SOIL for High-dimensional Linear Regression

1. Background

Variable importance measure is one constrictive approach to improve the reliability and reproducibility in choosing models for high-dimensional linear regression. It can help to identify which variables are essential and which are not for further estimation and prediction. The benefits of using it include saving time and cost due to the reduction of variables to be considered, helping decision-makers gain a more comprehensive understanding, providing a ranking of variables to balance the model selection and model averaging procedure, etc.

There are many importance measures with parametric models available now. One of the methods is based on a final selected model like t-test value and p-value, but it could lead to the problem of "winner takes all" since unselected variables would have zero importance. There is another approach based on the R^2 decomposition, and some scholars have proposed several methods such as dominance analysis (Budescu 1993), information criterion-based method (Theil & Chung 1988), and hierarchical partitioning (Chevan & Sutherland 1991).

There are also available non-parametric methods, and the random forest is the one that has been used in many fields for regression and classification. Although the random forest has a much lower overfitting risk than decision trees, this problem still exists. Due to the large number of trees considered, the computation may be far more complex than other algorithms.

In the paper, the author proposed a new approach called SOIL, and it has some beneficial features or advantages which can improve the procedure of data analysis. SOIL involves many candidate models for consideration which makes it more objective and reliable. It could avoid bias by using independent external weighting. Moreover, the SOIL importance measure could be naturally used for finding the best model since more critical variables for sparse modeling will receive higher importance values.

All in all, SOIL could provide more information than the existed measures for data analysis and is helpful in different steps of the modeling procedure. Therefore, it could improve the reliability and reproducibility of data analysis when variable selection is needed.

2. General Methodology

The methodology of Sparsity Oriented Importance Learning (SOIL) for high-dimensional regression models is shown below.

Let $\boldsymbol{X} = (X_1, \dots, X_p)$be the $n \times p$ design matrix, where $X_j = (x_{1j}, \dots, x_{nj})^T$, $j = 1, \dots, p$. And let $\boldsymbol{y} = (y_1, \dots, y_n)^T$ be the response vector. Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta}^* + \epsilon$,

where $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^T$ is a p -dimensional vector of true underlying model that generates the data and ϵ is the vector of n independent errors.

There are two components in SOIL importance: a set of candidate models and a dependable external weighting method on the models. Firstly, a collection of models needs to be obtained, which is referred to $\boldsymbol{A} = \{\mathcal{A}_k\}_{k=1}^K$. If the number of the predictors p is small, $\boldsymbol{A}$ could be a full list of all-subset models. When p is large, a group of models could be obtained using high-dimensional variable selection procedures (Lasso, Adaptive Lasso, SCAD and MCP, etc.). After getting the candidate models $\mathcal{A}_k, k = 1, \dots, K$, we refer $\boldsymbol{w} = (w_1, \dots, w_K)^T$ as the corresponding weighting vector estimated from the data. For the j-th variable, $j \in \{1, \dots, p\}$, the SOIL importance measure is defined as the sum of weights of the candidate models $\mathcal{A}_k$ that contains the j-th variable.

$$\textbf{SOIL Importance}: S_j \equiv S(j; \boldsymbol{w}, \boldsymbol{A}) = \sum_{k=1}^{K} w_k I(j \in \mathcal{A}_k)$$

2.1 Theoretical properties

Several properties need to be satisfied for consistency of the SOIL importance measure. The first two properties for the weighting vector $\boldsymbol{w}$ are *weak consistency* and *consistency*, ensuring that the weighting is concentrated enough around the true model to different degrees. The definitions are shown below:

- The weighting vector $\boldsymbol{w}$ is *weakly consistent* if
$$\frac{\sum_{k=1}^{K} w_k |\mathcal{A}^k \nabla \mathcal{A}^*|}{r^*} \xrightarrow{p} 0, \qquad as\ n \to \infty$$
- The weighting vector $\boldsymbol{w}$ is *consistent* if
$$\sum_{k=1}^{K} w_k |\mathcal{A}^k \nabla \mathcal{A}^*| \xrightarrow{p} 0, \qquad as\ n \to \infty$$

where ∇ denotes the symmetric difference of two sets and $|\cdot|$ denotes number counting. The r^* in the denominator of the weak consistency equation makes it more likely to be satisfied than consistency property when the dimension of the true model is allowed to increase as the sample size increases.

The author also provided two theorems under the assumption of the two properties.

- When weighting vector w is weakly consistent, we have:
$$\frac{\sum_{j \in \mathcal{A}^*} S_j}{r^*} \xrightarrow{p} 1, \qquad \frac{\sum_{j \notin \mathcal{A}^*} S_j}{r^*} \xrightarrow{p} 0, \qquad as\ n \to \infty$$
- When the weighting w is consistent, we have:
$$\min_{j \in \mathcal{A}^*} S_j \xrightarrow{p} 1, \qquad \max_{j \notin \mathcal{A}^*} S_j \xrightarrow{p} 0\,, \qquad as\ n \to \infty$$

Under the assumption that the weighting is weakly consistent, the sum of the SOIL importance of the true variables will tend to r^*, the size of the true model. The sum of the SOIL importance of unimportant variables will tend to zero as the sample size goes to infinity. Similarly, if the weighting is consistent, the sum of the SOIL converges to one for true variables and converges to zero for variables excluded by the true model.

Many different methods could provide the weighting vector for the candidate models. For example, a weighting method is based on information criterion (e.g., AIC and BIC). A weighting strategy by data splitting and cross-assessment is called the adaptive regression by mixing (ARM). Also, the weighting by Bayesian model averaging (BMA). Under the specific condition, these methods could give the consistent weights $\boldsymbol{w}$.

Usually, the consistency of the weighting method is proved when all subset models are considered. However, when p is large, it is computationally infeasible to include all the variables; therefore, a screening method is introduced to reduce the number of variables. The consistency of SOIL importance is proved as:

- (Path-consistent). A method is called path-consistent if
$$P(A^* \in \triangle) \to 1, as\ n\ \to \infty,$$
where $\triangle$ denotes the whole solution paths produced by the method.
- Under the assumption that the weighting $\boldsymbol{w}$ on the all-subset candidate models A is consistent, as long as at least one method is path-consistent, we have
$$\min_{j \epsilon A^*} S(j; w', A') \xrightarrow{p} 1, \quad \max_{j \notin A^*} S(j; w', A') \xrightarrow{p} 0, \quad as\ n \to \infty$$
where $\boldsymbol{w}'$ is the renormalized weighting on A', which is the collection of models using union of solution paths.

A threshold value $c \in (0,1)$ could be set up for the variable importance for applications so that only the variables' importance greater than c will be kept. Let $\mathcal{A}_c = \{j: S_j > c\}$ be the model selected by this threshold value, and it has the property below.

- For any threshold $c \in (0,1)$, denote $\overline{\mathcal{A}_c} = \{j \in \mathcal{A}^*: S_j \leq c, j = 1, \dots, p\}$, $\underline{\mathcal{A}_c} = \{j \notin \mathcal{A}^*: S_j > c, j = 1, \dots, p\}$, then if w is weak consistent, we have
$$\frac{\left|\overline{\mathcal{A}_c}\right|}{r^*} \xrightarrow{p} 0, \qquad \frac{\left|\underline{\mathcal{A}_c}\right|}{r^*} \xrightarrow{p} 0, \qquad as\ n \to \infty$$

It means for any c, the number of the true variables missed by $\mathcal{A}_c$ and those over-selected variables in $\mathcal{A}_c$ will be relatively small when the sample size increases.

2.2 Candidate model

Based on their study, there are two approaches to choose the model for SOIL importance: Using a complete collection of all subset models and using tools for high-dimensional penalized regression.

The complete collection of all-subset models' approach is expressed as:

$$A = \left\{\emptyset, \{j_1\}, \dots, \{j_p\}, \{j_1, j_2\}, \{j_1, j_3\}, \dots, \{j_1, \dots, j_p\}\right\} \text{ where } j_1, \dots, j_p \in \{1, \dots, p\}$$

And the other approach is using tools for high-dimensional penalized regression when p>>n

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \sum_{j=1}^{p} p_\lambda(\beta_j)$$

Where $p_\lambda(\cdot)$ is a non-negative penalty function with regularization parameter $\lambda \in (0, \infty)$, such as Lasso penalty $p_\lambda(u) = \lambda w|u|$ in the equation above, and nonconvex penalties including the smoothly clipped absolute deviation (SCAD) penalty or the minimax concave penalty (MCP).

The author first applies a high-dimensional model selection method: SCAD on the data to compute solution paths for a sequence of tuning parameters $\{\lambda_1, \dots, \lambda_L\}$. Let $\{\hat{\beta}^{\lambda 1}, \dots, \hat{\beta}^{\lambda L}\}$ be the estimate coefficients of L different regularization levels for the SCAD penalty and

$$A_{SCAD} = \{A^{\lambda 1}, A^{\lambda 2}, \dots, A^{\lambda L}\}$$

be the resulting estimated model, where $A^{\lambda l} \equiv supp(\hat{\beta}^{\lambda 1}) = \{j : \hat{\beta}_j^{\lambda l} \neq 0\}$. We then use the set A_{SCAD} as the set of candidate models.

Further increase the chance of capturing the true /best model by combining the resulting model from several different penalties, which forms a more extensive set of candidate model, e.g.

$$A = \{A_{Lasso}, A_{AdaptiveLasso}, A_{SCAD}, A_{MCP}\}$$

At least one candidate model in the solution paths must be the true model; then SOIL importance can work well and provide sensible weighting; even the rest are not close to the true model.

2.3 Weighting

There are two kinds of weighting methods focused on this paper: ARM weighting, a weighting strategy by data splitting, and BIC Weighting by BIC or modified BIC information criterion (BIC-p) for high dimensional data.

Yang & Barron (1998) pointed out that when there are exponentially many models, the model complexity should be considered, which can also be interpreted as the prior probability for the model.

- **Weighting using ARM with non-uniform priors.**

The ARM weighting method randomly split the data $D = \{(x_i, y_i)\}_{i=1}^{n}$ into a training set D_1 and a test set D_2 of equal size. Then the regression model trained on D_1 is used for prediction on D_2. Then the weights $w = (w_1, \dots, w_k)^T$ can be computed based on this prediction.

Algorithm 1 The procedure of the ARM weighting for the regression case.

- Randomly split $\mathbf{D}$ into a training set $\mathbf{D}_1$ and a test set $\mathbf{D}_2$ of equal size.
- For each $\mathcal{A}^k \in \boldsymbol{A}$, fit a standard linear regression of y on $\mathbf{x}_s^{(k)}$ using the training set $\mathbf{D}_1$ and get the estimated $\widehat{\boldsymbol{\beta}}_s^{(k)}$ and $\widehat{\boldsymbol{\sigma}}_s^{(k)}$.
- For each $\mathcal{A}^k$, compute the prediction $\mathbf{x}_s^{(k)\intercal}\widehat{\boldsymbol{\beta}}_s^{(k)}$ on the test set $\mathbf{D}_2$.
- Compute the weight w_k for each candidate model:

$$w_k = \frac{e^{-\psi C_k}(\widehat{\boldsymbol{\sigma}}_s^{(k)})^{-n/2}\prod_{i\in\mathbf{D}_2}\exp(-(\widehat{\boldsymbol{\sigma}}_s^{(k)})^{-2}(y_i - \mathbf{x}_{s,i}^{(k)\intercal}\widehat{\boldsymbol{\beta}}_s^{(k)})^2/2)}{\sum_{l=1}^{K} e^{-\psi C_l}(\widehat{\boldsymbol{\sigma}}_s^{(l)})^{-n/2}\prod_{i\in\mathbf{D}_2}\exp(-(\widehat{\boldsymbol{\sigma}}_s^{(l)})^{-2}(y_i - \mathbf{x}_{s,i}^{(l)\intercal}\widehat{\boldsymbol{\beta}}_s^{(k)})^2/2)},$$

  for $k = 1, \dots, K$, where $C_k = s_k \log\frac{e \cdot p}{s_k} + 2\log(s_k + 2)$.
- Repeat the steps above (with random data splitting) L times to get $w_k^{(l)}$ for $l = 1, \dots, L$, and get $w_k = \frac{1}{L}\sum_{l=1}^{L} w_k^{(l)}$.

- **Weighting using information criteria with non-uniform priors**

Define $I_k^{BIC} = -2(\log\ell_k + s_k \log n)$ as the BIC information criterion, where ℓ_k is the maximized likelihood for model k and s_k denotes the number of non-constant predictors. Then weight w_k for model $A^k \in \boldsymbol{A}$ is computed by

$$w_k = \exp\left(-\frac{I_k}{2} - \psi C_k\right) \Big/ \sum_{l=1}^{K} \exp\left(-\frac{I_l}{2} - \psi C_l\right)$$

3. Simulation

The author compared SOIL importance measures using ARM and BIC-p weighting with LMG, RFI1, and RFI2 to emphasize SOIL properties. LMG is the relative importance measure by averaging over all possible orderings for R^2 decomposition (Lindeman et al. 1980). Breiman (2001) proposed RFI1 and RFI2, which are importance measures in random forests. For the choice of the prior ψ, the author used 0.5 to avoid cherry-picking. In this paper, the comparison of variable importance measures based on Gaussian and Binomial case with different settings of sample sizes, dimensions, and feature correlations, and the simulation was repeated 100 times to increase the accuracy of the results.

Table 2 below shows the summary of the properties of different importance measures. The SOIL-ARM and SOIL-BIC-p almost have all the characteristics listed except pure relativeness and non-parametricness. Each characteristic will be discussed in detail.

	SOIL-ARM	SOIL-BIC-p	LMG	RFI1	RFI2
Inclusion/Exclusion	✓	✓			
Tuning in to information	✓	✓			
Robustness to feature correlation	✓	✓			
Robustness against confuser	✓	✓			
Sensitivity to high-order terms	✓	✓			
Pure relativeness			✓	✓	✓
Order preserving	✓	✓			
High-dimensionality	✓	✓		✓	✓
Non-parametricness				✓	✓
Non-negativity	✓	✓	✓		✓

Table 2: Comparison of the characteristics for the importance measures. A "✓" indicates that a specified method has the given property. A blank space indicates the absence of a property.

- **Inclusion/exclusion**

These two properties address the issue if an important measure can accurately give the "true" variable in the best model. As mentioned previously in the paper, SOIL gives large importance to the true variables and give zero to those unimportant variables. All the examples also proved that SOIL-ARM and SOIL-BIC-p have inclusion and exclusion properties.

- **Tuning in to information**

For high-dimensional data, sparsity is a reluctant acceptance due to the limitation of information. The optimal sparsity should depend on the sample size and noise level. A good importance measure should have this property which means it should change when more information is provided. The author showed that only SOIL importances reacted to the increase in sample size.

- **Robustness to feature correlation**

From the results, only SOIL-ARM and SOIL-BIC-p have the property of robustness against noise increase and higher feature correlation.

- **Robustness against confuser**

An importance measure should assign approximately zero to the confuser for sparse modeling. The examples indicated that SOIL importances are much more robust than the other three measures.

- **Sensitivity to higher-order term**

The simulation showed that both ARM and BIC-p could select true main-effect variables and true higher-order terms while LMG, RFI1, and RFI2 failed to select those variables when higher-order terms are included.

- **Pure relativity**

This property means the value itself does not have a valuable meaning individually. It is not preferred for the importance measure to have this property since the cases of equal importance

and equal unimportance are undifferentiable. The examples concluded that LMG, RFI1, and RFI2 could not provide the importance of each variable on its own.

- **Order Preserving**

For the true variables with not too high correlations with others, it may be natural to expect the ones with larger coefficients to have larger importance. Moreover, the true variables should have larger importance compared to the noisy ones. Overall, SOIL-BIC-p and SOIL-ARM perform the order-preserving properties in all the cases.

- **High dimensionality**

SOIL-BIC-p, SOIL-ARM, RFI1, and RFI2 can work for high dimensional data when p > n; however, LMG does not.

- **Non-negativity**

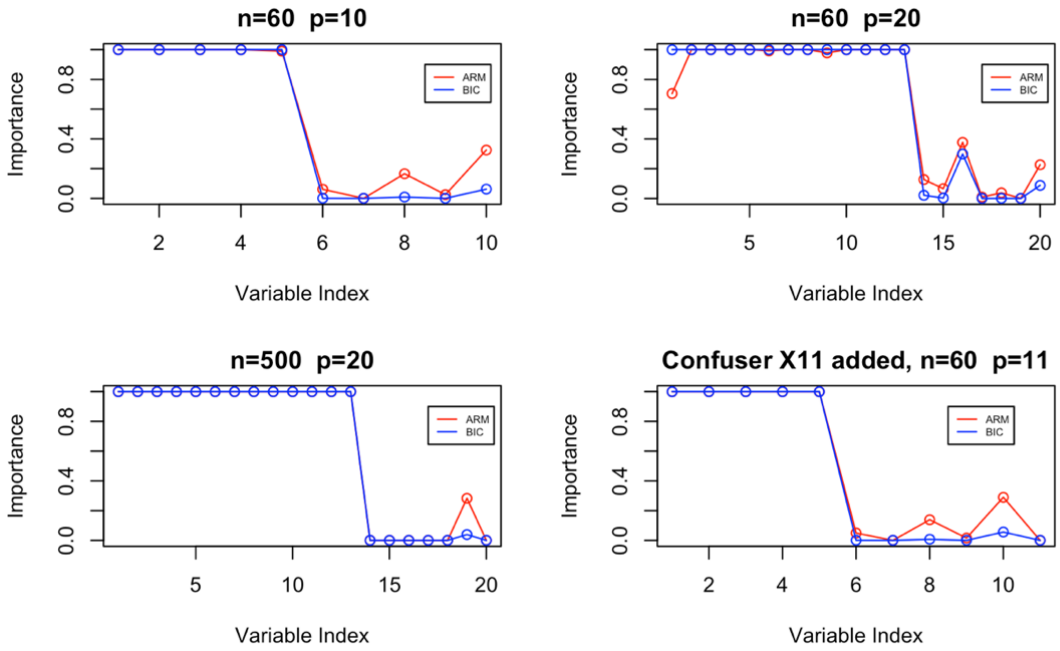
SOIL-BIC-p, SOIL-ARM, LMG, and IMG2 always produce non-negative importance value. However, RFI1 does not satisfy this criterion.

- **Non-parametricness**

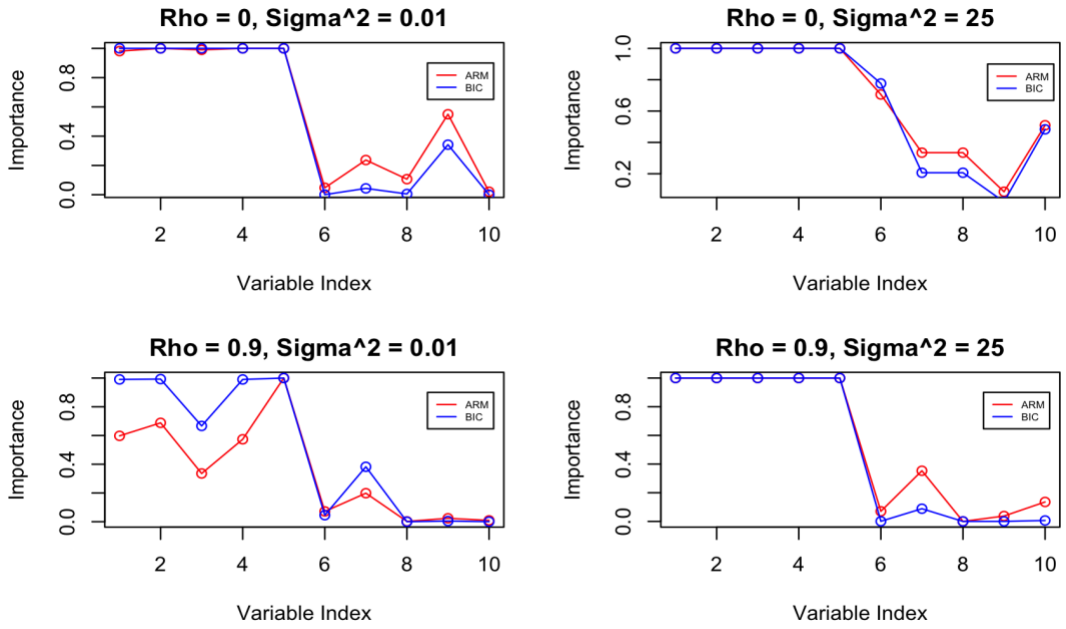
Among all these importance measures, only the RF1 and RF2 are not limited to parametric modeling.

4. Method Testing (SOIL)

We tested SOIL importance measure under several situations with different sample sizes and the number of variables. We also tested the SOIL properties of robustness against confuser and robustness to feature correlation. The figures below show the results of our simulations.



The top two figures show that no matter how the number of predictor "p" changes, SOIL-ARM and SOIL-BIC-p always have inclusion and exclusion properties. When the sample size increases, both the SOIL-ARM and SOIL-BIC-p react well to the much-increased information. In case of a confuser x_{11} added, SOIL importance measures show robustness to the confuser, since the confuser is not assigned to the true model as shown in figure bottom-right.



To test the robustness to feature correlation, we considered four cases including low feature correlation with low noise ($\rho = 0$, $\sigma^2 = 0.01$), low feature correlation with high noise ($\rho = 0$, $\sigma^2 = 25$), high feature correlation with low noise ($\rho = 0.9$, $\sigma^2 = 0.01$), and high feature correlation with high noise ($\rho = 0.9$, $\sigma^2 = 25$). We can see that even when the feature correlation and/or the noise is high, the SOIL importance measure can still accurately select the true variables. (The code that generated these figures could be found in the supplementary documents).

Our simulation on SOIL shows consistency with the results provided by the author in the paper, confirming that SOIL importance can indeed do a good job on different scenarios. Also, for applying the variable importance measure to the real data examples provided in the paper, SOIL showed relatively better performance than the other measures with large p or small p. For example, the top 10 variables from the Bardet dataset were chosen using ARM, BIC-p, RFI1, and RFI2. The ARM and BIC-p can get the same ranking of the top four variables in different runs, whereas RFI1 and RFI2 are unstable due to different rankings each time. Besides, SOIL cannot be perfect in all situations. For using cross-examination for the real data to test the measures, some property did not hold exactly in the home game of RFI1 or RFI2. However, although SOIL might miss subtle variables in the true model when p is small, it could rarely assign an unimportant variable as important.

5. Conclusion/Contribution

The goal of variable importance is to find the important variables for explaining or predicting the response. There are several challenges the author faced when they proposed the SOIL approach. First, the different goals may require different importance measures when importance depends on analysis and application. Second, the choice of parametric and non-parametric models that the importance should be based on is another problem the author needs to solve. Third, the author also needs to consider whether comparing different variables or values has its own meaning.

The proposed new variable importance measure: SOIL is formed by model combination for considering more than a single model, then providing an understanding of all the variables instead of a single “important” variable given a single model. Meanwhile, the SOIL approach has several wanted features such as exclusion/inclusion, order-preserving, and robustness in several aspects.

Besides the Statistical part, some benefits help people to solve the real word puzzle. The SOIL approach saves time and cost in data analysis and helps decision-makers to understand the underlying data process rather than trust any single model and gain the ability to change and replace variables.

Through our own simulation work, we have shown that SOIL-ARM and SOIL-BIC-p have the ability to identify the importance of variables in the true model. Overall, SOIL-ARM and SOIL-BIC-p have shown better performance compared to other existing methods.

6. Suggestions

During the reading and studying of this paper, we found several typos appeared in the paper, such as spelling mistakes "weighing" all over the paper. Also, the author wrongly referred the section 4 as 3.2 on page 5. This might lead to a misunderstanding of the content.

In general, the paper did well on the structure, content, layout, simulation, and explanations. Moreover, the SOIL approach proposed by the author dramatically improved the performance of selecting variables and model building. We hope the SOIL method can be widely used in both real-life problems and academic studies.

References

Ando, T., & Li, K. (2014). A Model-Averaging Approach for High-Dimensional Regression. *Journal of the American Statistical Association, 109*(505), 254-265. Retrieved April 16, 2021, from http://www.jstor.org/stable/24247152

Breiman, L. (2001), 'Random forests', *Machine Learning* 45(1), 5–32.

Budescu, D. V. (1993), 'Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression', *Psychological Bulletin* 114(3), 542.

Chevan, A. & Sutherland, M. (1991), 'Hierarchical partitioning', *The American Statistician* 45(2), 90–96.

Theil, H. & Chung, C. (1988), 'Information-theoretic measures of fit for univariate and multivariate linear regressions', *The American Statistician* 42(4), 249–252.

Yang, Y. & Barron, A. R. (1998), 'An asymptotic property of model selection criteria', *IEEE Transactions on Information Theory* 44(1), 95–116.