Internship Summary: Development of novel statistical methodologies for identifying predictive biomarker signatures

Wenxuan Deng

Location: Takeda Pharmaceuticals, 35 Landsdowne St, Cambridge, MA, 02139.

Biomarker signatures that are associated with clinical outcomes can be divided into two groups: predictive and prognostic. Prognostic biomarker signatures are associated with the outcome irrespective of treatment, while a predictive biomarker signature is associated with outcome in the presence of a therapeutic intervention. Developing predictive biomarker signatures allows for the identification of subgroups of patients who better respond to treatment, which benefits patients and makes FDA approval easier. A widely-used approach uses a traditional machine learning (ML) algorithm such as Elastic Net or Random Forest to develop a biomarker signature from the treatment arm of a clinical trial.

In my internship, we proposed a new method, Predictive Effects Net (PEN), based on group lasso and special hierarchical structure for figuring out predictive effects. The new approach takes predictive biomarkers as interaction effects between treatment and biomarkers. Our linear model is as the following:

$$Y = X_0\beta_0 + X_T\beta_\tau + X_1\beta_1 + X_T \otimes X_1\beta_2 + \epsilon$$

Where X_0 is the baseline variables, X_T is the treatment variable, X_1 is the high dimensional design matrix of genes, i.e. gene expression levels, SNP and mutations, and $X_T \otimes X_1$ is the interaction between genes and treatment. $\beta = (\beta_0, \beta_\tau, \beta_1, \beta_2)$ is the corresponding coefficients.
ϵ is random error.

The loss function is:

$$\min_\beta f(\beta | Y, X_0, X_T, X_1) + g(\beta)$$

$$g(\beta) = \lambda_1 \sum_i \phi_i |\beta_2^{(i)}| + \lambda_1 \sum_i \psi_i \sqrt{(\beta_1^{(1)})^2 + (\beta_2^{(1)})^2} + \lambda_2(\| \beta_1 \|_2^2 + \| \beta_2 \|_2^2)$$

Where $\beta = (\beta_0, \beta_\tau, \beta_1, \beta_2)$ is the parameter, and $f(\beta | Y, X_0, X_T, X_1)$ is L-2 loss function. When the model is the ordinary linear model, the L-2 loss function is $\| Y - (X_0\beta_0 + X_T\beta_\tau + X_1\beta_1 + X_T \otimes X_1\beta_2) \|^2$. Penalty function $g(\beta)$ can construct a complex hierarchical selection of β_1 and β_2, that nonzero β_2 is a sufficient but not necessary condition for nonzero β_1. λ_1 and λ_2 are regularization parameters.

To show PEN has an supreme performance, we also conducted experiments in different scenarios and comparisons with several other variable selection methods.