Variance-stabilizing transformation for DESeq

For parametrized dispersion fit

This file describes the variance stabilizing transformation (VST) used by DESeq when parametric dispersion estimation is used.
This is a *Mathematica* notebook. The file *vst.pdf* is produced from *vst.nb*.

When using *estimateDispersions* with *fitType="parametric"*, we parametrize the relation between mean μ and dispersion α with two constants a_0 and a_1 as follows:

$$\alpha \;=\; \mathbf{a_0 + a_1} \;/\; \mu$$

$$a_0 + \frac{a_1}{\mu}$$

In the package, a_0 is called the *asymptotic dispersion* and a_1 the *extra-Poisson factor*.

The variance is hence

$$\mathbf{v} \;=\; \mu + \alpha\,\mu^2 \;\texttt{// Expand}$$

$$\mu + \mu^2 a_0 + \mu\, a_1$$

A variance stabilizing transformation (VST) is a transformation u, such that, if X is a random variable with variance-mean relation v, i.e.,$\mathrm{Var}(X) = v(\mathrm{E}(X))$, then $u(X)$ has stabilized variance, i.e., is homoskedastic.

A VST u can be derived from a variance-mean relation v by $u(x) \;=\int^{x} \frac{d\mu}{\sqrt{v(\mu)}}$.

Hence, we can get a general VST with

$$\mathbf{u_0} = \texttt{Integrate}\left[\frac{1}{\sqrt{\mathbf{v}}}, \{\mu, 0, x\}, \texttt{Assumptions} \to \{a_0 > 0, a_1 > 0, x > 0\}\right]$$

$$\frac{\mathrm{Log}\left[\frac{1+2\,x\,a_0+a_1+2\sqrt{x\,a_0\,(1+x\,a_0+a_1)}}{1+a_1}\right]}{\sqrt{a_0}}$$

If u_0 is a VST, then so is $u(x) = \eta\, u_0(x) + \xi$. Hence, this here is a VST, too:

$$\mathbf{u} \;=\; \eta\, \mathbf{u_0} \;+\; \xi$$

$$\xi + \frac{\eta\, \mathrm{Log}\left[\frac{1+2\,x\,a_0+a_1+2\sqrt{x\,a_0\,(1+x\,a_0+a_1)}}{1+a_1}\right]}{\sqrt{a_0}}$$

We will now choose the parameters η and ξ such that our VST behaves like $\log_2$ for large values. Let us first look at the asymptotic ratio of the two transformations:

$$\texttt{Limit}[\mathbf{u} \,/\, \texttt{Log}[2, x], x \to \infty, \texttt{Assumptions} \to \{a_0 > 0, a_1 > 0, x > 0\}]$$

$$\frac{\eta\, \mathrm{Log}[2]}{\sqrt{a_0}}$$

Hence, if we set η as follows, both tranformations have asymptotically the ratio 1.

$$\eta = \frac{\sqrt{a_0}}{\mathtt{Log[2]}}$$

$$\frac{\sqrt{a_0}}{\mathtt{Log[2]}}$$

We also want the difference to vanish for large values:

$$\mathtt{Limit[u - Log[2, x], x \to \infty, Assumptions \to \{a_0 > 0, a_1 > 0, x > 0\}]}$$

$$\xi + \frac{\mathtt{Log}\left[\frac{4 a_0}{1+a_1}\right]}{\mathtt{Log[2]}}$$

So, we set

$$\xi = -\frac{\mathtt{Log}\left[\frac{4 a_0}{1+a_1}\right]}{\mathtt{Log[2]}}$$

$$-\frac{\mathtt{Log}\left[\frac{4 a_0}{1+a_1}\right]}{\mathtt{Log[2]}}$$

Check that both limits are now correct:

$$\mathtt{Limit[u / Log[2, x], x \to \infty, Assumptions \to \{a_0 > 0, a_1 > 0, x > 0\}]}$$

$$1$$

$$\mathtt{Limit[u - Log[2, x], x \to \infty, Assumptions \to \{a_0 > 0, a_1 > 0, x > 0\}]}$$

$$0$$

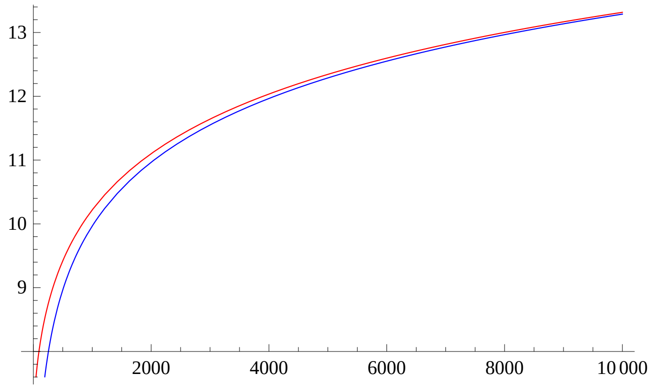
Hence, we arrive at this VST:

$$\mathtt{FullSimplify[u, Assumptions \text{ -> } \{a_0 > 0, a_1 > 0, x > 0\}]}$$

$$\frac{\mathtt{Log}\left[\frac{1+2\,x\,a_0+a_1+2\sqrt{x\,a_0\,(1+x\,a_0+a_1)}}{4\,a_0}\right]}{\mathtt{Log[2]}}$$

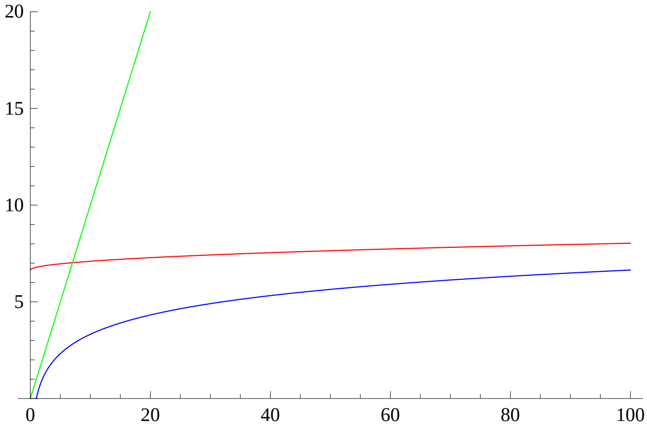
This VST (red) now behaves asymptotically as $\log_2$ (blue), shown here for typical values for a_0 and a_1.

$$\mathtt{Plot[\,\{u\ /.\ \{a_0 \to .01, a_1 \text{ -> } 3\}, Log[2, x]\}, \{x, 0, 10\,000\}, PlotStyle \to \{Red, Blue\}]}$$



For small values, however, the VST (red) compresses the dynamics much more dramatically than the logarithm (blue) and the identity (green). This reflects that the strong Poisson noise makes differences uninformative for small values.

```
Plot[ {u /. {a0 → .01, a1 -> 3}, Log[2, x], x},
 {x, 0, 100}, PlotStyle → {Red, Blue, Green}, PlotRange → {0, 20}]
```



A template for the R code in the function:

```
CForm[FullSimplify[u, Assumptions -> {a0 > 0, a1 > 0, x > 0}]] /.
 {a0 → asymptDisp, a1 → extraPois, x → q}
```

```
Log((1 + extraPois + 2*asymptDisp*q +
       2*Sqrt(asymptDisp*q*(1 + extraPois + asymptDisp*q)))/
     (4.*asymptDisp))/Log(2)
```

For local dispersion fit

In case of a local dispersion fit, the variance-stabilizing transformation $u(x) = \int^x \frac{d\mu}{\sqrt{v(\mu)}}$ is obtained by numerical integration of the fitted mean-dispersion relation $v(\mu)$ (by adding up along a asinh-spaced grid and a fitting a spline). Then, the scaling parameters η and ξ (see above) are chosen such that the VST is equal to $\log_2$ for two large normalized count values (for which the 95- and the 99.9-percentile of the sample-averaged normalized count values are used.)