High Dimensional Data and Statistical Learning

Lecture 4: Subset Selection and L_0 Penalization

Weixing Song

Department of Statistics
Kansas State University

Outline

Outline

High-dimensional Sparse Models

We have demonstrated that when the number of covariates p is large compared with the sample size n, estimation and prediction using the full model of all p covariates may not perform well or can even fail due to:

- the noise accumulation, high collinearity and spurious correlation,
- lack of model identifiability and interpretability,
- computational difficulty and instability.

Ridge regression addresses some of the issues, but it does not perform variable selection. It can be advantageous to impose some sparse model structure and generate a sequence of sparse candidate models using, e.g., variable selection methods.

Best Subset Regression

A natural idea of producing sparse models is the best subset regression/selection.

The best subset regression/selection has two steps:

(a). For each submodel $M \subset \{1, 2, \ldots, p\}$ consisting of covariates X_j's with indices in M, apply a model fitting procedure such as the LSE or MLE to the submodel (recall the empirical risk minimization).

(b). Select the best submodel among the fitted ones using a model selection criterion such as the AIC, BIC, ect..

The idea of best subset regression is appealing in producing the best sparse model but one has to fit a total of 2^p sparse models, that is, all submodels of $\{1, 2, \ldots, p\}$.

It is infeasible to implement the method even in moderate dimensions in tens; its computational complexity grows exponentially with the dimensionality p.

The method suffers from the instability due to the sampling variability and discontinuity; it can happen that different models can be selected both with significant probability.

Ideally we want a modeling procedure to produce a unique, appealing model with significant probability, say, with asymptotic probability one.

We want to develop effective sparse modeling procedures that retain appealing theoretical properties of the best subset selection and can be fitted with computationally efficient algorithms.

We also want the estimated model to be stable.

We will discuss some recent developments in high-dimensional variable selection that aim at these goals in this course.

In many applications, we assume (in correctly specified models) that the conditional distribution of the response variable Y given the covariate vector $X = (X_1, X_2, \ldots, X_p)^T$ depends on X only through a form of the linear form $\boldsymbol{\beta}^T X$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, a p-vector of regression coefficients.

Some β_j's are assumed to be zeros and we refer to the corresponding covariates as the noise variables.

The covariates with nonzero β_j's are referred to as the true (important) variables.

Variable selection aims to identify all important variables and provide efficient estimation of their coefficients.

Variable Selection vs. Model Selection

More generally, assume that the data are generated from the true density function $f_{\boldsymbol{\theta}_0}$ with parameter vector $\boldsymbol{\theta}_0 = \{\theta_1, \ldots, \theta_p\}^T$.

Often times we are uncertain about the true density, but more certain about a large family of models $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}_0$ is a nonvanishing subvector of the p-dimensional parameter vector $\boldsymbol{\theta}$.

The problems of how to estimate the dimension of the model and compare models of different dimensions naturally arise in many statistical applications, including time series modeling.

These problems are referred to as model selection in the literature.

In short, variable selection is concerned with how to build a sequence of good sparse candidate models, while model selection is concerned with how to compare those models.

Outline

There are two classical principles of model selection: the Kullback-Leibler (KL) divergence principle and the Bayesian principle.

Akaike (1973, 1974) proposed to choose a model that minimizes the KL divergence of the fitted model from the true model, or equivalently maximize the expected log-likelihood.

KL divergence of the density f from the density g can be fitted as

$$I(g;f) = E_g\left[\log\frac{g(z)}{f(z)}\right] = \int [\log g(z)]g(z)\,dz - \int [\log f(z)]g(z)\,dz.$$

Model Selection Criteria: AIC

Akaike (1973) considered the MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$ of the parameter vector $\boldsymbol{\theta}$ and showed that up to an additive constant, the KL divergence of the fitted model from the true model can be asymptotically expanded as

$$-l_n(\hat{\boldsymbol{\theta}}) + \dim(\hat{\boldsymbol{\theta}}) = -l_n(\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{p} I(\hat{\theta}_j \neq 0),$$

where $l_n(\boldsymbol{\theta})$ is the log-likelihood function, $\dim(\boldsymbol{\theta})$ denotes the dimension of the model.

This asymptotic expansion leads to the Akaike information criterion (AIC) for comparing models:

$$\mathrm{AIC}(\hat{\boldsymbol{\theta}}) = -2l_n(\hat{\boldsymbol{\theta}}) + 2\|\hat{\boldsymbol{\theta}}\|_0.$$

Model Selection Criteria: BIC

A typical Bayesian model selection procedure is to first give nonzero prior probability α_M on each model M, and then prescribe a prior distribution μ_M for the parameter vector in the corresponding model. See Schwarz (1978).

The Bayesian principle of model selection is to choose the most probable model a posteriori; that is, to choose a model that maximizes the log-marginal likelihood or the Bayes factor

$$\log \int \alpha_M \exp[l_n(\boldsymbol{\theta})] d\mu_M(\boldsymbol{\theta}).$$

Schwarz (1978) took a Bayesian approach with prior distributions that have nonzero prior probabilities on some lower dimensional subspaces of $\mathbb{R}^p$ and showed that the negative log-marginal likelihood can be asymptotically expanded as

$$-l_n(\hat{\boldsymbol{\theta}}) + \frac{\log n}{2} \|\hat{\boldsymbol{\theta}}\|_0,$$

where $l_n(\hat{\boldsymbol{\theta}})$ is the maximum log-likelihood.

This asymptotic expansion leads to the Bayesian information criterion (BIC) for comparing models:

$$\text{BIC}(\hat{\boldsymbol{\theta}}) = -2l_n(\hat{\boldsymbol{\theta}}) + \log(n)\|\hat{\boldsymbol{\theta}}\|_0.$$

For the normal linear regression model,

$$\text{AIC} = \frac{1}{ns^2}(RSS_d + 2ds^2),$$

and

$$\text{BIC} = \frac{1}{n}(RSS_d + ds^2 \log(n)),$$

where RSS_d is the residual sum of squares (RSS) of the linear regression with d predictors ($d \leq p$), s^2 is the MSE of the full model.

L_0-Penalized Likelihood

AIC and BIC suggest a unified approach to variable selection and model selection: choose a parameter vector $\boldsymbol{\theta}$ that minimizes the penalized log-likelihood

$$-l_n(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0, \tag{1}$$

where the L_0-norm $\|\cdot\|_0$ denotes the number of nonzero components and $\lambda \geq 0$ is a regularization parameter.

Given $\|\boldsymbol{\theta}\|_0 = m$, the solution to problem (1) is the best subset that has the largest maximum likelihood among all subsets of size m.

The model size m is then chosen to maximize (1) among p best subsets of sizes m, $1 \leq m \leq p$.

The L_0-penalized likelihood method is equivalent to the best subset selection.

The computation of the L_0-regularization problem (1) is a combinatorial problem with NP-complexity.

L_0-Penalized Empirical Risk Minimization

More generally, we have a unified approach of L_0-penalized empirical risk minimization for variable selection and model selection:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \hat{R}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0 \right\}, \tag{2}$$

where $\hat{R}(\boldsymbol{\theta})$ is the empirical risk function, which could be:

- The negative log-likelihood loss: equivalent to L_0-penalized likelihood.
- Squared error (quadratic) loss: L_0-penalized least squares.
- .....

Many model selection methods amount to the L_0-regularization problem (2) with different choices of the regularization parameter λ.

Connections to Other Model Selection Criteria

Let RSS_d be the residual sum of squares (RSS) of the best subset with d variables. Then

- $C_p = RSS_d/s^2 + 2d - n$ in Mallows (1973) corresponds to $\lambda = 1$, where s^2 is the MSE of the full model.
- The adjusted R^2 given by

$$R^2_{\text{adj}} = 1 - \frac{(n-1)RSS_d}{(n-d)SST}$$

also amounts to the L_0-regularization problem, where SST is the total sum of squares.

To see this, note that maximizing R^2_{adj} is equivalent to minimizing $\log(RSS_d/(n-d))$.

By $RSS_d/(n-d) \approx \sigma^2$ (the error variance), we have

$$n \log \frac{RSS_d}{n-d} \approx \frac{RSS_d}{\sigma^2} + d + n(\log \sigma^2 - 1).$$

This shows that the adjusted R^2 method is approximately equivalent to the L_0-regularization problem (2) with $\lambda = 1/2$.

Other examples include the generalized cross-validation (GCV) given by $RSS_d/(1-d/n)^2$, cross validation (CV) (Stone, 1974), and risk inflation factor (RIC) (Foster and George, 1994).

Properties of L_0-Regularization Methods

A comprehensive theory on risk bounds for model selection using the L_0-regularization method was presented in Barron et al. (1999).

An upper bound on the prediction risk was established. It shows that the tradeoff between the approximation error (model bias) and the price we pay in searching over a large family of models.

These results were for the squared error loss. It is challenging to derive results for other losses such as the L_q-loss $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_q$, $q > 0$, and obtain variable selection properties.

Computational challenges of L_0-Regularization Methods

Although L_0-regularization methods have appealing risk properties, the computation is infeasible in high-dimensional statistical endeavors due to its nature of combinatorial optimization (2).

It is infeasible to implement even in moderate dimensions in tens.

The computational difficulty comes from the discontinuity and nonconvexity of the L_0-penalty function $\lambda\|\boldsymbol{\beta}\|_0$.

A natural idea is to replace the L_0-penalty function with some continuous or convex penalty function.

Penalized Empirical Risk Minimization

Consider a continuous or convex relaxation of the L_0-regularization method

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \hat{R}(\boldsymbol{\beta}) + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \right\}, \tag{3}$$

where $\hat{R}(\boldsymbol{\beta})$ is the empirical risk function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and $p_\lambda(t)$, $t \geq 0$, is a nonnegative penalty function indexed by the regularization parameter $\lambda \geq 0$ with $p_\lambda(0) = 0$.

With an appropriately chosen penalty function, we hope to simultaneously select important variables and estimate their associated regression coefficients.

Choice of Penalty Function

The L_q-penalty $p_\lambda(t) = \lambda t^q$ for $0 < q \leq 2$ in the bridge regression in Frank and Friedman (1993), which bridges the best subset selection (L_0-regularization) and ridge regression (L_2-regularization), including the L_1-penalty as a special case.

The non-negative garrote introduced in Breiman (1995) for shrinkage estimation and variable selection.

The L_1-penalized least squares method was called the Lasso in Tibshirani (1996), and it is now collectively referred to as L_1-penalized empirical risk minimization method.

There are many other choices of the penalty function to be introduced later; e.g., the SCAD in Fan and Li (2001), adaptive lasso in Zou (2006), group lasso in Yuan and Lin (2006), etc.

Desirable Properties of Penalty Functions

Fan and Li (2001) advocated penalty functions that give estimators with three properties:

Sparsity: The resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity.

Unbiasedness: The resulting estimator is nearly unbiased, especially when the true coefficient β_j is large, to reduce model bias.

Continuity: The resulting estimator is continuous in data to reduce instability in model prediction.

It is desirable to have $p'_\lambda(0+) > 0$ to ensure the sparsity of the regularized estimate; see their paper for more insights.

Summary

The two standard techniques for improving the OLS estimates, subset selection and ridge regression, both have drawbacks.

- Subset selection provides interpretable models but can be extremely variable because it is a discrete process. Small changes in the data can result in very different models being selected.
- Ridge regression is a continuous process that shrinks coefficients and hence is more stable: however, it does not set any coefficients to 0 and hence does not give an easily interpretable model.

But, we can do better with appropriately chosen penalty function!

Outline

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6), 716-723.
- Barron, A., L. Birgé, and P. Massart (1999, February). Risk bounds for model selection via penalization. *Probability Theory and Related Fields,* 113(3), 301-413
- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics,* 37(4), 373-384.
- Breiman, L. (1996, December). Heuristics of instability and stabilization in model selection. *The Annals of Statistics,* 24(6), 2350-2383.
- Candès, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numerica,* 15, 257-325.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association,* 96(456), 1348-1360.

- Foster, D. P. and E. I. George (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics,* 22(4), 1947-1975.
- Frank, I. E. and J. H. Friedman (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics,* 35(2), 109-135.
- Mallows, C. L. (1973). Some comments on C_p. *Technometrics,* 15, 661-675.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* 6(2), 461-464.
- Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika,* 61(3), 509-515.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B), 58, 267-288.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B,* 68 (1), 49-67.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association,* 101, 1418-1429.