Empirical likelihood tests for nonparametric detection of differential expression from RNA-seq data

Article

Supplemental Material

Thorne, T. (2015) Empirical likelihood tests for nonparametric detection of differential expression from RNA-seq data. Statistical Applications in Genetics and Molecular Biology, 14 (6). pp. 575-583. ISSN 2194-6302 doi: 10.1515/sagmb-2015-0095 Available at https://centaur.reading.ac.uk/73955/

Published version at: http://dx.doi.org/10.1515/sagmb-2015-0095
To link to this article DOI: http://dx.doi.org/10.1515/sagmb-2015-0095

Publisher: De Gruyter

Supplementary material

Thomas Thorne[*], thomas.thorne@ed.ac.uk,
School of Informatics, University of Edinburgh

November 6, 2015

1 Empirical likelihood

For n i.i.d one dimensional observations $x_1, \ldots, x_n$ the empirical likelihood (Owen, 1988) can be defined as

$$f(x) = \prod_{i=1}^{n} p_i, \tag{1}$$

where we assign each observation a weight p_i, and constrain these such that $\sum_{i=1}^{n} p_i = 1$, $\forall i, 0 \leq p_i \leq 1$. Focussing on the empirical likelihood for the mean μ of our observations x_i, we simply require that

$$\sum_{i=1}^{n} p_i x_i = \mu. \tag{2}$$

Then we have three constraints, and aim to find the the p_i that maximise the empirical likelihood $f(x)$ under these constraints. Fortunately by using Lagrange multipliers we can find the optimal p_i by solving a one dimensional root finding problem. Defining

$$G = \sum_{i=1}^{n} \log(np_i) - n\lambda \sum_{i=1}^{n} p_i(x_i - \mu) + \gamma \left(\sum_{i=1}^{n} p_i - 1 \right), \tag{3}$$

and taking the partial derivative with respect to p_i, applying the method of Lagrange multipliers (Owen, 2001) we have

$$\frac{\partial G}{\partial p_i} = \frac{1}{p_i} - n\lambda(x_i - \mu) + \gamma = 0, \tag{4}$$

and we can solve for γ by considering

$$\sum_{i=1}^{n} p_i \frac{\partial G}{\partial p_i} = 0 \tag{5}$$

$$\sum_{i=1}^{n} (1 - n\lambda p_i(x_i - \mu) + p_i\gamma) = 0 \tag{6}$$

$$n + \gamma = 0, \tag{7}$$

[*] to whom correspondence should be addressed

since we know $\sum_{i=1}^{n} p_i(x_i - \mu) = 0$. Then substituting $\gamma = -n$ into equation 4 we have

$$\begin{aligned} \frac{1}{p_i} - n\lambda(x_i - \mu) - n &= 0 \quad (8) \\ p_i &= \frac{1}{n\lambda(x_i - \mu) + n}, \quad (9) \end{aligned}$$

and so p_i depends only on solving equation 4 for λ. We know that

$$\begin{aligned} \sum_{i=1}^{n} p_i(x_i - \mu) &= 0 \quad (10) \\ \sum_{i=1}^{n} \frac{(x_i - \mu)}{n\lambda(x_i - \mu) + n} &= 0, \quad (11) \end{aligned}$$

and so we can solve for λ for a given value of μ using a univariate root finding algorithm. Then using equation 9 we can find the p_i and calculate the empirical likelihood in equation 1.

1.1 Euclidean likelihood

The Euclidean likelihood (Baggerly, 1998) defines the log likelihood as

$$\log f(x|\mu) = -\frac{1}{2}\sum_{i=1}^{n}(np_i - 1)^2, \quad (12)$$

with the constraints $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i x_i - \mu = 0$. Again we apply the method of Lagrange multipliers (Owen, 2001)

$$G = -\frac{1}{2}\sum_{i=1}^{n}(np_i - 1)^2 - n\lambda\sum_{i=1}^{n} p_i(x_i - \mu) + \gamma\left(\sum_{i=1}^{n} p_i - 1\right), \quad (13)$$

and setting the partial derivative of G with respect to p_i to zero we have

$$\begin{aligned} \frac{\partial G}{\partial p_i} = n(1 - np_i) - n\lambda(x_i - \mu) + \gamma &= 0 \quad (14) \\ \frac{1}{n}\sum_{i=1}^{n}\left(n(1 - np_i) - n\lambda(x_i - \mu) + \gamma\right) &= 0 \quad (15) \\ -n\lambda(\bar{x} - \mu) + \gamma &= 0. \quad (16) \end{aligned}$$

Substituting $\gamma = n\lambda(\bar{x} - \mu)$ back into equation 14

$$\begin{aligned} n(1 - np_i) - n\lambda(x_i - \mu) + n\lambda(\bar{x} - \mu) &= 0 \quad (17) \\ p_i &= \frac{1}{n}\left(1 - \lambda(x_i - \bar{x})\right). \quad (18) \end{aligned}$$

Given that $\sum_{i=1}^{n} p_i(x_i - \mu) = 0$, we can substitute equation 18 to give

$$\sum_{i=1}^{n} \frac{(x_i - \mu)}{n} (1 - \lambda(x_i - \bar{x})) = 0 \tag{19}$$

$$\bar{x} - \mu - \sum_{i=1}^{n} \frac{\lambda}{n}(x_i - \mu)(x_i - \bar{x}) = 0 \tag{20}$$

$$\bar{x} - \mu - \sum_{i=1}^{n} \frac{\lambda}{n}(x_i - \bar{x})(x_i - \bar{x}) = 0 \tag{21}$$

$$\bar{x} - \mu - \lambda s = 0, \tag{22}$$

where s is defined as $s = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$. Substituting λ into equation 18 we have

$$p_i = \frac{1}{n}\left(1 - \frac{1}{s}(\bar{x} - \mu)(x_i - \bar{x})\right), \tag{23}$$

and substituting p_i into equation 12 we arrive at

$$\log f(x|\mu) = -\sum_{i=1}^{n}\left(\frac{1}{s}(\bar{x} - \mu)(x_i - \bar{x})\right)^2 \tag{24}$$

$$= -\frac{1}{s^2}(\bar{x} - \mu)^2 \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \tag{25}$$

$$= -\frac{1}{s}n(\bar{x} - \mu)^2, \tag{26}$$

References

Baggerly, K. A. (1998): “Empirical likelihood as a goodness-of-fit measure,” *Biometrika*, 85, 535–547.

Owen, A. B. (1988): “Empirical likelihood ratio confidence intervals for a single functional,” *Biometrika*, 75, 237–249.

Owen, A. B. (2001): *Empirical Likelihood*, CRC Press.