Subset selection, shrinkage method and SINDy

Yangxinyu Xie

Summer 2019

Last updated June 10, 2019

1. Backgrounds

❖

To train an efficient regression function f, we are especially interested in minimising the residual sum of squares

$$RSS(f) = \sum_{i=1}^{N}(y_i - f(x_i))^2 \tag{1.0.1}$$

One restricted set of solutions, usually referred to as linear regression, assumes that the "true" regression function $\mathbb{E}[Y|X]$ is linear, or that linear model is a reasonable approximation of the underlying "true" model.

1.1 Singular Value Decomposition and Principal Components ❖

Suppose $X \in \mathbb{R}^{M\times N}$ with $\mathbf{rank}(X) = r$, we can factor X as

$$X = UDV^T \tag{1.1.1}$$

where $U \in \mathbb{R}^{M\times r}, U^TU = I, V \in \mathbb{R}^{N\times r}, V^TV = I$ and $D = \mathbf{diag}(d_1, ..., d_r)$ with $d_1 \geq d_2 \geq ...d_r \geq 0$. Such a factorisation is called *singular value decomposition* of X. If one or more $d_i = 0$, then X is singular. The columns of U are called *left singular vectors*, the columns of V are *right singular vectors* and the numbers d_i are the *singular values*. Note that the eigenvalue decomposition is

$$X^TX = VD^2V^T = \begin{bmatrix} V & \tilde{V} \end{bmatrix} \begin{bmatrix} D^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V & \tilde{V} \end{bmatrix}^T \tag{1.1.2}$$

where $\tilde{V}$ is any matrix for which $\begin{bmatrix} V & \tilde{V} \end{bmatrix}$ is orthogonal. Note that the sample variance of X is $S = X^TX/N$ and we call $z_i = Xv_i$ principal components. The variance of the first principal component, denoted by z_1, is $Var(z_1) = d_1^2/N$ and hence the variance explained by the first principal component is $d_1^2/\sum d_i^2$. Since $u_1d_1 = z_1$, we call u_1 the *normalised first principal component*. We also defined the *pseudo-inverse* of X as

$$X^\dagger = VD^{-1}U^T \tag{1.1.3}$$

If $\mathbf{rank}(X) = N$, we have $X^\dagger = (X^TX)^{-1}X^T$.

1.2 Least Squares ❖

The least square regression seeks to minimise the Euclidean distance from the response variables $\mathbf{y}$ to the hyperplane spanned by predictor variables X,

$$\min \quad RSS(\beta) = ||\mathbf{y} - \beta_0 - X\beta||_2^2 \tag{1.2.1}$$

which is an unconstrained quadratic program. The least square solution is $\beta = X^\dagger\mathbf{y}$, which coincides with the maximum likelihood solution. Hence, the prediction $\hat{\mathbf{y}} = X\beta = XX^\dagger\mathbf{y}$. Note that the matrix $XX^\dagger = X(X^TX)^{-1}X^T$ is sometimes referred as the *hat matrix*. Using the singular value decomposition, we can write $\hat{\mathbf{y}} = X\beta = UU^T\mathbf{y}$.

1.3 The Lagrangian ❖

Given an optimisation problem in the standard form

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t} \quad & f_i(x) \leq 0, i \in [m] \end{aligned} \tag{1.3.1}$$

where $x \in \mathbb{R}^n$. We define the *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \tag{1.3.2}$$

The vectors λ is called the *Lagrange multiplier vectors* associated with the problem 1.3.1, and in the context of statistics, usually called the *tuning parameter*. The *Lagrange dual function* is the minimum value of the Lagrangian and it yields the lower bound of the optimal value p^* of problem 1.3.1. For further details, we refer the reader to [BV04].

2. Subset Selection

❖

Least square estimates tend to have low bias but high variance, especially when multicollinearity exists among predictor variables. To improve prediction accuracy, we can sometimes trade a tiny increase in bias for a substantial decrease in variance by setting some β_i to zero or shrinkage methods. The intuition is that some predictors may be excessive, in other words, highly correlated with other predictors and thus produce little predictive power. The general subset selection algorithm takes the following three steps:

1. Let $\mathcal{M}'$ be the initial model.
2. For each $k \in [p]$ where p is the number of predictors:

   we fit a set of models with k predictors, and choose the model $\mathcal{M}_k$ with the lowest RSS. In the case of logistic regression, we choose *deviance*, which is the $-2\times$maximised log-likelihood.
3. Out of all the models $\mathcal{M}', \mathcal{M}_1, ..., \mathcal{M}_p$, we choose the optimal model using cross validation or theoretical information criteria.

2.1 Subset Selection ❖

The *best subset selection* choose a model with no predictors as the initial model and for each k, it fits a set of $\binom{n}{k}$ models with k predictors. However, this approach is computationally heavy as 2^p models are to be examined.

The *forward-stepwise subset selection* is a **greedy** procedure that chooses a model with no predictors as the initial model and for each k, it chooses one predictor that has not yet been chosen in the $k-1$ predictors and adds it into $\mathcal{M}_{k-1}$. The tradeoff of the computational advantage is that this method will not necessarily get a model as optimal as the one obtained by the best subset selection.

The *backward-stepwise subset selection* is another greedy procedure that starts with a full model and decreasingly for each k, it chooses deletes one predictor that has the least impact to the change of RSS. It usually takes longer for backward stepwise selection to reach minimum error.

The *forward-stagewise subset selection* is similar to the forward-stepwise selection, but more constrained. It starts out with an intercept equals to $\bar{y}$ and for each step it seeks the predictor that is most correlated with the residual and stops until no further correlation is found. This method is usually inefficient in comparison with stepwise selection methods.

2.2 Loss Function, C_p and AIC ❖

A generalisation of sum of squared residuals is called *loss function*, which can be arbitrarily defined. Two common loss functions for a given regression training set $\mathcal{T}$ are

$$\begin{aligned} \text{squared error} \quad & L(\mathbf{y}, \hat{f}(X)) = (\mathbf{y} - \hat{f}(X))^2 \\ \text{absolute error} \quad & L(\mathbf{y}, \hat{f}(X)) = |\mathbf{y} - \hat{f}(X)| \end{aligned} \tag{2.2.1}$$

In the case of classification, we have a set $\mathcal{G}$ with K different categories of elements and $p_k(X) = \mathbb{P}(G = k|X)$

$$\begin{aligned} \text{0-1 loss} \quad & L(G, \hat{G}(X)) = I(G \neq \hat{G}(X)) \\ \text{deviance} \quad & L(G, \hat{p}(X)) = -2\sum_{k=1}^{K} I(G = k) \log \hat{p}_k(X) = -2 \log \hat{p}_G(X) \end{aligned} \tag{2.2.2}$$

The *training error* is defined as the average loss $\overline{err} = \sum_{i=1}^{N} L(\mathbf{y}, \hat{f}(x_i))/N$ and the *test error* is defined as $Err_{\mathcal{T}} = \mathbb{E}[L(\mathbf{y}, \hat{f}(X))|\mathcal{T}]$. The *in sample error* assumes that the input vector X^0 coincides with the training

input. That is

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Y^0}[L(Y^0, \hat{f}(x_i^0))|\mathcal{T}] \tag{2.2.3}$$

where we have N new observations at each of the training point $x_i, i \in [N]$. The *optimism* is defined as the difference between the in sample error and the training error, $\text{op} = Err_{in} - \overline{err}$. It can be shown that generally the expected optimism is

$$\frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i) \tag{2.2.4}$$

if the model is a linear fit with d *effective* predictors, then $\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i) = d\sigma^2$. The *Mallows' C_p Criterion* takes the form

$$C_p = \overline{err} + 2 \cdot \frac{d}{N}\hat{\sigma}^2 \tag{2.2.5}$$

which is an estimate of the in sample error when squared loss is applied. An alternative form of the C_p criterion is

$$C_p' = \frac{SSE}{MSE} - (N - 2d) \tag{2.2.6}$$

which can be shown to an estimate of the total mean squared error divided by σ^2, denoted by Γ_p

$$\begin{aligned}\Gamma_p' &= \frac{1}{\sigma^2}\sum_{i=1}^{N}\{[(\mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i)]^2 + \text{Cov}(\hat{y}_i, y_i)\} \\ &= \frac{1}{\sigma^2}\{\sum_{i=1}^{N}[(\mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i)]^2 + d\sigma^2\} \\ &= \frac{1}{\sigma^2}\mathbb{E}(SSE) - (N - 2d)\end{aligned} \tag{2.2.7}$$

where $\mathbb{E}(SSE) = [\sum_{i=1}^{n}(\mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i)]^2 + (n-d)\sigma^2$. Dropping the constant term and multiply by σ^2/N, we obtain the equivalent form

$$\Gamma_p = \mathbb{E}(\overline{err}) - 2\frac{d}{N}\sigma^2 \tag{2.2.8}$$

For further discussions about C_p criterion, we refer the reader to [KNN+05]. The *Akaike information criterion* is a similar estimate for the in sample when a log-likelihood loss is applied. As an analogy to 2.2.8,

$$-2\mathbb{E}[\log \mathbb{P}_{\hat{\theta}}(y)] \approx -\frac{2}{N}\mathbb{E}[\text{maximised log likelihood}] + 2\frac{d}{N} \tag{2.2.9}$$

where maximised log likelihood is estimated by $\sum_{i=1}^{N}\log \mathbb{P}_{\hat{\theta}}(y_i)$. Hence, for logistic regression model with a binomial log-likelihood,

$$\text{AIC} = -\frac{2}{N}\log \text{lik} + 2\frac{d}{N} \tag{2.2.10}$$

Another equivalent form with squared loss is

$$\text{AIC} = N\log SSE - N\log N + 2d \tag{2.2.11}$$

2.3 Schwarz' Bayesian Criterion ❖

Suppose we have a set of candidate models $\mathcal{M}_m, m \in [M]$ and corresponding model parameters θ_m and wish to choose a best model. Assuming the prior distribution $\mathbb{P}(\theta_m|\mathcal{M}_m)$ is given, the posterior probability is

$$\begin{aligned}\mathbb{P}(\mathcal{M}_m|\mathcal{T}) &\propto \mathbb{P}(\mathcal{M}_m)\mathbb{P}(\mathcal{T}|\mathcal{M}_m) \\ &\propto \mathbb{P}(\mathcal{M}_m)\int \mathbb{P}(\mathcal{T}|\theta_m, \mathcal{M}_m)\mathbb{P}(\theta_m)d\theta_m\end{aligned} \tag{2.3.1}$$

Note that $\mathcal{T}$ represents the training data. Using Laplace approximation to the integral,

$$\mathbb{P}(\mathcal{T}|\mathcal{M}_m) = \mathbb{P}(\mathcal{T}|\theta_m, \mathcal{M}_m) - \frac{d_m}{2}\log N + O(1) \tag{2.3.2}$$

Since we usually assume the prior to be uniform, $\mathbb{P}(\mathcal{M}_m)$ can be then viewed as constant, hence, by taking the loss function $-2\log\mathbb{P}(\mathcal{T}|\hat{\theta}_m, \mathcal{M}_m)$, we obtain the *Schwarz' Bayesian Criterion*

$$\text{BIC} = -2\log\text{lik} + d\log N \tag{2.3.3}$$

Note that a model with lowest BIC has the largest posterior probability. Using the squared loss function, we have

$$\text{BIC} = \frac{N}{\sigma^2}[\mathbb{E}(\overline{err}) - \log N\frac{d}{N}\sigma^2] \tag{2.3.4}$$

or its equivalent form,

$$\text{BIC} = N\log SSE - N\log N + \log Nd \tag{2.3.5}$$

Note that for $N > 8$, the BIC tends to favour more parsimonious models than AIC.

3. Shrinkage Models

❖

Subset selection is a discrete process - it either retains one predictor or discard it. However, dropping variables discretely can introduce an unsuccessful variance reduction for although one predictor is correlated with others, it may still attain a considerable amount of predictive power. Shrinkage method, by introducing a constraint on the parameters, can make the process more continuous.

3.1 Ridge Regression ❖

Ridge regression put an explicit constraint on the size of the parameters and the quadratic optimisation problem in 1.2.1 thus becomes

$$\begin{aligned} \min \quad & ||\mathbf{y} - \beta_0 - X\beta||_2^2 \\ s.t \quad & ||\beta||_2^2 \leq t \end{aligned} \tag{3.1.1}$$

To see the continuousness of the process, we observe that as the constraint imposes that as t shrink toward zero, some coefficients β_i will be shrunk toward zero as well. Writing the problem in its Lagrangian form,

$$\hat{\beta}^{ridge} = \arg\min_{\beta}[||\mathbf{y} - \beta_0 - X\beta||_2^2 + \lambda||\beta||_2^2 - \lambda t] \tag{3.1.2}$$

Note that the right most term λt is usually dropped. The solution to the quadratic program is obtained by

$$\hat{\beta}^{ridge} = (X^T X + \lambda \mathbf{I})^{-1} X^T \mathbf{y} \tag{3.1.3}$$

Using the singular value decomposition, we obtain

$$X\hat{\beta}^{ridge} = X(X^T X + \lambda \mathbf{I})^{-1} X^T \mathbf{y} = \sum_{i=1}^{p} u_i \frac{d_i^2}{d_i^2 + \lambda} u_i^T \mathbf{y} \tag{3.1.4}$$

Since d_i represents the predictive power of a parameter, we see from this from that for β_i with low d_i, it will have greater shrinkage as a result of $d_i^2/(d_i^2 + \lambda)$.

3.2 Lasso and Generalisation ❖

A variation of ridge regression is *Lasso*, which imposes a different kind of constraint on β

$$\begin{aligned} \min \quad & ||\mathbf{y} - \beta_0 - X\beta||_2^2 \\ s.t \quad & \sum_{i=1}^{p} |\beta| \leq t \end{aligned} \tag{3.2.1}$$

or equivalently the unconstrained quadratic programming

$$\min \quad ||\mathbf{y} - \beta_0 - X\beta||_2^2 + \lambda \sum_{i=1}^{p} |\beta| \tag{3.2.2}$$

In fact, we can generalise the optimisation problem as

$$\min \quad ||\mathbf{y} - \beta_0 - X\beta||_2^2 + \lambda \sum_{i=1}^{p} |\beta|^q \tag{3.2.3}$$

for some $q > 0$. *Elastic-net* further explores the flexibility of the penalty term. In fact, it is a compromise of Ridge and Lasso.

$$\min \quad ||\mathbf{y} - \beta_0 - X\beta||_2^2 + \lambda \sum_{i=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|) \tag{3.2.4}$$

4. Subset Selection, Shrinkage Models and SINDy

❖

4.1 Observations on Subset Selection and Shrinkage Models ❖

Consider the simplest case when the column axis of X are all orthonormal. We can obtain the solutions of subset selection and shrinkage methods as below,

1. Best subset selection with threshold $\gamma > 0$

$$\hat{\beta}_j \cdot I(|\beta_j \geq \gamma|) \tag{4.1.1}$$

2. Ridge regression

$$\frac{\hat{\beta}_j}{1+\lambda} \tag{4.1.2}$$

3. Lasso regression

$$\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+ \tag{4.1.3}$$

where subscript + denote the positive entries of the vector.

Note that all three methods simply apply a transformation to the least square solution $\hat{\beta}_j$. For further discussions about the non-orthonormal situations, we refer the readers to [TRJ09]. Due to the computational complexity of the best subset selection methods, which aims to penalise the Loss function with l_0 norm $||\beta||_0$ (the number of non-zero entries in β), we tend to choose ridge regression or Lasso which relaxes the penalty with l_2 or l_1 norm [MB06]. However, higher order penalty gives biased estimates of β and penalises the large coefficients more severely than the smaller coefficients. This can easily lead to a suboptimal selections of variables if large λs are chosen [BKM$^+$16]. Also, for Lasso to perform well in subset selection, strong assumptions are to be made and can be infeasible in reality [MB10].

4.2 SINDy ❖

Sparse Identification of Nonlinear Dynamics (SINDy) [BPK16] applies a compromise between higher order penalised regression and best subset selection. In fact, the algorithm first uses ridge or Lasso regression with relatively small λ to obtain a "safe" set of features from the original feature space and then apply best subset selection with information criterion to look for an optimal model [MKBP17]. However, since the appropriate choice of λ is unknown, the tendency of choosing a small λ can still make the selection process computationally expensive. Especially in situations where the experiment data are small-sized and noisy, with a wide range of candidate features, the performance of ridge and Lasso are not stable [NDS$^+$19]. Relaxed Lasso [Mei07] and Trimmed Lasso [BCM17] has also been proposed as remedies as for the bias introduced by classical Lasso regression. However, no thorough investigations of the two proposals have been conducted in the context of limitedly sized, noisy experiment data.

Advances in mixed integer optimisation (MIO) gives a hint to the reduction of computational expense. The MIO modified l_0 subset selection adds an additional polyhedral constraint to β:

$$\begin{aligned} \min \quad & \text{Loss Function} \\ \text{s.t.} \quad & ||\beta||_0 \leq k, A\beta \leq b \end{aligned} \tag{4.2.1}$$

given that the polyhedral constraints may be given from the background knowledge. Another approach to decreasing the computational complexity of l_0 penalised regression is descent methods [HM18]. However, these two approaches are relatively new with only limited reviews and experiments in the statistical community [MRD17].

Another significance of the SINDy algorithm is that the additive assumption of the underlying model in classical statistics does not necessarily hold in this context. In fact, [MBPK16] proposes implicit SINDy which aims to identify the underlying models that take the form

$$f(x) = \frac{f_N(x)}{f_D(x)} \tag{4.2.2}$$

where both f_N and f_D are additive polynomials. The method uses the fact that

$$f_N(x) - f_D(x)dx = 0 \tag{4.2.3}$$

to recover the additive assumption by constructing an extended starting library.[1]. Generalising this idea, underlying models can take more complicated form that requires new methods of estimation.

[1]The suggestion in [MBPK16] that implicit SINDy has a linkage to information criterion seems rather unlikely other than the connection between subset selection and penalised regression as mentioned in section 4.1

References

[BCM17] Dimitris Bertsimas, Martin S Copenhaver, and Rahul Mazumder. The trimmed lasso: Sparsity and robustness. *arXiv preprint arXiv:1708.04527*, 2017.

[BKM+16] Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

[BPK16] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[HM18] Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454*, 2018.

[KNN+05] Michael H Kutner, Christopher J Nachtsheim, John Neter, William Li, et al. *Applied linear statistical models*, volume 5. McGraw-Hill Irwin Boston, 2005.

[MB06] Nicolai Meinshausen and Peter Bühlmann. Variable selection and high-dimensional graphs with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[MB10] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[MBPK16] Niall M Mangan, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):52–63, 2016.

[Mei07] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[MKBP17] Niall M Mangan, J Nathan Kutz, Steven L Brunton, and Joshua L Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2204):20170009, 2017.

[MRD17] Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the snr is low. *arXiv preprint arXiv:1708.03288*, 2017.

[NDS+19] John Nardini, Allison Duprey, Fanuel Sisay, Natasha Stewart, and Yangxinyu Xie. Along the sindy frontier. `https://github.com/Xieyangxinyu/SINDy`, 2019.

[TRJ09] Hastie Trevor, Tibshirani Robert, and Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer, 2009.