Multivariate Statistics in Ecology and Quantitative Genetics

Analyzing RNA-Seq gene expression data

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

14. Juli 2016

DESeq2

https://bioconductor.org/packages/release/bioc/html/DESeq2.html/

M.I. Love, W. Huber, S. Anders (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.
Genome Biology **15**:550.
http://dx.doi.org/10.1186/s13059-014-0550-8

Install in R with:

```
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
```

DESeq2 basic assumptions

i: label of gene (row in read count table)

j: label of sample (column in read count table and rows in sample data table "colData")

K_{ij}: read counts for gene i in sample j.

DESeq2 basic assumptions

i: label of gene (row in read count table)

j: label of sample (column in read count table and rows in sample data table "colData")

K_{ij}: read counts for gene i in sample j.

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i),$$

that is, K_{ij} comes from a negative binomial distribution with mean μ_{ij} and dispersion parameter α_i (depends only on gene), that is

$$\text{Var}(K_{ij}) = \mu_{ij} + \alpha_i \cdot \mu_{ij}^2.$$

DESeq2 basic assumptions

i: label of gene (row in read count table)

j: label of sample (column in read count table and rows in sample data table "colData")

K_{ij}: read counts for gene i in sample j.

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i),$$

that is, K_{ij} comes from a negative binomial distribution with mean μ_{ij} and dispersion parameter α_i (depends only on gene), that is

$$\text{Var}(K_{ij}) = \mu_{ij} + \alpha_i \cdot \mu_{ij}^2.$$

Furthermore:

$$\mu_{ij} = s_j \cdot q_{ij} \qquad \text{and} \qquad \log_2(q_{ij}) = \beta_{i0} + \beta_{i1} x_{j1} + \beta_{i2} x_{j2} + \cdots + \beta_{ik} x_{jk},$$

where the $x_{\cdot r}$ are columns of the colData table (and indicator variables for levels of factors or interaction terms).

Role of s_j and its estimation

Usually s_j accounts for sample-specific coverage but can be replaced by user-specified s_{ij} that depends on gene (or e.g. its GC content) and sample.

Usually, s_j is set to the median of K_{ij} normalized by their geometric mean:

$$s_j = \text{median}_i \frac{K_{ij}}{\left(\prod_{j'=1}^{m} K_{ij'}\right)^{1/m}},$$

where i for which one of the $K_{ij'}$ (and thus the denominator) is 0 are excluded. That is, genes are only considered if they are expressed in all samples.

Bayesian approach for dispersion parameter α_j

A log-normal prior is used for α_j:

$$\log \alpha_i \sim \mathcal{N}(\log \alpha_{\text{tr}}(\bar{\mu}_i), \sigma_d^2)$$

with

$$\bar{\mu}_i = \frac{1}{m} \sum_j \frac{K_{ij}}{s_j} \qquad \text{and} \qquad \alpha_{\text{tr}}(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + \alpha_0.$$

(Procedure to estimate α_i is rather complicated with several optimization steps.)

Priors/regularization for coefficients β_{ir}

For some genes only little data is available. This can lead to overfitting by choosing large values of β_{ir} that may partly cancel each other. Therefore, small (absolute) values of β_{ir} are preferred by using a prior:

$$\beta_{ir} \sim \mathcal{N}(0, \sigma_r^2)$$

This prior, however, is not reflecting the prior beliefs of the user. Instead, σ_r is adapted to the data.

Priors/regularization for coefficients β_{ir}

For some genes only little data is available. This can lead to overfitting by choosing large values of β_{ir} that may partly cancel each other. Therefore, small (absolute) values of β_{ir} are preferred by using a prior:

$$\beta_{ir} \sim \mathcal{N}(0, \sigma_r^2)$$

This prior, however, is not reflecting the prior beliefs of the user. Instead, σ_r is adapted to the data.

Applying then MAP to estimate β_{ir} corresponds to *iteratively weighted ridge regression* in frequentistic statistics.