0.1 Simple statistics for differential proportionality

We start by introducing a short-hand notation allowing to denote projections of log-ratios of two vectors $\mathbf{x}, \mathbf{y}$ onto a subset of size k of the vectors' n components:

$$\mathrm{L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k} := \left(\log \frac{x_1}{y_1}, \ldots, \log \frac{x_k}{y_k}\right). \tag{1}$$

Equivalently, the log-ratio mean (LRM) and variance (LRV) evaluated on this subset are denoted by $E(\mathrm{L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k})$ and $\mathrm{var}(\mathrm{L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k})$ respectively. Let us now assume we have a natural partition of our n samples into two subsets (conditions) of experimental replicates of sizes k and $n-k$. To avoid clutter, we drop $\mathbf{x}, \mathbf{y}$ from the notation in the following equation. It is well known that variance evaluates to

$$\begin{aligned}
\mathrm{var}(\mathrm{L}_{1,\ldots,n}) &= E(\mathrm{L}^2_{1,\ldots,n}) - E^2(\mathrm{L}_{1,\ldots,n}) \\
&= \frac{kE(\mathrm{L}^2_{1,\ldots,k}) + (n-k)E(\mathrm{L}^2_{k+1,\ldots,n})}{n} - \frac{\left(kE(\mathrm{L}_{1,\ldots,k}) + (n-k)E(\mathrm{L}_{k+1,\ldots,n})\right)^2}{n^2} \\
&= \frac{kE^2(\mathrm{L}_{1,\ldots,k}) + (n-k)E^2(\mathrm{L}_{k+1,\ldots,n})}{n} + \frac{k\mathrm{var}(\mathrm{L}_{1,\ldots,k}) + (n-k)\mathrm{var}(\mathrm{L}_{k+1,\ldots,n})}{n} \\
&\qquad - \frac{\left(kE(\mathrm{L}_{1,\ldots,k}) + (n-k)E(\mathrm{L}_{k+1,\ldots,n})\right)^2}{n^2} \\
&= \frac{k(n-k)}{n^2}\left(E(\mathrm{L}_{1,\ldots,k}) - E(\mathrm{L}_{k+1,\ldots,n})\right)^2 + \frac{k\mathrm{var}(\mathrm{L}_{1,\ldots,k}) + (n-k)\mathrm{var}(\mathrm{L}_{k+1,\ldots,n})}{n}. \qquad (2)
\end{aligned}$$

This is the well-known decomposition into between-group variance (first term) and within-group variance (second term) known from ANOVA. Note that the variances are here defined as the biased estimators (so the sum of squares are divided by k rather than $k-1$, with k the number of summands). As will be seen from the discussion below, differential proportionality can be studied relative to LRV and and there is no need for evaluation of the total size of LRV (which is a problem when studying proportionality across all the samples). If we divide (2) by $\mathrm{var}(\mathrm{L}_{1,\ldots,n})$, we obtain the various proportions of (weighted) group variances and of the between-group variance to the overall variance. For illustration, this is visualized as a ternary diagram in the upper left panel of Figure 1. The proportion of within-group variance with respect to overall variance is thus a function of the three LRVs:

$$\vartheta(\mathbf{x},\mathbf{y}) = \frac{k\mathrm{var}\ \mathrm{L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k} + (n-k)\mathrm{var}\ \mathrm{L}^{\mathbf{x},\mathbf{y}}_{k+1,\ldots,n}}{n\mathrm{var}\ \mathrm{L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,n}}. \tag{3}$$

Conveniently, ϑ is a number between zero and one. When approaching zero it indicates that the total LRV is explained by the squared difference in group LRMs (see upper right panel of Figure 1). A large enough difference means that scatter plots of $\mathbf{y}$ vs. $\mathbf{x}$ will have different slopes depending on the condition the samples come from. This case is thus characterized by tissue-specific proportionality factors (group LRMs). We call this type of differential proportionality *disjointed* proportionality here. We can use ϑ for testing this property on our vector pairs and evaluate its significance using a simple permutation test for an estimate of the false discovery rate. Alternatively, a classical test-statistic known from one-way ANOVA with two groups is the squared t-statistic F. It is related to ϑ by

$$F = (n-2)\frac{(1-\vartheta)}{\vartheta}. \tag{4}$$

This statistic can be used to do a classical F-test of the null hypothesis of equal group (population) LRMs under standard ANOVA assumptions.

We have seen that disjointed proportionality describes pairs where between-group variance constitutes the major part of their LRV. Another type of differential proportionality can be defined for those pairs where one of the group LRVs dominates the total LRV. A scatter of $\mathbf{y}$ vs. $\mathbf{x}$ will then show proportionality for samples in one condition but no correlation for the other condition. We

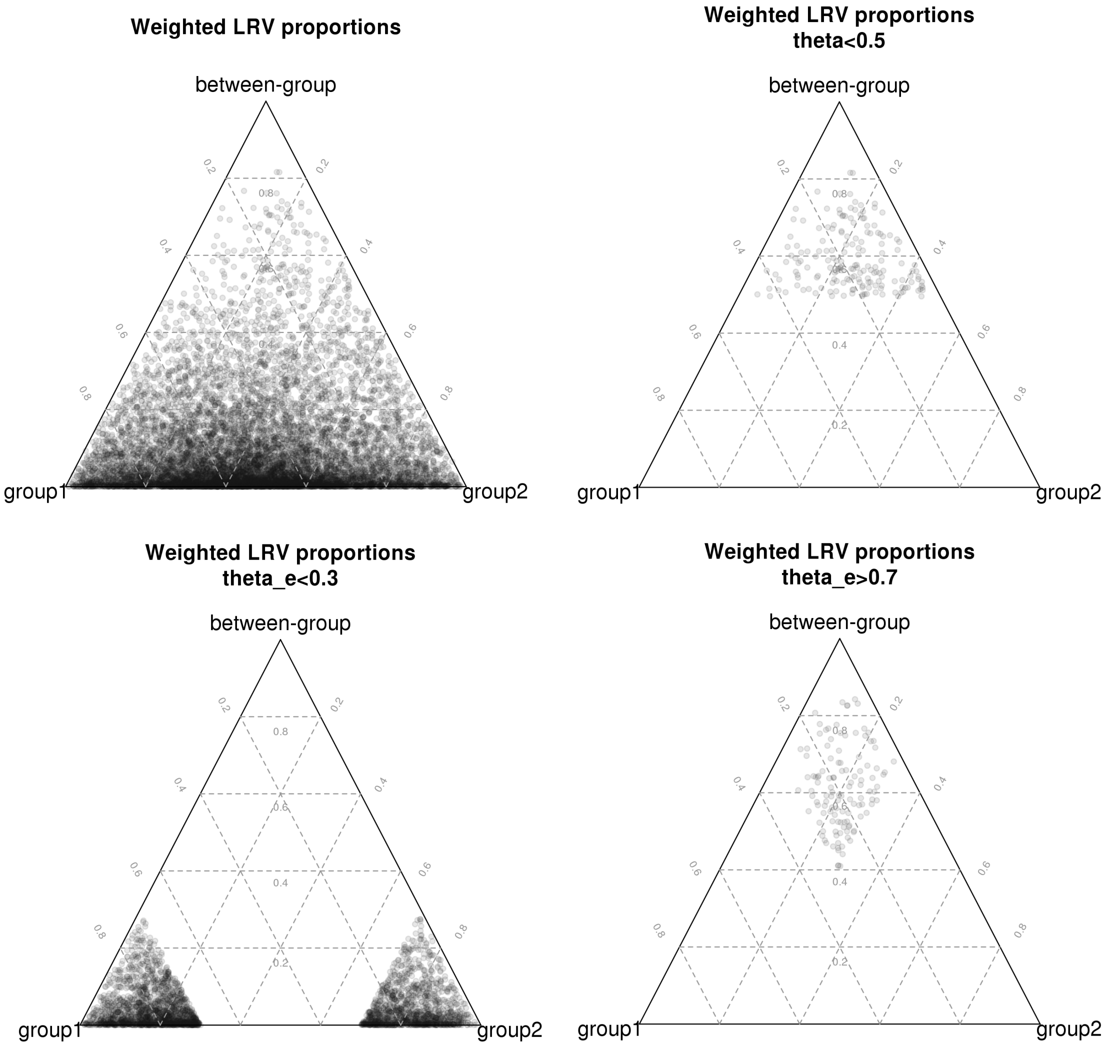

Figure 1: Decomposition of log-ratio variance into (weighted) group variances and between-group variance shown in a ternary diagram. The data shown come from 10,000 randomly sampled gene pairs with expression in 54 cerebellum (group 1) and 44 cortex (group 2) samples in human. *Upper left:* The 10,000 dots corresponding to LRVs of each gene pair. *Upper right:* Gene pairs fulfilling $\vartheta < 0.5$ (disjointed proportionality). *Lower left:* Gene pairs fulfilling $\vartheta_{\mathrm{e}} < 0.3$ (emergent proportionality). *Lower right:* Gene pairs fulfilling $\vartheta_{\mathrm{e}} > 0.7$. A cut-off from below induces a cut-off on ϑ and an additional restriction on the difference between weighted group variances.

will call this type of proportionality *emergent* to distinguish it from disjointed proportionality. In complete analogy to the definition of ϑ, from (2) we get

$$\vartheta_1(\mathbf{x},\mathbf{y}) = \frac{n\text{var L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,n} - k\text{var L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k}}{n\text{var L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,n}}, \tag{5}$$

as the proportion of the sum of between-group variance and the LRV of group 2 to the total LRV. Small values of ϑ_1 indicate that the LRV of group 1 constitutes the major part of the total LRV, which is our defining feature of emergent proportionality in group 2. A convenient measure for detecting emergent proportionality regardless of group can be defined as

$$\vartheta_\text{e}(\mathbf{x},\mathbf{y}) = 1 - \frac{\max\left(k\text{var L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,k}, (n-k)\text{var L}^{\mathbf{x},\mathbf{y}}_{k+1,\ldots,n}\right)}{n\text{var L}^{\mathbf{x},\mathbf{y}}_{1,\ldots,n}}, \tag{6}$$

of which a cut-off from above will give us the a set of pairs that are proportional in just one of the two conditions (see lower left panel of Figure 1). Let us now look at the relationship between ϑ_e and ϑ. Note that we have

$$\vartheta_\text{e} = 1 - \vartheta + \frac{\min\left(k\text{var L}_{1,\ldots,k}, (n-k)\text{var L}_{k+1,\ldots,n}\right)}{n\text{var L}_{1,\ldots,n}}. \tag{7}$$

It follows that

$$1 - \vartheta \leq \vartheta_\text{e} \leq 1 - \vartheta/2, \tag{8}$$

with the equality $1 - \vartheta = \vartheta_\text{e}$ holding if one of the group LRVs vanishes and $\vartheta_\text{e} = 1 - \vartheta/2$ in the case of equality of weighted group LRVs $k\text{var L}_{1,\ldots,k} = (n-k)\text{var L}_{k+1,\ldots,n}$. It transpires that ϑ_e can be used to study both types of differential proportionality since large values of it enforce small ϑ. For this, a second cut-off on ϑ_e, this time from below, needs to be determined. However, note that a cut-off $\vartheta_\text{e} > C$ would enforce a somewhat stricter definition on disjointed proportionality, where the induced cut-off $\vartheta < 2(1-C)$ can only be attained for equality of weighted group LRVs, a condition that is relaxed when going further down with ϑ. In fact, cut-offs from below on ϑ_e cut the upper corner of the ternary diagram with two lines that yield a diamond shape as opposed to the triangle that results from a cut-off on ϑ (see lower right panel of Figure 1).