Gene set tests & correlation

Michael Love
July 2018

Gene set tests: competitive vs self-contained

From *CAMERA* publication:

> “… **self-contained** gene set tests examine a set of genes in their own right without reference to other genes in the genome, whereas **competitive** gene set tests compare genes in the test set relative to all other genes.”

Gene set tests: competitive vs self-contained

From *CAMERA* publication:

> "**Self-contained** tests are of interest for assessing the relevance of an individual biological process to the experiment at hand, whereas the **competitive** tests focus more on distinguishing the most important biological processes from those that are less important. Competitive tests are overwhelmingly more commonly used in the genomic literature."

Gene set tests with correlation

(from **PH525x notes** and **Barry, Nobel and Wright, 2008**)

Consider the average t-statistic from N genes in a set G:

$$\bar{t} = \frac{1}{N} \sum_{i \in G} t_i$$

This statistic $\bar{t}$ combines the information about DE from the set and might be a useful test statistic.

Gene set tests with correlation

Under the null hypothesis, the t have mean 0. If the t are independent then $\sqrt{N}\bar{t}$ has standard deviation 1 and is approximately normal:

$$\sqrt{N}\bar{t} \sim N(0,1)$$

This comes from the following decomposition of the variance:

$$\begin{aligned}
\mathrm{Var}(\bar{t}) &= \frac{1}{N^2}\mathrm{Var}(t_1 + \cdots + t_N) \\
&= \frac{1}{N^2}(\mathrm{Var}(t_1) + \cdots + \mathrm{Var}(t_N)) \\
&= \frac{1}{N}
\end{aligned}$$

Gene set tests with correlation

Now consider the case that the test statistics t_i in a gene set are not independent but have correlation ρ *under the null hypothesis*.

$$\bar{t} = \frac{1}{N} \sum_{i \in G} t_i$$

$$\operatorname{corr}(t_i, t_{i'}) = \rho, \quad i, i' \in G$$

The variance of the average *t*-statistics will be:

$$
\begin{aligned}
\mathrm{Var}(\bar{t}) &= \frac{1}{N^2}\,\mathrm{Var}((1\,\ldots\,1)(t_1\,\ldots\,t_N)')\\
&= \frac{1}{N^2}\,(1\,\ldots\,1)\begin{pmatrix} 1 & \rho & \ldots & \rho & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \ldots & \rho & 1 \end{pmatrix}(1\,\ldots\,1)'\\
&= \frac{1}{N^2}\,\{N + (N-1)N\rho\}\\
&= \frac{1}{N}\,\{1 + (N-1)\rho\}
\end{aligned}
$$

Variance inflation with correlation

So the variance inflation factor (VIF) comparing the independent case to the case with correlation is:

$$VIF = 1 + (N - 1)\bar{\rho}$$

So the increased width (standard deviation) of the null distribution for a gene set with 20 genes and average correlation 0.1 will be:

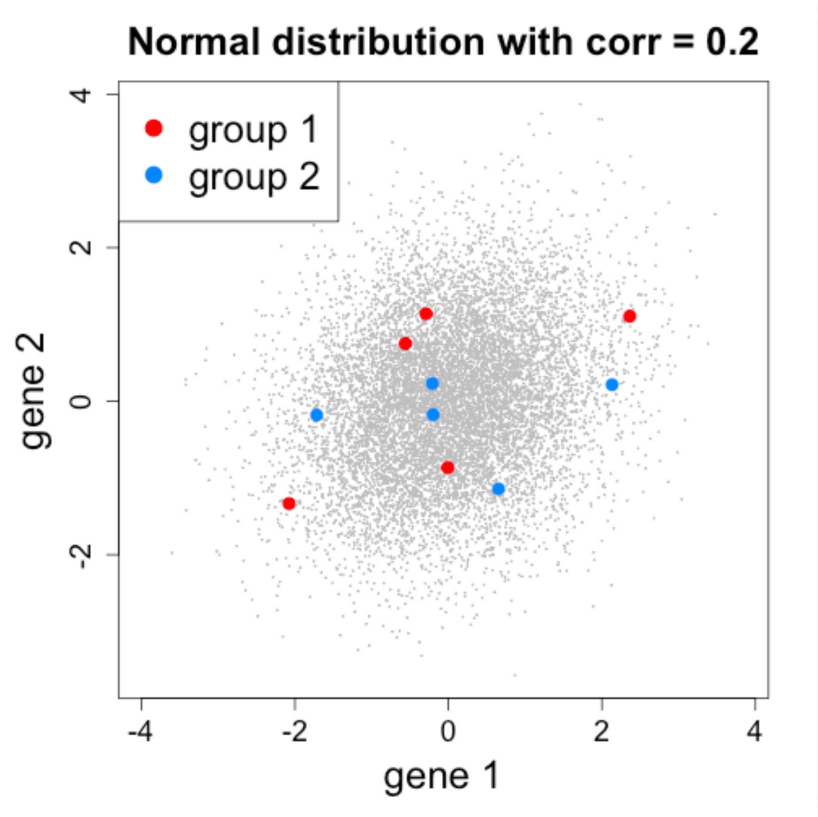
```
sqrt(1 + 19 * 0.1)
```

```
[1] 1.702939
```

This VIF is approximately true also for testing the set statistics against the complement: the genes not in the set (see Barry, Nobel and Wright 2008).

Test statistic vs expression correlations

Here, the expression of 5 samples vs 5 samples, no difference in expression across group but a correlation of gene expression.



Test statistic vs expression correlations

If the test statistic T is a linear form of the data X (e.g. log fold change), then:

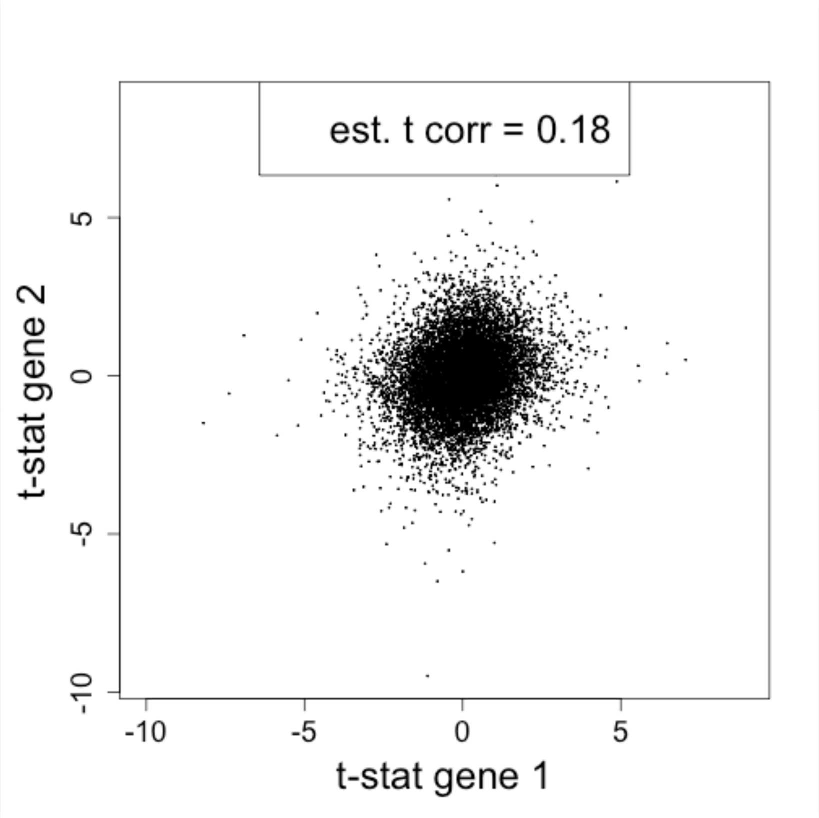
$$\rho^T_{i,i'} = \rho^X_{i,i'}$$

For t-test, the relationship is monotone, approximately linear and:

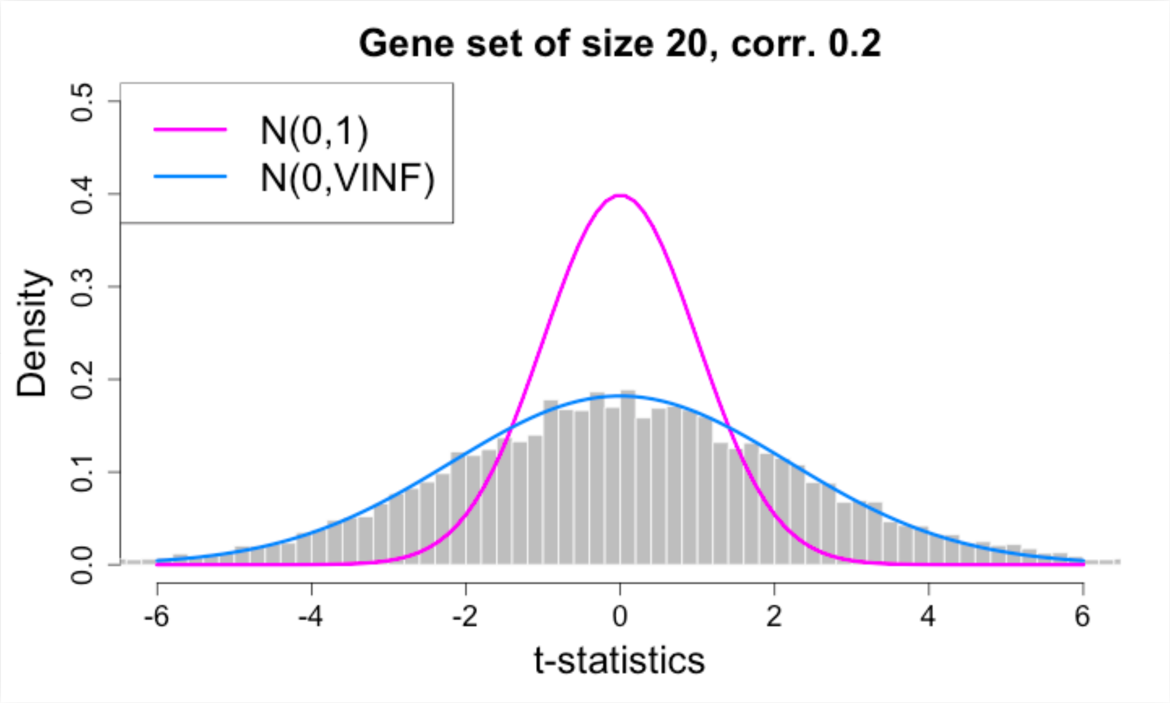
$$\rho^T_{i,i'} \approx \rho^X_{i,i'}$$

(Barry, Nobel and Wright, 2008)

Simulate expression correlation of 0.2

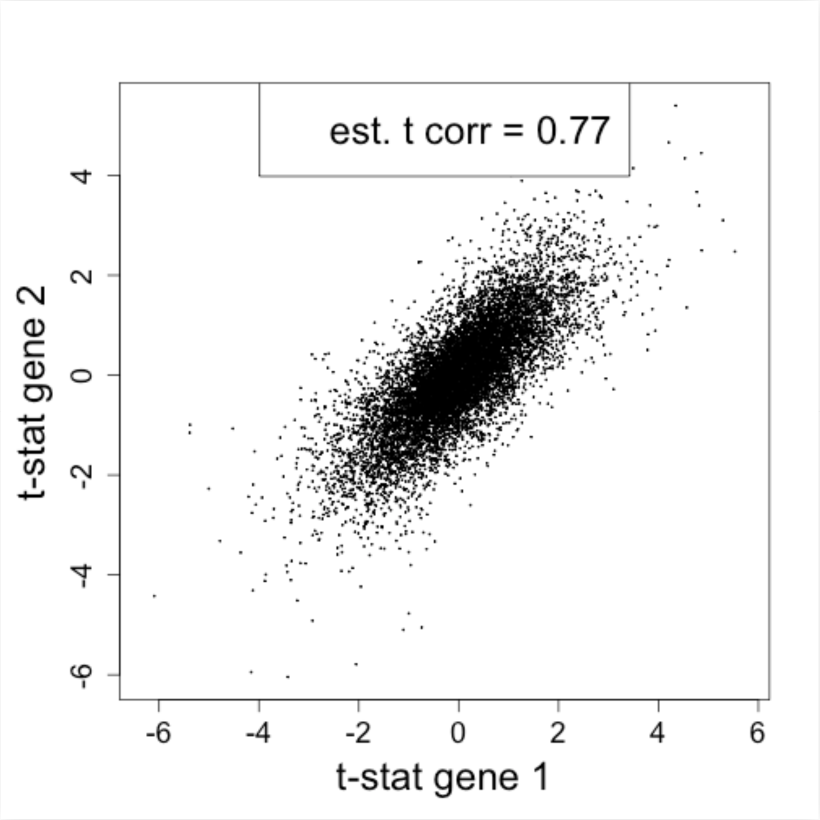


Distribution of t-statistics

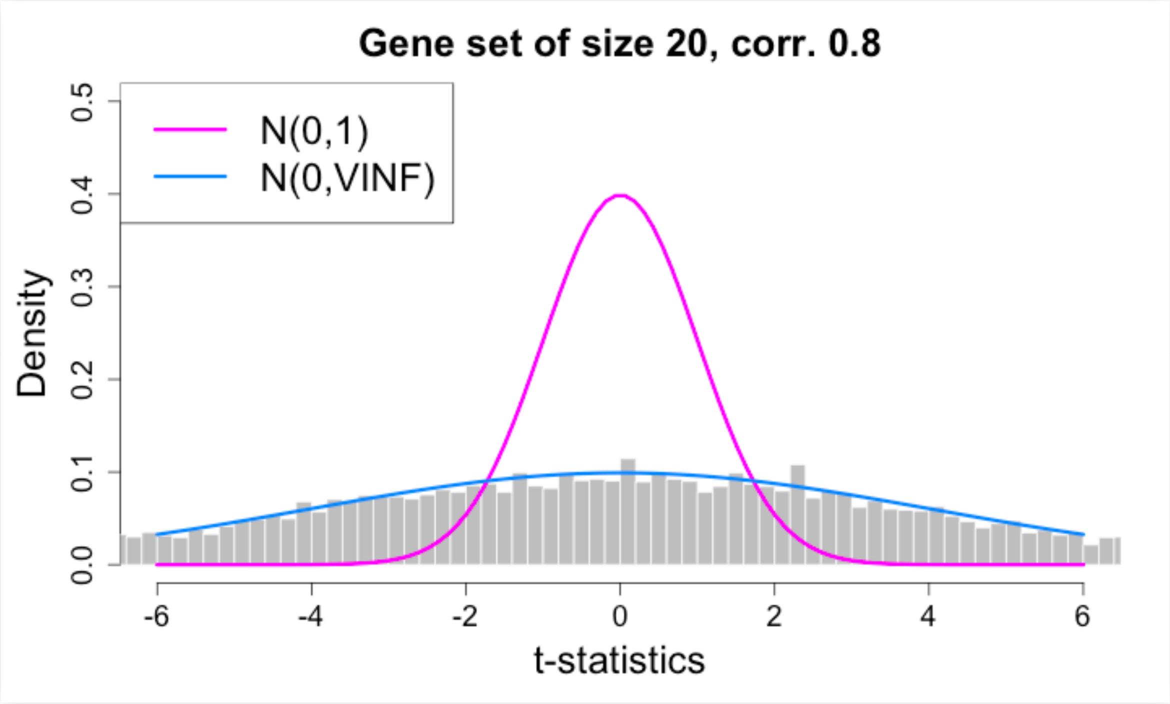


27% of the simulated t-statistics are outside of the center 99% of the $N(0, 1)$ distribution.

Again, with correlation of 0.8



Distribution of t-statistics (corr = 0.8)



54% of the simulated t-statistics are outside of the center 99% of the $N(0, 1)$ distribution.

Intuition

- Suppose expression is correlated within a gene set under the null
- By chance, for some gene, the expression could be high for the group 2 samples, and low for group 1 samples
- The t-statistic will be large and positive for this gene
- Because expression is correlated across genes in the set, other genes will likely see the same pattern
- t-statistics will be correlated within the set

Why would we see null correlations?

Where do expression correlations *under the null* come from? My guesses in order of importance:

- uncorrected “batch effects”
  - library preparation
  - RNA extraction
  - biological: cell-type composition
- large scale amplifications in cancer
- gene regulatory networks

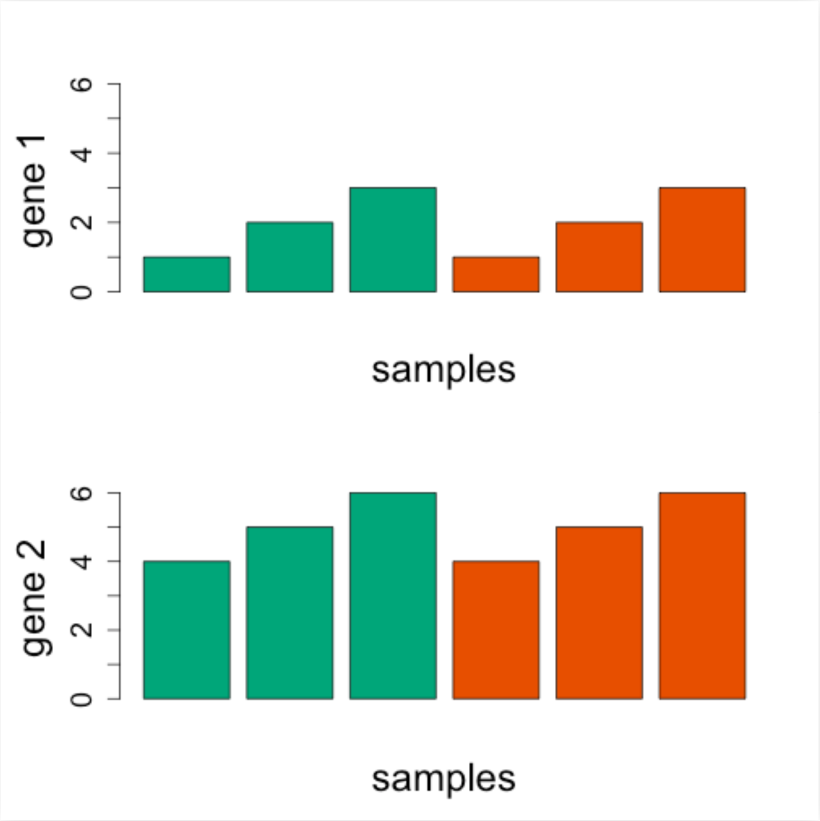
CAMERA (Wu and Smyth 2012)

for **C**orrelation **A**djusted **ME**an **RA**nk gene set test, available in the limma package.

- estimating the inter-gene correlation from the data
- using it to adjust the gene set test statistic
- suitable for any experiment that can be represented by genewise linear models
- a competitive gene set test

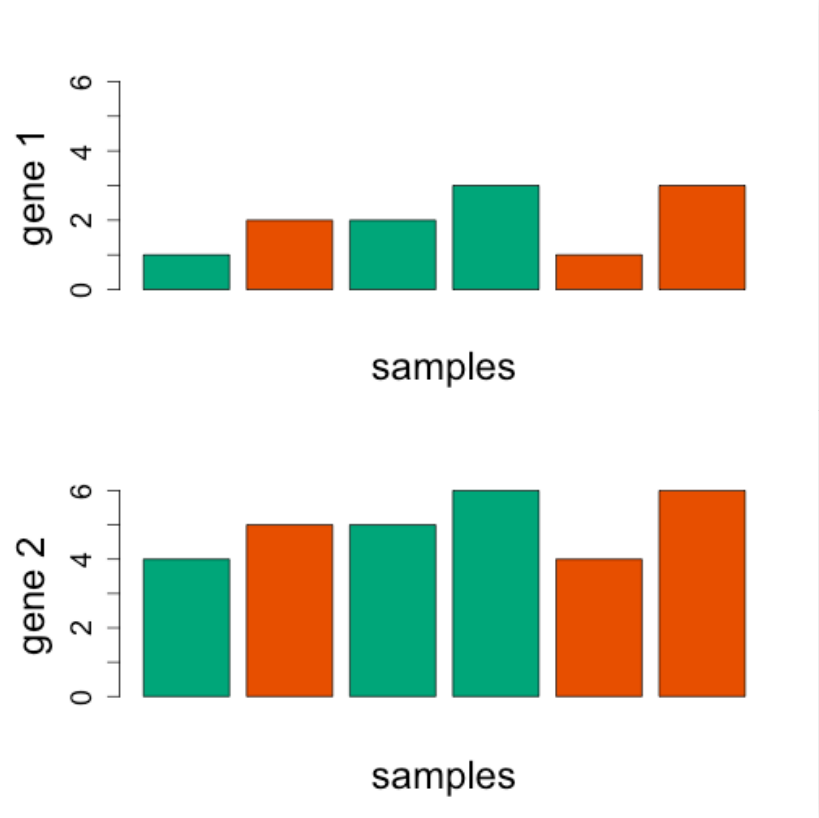
Permutations

Assume the null: no differences across condition, although gene-gene correlation are present



Permutations

Assume the null: no differences across condition, although gene-gene correlation are present



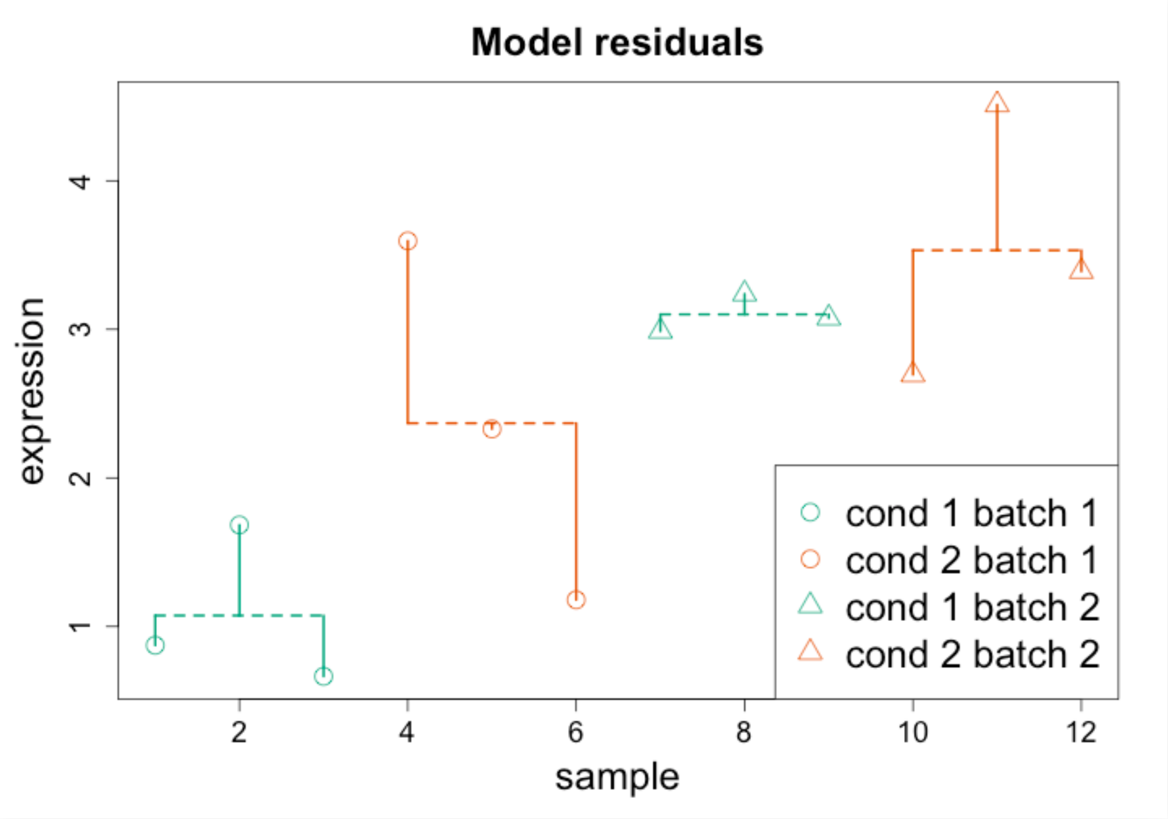
Permutations

Has limitations:

- only if samples are exchangeable (no batch effects)
- not easy to implement for complex designs (although see SAMseq in the **samr** package for strategies)
- requires sufficient samples for small p-values (although see **Larsen and Owen** for moment-based trick)

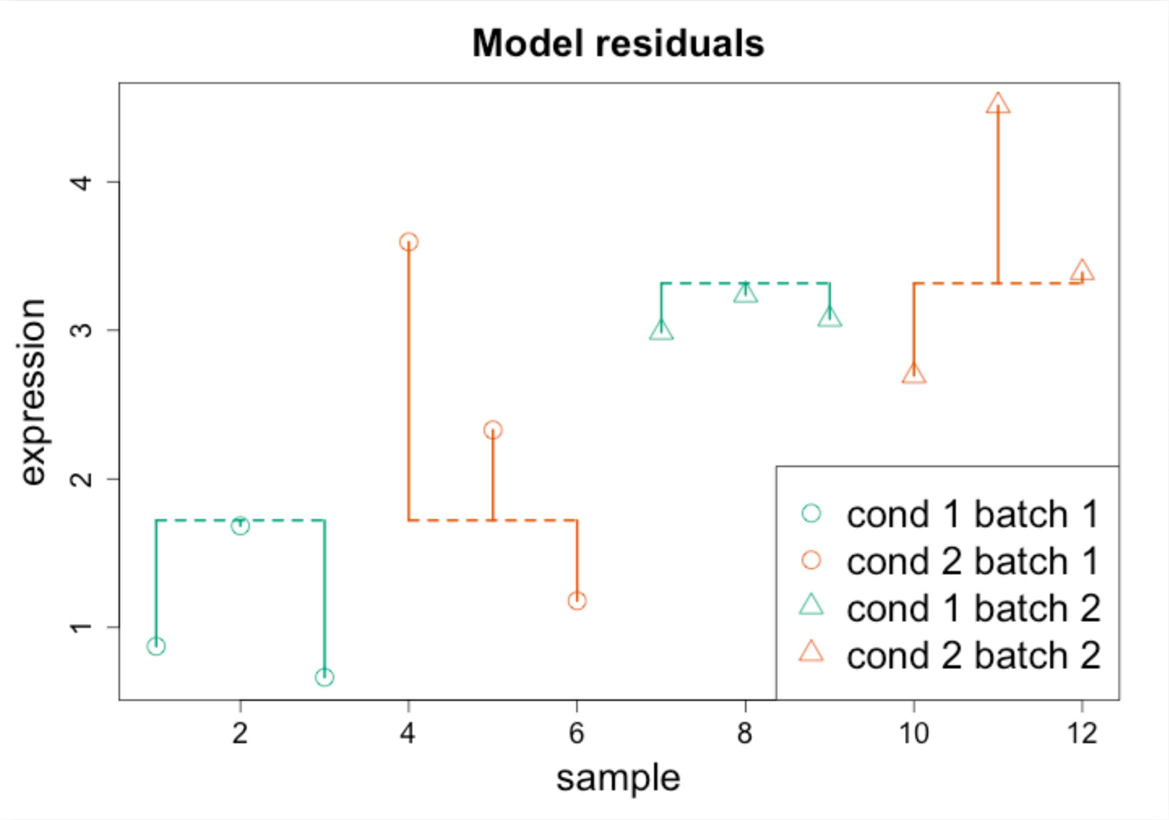
Approach using residuals

Suppose we have a 2 condition experiment with 2 batches:



Approach using residuals

Remove design matrix columns not involving the null hypothesis:



ROAST (Wu et al. 2010)

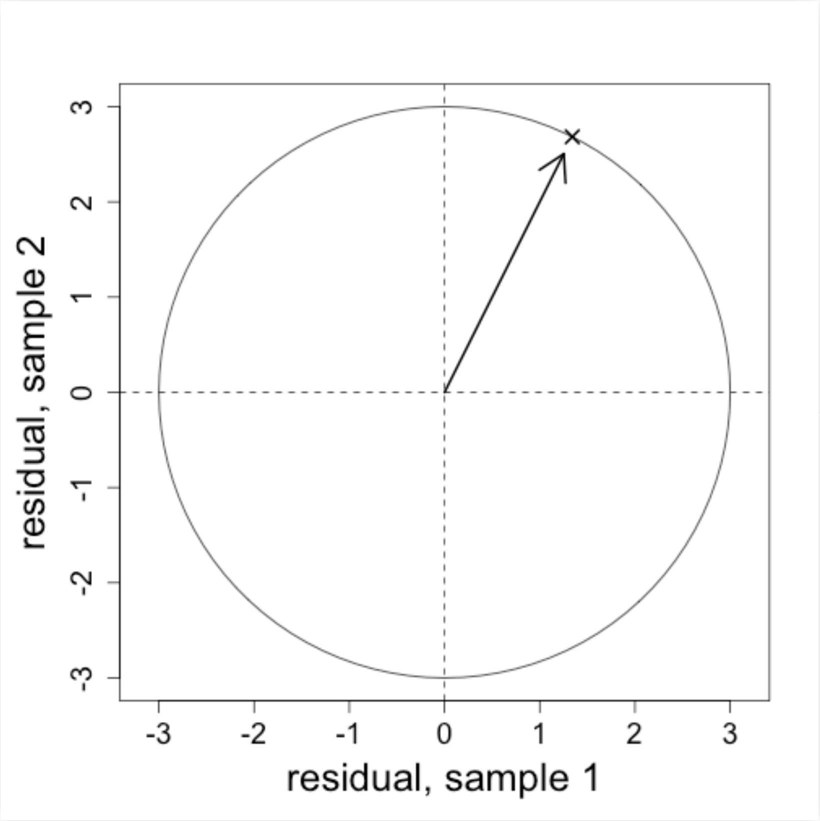
The ROAST method available in the limma package:

Under the null hypothesis (and assuming a linear model) the residuals are independent and identically distributed $N(0, \sigma_g^2)$.

We can *rotate* the residual vector for each gene in a gene set, such that gene-gene expression correlations are preserved.

What does residual rotation look like?

Like this diagram but around an n-sphere (n, the number of samples).



ROAST (Wu et al. 2010)

Repeat 10,000 times:

1. rotate the residual vector from each gene in the set using the same rotation
2. create new data, preserving the gene-gene correlations
3. compute test statistics for the rotated data for each gene and compute the gene set statistic

Lastly compare the original gene set statistic to the null distribution from 1-3.

Pros: fast and efficient, fits with any linear model. Designed for testing a single gene set (self-contained), for testing many see **mROAST** (also self-contained).

Summary

1. "**Self-contained** tests are of interest for assessing the relevance of an individual biological process"
2. "**Competitive** tests focus more on distinguishing the most important biological processes from those that are less important"
3. Gene-gene correlations inflate the null distribution of gene set statistics
4. This inflation factor can be directly calculated from the data (CAMERA)
5. Rotating residuals can also be used to generate a null which incorporates correlations (ROAST)