The math of DBC

Scott Olesen

The state of the problem

The mathematically critical part of DBC is deciding whether or not two OTUs are distributed identically. The original paper mentions two possibilities:

1. The χ^2 test. This has good theoretical support, since it tests exactly the question of whether two things are independent or not, but it runs into two problems. First, when counts are small, the asymptotic (and easy-to-compute) χ^2 statistic is not accurate, so the p-value needs to be simulated empirically, which is computationally expensive. Second, when counts are large, the χ^2 test seems to be *too* restrictive: the sorts of variations that we don't find unusual in microbiome data are construed as true differences by the statistic.
2. The Jensen-Shannon divergence. Cutting off at some particular JSD works well for large counts: it accords better with the sorts of differences we would consider meaningful for microbiome data. However, because the JSD works with proportions and not counts, it tends to be overly sensitive when the counts for one OTU are small. For example, if one OTU has only one count in one sample, the JSD treats that the same as if that OTU has a million counts in only that sample.

In informal discussions, I learned that:

- Correlation coefficients (e.g., Kendall τ) have similar strengths and weaknesses as the JSD: they perform well when the number of counts is small but tend to be overly sensitive when the counts are small.
- "De-blurring" or "cluster-free filtering" (which I think means de-noising?) might be productive ways to reduce the number of comparisons required.

My proposed solution

I found that, for a few example cases, the asymptotic *likelihood ratio test*, which is quick to compute, gives results similar to the simulated χ^2 test. In the few examples I tested, the p-value for the likelihood ratio test were around five times

greater than the p-value from the simulated χ^2 test. I consider this a push in the right direction, since the simulated χ^2 test tended to be too sensitive.

Definitions

There are N samples. In sample i, the number of reads assigned to the first, more abundant OTU is a_i; the other OTU has b_i. Define also $A = \sum_{i=1}^{N} a_i$ and similarly B.

Formulation of the null and alternative hypotheses

The alternative hypothesis is that the two OTUs are distributed differently, that is, that each of the a_i and b_i are all drawn from different random variables. Technical replicates from sequence data seem to be well-modeled by Poisson random variables[1], so I formulate this hypothesis as

$$H_1 : a_i \sim \text{Poisson}(\lambda_{\text{a}i}) \text{ and } b_i \sim \text{Poisson}(\lambda_{\text{b}i}),$$

where there are no constraints on the relationships between the Poisson parameters.

The null model asserts that there is some relationship between the two OTUs, specifically that second OTU is distributed "the same as" the first one, just rescaled to some lower abundance. I articulate this as

$$H_0 : a_i \sim \text{Poisson}(\lambda_i) \text{ and } b_i \sim \text{Poisson}(\sigma\lambda_i),$$

that is, that the a_i are all free to be distributed differently from one another, but each b_i is constrained to be distributed according a Poisson random variable that has the same mean as a_i's random variable, just rescaled by some common scaling factor σ. (Because the second OTU is less abundant, we expect that $0 < \sigma < 1$.)

Maximum likelihood of models

The likelihood ratio test will require the maximum likelihood with respect to those parameters. Not surprisingly, this requirement implies that $\lambda_{\text{a}i} = a_i$ and $\lambda_{\text{b}i} = b_i$, i.e., that the best estimates for the Poisson parameters are just the single number that distribution produces. For the null hypothesis, we find that $\sigma = \frac{B}{A}$, i.e., the scaling factor is just the ratio of the total counts for the two OTUs, and

$$\lambda_i = \frac{A}{A+B}(a_i + b_i).$$

[1] McMurdie and Holmes, 2014. doi:10.1371/journal.pcbi.1003531

Inserting these variables shows that the log likelihoods are:

$$\mathcal{L}_0 = -(A+B) + \sum_i \left(a_i \ln a_i + b_i \ln b_i - \ln a_i! - \ln b_i!\right)$$

$$\mathcal{L}_1 = -(A+B) + A\ln A + B\ln B - (A+B)\ln(A+B) + \sum_i \left[(a_i+b_i)\ln(a_i+b_i) - \ln a_i! - \ln b_i!\right]$$

The difference between the log can be conveniently written in terms of a helper function

$$f(\boldsymbol{x}) \equiv \sum_i x_i \ln x_i - \left(\sum_i x_i\right) \ln\left(\sum_i x_i\right)$$

so that

$$\mathcal{L}_1 - \mathcal{L}_0 = f(\boldsymbol{a}+\boldsymbol{b}) - f(\boldsymbol{a}) - f(\boldsymbol{b}).$$

The statistic

The likelihood ratio test uses the statistic $\Lambda = -2\left(\mathcal{L}_1 - \mathcal{L}_0\right)$, which is distributed according to a χ^2 distribution with $(N-1)$ degrees of freedom. (This is the difference in the number of parameters in the two models: the alternative has $2N$, i.e., one for each OTU and sample, and the null has $N+1$, one for each sample and the scaling factor σ.) The cumulative distribution function of χ^2 at Λ is easy to compute.

Comparisons with other solutions

The χ^2 test

The "vanilla" Pearson's χ^2 test has a statistic

$$\chi^2 = \sum_i \sum_t \frac{(O_{ti} - E_{ti})^2}{E_{ti}},$$

where t stands for the two taxa (a and b), O_{ti} is the observed number of counts in that cell of the table, and E_{ti} is the expected number of counts. The expected counts are computed using the marginals, e.g.,

$$E_{ai} = (a_i + b_i)\frac{A}{A+B}.$$

Plugging in these values gives

$$\chi^2 = (A+B)\sum_i \left(\frac{a_i}{a_i+b_i}\frac{a_i}{A} + \frac{b_i}{a_i+b_i}\frac{b_i}{B} - 1\right),$$

which does not bear any immediate obvious relationship to the other statistic Λ.

JSD

Defining $m_i = \frac{1}{2}(a_i + b_i)$, then the JSD between the two taxa is

$$\mathrm{JSD} = \frac{1}{2}\sum_i a_i \log\frac{a_i}{m_i} + \frac{1}{2}\sum_i b_i \log\frac{b_i}{m_i},$$

which simplifies to

$$\mathrm{JSD} = \sum_i \left[\frac{1}{2}\left(a_i \log a_i + b_i \log b_i\right) + (a_i + b_i)\log(a_i + b_i)\right].$$

This bears a greater resemblance to the equation for Λ, excepting some factors of two and the "sum" terms in $f(\boldsymbol{x})$.