Volume 22, Number 1, February/March 2016

Institute for Health and Society, Medical College of Wisconsin

MEDICAL COLLEGE OF WISCONSIN

Datum

NEWSLETTER

Division of Biostatistics

ADVANCING HEALTH THROUGH RESEARCH AND DISCOVERY CTSI

Modern variable selection techniques

Kwang Woo Ahn, PhD, Division of Biostatistics, MCW

Advancements in genetics enable researchers to produce enormous amounts of data called “high dimensional data.” One of the statistical challenges in high dimensional data is how to select important variables when the number of variables (p) is much larger than the number of individuals (n). For example, it is common to see >10,000 genes with only a couple of hundred subjects.

Classical variable selection methods include forward selection, backward elimination, and stepwise selection. The names are tied with the direction of the significant variable search. Forward selection starts with no selected variables. During subsequent steps, it evaluates if each candidate variable improves some chosen statistical criterion given previously selected variables, and adds the variable that improves the criterion most. It repeats these steps until none of the remaining variables improves the criterion. On the other hand, backward elimination starts with the full model, that is, all the variables. At each step, it removes a variable that is least important and does not meet the criterion. Stepwise selection is the combination of forward selection and backward elimination. These classical methods are intuitive and work properly in many regression models, so they are well accepted in practice. However, some of the classical variable selection techniques such as backward elimination and backward stepwise selection cannot be used when p is greater than n. On the other hand, forward selection and forward stepwise selection may work even when p is bigger than n, but it may be computationally inefficient. Another issue is that many of the most popular criteria to determine whether a candidate variable is selected are test-based-statistics including Wald test, score test, and F-test. Investigators typically set some cutoff for p-value to use these selection criteria. Popular cutoffs are 0.01 and 0.05. Using these criteria inflates Type I error rate, which is the probability of incorrectly rejecting the null hypothesis when it is indeed true. This Type I error rate increases every time more tests are added. Thus, if the chosen variable selection procedure used many steps to find significant variables, there would be a very high chance that some of the selected variables may not be truly significant.

Modern variable selection methods including lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and bridge estimator (Huang et al., 2008) received a lot of attention due to the demands of high dimensional data analysis. Each method has a different penalty term in its penalized estimation function. So they are called penalty-driven methods. Although each method has a different mathematical form, all of them can shrink a large number of regression coefficients to zero, which implies that those predictors are "not significant." They assume that the true significant variables are sparse. In addition, they work even when p is greater than n. Thus, they may be used when researchers are interested in finding a small number of genes to predict outcomes of interest. All of these methods are designed to select individual variables such as genes. However, scientists are often interested in group variables, e.g., SNPs and gene pathways. Various group variable selection techniques were proposed to handle them. They include group lasso (Yuan and Lin, 2006) and group bridge (Huang et al., 2009). These penalty-driven methods turned out to be quite flexible so that they could be extended to more complicated biological mechanisms. For example, Friedman et al. (2008) proposed graphical lasso to study gene networks. Tibshirani et al. (2005) developed the fused lasso to identify important genes when the order of genes matters. In general, the penalty-driven methods are computationally more efficient compared to classical variable selection methods. In addition, they do not rely on p-values in selecting variables; therefore, they do not suffer from inflated Type I error rates. The downsides of these techniques are i) One may have a hard time to find software to implement recent techniques; ii) Even if software is available, non-statisticians may have a difficult time using it; and iii) Estimates from the penalty-driven methods may be somewhat biased due to the use of penalty terms especially when the sample size is small. To obtain unbiased estimates, one might have to re-fit regular models with variables selected by penalty-driven methods.

Variable selection is one of the most active research areas in statistics. Although classical variable selection methods can still be used for common regression settings, they may not be applicable for high dimensional data and complicated biological settings. Recently penalty-driven methods received a lot of attention due to the increasing availability of genetics data. Some of these methods can be used even when p is much larger than n. Furthermore, they are flexible enough to accommodate many complex biological settings.

References

1. Fan, J. and Li, R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348—1360.
2. Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9, 432-441.
3. Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. The Annals of Statistics, 36, 587-613.
4. Huang, J., Ma, S., Xie, H. L. and Zhang, C.-H. (2009). A group bridge approach for variable selection. Biometrika, 96, 339-355.
5. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58, 267-288.
6. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society Series B, 67, 91-108.
7. Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B, 68, 49-67.