Modified Akaike Information Criterion for Quantifying Regularization in Nonlinear Regression

Sungjun Eom

January 11, 2025

Contents

1 Introduction

Model selection is a crucial step in statistical modeling. It is about balancing the trade-off between model complexity and goodness-of-fit. One of the most widely used criteria is the Akaike Information Criterion (AIC). It calculates the likelihood penalized by the number of parameters to discourage overfitting, or over-parameterization. Smaller AIC value is assumed to be better. On the other hand, in modern applications such as deep learning, the models often contain a large number of parameters so that models constructed by deep learning are always considered over-parameterized. Although they are over-parameterized, the deep learning models generalize well because of the regularization techniques, such as dropout, L1 (LASSO) and L2 (ridge) regularization. If one tries to calculate the AIC for these highly over-parameterized models, it always gives a high score, meaning that the models are over-parameterized. In this respect, the AIC is not a suitable measure for models which are over-parameterized but generalize well. AIC does not inherently account for regularization, as it penlizes models based purely on the number of parameters rather than their effective complexity after regularization.

This research aims to address this gap by proposing a modification to AIC that incorporates the effects of regularization. Specifically, we introduced the effective number of parameters whose absolute values exceed a certain threshold. This approach reflects the fact that regularization can reduce the number of active parameters in a model. By introducing the effective number of parameters into AIC, we want to improve its applicability in regularized models. We named this modified AIC as Effective Akaike Information Criterion (EAIC) meaning that it reflects the "effective" number of parameters.

The objective of this study is to validate the theory of EAIC through empirical analysis. We will investigate the properties of EAIC with various nonlinear regression techniques, such as polynomial regression, domain-knowledge based nonlinear regression, and deep learning. By incorporating regularization into AIC, this research seeks to provide a more robust tool for model selection while considering regularization.

2 Related Work

2.1 Akaike Information Criterion

The AIC has long been a pivotal tool in model selection, particularly in contexts where statistical inference and predictive accuracy are crucial. Developed by Akaike [1], AIC provides a method for

balancing the trade-off between model complexity and goodness-of-fit. It operates under the principle of parsimony, where the aim is to select a model that adequately explains the data with the fewest parameters. AIC is calculated as a combination of the likelihood of the model and a penalty term based on the number of estimated parameters (1).

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \tag{1}$$

$\hat{L}$ denotes the maximized value of the likelihood function and k denotes the number of parameters. This formulation effectively discourages over-fitting by penalizing models with excessive complexity, thus aligning well with parsimony. Over the years, AIC has been extended and adapted for various modeling approaches. Adjustments such as AICc, a corrected version of AIC, have been developed to address biases in small smaple sizes [4]. These extensions highlight the flexibility of AIC in different statistical environments.

In highly parameterized models, such as deep learning, traditional AIC may fall short. This is because AIC inherently assumes that model complexity is directly correlated with the number of parameters, which does not account for the regularization effects that can significantly alter the active dimensional complexity of a model.

2.2 Regularization

Regularization techniques have become essential in modern statistical modeling and machine learning when dealing with high-dimensional data. These methods aim to prevent overfitting by introducing additional constraints on the model parameters. Two famous forms of regularization are L1 regularization (LASSO) and L2 regularization (ridge regression). [5] These constraints are incorporated into objective with a Lagrangian framework. In LASSO, the penalty term is equal to the absolute value of the coefficients, and the optimization problem can be expressed as below.

$$\mathcal{L}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In this case, the Lagrangian incorporates the regularization constraint which forces sparsity in the parameter estimates. The presence of the absolute value penalty effectively pushes some coefficients exactly to zero, making LASSO particularly useful for feature selection.

Ridge regression introduces a penalty based on the squared values of the coefficients.

$$\mathcal{L}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

In this formulation, the Lagrangian framework forces the coefficients to have small values. L2 regularization does not produce exact zeros but shrinks all parameter estimates toward zero, leading to a smoother model. This approach is particularly advantageous in situations where multicollinearity exists among the predictors, as it helps distribute the influence of correlated variables more evenly.

2.3 Effective Number of Parameters

In over-parameterized models including neural networks, since not all parameters contribute to the model's predictability, there has been attempts to count the effective number of parameters. Recent work by [2] addresses this concept by examining the phenomenon called Double Descent. Traditional statistical learning has long been guided by a U-shaped curve relating model's complexity and generalization error. However, in over-parameterized models, a second descent in test error has been observed as parameter counts continue to increase. The authors of [2] examine this by redefining the complexity through the effective number of parameters, not the true number of parameters. Specifically, they demonstrate that while raw parameter counts grow, not all parameters increase model's complexity. They propose Generalized Effective Parameter measure for model complexity in over-parameterized models. Their definition is based on the variance of parameters with some weights called smoother.

From a theoretical perspective, [6] expand on the concept of effective parameters by introducing a general framework for Mean Integrated Squared Error (MISE) in large neural networks. They identify two distinct behaviors—double descent and monotonic MISE reduction—depending on model design and regularization strategies. Additionally, [8] focus on sparse double descent, where L1 regularization and sparsity induce a nuanced complexity-generalization relationship. These studies collectively underscore the importance of redefining complexity metrics to better reflect the impact of regularization and parameter sparsity.

Building on these foundations, our work introduces the Effective Akaike Information Criterion (EAIC), a practical adaptation that explicitly incorporates effective parameter counts. By defining a threshold for parameter significance, we hope that EAIC will bridge theoretical insights with model selection considering regularization, providing a practical tool for evaluating models across diverse settings, including highly parameterized neural networks and traditional regression models.

3 Methodology

3.1 Effective Akaike Information Criterion

In traditional AIC, the complexity is purely a function of the total number of parameters denoted as k. However, in regularized models, not all parameters contribute equally to the model's predictive ability. Regularization techniques such as L1 and L2 induce sparsity or shrink the value of parameters to reduce model's complexity. Therefore, the true "effective" complexity of the model is less than or equal to the number of total parameters.

EAIC is derived from an additive combination of AIC and a negative value of measure l which quantifies how much regularized the model is.

$$\text{EAIC} := \text{AIC} - 2l(g) \tag{2}$$

$$= 2(k - l(g)) - 2\ln(\hat{L}) \tag{3}$$

$$= 2k_{\text{eff}} - 2\ln(\hat{L}) \tag{4}$$

l is a function that gives the number of regularized parameters below a threshold h as an output where g is an approximating model. The first two terms in (3) can be combined together then become k_{eff} in (4) which refers to the effective number of parameters. Then EAIC is reformulated by replacing k in AIC with k_{eff}, representing only those parameters that have an absolute value exceeding a predefined threshold h. These are considered to be the parameters that are still meaningful to the model. This reflects the number of parameters whose estimated values significantly contribute to the model after regularization has been applied, distinguishing between active and inactive parameters.

A formal definition of k_{eff} is below,

$$k_{\text{eff}} = \sum_{i=0}^{k} I(|\beta_i| > h) \tag{5}$$

where $I(\cdot)$ denotes an indicator function. This approach ensures that the model complexity, as measured by EAIC, better reflects the true number of effective parameters after regularization, taking into account their contribution to the model's performance.

The choice of the threshold h is a crucial component of the EAIC. One may consider the scale of independent variables for determining h, since the unit of independent variables determines the magnitude of slope with the analogy of simple linear regression. When the independent variables are many,

meaning that the dimensions of input are high, the data often normalized to have values between -1 and 1 or similar. To consider the scale-variant nature of parameters shown here, the threshold h is determined as a fraction of or a multiplication r to the empirical deviation σ of parameters, $h = r\sigma$, resulting in the number of statistically significant parameters. Typical choices for r are listed below.

- $h = 0.1\sigma$: This would be a relatively strict threshold, meaning only parameters that deviate significantly from zero.
- $h = 1.96\sigma$: 95% confidence interval as a threshold implying that how significant a single parameter is when assumed for it to follow Gaussian distribution.

The introduction of effective parameters enables for statisticians to compare models with same parameter space but with different values of parameters. This happens in the nonlinear regression having a non-convex objective, leading multiple optima.
In the demonstration of EAIC with empirical analysis that follows, we will apply the EAIC to various regularized regression models that the Regression Analysis 2 (회귀분석2) covered so far.

4 Dataset

4.1 Abalone Dataset

The Abalone dataset from UCI Machine Learning Repository [7] is a well-known dataset used primarily for regression analysis. The goal is to predict the age of abalone from physical measurements. This dataset has been used in predictive modeling research. The properties of this dataset are follows.

- **Sex (Nominal)**: The gender of the abalone (M: Male, F: Female, I: Infant).
- **Length (Continuous)**: The longest shell measurement in millimeters.
- **Diameter (Continuous)**: The diameter of the shell, perpendicular to the length.
- **Height (Continuous)**: The height of the shell, with meat inside, in millimeters.
- **Whole weight (Continuous)**: Total weight of the abalone, in grams.
- **Shucked weight (Continuous)**: Weight of the abalone's meat, in grams.
- **Viscera weight (Continuous)**: Gut weight after bleeding, in grams.
- **Shell weight (Continuous)**: Weight of the shell after being dried, in grams.

- **Rings (Integer)**: The number of rings, which serves as the target variable for prediction. The age of an abalone is usually estimated as `Rings + 1.5`.

4.2 MNIST

In addtion to the Abalone dataset, we will utilize the MNIST dataset [3] which is widely recognized benchmark for image classification tasks. We demonstrate the effectiveness of the EAIC in the context of highly nonlinear regression models with non-convex multiple optima objectives. The MNIST dataset consists of images of handwritten digits (0-9), providing a rich and complex challenge for classification. The objective is to classify each image into one of the ten digit categories. By leveraging this dataset, we aim to illustrate how EAIC can quantify the generalizability of deep neural networks, which are often characterized by complex architectures and a large number of parameters. Evaluating the models using EAIC will allow us to gain insights into their effective complexity and predictive accuracy, particularly in the face of the challenges posed by nonlinear relationships and potential overfitting in deep learning scenarios.

5 Results

In this section, we demonstrate our proposed evaluation metric for model selection with two different cases, which are abalone dataset and MNIST. Each case represents two different situations in regression analysis. The first case, the abalone dataset, is typically considered to be suitable for traditional linear regression analysis. On the other hand, MNIST is recently developed and suitable for nonlinear regression because of its high nonlinearity.

5.1 Case 1: Ablaone

5.1.1 Linear regression with and without regularization

In this section, we present a comparison between the Akaike Information Criterion (AIC) and the Effective Akaike Information Criterion (EAIC) in selecting regression models with better generalizability using the Abalone dataset. The goal is to evaluate whether EAIC provides a better measure for model selection in regularized models compared to the traditional AIC.
We conducted an analysis using three different models:

- **LASSO Regression** Linear regression with L1 regularization.

- **Ridge Regression** Linear regression with L2 regularization.
- **Linear Regression** Standard linear regression without regularization.

The Abalone dataset was used for this analysis, excluding the nominal variable `Sex` to avoid issues with non-numeric data. The dataset was split into training and test sets, with 80% of the data used for training and the remaining 20% for testing. The threshold parameter r for EAIC was set to 0.1 and the lambda parameter r for regularization was also set to 0.1

The results of the analysis are summarized in Table 1, which presents the AIC, EAIC, and test Mean Squared Error (MSE) for each model.

Model	**AIC**	**EAIC**	**Test MSE**
LASSO	15084.89	15076.89	5.7690
Ridge	14956.28	14952.28	5.6247
Linear	14825.98	14823.98	5.3125

Table 1: AIC and EAIC values for different models

From the results, the Linear Regression model achieved the lowest AIC, EAIC, and test MSE, indicating that it is the best model according to both information criteria and actual performance on unseen data. This outcome is expected due to the simplicity of the dataset and the absence of high-dimensional features that might necessitate regularization.

To assess which criterion better predicts generalizability, we computed the correlation between the information criteria and the test MSE.

Metric	**Correlation(%)**
Correlation between AIC and Test MSE	97.8983
Correlation between EAIC and Test MSE	97.9921

Table 2: AIC and EAIC values for different models

In Table 2, the slightly higher correlation between EAIC and test MSE suggests that EAIC may be marginally better at representing generalizability. This is anticipated since EAIC adjusts for the effective number of parameters, potentially providing a more accurate penalty for model complexity in regularized models.

5.2 Case 2: MNIST

In this section, we demonstrate the application of the Effective Akaike Information Criterion (EAIC) on the MNIST dataset, a benchmark for evaluating image classification models. MNIST consists of

70,000 grayscale images of handwritten digits (0-9), each of size 28×28, divided into 60,000 training samples and 10,000 test samples. The dataset represents a complex nonlinear problem, making it ideal for testing the robustness of EAIC in selecting models with a large number of parameters.

5.2.1 Models Evaluated

We analyzed the following models for this case:

- **Shallow Neural Network (SNN)**: A single-hidden-layer neural network with 128 neurons. This model serves as a baseline for comparison.
- **Deep Neural Network (DNN)**: A multilayer perceptron with three hidden layers, each containing 512 neurons. This model represents a more complex architecture with a larger number of parameters.
- **Deep Neural Network with Regularization (DNN-Reg)**: The same architecture as DNN but with L2 regularization ($\lambda = 0.01$) applied to the weights to prevent overfitting.

5.2.2 Experimental Setup

The models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 64. The activation function used was ReLU for all hidden layers, and the output layer employed softmax for classification. Training was conducted for 10 epochs.

For EAIC calculation, the threshold h for determining the effective number of parameters (k_{eff}) was set to $h = 0.1\sigma$, where σ is the standard deviation of the parameter values. This threshold was chosen to balance the sensitivity of EAIC to parameter significance.

5.2.3 Results and Analysis

Table 3 summarizes the results of the evaluation, including test accuracy, AIC, EAIC, total parameters, and effective parameters for each model.

Model	Test Accuracy (%)	AIC	EAIC	Total Params	Effective Params
SNN	95.17	222926.91	204022.91	101770	92318
DNN	97.04	493637.43	453629.43	242762	222758
DNN-Reg	96.99	494638.56	451640.56	242762	221263

Table 3: Model performance metrics including test accuracy, AIC, EAIC, total parameters, and effective parameters.

Table 4 presents the correlation between the test accuracy and both AIC and EAIC. The correlation indicates how well these criteria predict the generalizability of the models.

Metric	**Correlation (%)**
Correlation between AIC and Test Accuracy	99.96
Correlation between EAIC and Test Accuracy	99.99

Table 4: Correlation analysis between AIC/EAIC and test accuracy.

6 Conclusion

In this paper, we introduced the EAIC as a modification to the traditional AIC for models incorporating regularization. While the AIC has proven to be a valuable tool for model selection, it does not account for the effects of regularization techniques that can reduce the true complexity of a model by shrinking or sparsifying parameters. The EAIC addresses this limitation by introducing the concept of the effective number of parameters k_{eff} which reflects only those parameters that have a significant contribution to the model after regularization. By adjusting the threshold h, the EAIC can balance model complexity and goodness-of-fit more efficiently than traditional AIC, especially in over-parameterized and regularized models such as deep neural networks. However, we also found it out that in the traditional regression where over-parameterization does not occur, the AIC still be considerable.

References

[1] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[2] Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[4] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[5] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[6] Marco Morucci and Arthur Spirling. Model complexity for supervised learning: Why simple models almost always work best, and why it matters for applied research. Technical report, Working Paper, 2024.

[7] Nash, Warwick, Sellers, Tracy, Talbot, Simon, Cawthorn, Andrew, Ford, and Wes. Abalone. UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C55C7W.

[8] Ya Shi Zhang. Manipulating sparse double descent. *arXiv preprint arXiv:2401.10686*, 2024.