Mean-variance modeling

Differential Expression

Kasper D. Hansen

Fall 2023

RNA-seq data

We will now turn to a number of issues related to RNA-seq data analysis

1. Variance shrinkage. This is useful whenever we do high-dimensional analysis (many genes) and have few replicates.
2. Handling the count aspect of RNA-seq data. Specifically the mean-variance relationship.

Variance, many genes, t-tests

Under the assumption of equal variance (together with the minimal assumptions), we have

$$\text{sd}(D_g) = \sigma_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where σ_g is the standard deviation of gene g.

It is hard to estimate standard deviations (variances) with small sample sizes, and in the t-test, this quantity is in the denominator: a potential problem!

In practice, this is a huge issue when we have small sample size and many genes

This model assumes we have a gene-specific variance (which makes sense), but which is hard to estimate because we have few samples.

An alternative model – which is clearly biologically unrealistic – is to assume that all genes have the same common variance. In this case, we have a large amount of data to estimate this parameter. This is an example of bias-variance trade-of:

Assumption	Consequence
gene specific variance	low bias, high variance
common variance	high bias, low variance

Variance shrinkage, many genes, t-tests

Consider a quantity like

$$\hat{\sigma}^2_{\text{shrink}} = \frac{d_0 s^2 + d_g s_g^2}{d_0 + d_g}$$

Here, we shrink the estimates of the gene-specific variances towards a common variance, trying to achieve a balance between bias and variance.

How do we pick d_0, d_g? One approach to this is using Empirical Bayes on the variance.

Using these shrunken variances gives us a **moderated** t-statistic, which has been **extremely** useful in genomics. We are "borrowing information across genes". In genomics we tend to do this **only** for the variances; one could imagine doing something along these lines for the mean parameters as well.

Counting and Poisson distributions

The most basic statistical distribution representing "counting" is the Poisson distribution.

Marioni (2008 Genome Res) and Bullard (2010 BMC Bioinformatics) established that if you sequence the exact same RNA-seq library multiple times, you get Poisson variation between the sequencing runs.

This result also holds for other type of sequencing (DNA, ChIP etc). This implies that the technical variation introduced by *sequencing* (excluding sample handling, library preparation etc) can be modeled as Poisson.

More precisely, when we sequence the same library several times to get counts $Y_1, \ldots, Y_n$, the model

$$Y_i \sim \text{Poisson}(L_i \lambda)$$

It is critical that it is *the exact same library*. This result does not mean that RNA-seq data is Poisson distributed.

Facts about the Poisson distribution

The standard model for sequencing is therefore

$$Y \sim \text{Poisson}(L\lambda)$$

where L is a known library size and λ is the unknown "expression" level of whatever you're counting. L is used to be able to compare between runs with different sequencing depth.

For a Poisson distribution with deterministic λ, the following holds

$$E(Y) = L\lambda, \quad \text{Var}(Y) = L\lambda, \quad \text{sd}(Y) = \sqrt{L\lambda}$$

In other words, the mean is the same as the variance. We have a mean-variance relationship.

Does the variance increase with the mean?

If the mean is equal to the variance, it seems as if the variance gets bigger with the mean.

The result is different for log-transformed counts. Here, the variance will *decrease* as the mean increases.

For log-transformed data, we would expect a decrease (roughly following $1/x$) stabilizing at a level reflecting biological variance (details left out).

What about RNA-seq data

We now assume that the relative expression levels λ is a random variable. This random variable incorporates

1. Technical variation beyond sequencing variation, such as library preparation, sample isolation.
2. Biological variation

Let us use the symbol Z for the random variable. Then, we know that

$$Y \sim \text{Poisson}(LZ)$$

Then we have

$$E(Y) = E(LZ), \qquad V(Y) = E(LZ) + V(LZ) = E(LZ) + CV(Z)/E^2(Z)$$

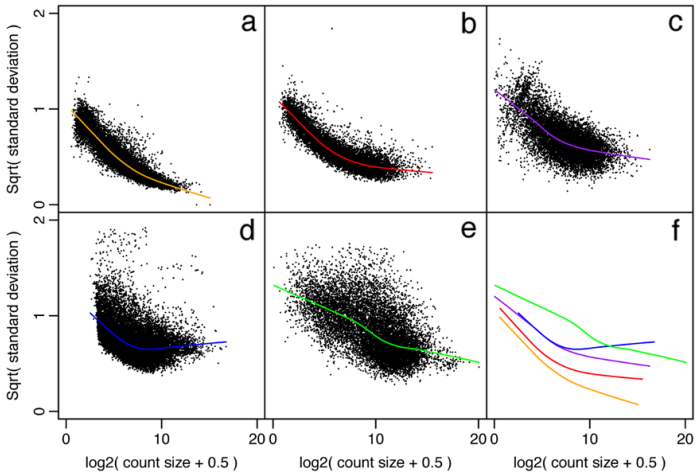
in other words, we have **over-dispersion** compared to the Poisson, because we have a second term in the formula. The term $CV(Z)$ (the coefficient of variation for the unobserved variable Z) is typically called the dispersion (although it can sometimes depend a bit on the exposition).

Negative binomial model

A negative binomial model is a special case of the previous page, where we assume Z has a Gamma distribution.

But just think of it as a count model which allows for overdispersion. For each gene, we have a dispersion parameter which you can think of as being analogue to a gene-specific variability term (but on a different scale)

Mean-variance in RNA-seq data



a: SEQC
b: BL6 mice
c: simulation
d: LCLs
e: fruit fly development

Models

There are 2 roads to modeling mean-variance relationship.

- Use models built on the negative binomial model (DESeq2, edgeR)
- Use weighted linear models (limma-voom)

Experiments show that they are very similar in performance.

SizeFactors

For RNA-seq data we have our size factors L where we typically get the relative expression by

$$Y/L = \hat{\lambda}$$

At first, it seems sensible to let L be the sequencing depth (which can be measured in multiple ways). However, we have found that this sometimes gives misleading results (best presented in Robinson et al (2010) Genome Biology), especially when you compare two groups of samples with large changes in expression such as two tissue types.

There are multiple alternatives, which all have similar performance

- TMM (trimmed mean of m-values)
- Upper quartile (UQ)
- median ratio method from DESeq2