Midterm Proposal

Subgroup Method on Compositional Data

Xuezhixing Zhang

March 16, 2022

1 Motivation

Microbiome analysis is always a popular topic in biostatistic. Although variable selection is one of the most relevant tasks in this field, researchers still keep using methods that may ignore the compositional nature of microbiome data. In [Lin et al., 2014], the author proposed a regression method with linear log-contrast model by adding a zero-sum constraint on coefficients and applying the $\mathcal{L}_1$ regularization approach.

Besides variable selection, subgroup analysis is also an important part in microbiome analysis. [Ma and Huang, 2015] proposed a concave pairwise fusion approach to identify homogeneous subgroups. $\mathcal{L}_1$ penalty, smoothly clipped abbsolute deviation penalty (SCAD) [Fan and Li, 2001] and the minimax concave penalty (MCP) [Zhang, 2010] were compared in this article.

So the idea is to combine these two ideas and propose a subgroup method on compositional data by adding a fused penalty on intercept in models from [Lin et al., 2014]. This method is easier than creating subgroups based on slopes and hopefully can be done in 6 weeks. This is an applied project, I will try to minimize the objective function and do some simulations based on this algorithm after the model was constructed.

2 Description

2.1 Optimization Problem

Suppose we have an $n \times p$ matrix $X = (x_{ij})$ of covariates. Then by applying the log-ratio transformation model to these covariates, we will have the linear log-contrast model

$$y = Z\beta^* + \epsilon, \quad \sum_{j=1}^{p} \beta_j = 0 \tag{1}$$

where $Z = (z_1, ..., z_p) = (\log x_{ij})$ is $n \times p$ designed matrix and $\beta = (\beta_1, ..., \beta_p)^T$ the corresponding regression coefficients.

The original optimization problem proposed in [Lin et al., 2014] is

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2n}||y - Z\beta||_2^2 + \lambda||\beta||_1 \\ \text{s.t.} \quad & \sum_{j=1}^{p} \beta_j = 0 \end{aligned} \tag{2}$$

where $\beta = (\beta_1, ..., \beta_p)^T$, $\lambda > 0$ is the regularization parameter.

Based on the fused penalty methods proposed by [Ma and Huang, 2015], we can then get our optimization problem by adding a fused penalty on the intercept.

$$\begin{aligned} \min_{\mu,\beta,\eta} \quad & \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mu_i - z_i^T\beta)^2 + \sum_{i<j} p_\gamma(|\eta_{ij}|,\delta) + \lambda||\beta||_1 \\ \text{s.t.} \quad & \sum_{j=1}^{p}\beta_j = 0, \quad \mu_i - \mu_j - \eta_{ij} = 0 \end{aligned} \tag{3}$$

where $\eta_{ij} = \mu_i - \mu_j$ is the new parameter introduced by [Ma and Huang, 2015] to simplify the optimization problem. The penalty term $p_\gamma(|\eta_{ij}|,\delta)$ can be $\mathcal{L}_1$, SCAD and MCP.

The MCP has the form

$$p_\gamma(|\eta_{ij}|,\delta) = \delta\int_0^{|\eta_{ij}|}(1 - \frac{x}{(\gamma\delta)})_+ dx, \quad \gamma > 1 \tag{4}$$

and the SCAD penalty is

$$p_\gamma(|\eta_{ij}|,\delta) = \delta\int_0^{|\eta_{ij}|} min\{1, \frac{(\gamma - \frac{x}{\delta})_+}{(\gamma - 1)}\}dx, \quad \gamma > 2 \tag{5}$$

where γ is a paraneter that controls the concavity of the penalty function. Both penalties will converge to $\mathcal{L}_1$ penalty as $\gamma \to \infty$.

2.2 Optimization Technique

Refer to the methods in [Ma and Huang, 2015], I plan to use the augmented Lagrangian method (ALM) to rewrite this optimization problem and derive an alternating direction method of multipliers (ADMM) algorithm for our approach. If time permits, I will try all three penalties mentioned above. Otherwise I'll just try MCP since it gave a better result in [Ma and Huang, 2015].

2.3 Simulation Settings

I plan to use the modified Bayesian Information Criterion (BIC) [Wang et al., 2015] to minimize the hyper-parameters δ and λ and use a fixed γ to conduct the simulations. The simulation results may include:

1. Solution paths for $\mu_1, ..., \mu_n$ against δ.
2. The mean, median and stand error (s.e.) for the estimated group numbers.
3. The bias for the estimates of our true intercepts α_1 and α_2.
4. The estimated p values under the null hypothesis $\alpha_1 = \alpha_2$.

3 Expected Results

In the final report, I plan to give a detailed description on the optimization problem and propose the computation algorithm by ADMM. Simulation results mentioned above will also be presented in the final report.