Statistical methods for data integration

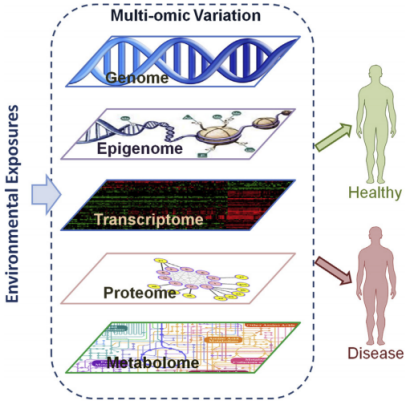
Ricard Argelaguet
ricard@ebi.ac.uk

May 25, 2021

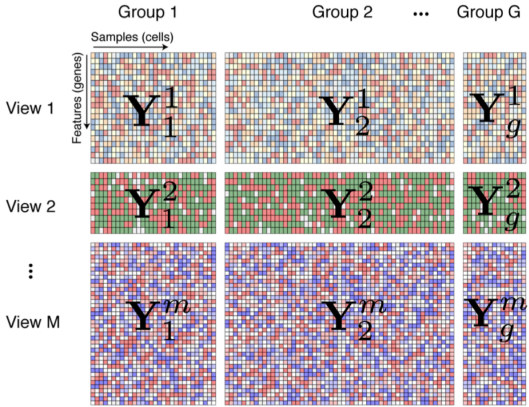
European Bioinformatics Institute (EMBL-EBI)
University of Cambridge

Why multi-omics?

The integrative analysis of diverse data modalities in a systems biology approach will capture better the molecular phenotypic varaition of biological systems

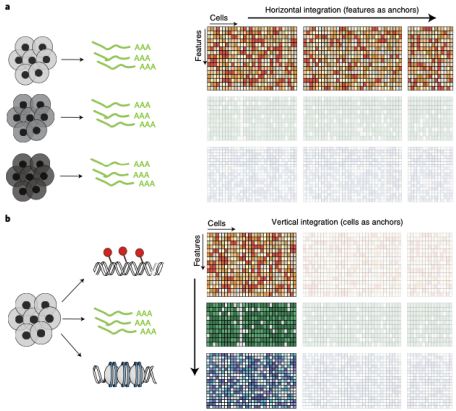


Abstraction of a multi-omics experimental design



Strategies for multi-omics data integration

The first step is to choose the anchoring unit for the integration



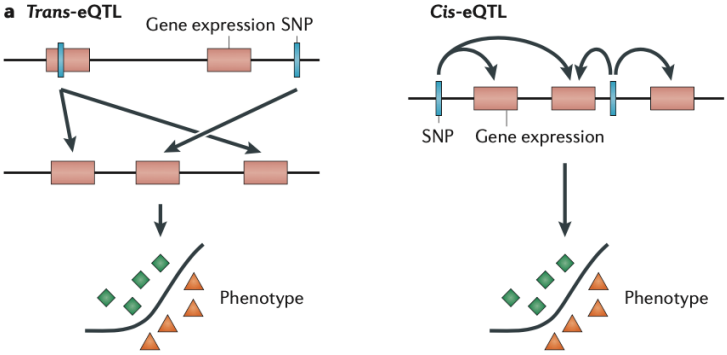
Strategies for multi-omics data integration

Two general strategies for vertical integration (multi-omics data derived from the same set of samples);

- **Local analysis**: test for marginal associations between features from different molecular layers. Generally supervised.
- **Global analysis**: exploit the dependencies between the features to construct a mathematical representation of the data. Generally unsupervised.

Local analysis

The most prominent examples of local analysis are quantitative trait loci mapping (GWAS and eQTLs)[1]:

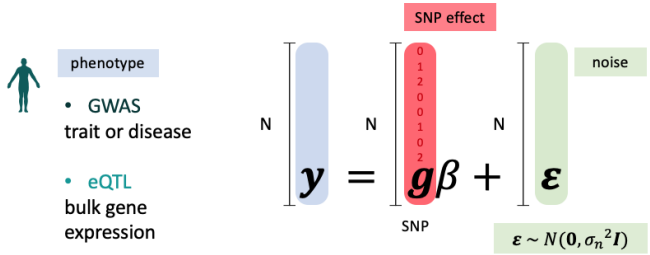


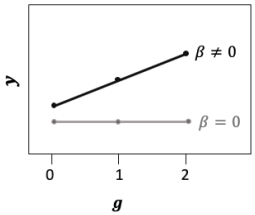
[1] Ritchie, M. D. et al. "Methods of integrating data to uncover genotype–henotype interactions" *Nature Reviews Genetics* 2015.

eQTL: expression Quantiative Trait Loci

Local analysis

Local analysis is typically done using (generalised) linear models





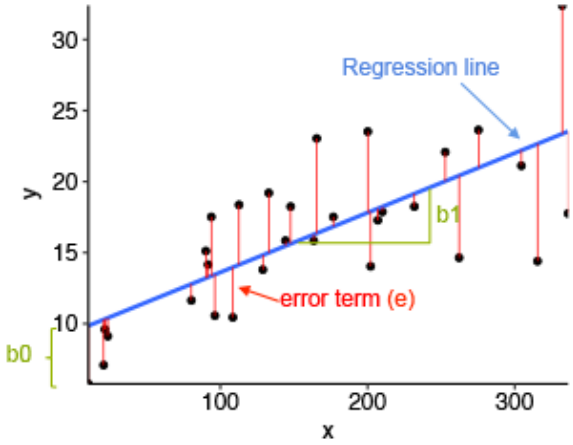
Test:

H0: $\beta = 0$

H1: $\beta \neq 0$

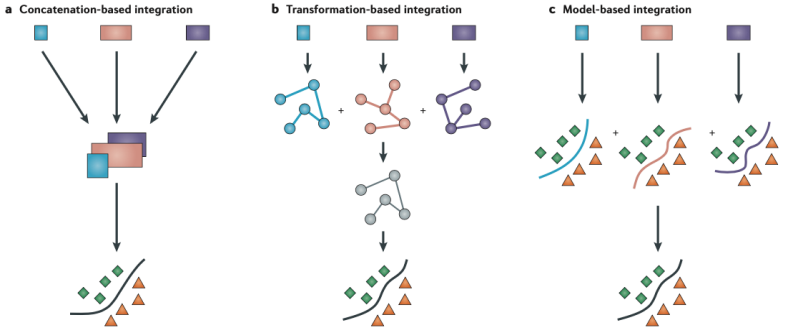
GWAS: Genome-Wide Association Study (GWAS)
eQTL: expression Quantiative Trait Loci

Local analysis



Global analysis

In global analysis the aim is to exploit the relationship between all features to create useful mathematical representations[2]



[2]Ritchie, M. D. et al. "Methods of integrating data to uncover genotype–henotype interactions" *Nature Reviews Genetics* 2015.

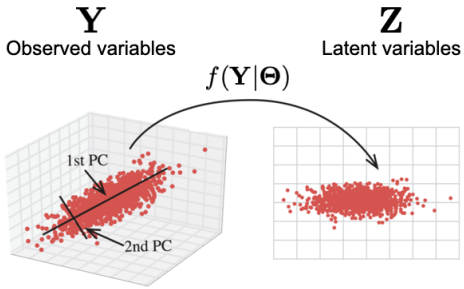
Global analysis

Challenges in (global) multi-omics data integration:

- Data collected using different techniques (i.e. data modalities) generally exhibit heterogeneous statistical properties
- Large amounts (and different patterns) of missing values
- Overfitting
- Undesired sources of heterogeneity
- Complexity of the data requires unsupervised interpretable approaches

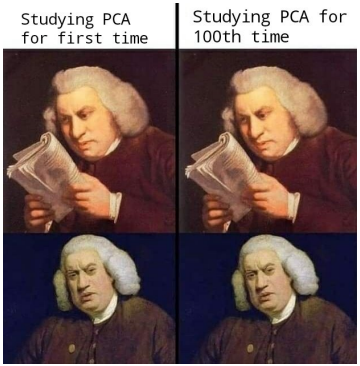
Latent variable models

Given a dataset $\mathbf{Y}$ of N samples and D features, latent variable models exploit the dependencies between the features to reduce the dimensionality of the data. The mapping from the high-dimensional to the low-dimensional space is performed via a function $f(\mathbf{Y}|\mathbf{\Theta})$:



Principal component analysis (PCA)

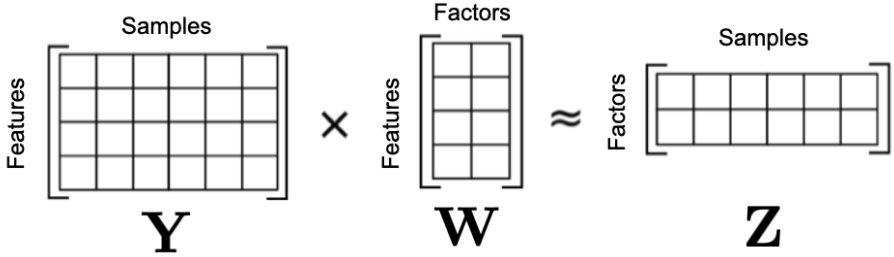
Principal Component Analysis (PCA) is the most popular technique for dimensionality reduction.



Credit to Raunak Joshi

Principal component analysis (PCA)

PCA defines $f(\mathbf{Y}|\mathbf{\Theta})$ to be a linear transformation via a matrix $\mathbf{W} \in \mathbb{R}^{D \times K}$ that maps the observations $\mathbf{Y} \in \mathbb{R}^{N \times D}$ onto the latent space $\mathbf{Z} \in \mathbb{R}^{N \times K}$.



Mathematical derivation of PCA: maximum variance formulation

The aim in PCA is to infer the matrix $\mathbf{W}$ such that the variance of $\mathbf{Z}$ (the projected data) is maximised. If we consider a single latent factor, the variance of the projected data is:

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{z}_n - \hat{\mathbf{z}})^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n^T\mathbf{w} - \hat{\mathbf{y}}^T\mathbf{w})^2\end{aligned}$$

where $\hat{\mathbf{y}}$ is a vector with the feature-wise means. If we center the data this simplifies to:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n^T\mathbf{w})^2$$

Mathematical derivation of PCA: maximum variance formulation

A bit of algebra allows us to define this equation in terms of the (centered) data covariance matrix: $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^T$:

$$
\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{n=1}^{N} (\mathbf{y}_n^T \mathbf{w})^T (\mathbf{y}_n^T \mathbf{w}) \\
&= (\mathbf{w}^T \mathbf{y}_n)(\mathbf{y}_n^T \mathbf{w}) \\
&= \mathbf{w}^T (\mathbf{y}_n \mathbf{y}_n^T) \mathbf{w} \\
&= \mathbf{w}^T \mathbf{S} \mathbf{w}
\end{aligned}
$$

Mathematical derivation of PCA: maximum variance formulation

The optimisation problem to find the first latent variable could be defined as:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathbf{w}^T \mathbf{S} \mathbf{w}$$

(Q) Maximising this expression does not work, we need a constrain. Why?

Mathematical derivation of PCA: maximum variance formulation

The constrained optimisation problem can be defined as:

$$\hat{\mathbf{w}} = \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^T \mathbf{S} \mathbf{w}$$

It can be solved by introducing a Lagrange multiplier λ to enforce the constraint:

$$f(\mathbf{W}, \lambda) = \mathbf{w}^T \mathbf{S} \mathbf{w} + \lambda(1 - \mathbf{w}^T \mathbf{w})$$

By setting the derivative $\frac{\partial f(\mathbf{W},\lambda)}{\partial \mathbf{w}}$ to zero, we obtain the following equation:

$$\mathbf{S}\mathbf{w} = \lambda \mathbf{w}$$

which should be familiar (perhaps in this form $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$)?

Mathematical derivation of PCA: maximum variance formulation

Among all possible orthonormal basis, the one that maximises the projected variance corresponds to the basis defined by the eigenvectors of the covariance matrix **S**. These vector basis are called the principal components.

The corresponding eigenvalue λ corresponds to the variance σ (proof in the appendix).

Generalisation to multiple principal components

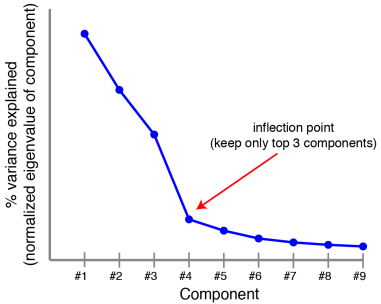
Most data can not be well-described by a single principal component. The k-th principal component can be found by subtracting from $\mathbf{Y}$ the reconstructed data by the previous $k-1$ principal components:

$$\hat{\mathbf{Y}} = \mathbf{Y} - \sum_{k=1}^{K} (\mathbf{z}_k \mathbf{w}_k^T)$$

and repeating the procedure above using the reconstructed covariance matrix $\hat{\mathbf{S}}$ as input

Finding the right number of Principal Components

Principal components are ranked by the amount of variance they capture in the original dataset, a scree plot can provide some sense of how many components are needed.



Problems of using PCA for multi-omics data integration

PCA is a great exploratory tool for single multivariate data sets, but it has important pitfalls in the analysis of multi-omics data:

- Does not generalise to an arbitrary number of data modalities.
- No natural way to combine different data modalities (binary data with continuous data).
- Cannot handle missing values.

Canonical correlation analysis

Canonical Correlation Analysis (CCA) is a simple extension of PCA to find linear components that capture correlations between **two** datasets[3].

Given two data matrices $\mathbf{Y}_1 \in \mathbb{R}^{N \times D_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times D_2}$ CCA finds a set of linear combinations $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ with maximal cross-correlation.

For the first pair of canonical variables, the optimisation problem is:

$$(\hat{\mathbf{u}_1}, \hat{\mathbf{v}_1}) = \underset{\|\mathbf{u}_1\|=1, \|\mathbf{v}_1\|=1}{\arg\max} \; corr(\mathbf{u}_1^T \mathbf{Y}_1, \mathbf{v}_1^T \mathbf{Y}_2)$$

[3] Hotelling, H. "Relations between two sets of variates" *Biometrika* 1936.