Regression Methods for GSEA

AJ Fagan

June 3, 2024

Gene Set Enrichment Analysis

Goal: from some gene-level data, determine which of some gene sets are of interest in the present data

GSEA in DE Analysis

- DE output: list of scores for each gene (insert example here)
- Fed as input into GSEA methods

Permutation Methods

- Construct binary scores for each gene
- Determine, for each gene set, if it is enriched for 1's compared to a random set of the same size, or to the remaining genes
- Simplest example is Fisher's Exact Test to see if a score of 1 is associated with inclusion in each gene set

Problem - Gene Set Overlap

Many gene sets, such as those from the Gene Ontology (GO), have considerable overlap. As a result, methods like those above will often return hundreds of highly correlated gene sets.

Solution - Multiset Methods

Instead of considering only one gene set at a time, methods like SetRank and the Rolemodel consider every gene set when determining significance of any other. This enables overlap to be addressed, reducing the sizes of generated lists substantially.

Problem - Gene-Gene Correlation

Another problem with permutation-based methods is that they fail to account for inter-gene correlation. As an example, if a gene set has 5 perfectly correlated genes found as hits, and 5 uncorrelated genes found as misses, this set should be treated closer to 1/6, then 5/10.

Solution - Multivariate Regression Methods

Methods such as ROAST utilize multivariate regression techniques to account for inter-gene correlation. These methods cannot operate solely on gene-level output of some pre-existing DE analysis, as such data would lose all inter-gene information.

Proposed Work

As far as I can find, no method currently exists that enables multiset, multivariate regression techniques that can adequately account for both the inter-gene correlation, and the gene set-overlap problem. Such a method would permit both multiset methods' cleanliness of returned gene sets, as well as the robustness to inter-gene correlation offered by the multivariate regression methodology.

Data Model

Let $g = 1, \ldots G$ denote the genes present in our analysis, and let $s = 1, \ldots S$ denote the gene sets. We model, for $i = 1, \ldots n$:

$$y_i = X_i\alpha + Z_i\beta + \varepsilon_i,$$

where

- y_i is the length G vector of gene expression data for sample i
- X_i is the size $G \times k_0$ design matrix for sample i under the null of non-differential expression
- Z_i is the size $G \times k_A$ design matrix for sample i under the alternative of differential expression
- and ε_i is an error term.

The Rolemodel

For each gene g, let $A_g \in \{0, 1\}$ denote that gene's "activity". Similarly, for each gene set s, let $T_s \in \{0, 1\}$ denote that gene set's "activity".

The Rolemodel asserts that, for each gene g,

$$A_g = 1 \iff \exists s, g \in s, T_s = 1.$$

Traditionally, it operates on a set of observed gene-level binary responses $\hat{A}_g$, informing the likelihood of $P(\hat{A}_g | A_g)$, forcing

$$P(\hat{A}_g | A_g = 1) > P(\hat{A}_g | A_g = 0).$$

From this, it constructs the posterior probability that $T_s = 1$ for each gene set s.

The Rolemodel - Expanded

To extend the Rolemodel to function in the multivariate regression context, we simply alter the role of A_g in the model. Here, we use the latent A_g to inform the prior on β_g:

$$\beta_g \sim (1 - A_g)F_0(\beta_g|\lambda_0) + A_g F_1(\beta_g|\lambda_1),$$

where $F_0(\beta_g|\lambda_0)$ is a null distribution, and $F_1(\beta_g|\lambda_0)$ is a distribution indicative of DE.

Example - Atovaquone vs DMSO

Design - three batches ($j = 1, 2, 3$), each containing:

- Three times ($t = 1, 8, 24$ hours)
- 3 treatments ($i =$ DMSO, IC50, IC75)
- 3 replicates ($k = 1, 2, 3$)

Goal:

Explain DE in terms of GO functions.

Example (cont'd)

Let

$$y_{itjk} = \mu + \alpha_j + \gamma_t + \beta_i + (\gamma\beta)_{it} + \varepsilon_{itjk},$$

where

- α_j indicates the batch mean for each gene,
- γ_t indicates the (discrete) time effect on each gene,
- β_i indicates the treatment effect on each gene, and
- $(\gamma\beta)_{it}$ indicates the time-treatment combination effect on each gene.

Then, under the null of non-DE, for each i and t,

$$\beta_i = (\gamma\beta)_{it} = 0.$$

Example - Data Likelihood

At this point, DE models such as DESeq2 or multiset GSEA models such as (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3954234/R2) employ some discrete data distribution (negative binomial and non-central hypergeometric, respectively) to model y_{itjk}.

However, such discrete distributions are unable to account for the covariance structure of the gene expression. Therefore, we employ log normalized-counts as our dependent variable, and model,

$$y_{itjk} - \hat{y}_{itjk} = \varepsilon_{itjk} \sim N(0, \Sigma).$$

Example - Priors

We model priors for our coefficients as

$$\begin{aligned}
\mu &\sim 1, \\
\alpha_j &\sim N(0, \sigma^2_{batch} I), \\
\gamma_t &\sim N(0, \sigma^2_{time} I), \\
\beta_i | A &\sim (1 - A) \times I(\beta = 0) + A \times N(0, \sigma^2_{treat} I), \\
(\gamma\beta)_{it} &\sim (1 - A) \times I(\beta = 0) + A \times N(0, \sigma^2_{interact} I),
\end{aligned}$$

and each $\sigma^2_{\cdot} \sim IG(0.001, 0.001)$.

However, the prior for Σ is a bit more tricky.

Example - Prior Variance

The human genome contains approximately 20_000 genes. Therefore, Σ would be approximately $20\_000 \times 20\_000$. Assuming Σ was populated with 64-bit floating point numbers, this matrix would occupy 3.2GB worth of RAM.