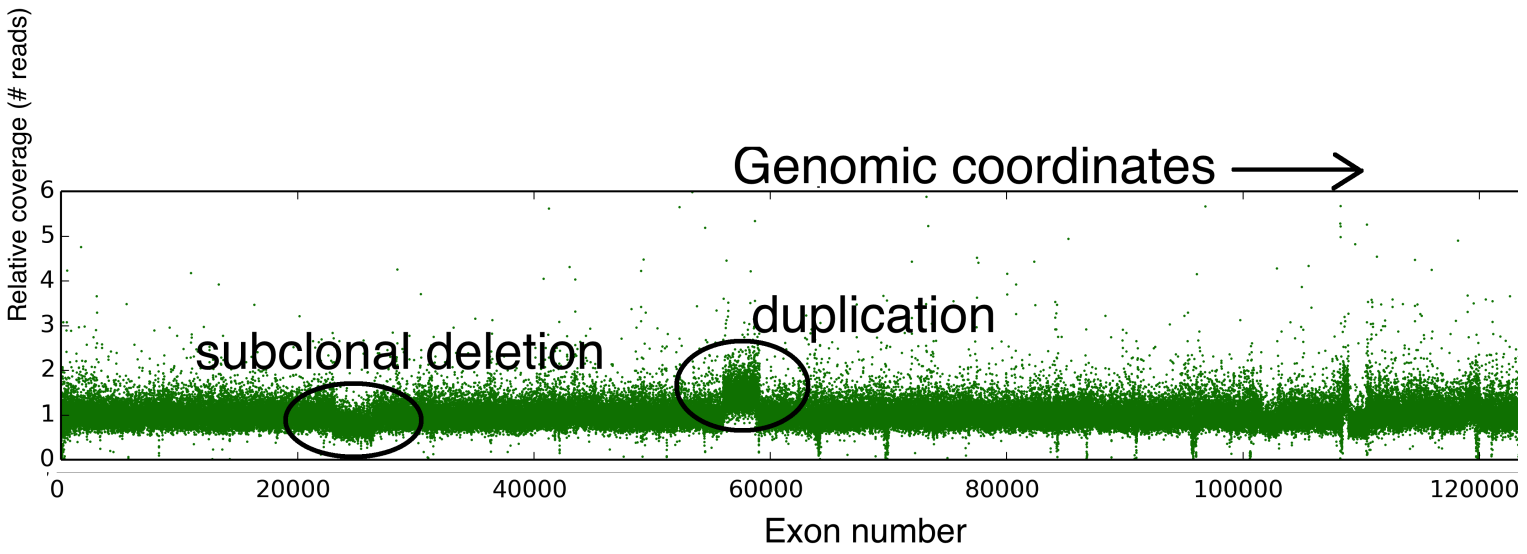
crash course in exome capture

- get DNA
- exon DNA hybridizes to baits, throw out remaining DNA
- sequence exon DNA

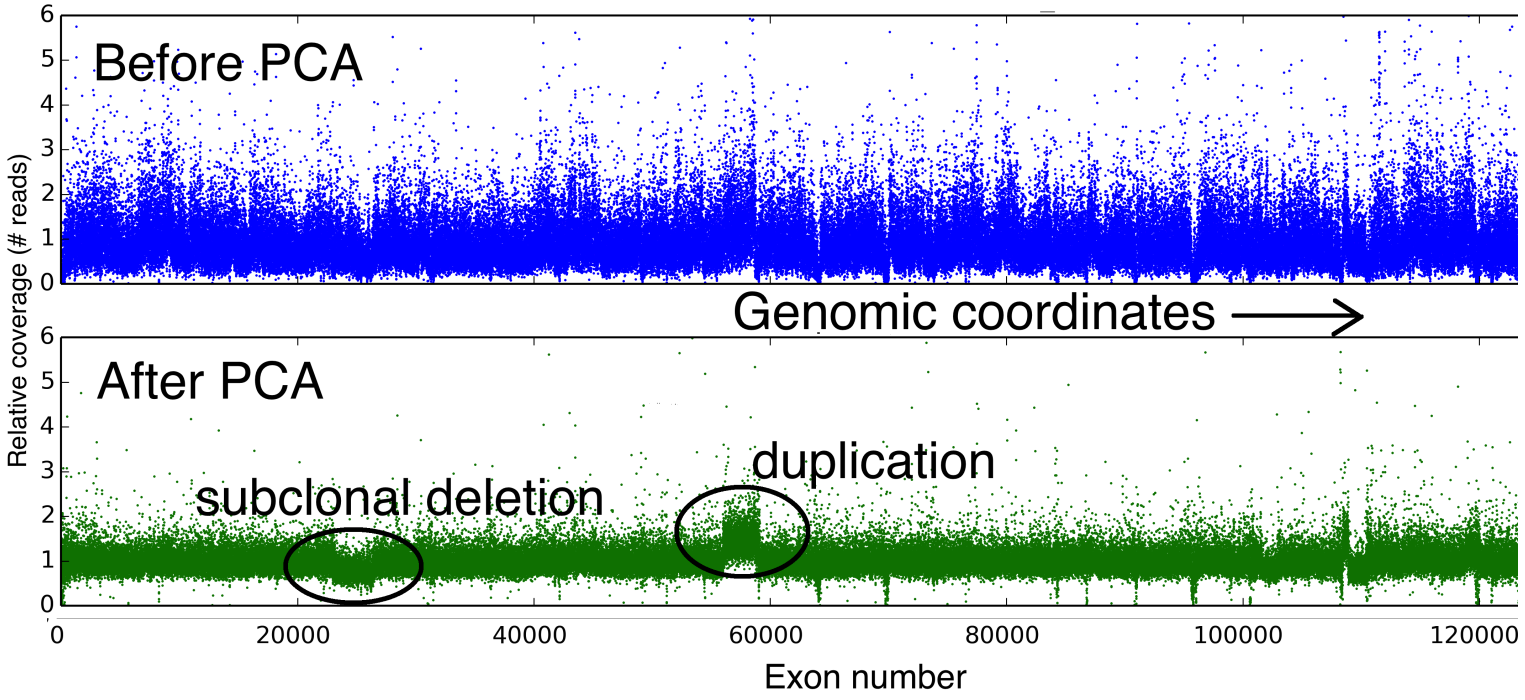
copy number variation

- align sequenced DNA to reference genome
- count number of reads from each exon
- more (less) reads implies duplication (deletion)

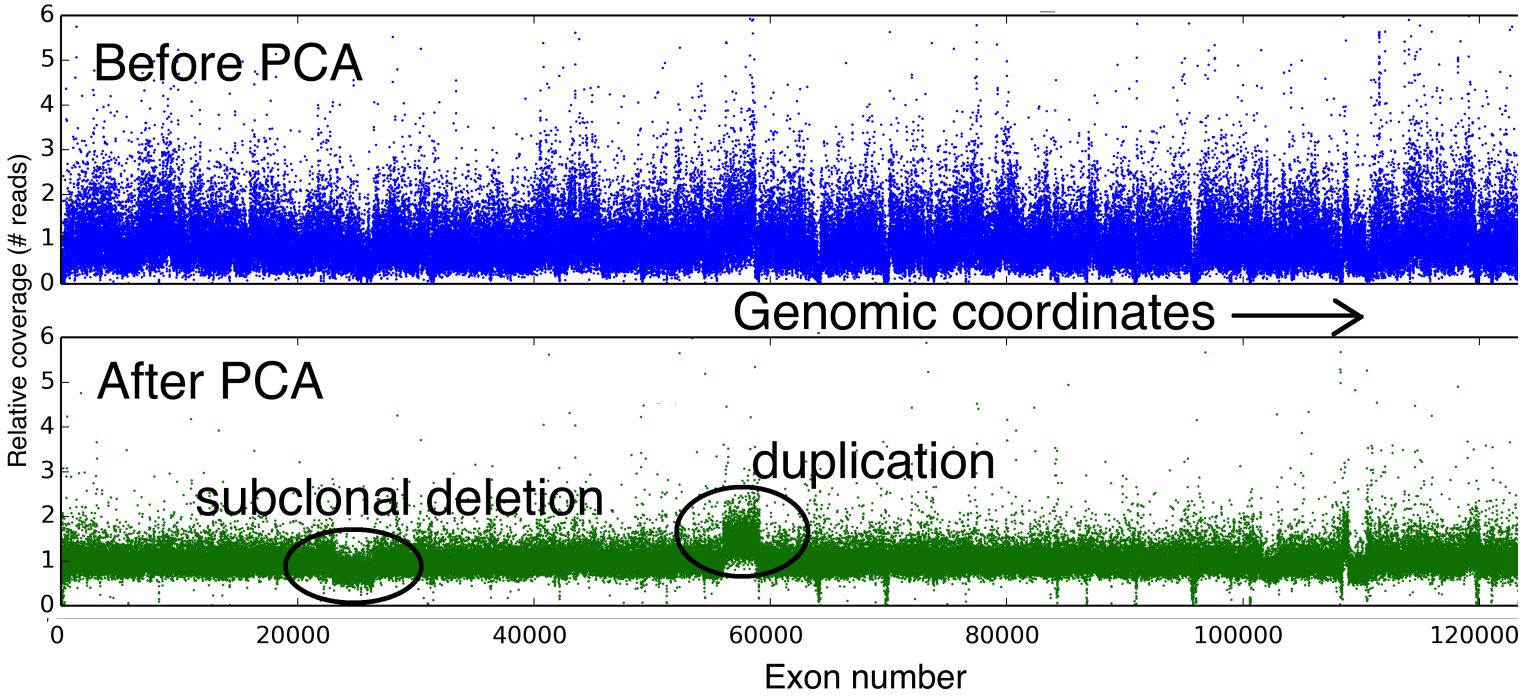
Copy number variation from exome capture



Copy number variation from exome capture



Copy number variation from exome capture



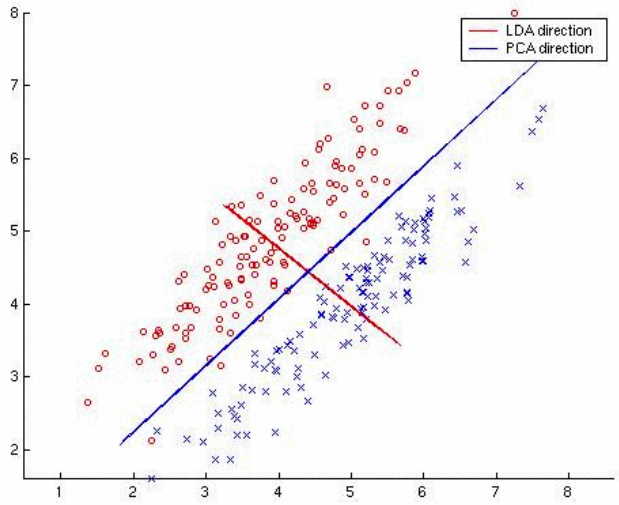
$$\mathbf{x} = \mu + \sum_k (\mathbf{v_k}^\top \mathbf{x})\mathbf{v_k} + \text{copy number signal}$$

$$\Rightarrow \text{copy number signal} = \mathbf{x} - \mu - \sum_k (\mathbf{v_k}^\top \mathbf{x})\mathbf{v_k}$$

PCs $\mathbf{v}$ come from *non-tumor* samples with no CNVs!

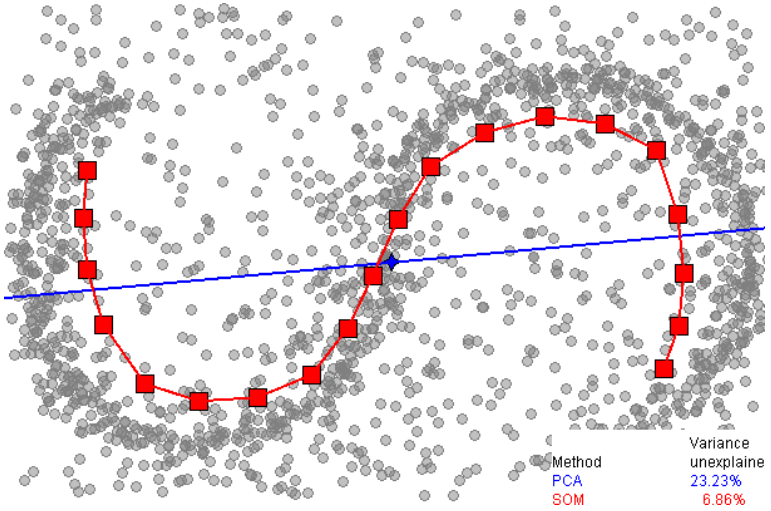
Pitfalls

- PCs might not be good for classification



Pitfalls

- PCs might not be good for classification
- Low-dimensional space might be non-linear



Pitfalls

- PCs might not be good for classification
- Low-dimensional space might be non-linear
- Non-issue: Σ is a big matrix. (Use iterative PCA, FastPCA, flashpca. . .)

Generalizations

- $\mathbf{x} = \mu + \sum c_k \mathbf{v_k} + \text{noise}$ is part of a larger model: probabilistic PCA.

Generalizations

- $\mathbf{x} = \mu + \sum c_k \mathbf{v_k} + \text{noise}$ is part of a larger model: probabilistic PCA.
- Don't like heuristics for choosing number of PCs to use: Bayesian PCA.

Generalizations

- $\mathbf{x} = \mu + \sum c_k \mathbf{v_k} + \text{noise}$ is part of a larger model: probabilistic PCA.
- Don't like heuristics for choosing number of PCs to use: Bayesian PCA.
- Data are not linear: nonlinear dimensionality reduction (tSNE, autoencoders, GPLVM, Isomap, SOM. . .)

Equations

Find the direction (unit vector) $\mathbf{v}$ of greatest variance. Projection of $\mathbf{x}$ is $\mathbf{x}^\top \mathbf{v}$.

$$\begin{aligned}\sigma^2 =& \frac{1}{N}\sum_n \left(\mathbf{x_n}^\top \mathbf{v} - \mu^\top \mathbf{v}\right)^2 = \frac{1}{N}\sum_n \left((\mathbf{x_n} - \mu)^\top \mathbf{v}\right)^2 \\ =& v^\top \frac{1}{N}\sum_n (\mathbf{x_n} - \mu)(\mathbf{x_n} - \mu)^\top \mathbf{v} = \mathbf{v}^\top \Sigma \mathbf{v}\end{aligned}$$

Set $\nabla_{\mathbf{v}} = 0$ with Lagrange multiplier for $\mathbf{v}^\top \mathbf{v} = 1$:

$$\nabla_{\mathbf{v}} \left(\mathbf{v}^\top \Sigma \mathbf{v} + \lambda(1 - \mathbf{v}^\top \mathbf{v})\right) = 0 \Rightarrow \Sigma \mathbf{v} = \lambda \mathbf{v}$$

Dotting with $\mathbf{v}^\top$ gives $\lambda = \lambda \mathbf{v}^\top \mathbf{v} = \mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} = \sigma^2$.