LEMMA - Brief Summary of the Linear Model

Bar, Booth, Schifano and Wells (2009) define the mixed effect mixture linear model:

$$y_{ijg} = \mu + \tau_i + \gamma_g + \psi_{ig} + \epsilon_{ijg}$$

where:

- y_{ijg} is the (normalized) response for gene g, individual j, in group i, where $g = 1, \ldots, G$, $j = 1, \ldots, n_{ig}$, $i = 1, 2$;
- μ is the overall mean;
- τ_i is the (fixed) effect of treatment group i, constrained such that $\tau_1 + \tau_2 = 0$;
- γ_g is the (random) effect of gene g, assumed to be distributed $N(0, \sigma^2_\gamma)$;
- ψ_{ig} is the (random) gene-treatment interaction for treatment group i and gene g; constrained such that $\psi_{1g} + \psi_{2g} = 0$, but where the difference $\psi_g \equiv \psi_{1g} - \psi_{2g}$ is assumed to be distributed $N(\psi, \sigma^2_\psi)$ for genes in the nonnull group, and identically 0 for genes in the null group.
- ϵ_{ijg} is the (random) gene-specific error for gene g, assumed to be distributed $N(0, \sigma^2_{\epsilon,g})$.

Further assume that $\sigma^2_{\epsilon,g} \sim$ i.i.d. IG(α, β) and define the latent indicator variables $\{b_g\}$ of nonnull status, such that $b_g \sim$ i.i.d. Bern(p_1), where p_1 is the prior probability of a gene being nonnull.

This model is referred to as the RR model - Random gene×treatment interaction (ψ_{ig}), Random gene-specific error variances ($\sigma^2_{\epsilon,g}$).

We implemented a generalization of this model, in which we assume that there are two nonnull groups – one in which genes from treatment group #1 are more expressed than genes from group #2, and one in which genes from treatment group #2 are more expressed than genes from group #1. In this generalized model, we define the latent indicator variables $\{b_{0g}\}, \{b_{1g}\}, \{b_{2g}\}$ with prior probabilities p_0, p_1, p_2, respectively, such that $p_0 + p_1 + p_2 = 1$ and

$$\begin{array}{ll} b_{0g} = 1 & \text{iff } \psi_g \equiv 0 \\ b_{1g} = 1 & \text{iff } \psi_g \backsim N(\psi, \sigma^2_\psi) \\ b_{2g} = 1 & \text{iff } \psi_g \backsim N(-\psi, \sigma^2_\psi). \end{array}$$

For estimation, consider the sufficient statistics:

- $s_g = \bar{y}_{1\cdot g} + \bar{y}_{2\cdot g}$
- $d_g = \bar{y}_{1\cdot g} - \bar{y}_{2\cdot g}$
- $m_g = \sum_i \sum_j^{n_{ig}} (y_{ijg} - \bar{y}_{i\cdot g})^2 / f_g$

where m_g are the mean squared errors, with $f_g = n_{1g} + n_{2g} - 2$ degrees of freedom.

The goal is to determine for each gene g whether it belongs to the null group (no gene-treatment interaction), or of of the two nonnull group (significant interaction). The term *treatment* is generic - it could literally be a treatment vs. control comparison, or any other pair of conditions, such as "older than X" vs. "younger than X", or "male" vs. "female", etc.

The LEMMA estimation approach and software uses data ($\{d_g\}$, $\{m_g\}$) to estimate parameters p_1, p_2, $\tau \equiv \tau_1 - \tau_2$, ψ, σ^2_ψ, α, and β. The parameters μ and γ_g are not estimated, as their effects cancel in the differences $\{d_g\}$. Based on the estimates, the software provided classifies genes as nonnull (being associated with the treatment) based on either local fdr values (Efron, 2005) or FDR-adjusted p-values (Benjamini and Hocberg, 1995).