Bioinformatics 2 - Lecture 2

Guido Sanguinetti

School of Informatics
University of Edinburgh

January 24, 2012

Problem formulation

- Often in biology we have samples of a certain quantity from two conditions
- E.g., we have measurements of a protein expression x in two cohorts (treated and control)
- We want to know whether the di?erences observed between the two populations are statistically signicant, so x may be used as a biomarker
- How can we do that since we know nothing of the distribution of x?

Terminology

- We start by assuming the *null hypothesis*: the two sets of samples have the same distribution
- The procedure to asses whether it holds is called *hypothesis testing*
- The samples in each set are assumed to be independent and identically distributed (no cousins in the set please!)
- The test is *unpaired* if the samples in the two sets are independent (e.g. different people form the two sets)
- The test is *paired* if the samples are dependent in the two sets (e.g. same people before/ after treatment, different algorithms on same data sets)
- The testing consists in computing a *test statistic* from the sample whose distribution is approximately known

Power and errors

- An important concept is the *power* of a statistical test, i.e. its ability to flag up (correctly)a deviation from the null hypothesis
- Conversely, it is important to define the type of errors one can make
- A *type I error* means incorrectly rejecting the null hypothesis (false positive, crying wolf)
- A *type II error* means being too conservative, i.e. accepting a wrong null hypothesis. It is the complement of the power: if β is the rate of type II errors, the power of the test is mathematically defined as $1 - \beta$

The t-test

- Under the assumption of normality (which can be checked using e.g. a Kolmogorov-Smirnov test), one uses *Student's* t-test
- The unpaired test statistic (assuming equal sample size and variance) is given by

$$t = \frac{\langle x \rangle_1 - \langle x \rangle_2}{\sqrt{\frac{var_1 + var_2}{n}}}$$

  where the subscript indicates empirical expectations taken in the two sets and n is the sample size
- t follows a Student t distribution with $n - 1$ degrees of freedom

The p-value

- One can then look up the probability of getting a value of t greater than the empirical one from the samples
- This is the p-value: the probability that the experiment would return a result at least as extreme under the null hypothesis
- Depending on the application, p-values of 0.05 or 0.01 are considered significant
- Notice that t grows as $\sqrt{n}$ so increasing n we get more and more statistically signicant results$\longrightarrow$ experimental design!

Multiple hypothesis testing

- Suppose instead of measuring a biomarker across two samples, you've done a high-throughput experiment, i.e. measured 20K genes' expression
- You want to use all this data to check whether the two conditions are different
- What do you do? Do you do independent tests for each gene and see whether any are differentially expressed? What's the obvious problem?
- This is an example of multiple hypothesis testing. A classic approach is to correct (e.g. *Bonferroni correction*, however this is very conservative)

Non-parametric testing

- The t test makes a parametric assumption, i.e. normality. What if it doesn't hold?
- A popular non-parametric test is the *Wilcoxon rank-sum test* (or Mann-Whitney test), which tests whether one sample is larger than the other
- The idea is that, if two samples are statistically the same, the ranking should come from a *uniform* distribution over the group of permutations
- The Wilcoxon rank-sum test has almost the same power as the t test under the normal assumption (~ 0.98)
- If the normal assumption is violated, the Wilcoxon rank-sum test can be several times more powerful and is more robust to outliers

Wilcoxon rank-sum test: algorithm

- Pool the data and rank them in ascending order
- Compute the rank-sum for the samples (sum the ranks), R_1 and R_2
- The U-statistic is obtained as

$$U_i = R_i - \frac{n_i(n_i - 1)}{2}$$

- Tables contain *critical values* of U by sample sizes: U_i either bigger or smaller than U_{crit} indicates significance (careful which way it goes)

Testing summary

- Statistical procedure to determine whether observed differences in samples are what is to be expected from random fluctuations
- Essential to determine the type of test (paired/ independent)
- For large data sets or for normally distributed samples (e.g. following a K-S test), one uses a t-test
- For smaller data-sets far from normality a non-parametric test such as Wilcoxon's rank-sum is probably better)

Examples

- I'll use R, a powerful statistical language, to work through these examples, but this is not essential
- You can get R from `www.rproject.org`, if you become a bioinformatician it will be your main language
- I'll demonstrate some testing (with tables) on some simulated data sets
- The source for tables is the web

Problem statement

- Data in biology often very high dimensional with very few samples
- E.g., in a cancer study, we could have 40 subjects and 10000 features (genes) each
- Find a suitable 2D projection of the data that highlights structure
- Determine the projection (not necessarily 2D) that identies the most relevant features
- In general, we seek to find the optimal projection from D (original) dimensions to Q (target) dimensions, based on a sample of N points

Principal Component Analysis

- A plausible assumption is that the interesting directions are the ones with the greatest variation
- The empirical covariance of a data set $\mathbf{x}_i$ with mean $\hat{\mu}$ is

$$\hat{\Sigma} = \frac{1}{N} \sum_i (\mathbf{x}_i - \hat{\mu}) (\mathbf{x}_i - \hat{\mu})^T$$

- The directions that maximise the projected variance satisfy

$$\Sigma V = \Lambda V$$

with Λ a diagonal matrix containing the Q largest eigenvalues of the empirical covariance

Least-squares fit of a subspace

- An equivalent way of looking at the problem is to find the (hyper)-plane which best interpolates the data (why?)
- So, we need to find D-dimensional vectors $\mathbf{v}_j$ $(j = 1, \ldots, Q)$ and scalars t_i^j $(i = 1, \ldots, N)$ such that the error function

$$\mathcal{E} = \sum_{i=1}^{N} \|\mathbf{x}_i - \sum_{j=1}^{Q} \mathbf{v}_j t_i^j\|^2$$

  is minimised
- We can rewrite the error function using the formula for the residual of the projection of a point onto a hyperplane

$$\mathcal{E} = \sum_{i=1}^{N} \|\mathbf{x}_i - \sum_{j=1}^{Q} \mathbf{v}_j \mathbf{v}_j \cdot \mathbf{x}_i\|^2$$

- Show that this yields the same formula as in the previous slide

Factor Analysis

- What we have just shown is that PCA is a special case of matrix factorisation, where the data matrix X $(D \times N)$ is decomposed as the product of V $(D \times Q)$ and projected *latent* points T $(Q \times N)$
- This suggests a probabilistic model for *Probabilistic PCA* (Tipping and Bishop 1998)

$$\mathbf{x} = V\mathbf{t} + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \quad \mathbf{t} \sim \mathcal{N}(0, I)$$

- More generally, by relaxing the spherical covariance requirement on $\mathbf{t}$ to a diagonal, we obtain *Factor Analysis*

Proofs and examples

- Let's demonstrate the meaning of PCA on some examples, again using R
- What do principal components tell us?
- What do the factors tell us?
- If we have time, let's prove some formulae of the above

Next week

- Next week we'll start thinking about networks and how to reconstruct them
- We'll see efficient methods for building networks based on correlations
- We will also introduce the important concept of conditional independence and a (slightly) more sophisticated way of reconstructing networks
- Early in the morning we will have a tutorial on probability review and hypothesis testing