Adjusting the proportion of non-differentially expressed genes

In order to make realistic power and sample size calculations the distribution of effect sizes needs to be known. A few authors [1–3] have proposed methods to estimate the distribution of effect size from pilot data, respectively using a deconvolution estimator, expectation-maximization algorithm or a spline-model. We use the deconvolution estimator [1]. The proportion of non-differentially expressed genes, $\pi_0$, is the first quantity to be estimated. The empirical density of the test statistics, $m$, is estimated from the set of test statistics from the pilot data.

The deconvolution involves solving the following equation for $\lambda$:

$$m(t) - \pi_0 \phi(t) = \int_{-\infty}^{+\infty} \phi(t - \theta \sqrt{N}) \lambda(\theta) d\theta,$$

where $\lambda$ is the density of effect sizes, $\phi$ represents the density of the test statistics under $H_0$ and $\theta$ represents the effect size. The left-hand side of Equation 1 expresses the difference between the observed density of test statistics $m$ and that of the assumed density of test statistics under $H_0$ (standard Normal, $\phi$), weighted by the proportion of non-differentially expressed genes. This linear combination of densities puts certain constraints on the value of $\pi_0$. The constrain can be formulated as:

$$\min_i \left( m(t_i) - \pi_0 \phi(t_i) \right) \geq 0 \Rightarrow \pi_0 \leq \min_i \left( \frac{m(t_i)}{\phi(t_i)} \right).$$

This constraint is used to adjust the value of $\pi_0$ in order to guarantee that the estimated density of effect size is continuous and non-negative.

References


Adjustments of density of effect sizes: The three curves solid(blue)-, short-dashed(pink)-, dotted(green)-lines represent respectively the adjusted density, a density truncated near zero and the unadjusted non-valid density. On the x-axis is the standardized effect size, and on the y-axis is the estimated density.