Likelihood calculations for *vsn*

Wolfgang Huber

April 25, 2007

This vignette contains some of the computations that underlie the numerical code of *vsn*. If you are a new user and looking for an introduction on how to **use** *vsn*, please refer to the vignette *Robust calibration and variance stabilization with vsn*, which is provided separately.

1 Likelihood for Incremental Normalization

Here, *incremental normalization* means that the model parameters $\mu_1, \ldots, \mu_n$ and σ are already known from a fit to a previous set of μarrays, ie a set of reference arrays. See Section 2 for the profile likelihood approach that is used if $\mu_1, \ldots, \mu_n$ and σ are not known and need to be estimated from the data. The latter is the approach that was presented in the initial publication on *vsn* [1] and implemented in versions 1.X of the package.

The probability of the data is

$$P(\text{data}) = \prod_{k=1}^{n} \int_{y_k^\alpha}^{y_k^\beta} dy_k \;\; p_{\text{Normal}}(h(y_k), \mu_k, \sigma) \;\; \frac{dh}{dy}(y_k), \tag{1}$$

where $h(y) \equiv h(y, a, b) = \operatorname{arsinh}(a + by)$,

$$\frac{dh}{dy} = \frac{b}{\sqrt{1 + (a + by)^2}},$$

and the integration is over a volume element of y-space. With

$$p_{\text{Normal}}(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

the likelihood is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{k=1}^{n} \exp\left(-\frac{(h(y_k)-\mu_k)^2}{2\sigma^2}\right) \frac{dh}{dy}(y_k),$$

and the negative log-likelihood

$$-LL = n\log\left(\sqrt{2\pi}\sigma\right) + \sum_{k=1}^{n}\left(\frac{(h(y_k)-\mu_k)^2}{2\sigma^2} - \log\frac{b}{\sqrt{1+(a+by_k)^2}}\right). \quad (2)$$

This is what we want to optimize as a function of a and b. The optimizer benefits from the derivatives. The derivative with respect to a is

$$\frac{\partial}{\partial a}(-LL) = \sum_{k=1}^{n}\frac{1}{\sigma^2}\frac{h(y_k)-\mu_k}{\sqrt{1+(a+by_k)^2}} - \frac{a+by_k}{1+(a+by_k)^2} \quad (3)$$

and with respect to b

$$\frac{\partial}{\partial b}(-LL) = -\frac{n}{b} + \sum_{k=1}^{n}\left(\frac{1}{\sigma^2}\frac{h(y_k)-\mu_k}{\sqrt{1+(a+by_k)^2}} - \frac{a+by_k}{1+(a+by_k)^2}\right)y_k \quad (4)$$

2 Profile Likelihood

The derivation of the negative log profile likelihood is given in reference [2]. It turns out that the expression is contains similarities to Equation (2), and the expression for its gradient is obtained similarly as that in Equations (3) and (4).

References

[1] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stablization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics*, 18:S96–S104, 2002. 1

[2] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, Vol. 2: No. 1, Article 3, 2003. http://www.bepress.com/sagmb/vol2/iss1/art3 2