Supplementary material for "Dissecting systems-wide data using mixture models: application to differentially expressed genes in the BRCA network"

Peter Svensson¹,², Renée X. de Menezes³, Ingela Turesson², Micheline Giphart-Gassler¹, Harry Vrieling*¹

¹Department of Toxicogenetics, Leiden University Medical Centre, P.O. Box 9503, 2300 RA Leiden, the Netherlands
²Department of Oncology, Radiology and Clinical Immunology, Academic Hospital, 751 85 Uppsala, Sweden
³Department of Medical Statistics, Leiden University Medical Centre, P.O. Box 9604, 2300 RA Leiden, the Netherlands

Email: Peter Svensson - p.svensson@lumc.nl; Renée X. de Menezes - r.x.menezes@lumc.nl; Ingela Turesson - ingela.turesson@onkologi.uu.se; Micheline Giphart-Gassler - m.giphart-gassler@lumc.nl; Harry Vrieling* - h.vrieling@lumc.nl;

*Corresponding author

1 Calculation of derivatives

The probability density function, \( f_p(p) \) is used as in

\[
P\{ p \in [0, \delta] \} = \int_{[0,\delta]} f_p(p) dp
\]  

which is related to the cumulative density function (probability function) as

\[
y = F_p(p) = \int_0^p f_p(u) du
\]

The mirror of the function is

\[
p = F^{-1}(y) = g(y)
\]

where \( F^{-1}(p) \) is the inverse of \( F(p) \). We are interested in the maximum of the second derivative of \( g(y) \), i.e. where the third derivative \( g'''(y) \) is zero:\

\[
g'''(y) = \frac{3f'(p)^2 - f(p)f''(p)}{f(p)^5} \Bigg|_{p=F^{-1}(y)}
\]

Under the condition \( g'''(y_0) = 0 \), we get

\[
f(p_0)f''(p_0) = 3f'(p_0)^2
\]

\[\text{1}\]
2 Exponential distributions

The density function arising from a mixture of distributions in general can be written \( f(p) = \sum_k \pi_k f_k(p) \) for \( f(p) \) consisting of \( k \) distributions, each affecting a proportion of \( \pi_k \) genes. \( f(p) = \frac{\lambda}{1-e^{-\lambda p}} e^{-\lambda p} \) where \( \lambda \geq 0 \) and \( 0 \leq p \leq 1 \). The distribution of the p-values of the null features, \( f_0(p) \), are uniformly distributed over \( p = [0, 1] \).

As \( f(p) \) is a sum of independent functions, the derivatives will also be sums of functions.

\[
f'(p) = \sum_k \pi_k f'_k(p), \quad f''(p) = \sum_k \pi_k f''_k(p)
\] (6)

At the point \( p_0 \) where \( g''(y_0) = 0 \) \( y_0 = F^{-1}(p_0) \). From equation (5) follows

\[
\pi_0 f''(p_0) = 2(1 - \pi_0)^2 - \sum_{i<j} \pi_i \pi_j (f_i(p_0)f_j(p_0))',
\] (7)

at the point \( p_0 \) where \( g''(y_0) = 0 \) \( y_0 = F^{-1}(p_0) \). If \( \lambda_i \ll \lambda_1 \) the second term becomes negligible. The roots of equation (7) will provide equations to determine \( \lambda_i \) and \( \pi_i \) for \( i \in [1, k] \). Depending on \( f(p) \) there may be several roots to equation (5).

In the special case where the p-values come from a mixture of null features with a uniform distribution, \( f_0(p) \), and a single population of alternative features following the exponential distribution \( f_1(p) \), then \( f(p) = \pi_0 + (1 - \pi_0)f_1(p) \). Equation (5) leads to

\[
f_1(p_0) = \frac{\pi_0}{2(1 - \pi_0)},
\] (8)

With \( f(p_0) = \frac{1}{2} \pi_0 \), we can read off an estimate of \( \pi_0 \) from the density function, as two thirds of \( f(p_0) \). There are several other ways described to estimate \( \pi_0 \) and each method will give slightly different results. We can estimate \( \lambda_1 \) and fully describe the distribution, \( f_1(p) \), making it possible to select any point as a threshold.
The height of the maximum of the second derivative is increasing with \( \lambda \) on a reasonable interval. For a single exponential distribution of alternative genes, 

\[
g''(y_0)|_{y_0=F(p_0)} = -\frac{e^{2\lambda p_0}}{27\lambda(1-e^\lambda)^2(1-\pi_0)^2}
\]

This means that \( g''(y_0) \) is an indicator for the overlap between the populations of genes.

### 3 Beta distributions

The Beta distribution is defined on the closed interval \( p = [0,1] \), and described by two parameters, \( \alpha \) and \( \beta \).

\[
f(p) = \frac{(1-p)^{\beta-1}p^{\alpha-1}}{B(\alpha, \beta)} = \frac{e^{(\beta-1)\log(1-p)+(\alpha-1)\log(p)}}{B(\alpha, \beta)} = \frac{e^{h(p)}}{B(\alpha, \beta)}
\]

with

\[
h(p) = (\beta - 1)\log(1-p) + (\alpha - 1)\log(p) \tag{14}
\]

\[
h'(p) = -\frac{\beta - 1}{1-p} + \frac{\alpha - 1}{p} \tag{15}
\]

\[
h''(p) = -\frac{\beta - 1}{(1-p)^2} - \frac{\alpha - 1}{p^2} \tag{16}
\]

If we assume \( f(p) \) to consist of a beta distribution of changed genes \( f_1 \) and a uniform distribution of unchanged genes, \( \pi_0 \), equation (5) leads to

\[
\pi_0 \left(h'_1(p_0)^2 + h''_1(p_0)\right) = (1-\pi_0)f_1(p_0) \left(2h'_1(p_0)^2 - h''_1(p_0)\right) \tag{17}
\]

If we restrict ourselves to explaining the distributions when \( \alpha = 1 \), the expression can be further simplified to

\[
f_1(p_0) = \frac{\pi_0(\beta - 2)}{(1-\pi_0)(2\beta - 1)}, \quad \beta \neq \frac{1}{2} \tag{18}
\]

For large values of \( \beta \), \( \frac{\beta - 1}{2\beta - 2} \approx 2 \), and we end up with the same expression as for the exponential distribution.