Prior work on regularization

Regularization as well as variable selection plays a major role in linear regression. A large number of proposals are made to overcome the limitation of least square regression. The early milestone in this area is ridge regression [1]. Subjected to a L2 penalty, ridge regression aims to minimize the sum of squared error. Ridge regression achieves better performance through bias-variance trade-off. However, the main drawback is that ridge regression is not a parsimonious model; it simply keeps all predictors in the model.

Another popular algorithm, lasso [2], was proposed in 1996. By incorporating the L1-norm for regularization, lasso performs optimally in high-dimensional, low-correlated settings, in terms of both prediction and parameter estimation. Unlike ridge regression, lasso does have sparse representation. People have figured out that, nevertheless, lasso performs poorly under three conditions: 1) Let p represent the number of features and n represent the number of observations. For p >n, lasso selects at most n variables before it saturates; besides, lasso is not well defined unless the bound on L1-norm of the coefficients is smaller than a certain value. 2) For a group of variables whose pairwise correlations are very high, lasso tends to select only one variable and does not care which one is selected. 3) For n>p case, when predictors have high correlations, the prediction performance of lasso is dominated by ridge regression.

In order to address the third problem of lasso, elastic net [3] was proposed in 2005 by adding the squared L2 norm, a strongly convex penalty term to the L1 norm in lasso. Elastic net performs similarly to lasso in scenario 1) and 2), but by encouraging grouping effect, has higher accuracy than lasso in scenario 3). To be more specific, in scenario 3), some features are highly correlated with each other and are associated with response; thus elastic net aims to perform less shrinkage on those subsets of features. In addition, similar to lasso, elastic net does both continuous shrinkage and automatic variable selection. Note that ridge regression has only continuous shrinkage but no automatic variable selection.

Group lasso [4] is another approach to implement the grouping effect to lasso. Group lasso divides predictors into group and penalizes the sum of L2-norm. The basic assumption of group lasso is known a priori that there are distinct groups or clusters among variables; it uses L2 penalty on coefficients within each of K known and non-overlapping groups. However, as many researchers pointed out, knowing the group in advance is not always possible.

On top of group lasso and elastic net, clustering algorithm is also involved to each of them. Cluster group lasso [5] seeks sets of correlated features with similar associations with the response. The idea is to first identify groups among features using hierarchical clustering and then apply group lasso. The basic assumption is that all correlated features have similar association with response. If this assumption fails to hold so that not all correlated features have a similar association with the response, then cluster elastic net (CEN) [6] would show its advantage since CEN seeks sets of correlated features with similar associations with the response. Besides, while elastic net shrinks all coefficients towards the origin, CEN selectively shrinks coefficients for highly-correlated variables towards each other.

[1] AE Hoerl, RW Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970

[2] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical*

*Society. Series B (Methodological)*, 58(1):267–288, 1996.

[3] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the*

*Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[4] F. Bach. Consistency of the group Lasso and multiple kernel learning. *The Journal of Machine*

*Learning Research, 9:1179–1225*, 2008.

[5] P. Buhlmann et al. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference Volume 143, Issue 11, November 2013, Pages 1835-1858*. 2013

[6] DM Witten et al. The Cluster Elastic Net for High-Dimensional Regression With Unknown Variable Grouping. *Technometrics. 2014 February 20; 56(1): 112–122. doi:10.1080/00401706.2013.810174*. 2014