Theoretical and sample variance of logarithmic dispersion estimates

Michael Love

October 14, 2014

1 Simulate Negative-Binomial-distributed data

The following simulates i.i.d. Negative-Binomial-distributed data, then uses *DESeq2* to obtain the gene-wise estimates of dispersion. The simulated values have a mean of $2^10 = 1024$, and true dispersion $\alpha = .05$ or .5. The variance of the finite log dispersion estimates is compared to the theoretical approximation discussed in the Methods section. A range of values for m and p is considered. When p is equal to 3, this involves the addition of another covariate, "group", in addition to the standard covariate "condition".

```
library("DESeq2")
ms <- rep(c(6, 8, 16), c(2, 4, 4))
ps <- rep(c(2, 3, 2, 3), c(4, 2, 2, 2))
alphas <- rep(c(0.05, 0.2), 5)

set.seed(1)
d <- data.frame()
for (i in seq_along(ms)) {
    m <- ms[i]
    p <- ps[i]
    alpha <- alphas[i]
    theorvar <- trigamma((m - p)/2)
    dds <- makeExampleDESeqDataSet(n = 4000, m = m, interceptMean = 10, interceptSD = 0,
        dispMeanRel = function(x) alpha)
    colData(dds)$group <- factor(rep(c("X", "Y"), times = m/2))
    design(dds) <- if (p == 2) {
        ~condition
    } else {
        ~group + condition
    }
    sizeFactors(dds) <- rep(1, ncol(dds))
    dds <- estimateDispersionsGeneEst(dds)
    disp <- mcols(dds)$dispGeneEst
    # exclude the dispersions which head to -Infinity
    samplevar <- var(log(disp[disp > 1e-07]))
    d <- rbind(d, data.frame(m = m, p = p, alpha = alpha, theorvar = theorvar,
        samplevar = samplevar))
}
```

```
library("xtable")
names(d) <- c("m", "p", "disp.", "theor. var.", "sample var.")
print(xtable(d, digits = c(0, 0, 0, 2, 3, 3)), include.rownames = FALSE)
```

m	p	disp.	theor. var.	sample var.
6	2	0.05	0.645	0.670
6	2	0.20	0.645	0.642
8	2	0.05	0.395	0.409
8	2	0.20	0.395	0.396
8	3	0.05	0.490	0.530
8	3	0.20	0.490	0.462
16	2	0.05	0.154	0.160
16	2	0.20	0.154	0.138
16	3	0.05	0.166	0.169
16	3	0.20	0.166	0.156

2 Session information

- R version 3.1.0 (2014-04-10), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: BiocGenerics 0.10.0, DESeq2 1.4.0, DESeq2paper 1.3, GenomeInfoDb 1.0.0, GenomicRanges 1.16.0, IRanges 1.21.45, LSD 2.5, MASS 7.3-31, RColorBrewer 1.0-5, Rcpp 0.11.1, RcppArmadillo 0.4.200.0, colorRamps 2.3, ellipse 0.3-8, gtools 3.3.1, schoolmath 0.4, xtable 1.7-3
- Loaded via a namespace (and not attached): AnnotationDbi 1.26.0, Biobase 2.24.0, DBI 0.2-7, RSQLite 0.11.4, XML 3.98-1.1, XVector 0.4.0, annotate 1.42.0, codetools 0.2-8, digest 0.6.4, evaluate 0.5.5, formatR 0.10, genefilter 1.46.0, geneplotter 1.42.0, grid 3.1.0, highr 0.3, knitr 1.5, lattice 0.20-29, locfit 1.5-9.1, splines 3.1.0, stats4 3.1.0, stringr 0.6.2, survival 2.37-7, tools 3.1.0