Functional Enrichment by Expectation Maximization

Nicholas Knoblauch

May 13, 2020

Contents

Introduction

Method

Model

Suppose we have gathered genetic data for a set of individuals to identify which genes are causally related to a disease or trait of interest. For each gene $g \in \{1 \dots G\}$, let the indicator variable $z_g = 1$ indicate that gene g is causaully related to our trait or disease of interest. We can summarise the evidence for and against our hypothesis that $z_g = 1$ using a bayes factor:

$$B_g = \frac{P(x_g|z_g = 1)}{P(x_g|z_g = 0)}$$

where x_g is the subset of the aforementioned genetic data corresponding to the gth gene.

Suppose further that we know a set of F functional annotations or properties for each of our G genes. Let $\mathbf{a}_g$ denote the length F vector of annotations for gene g, and $\mathbf{A}$ denote the matrix with F rows and G columns consisting of $\mathbf{a}_1 ... \mathbf{a}_G$ We define the vector $\boldsymbol{\beta}$ and the function $\pi(\boldsymbol{\beta}, \mathbf{a}_g)$ such that:

$$P\pi(\boldsymbol{\beta}, \mathbf{a}_g) = \frac{1}{1 + e^{-(\beta_0 + \sum_{f=1}^{F} A_{f,g}\beta_f)}} = (z_g = 1|\mathbf{a}_g, \boldsymbol{\beta})$$

We can compute the likelihood of a particular value of $\boldsymbol{\beta}$ as:

$$P(\mathbf{x}|\boldsymbol{\beta}, \mathbf{A}) = \prod_{g=1}^{G} P(x_g|\boldsymbol{\beta}) = \prod_{g=1}^{G} [\pi(\boldsymbol{\beta}, \mathbf{a}_g) P(x_g|z_g = 1) + (1 - \pi(\boldsymbol{\beta}, \mathbf{a}_g)) P(x_g|z_g = 0)]$$

By factorizing out the term $\prod_{g=1}^{G} P(x_g|z_g = 0)$ which does not depend on $\boldsymbol{\beta}$, we can express the likelihood for $\boldsymbol{\beta}$ in terms of $\mathbf{B}$:

$$P(\mathbf{x}|\boldsymbol{\beta}, \mathbf{A}) \propto \prod_{g=1}^{G} [\pi(\boldsymbol{\beta}, \mathbf{a}_g) B_g + (1 - \pi(\boldsymbol{\beta}, \mathbf{a}_g))]$$

Given a particular value of $\boldsymbol{\beta}$, we arrive at a new posterior:

$$P(z_g = 1|x_g, \boldsymbol{\beta}, \mathbf{a}_g) = \frac{\pi(\boldsymbol{\beta}, \mathbf{a}_g) B_g}{\pi(\boldsymbol{\beta}, \mathbf{a}_g) B_g + 1 - \pi(\boldsymbol{\beta}, \mathbf{a}_g)}$$

Algorithm

Our goal is both to estimate $\boldsymbol{\beta}$ as well as $P(Z_g = 1|\mathbf{a}_g, \boldsymbol{\beta}, x_g)$ for each gene. We use the EM algorithm [1] to obtain a maximum likelihood estimate for $\boldsymbol{\beta}$. If $\mathbf{z}$ is our latent variable then our complete data likelihood is:

$$P(\mathbf{x}, \mathbf{z}|\boldsymbol{\beta}, \mathbf{A}) = \prod_{g=1}^{G} [P(z_g|\boldsymbol{\beta}, \mathbf{a}_g) P(x_g|z_g)] = \prod_{g=1}^{G} \left[\pi(\boldsymbol{\beta}, \mathbf{a}_g)^{z_g} (1 - \pi(\boldsymbol{\beta}, \mathbf{a}_g))^{1-z_g} B_g^{z_g}\right]$$

From which we form the Q function

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}_n) = \sum_{g=1}^{G} [P(z_g = 1|x_g, \boldsymbol{\beta}_n, \mathbf{a}_g) \log (P(x_g, z_g = 1|\boldsymbol{\beta})) + P(z_g = 0|x_g, \boldsymbol{\beta}_n, \mathbf{a}_g) \log (P(x_g, z_g = 0|\boldsymbol{\beta}))] =$$

Where $u_g = P(z_g = 1|x_g, \boldsymbol{\beta_n}, \mathbf{a}_g) = \frac{\pi(\boldsymbol{\beta_n}, \mathbf{a}_g) B_g}{\pi(\boldsymbol{\beta_n}, \mathbf{a}_g) B_g + 1 - \pi(\boldsymbol{\beta_n}, \mathbf{a}_g)}$
The model fitting procedure proceeds as follows

- Start with an initial guess of $\boldsymbol{\beta}$
- Repeat until $Q(\boldsymbol{\beta}|\boldsymbol{\beta}_{n-1}) - Q(\boldsymbol{\beta}|\boldsymbol{\beta}_n) <$ tol:
  - Compute $P(z_g = 1|x_g, \boldsymbol{\beta_n}, \mathbf{a}_g)$
  - Maximize the likelihood for $\boldsymbol{\beta}$ using proportional logistic regression

Feature selection

The number of gene-level features one might include in such a model is very large. It is impossible, from both a computability and interpretability standpoint, to include all conceivable features in the model. A related but distinct issue is that of collinearity. As the number of features in the model increases, there is a higher probability that some subset of features will be collinear with one-another, which can complicate model-fitting. To avoid this issue, we employ a multi-stage model fitting procedure. In the first step, all single-feature-plus-intercept models are fit, and their p-value

We employed a forward selection procedure to construct a multivariate model. We begin the procedure by fitting all features in single-feature models. We test the significance of each model against an intercept-only model using the likelihood ratio test. From this set of univariate models multivariate models all of the nonsignificant (false discovery rate of 0.05) univariate features for each cancer type were removed from the analysis. In the case of KICH, no features were significant after multiple testing correction (minimum q value: 0.104 correspodning to GO:0071456, "cellular response to hypoxia"), and in the case of CHOL, only one feature was significant after multiple testing correction (GO:0008285 " negative regulation of cell population proliferation", q value: 0.00909). Significant features were then further pruned using a jaccard index similarity cutoff.

Jaccard index for binary feature overlap

The Jaccard index that measures the similarity of two sets. GO terms are binary features (gene is either a member of a GO term or it is not), and can be models as a set, where the genes in the GO terms are the elements of the set. The definition of the Jaccard Index is:

$$J(A,B) = \frac{|A \cap B|}{A \cup B} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

The strategy for feature selection works as follows: first take the most significant single-feature model, and then remove the most similar features to the selected feature (i.e the features with a jaccard similarity above 0.1). Then fit all 2 term models that include the most significant feature, select the feature with the highest significance when tested against the single feature model, and remove all features similar to the features in the selected model. This process of testing all available n feature models, against the (greedily) best $n-1$-feature model, and removing from the candidate pool features

similar to the chosen nth feature, is repeated until the most significant model in the n term vs $n-1$ comparison is not significant at $p < 0.05$ by the likelihood ratio test.

Validation against COSMIC Cancer Gene Census

The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) [2] is an effort to catalogue genes which contain mutations causally implicated in cancer.

Data

Prognostic

The Human Pathology Atlas is a dataset of 900,000 patient survival profiles across 17 types of cancersurvival data [3]

Gene-Level Summary Statistics

A set of Bayes factors from a study of cancer driver genes [4]. For each of 20 TCGA tumor types, roughly 20,000 genes were analyzed and the posterior probability that each gene (in each cancer type) was a causal gene was assessed and summarized via Bayes Factor.

Gene-Level Annotations

- Gene Ontology

  The "Biological Process" Gene Ontology [5] was downloaded from the Bioconductor package `GO.db` [6]. Of the 10,930 possible biological process gene ontology terms, the 2,198 terms that include 10 or more genes were analyzed, so as to reduce the multiple testing burden.

Software

Our method is distributed as a freely available R package FGEM cite:R .FGEM relies on the `SQUAREM` package to accelerate EM convergence [7] and the hot-path functions are implemented as C++ functions using the `RcppArmadillo` [8] package.

Results

Discussion

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[2] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, Oct 2018.

[3] Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, and Fredrik Ponten. A pathology atlas of the human cancer transcriptome. *Science*, 357(6352):eaan2507, 2017.

[4] Siming Zhao, Jun Liu, Pranav Nanga, Yuwen Liu, A. Ercument Cicek, Nicholas Knoblauch, Chuan He, Matthew Stephens, and Xin He. Detailed modeling of positive selection improves detection of cancer driver genes. *Nature Communications*, 10(1):3399, 2019.

[5] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 11 2018.

[6] Marc Carlson. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2020. R package version 3.11.0.

[7] Yu Du and Ravi Varadhan. SQUAREM: An R package for off-the-shelf acceleration of EM, MM and other EM-like monotone algorithms. *Journal of Statistical Software*, 92(7):1–41, 2020.

[8] Dirk Eddelbuettel and Conrad Sanderson. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014.