Machine learning and multinomal models: A simulation study

F. Chan [a], M. Harris [a], R. Singh [a] and W. Yeo [a]

[a] *School of Economics and Finance, Curtin University, Perth, Western Australia*
Email: weiern.yeo@postgrad.curtin.edu.au

Abstract: In recent times, the use of machine learning in academic research has become a more and more attractive option, an option that is becoming increasingly popular in the field of econometric research A significant amount of econometric literature has pitched the use of machine learning techniques for the purposes of improving research in roles such as variable selection or causal effect evaluation, in papers such as Athey and Imbens (2017) and Varian (2014). However, conversely, there is an equal amount of literature warning against blindly making use of these techniques without adjustments in papers such as Athey and Imbens (2019) and Mullainathan and Spiess (2017).

In order to attain a better understanding of how the estimators provided by some of these machine learning techniques behave compared to their more traditional counterparts, this paper conducts a Monte Carlo Simulation study between Shrinkage based techniques such as LASSO and the Elastic Net, against Logistical Regression in a sparse multinomial model environment, comparing both selection consistency and estimator variance between different sample sizes.

Overall, the results of this paper show that when dealing with sparse multinomial environments, of the methods tested, the Elastic Net provides the clearest advantage in selection consistency, both being able to mark strongly relevant variables as strongly relevant more often than LASSO and Logit, but also managing to avoid marking irrelevant variables as relevant. Moreover, when compared to its peers Elastic Net holds the greatest advantage when dealing with datasets with low samples. However, this advantage is not universal, as it holds a natural bias when assigning values to coefficients and its ability to pick out weakly relevant variables is comparable to that of Logit.

Keywords: *Econometrics, shrinkage, non-linear models, Monte Carlo simulation*

1 INTRODUCTION

In recent times, the use of machine learning in academic research has become a more and more attractive option, an option that is becoming more and more popular in the field of Econometric research A significant amount of econometric literature has pitched the use of machine learning techniques for the purposes of improving research in roles such as variable selection or causal effect evaluation, in papers such as Athey and Imbens (2017) and Varian (2014). However, conversely, there is an equal amount of literature warning against blindly making use of these techniques without adjustments in papers such as Athey and Imbens (2019) and Mullainathan and Spiess (2017).

In order to attain a better understanding of how the estimators provided by some of these machine learning techniques behave compared their more traditional counterparts, this paper conducts a Monte Carlo Simulation study between Shrinkage based techniques such as LASSO and the Elastic Net, against Logistical Regression in a sparse multinomial model environment.

2 BACKGROUND

A significant amount of econometric literature has pitched the use of machine learning techniques for the purposes of improving research. Varian (2014), for example, highlights a few machine learning techniques they think are beneficial to the econometric field, proposing using machine learning for variable selection, data manipulation and model selection. Athey and Imbens (2017), conversely, in their review on techniques evaluating the causal effects of policies, specifically highlight the use machine learning techniques have found evaluating both average and heterogeneous causal effects. Other similar econometric review papers include Kuan and White (1994), Gentzkow et al. (2017) and Chen (2021).

However, there is parallel literature warning econometric researchers of the danger of blindly applying machine learning techniques to problems without thought. A major theme of Athey and Imbens (2019)'s review notes that while there are cases where using off the shelf machine learning techniques would be sufficient, to adapt to the need of the econometric researcher machine learning techniques typically need careful adjustment and adaption to the specific problem they are trying to address. Similarly, Mullainathan and Spiess (2017) notes the danger in blindly accepting machine learning based predictions as inference, recording the field's focus on predicting $\hat{y}$ rather than accurate estimator values as a concern.

As there is an innumerable variety of machine learning techniques, of which only some of the most popular include the weights based Artificial Neural Networks proposed by McCulloch and Pitts (1943) and the decision tree based Random Forests by Breiman (2001) and the many variations thereof, this paper will primarily focus on one of them, regularisation, more commonly referred to as Shrinkage, due to the ties it has with econometric literature. For example, Shrinkage has found some use in the econometric field as a form of variable selection - where explanatory variables are subjected to regularisation to determine if they are relevant to the determinant, typically followed by placing the relevant variables into a chosen model such as basic OLS (for example, Caner et al. (2016), Shi (2016)).

2.1 Regularisation and Shrinkage Estimators

When stepwise regression fell out of more general use due to criticisms such as the tests being biased as demonstrated by Wilkinson and Dallal (1981), regularisation came in to fill the niche that it left behind. One of the most simple examples of which is Ridge Regression, defined below as an objective and a penalty function, where θ is a p x 1 vector of parameters, x denotes a k x 1 vector of explanatory variables and y is the determinant. The first half of the equation, $g(\theta; \mathbf{y}, \mathbf{x})$ represents the objective function and the second half $\lambda p(\theta; \alpha)$ is the penalty function.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) - \lambda p(\boldsymbol{\theta}; \boldsymbol{\alpha})$$

First introduced by Hoerl and Kennard (1970), ridge regression aims to reduce prediction error in the model through the elimination of irrelevant variables. Rather than testing all variables for relevance, ridge regression subjects the sum of all estimators to a size limit, with the aim of shrinking irrelevant variable parameters down to insignificance.

Naturally, there have been multiple variations of the shrinkage estimator developed. One of the most well known being the prior mentioned LASSO, which penalises estimators to an absolute size limit, introduced by

Tibshirani (1996). Other models proposed include the elastic net, first proposed by Zou and Hastie (2005), which uses both the shrinkage and LASSO estimator in conjunction with differing weights in order to make use of both of their strengths in evaluating estimators. Being some of the most popular and most basic shrinkage methods, an examination of these processes will both provide strong contributions to the literature by showing how the absolute size limit affects the produced estimators. Moreover, examination of these methods will provide a good base for comparison to more exotic, less used shrinkage estimators.

This paper aims to both contribute to the machine learning and the econometric literature by conducting simulation studies in order to get a better grasp of the advantages and disadvantages that shrinkage based methods can provide over their more traditional counterparts such as Logistical Regression in a non-linear environment.

3 Experiment Design

In order to test the ability of shrinkage against their more traditional counterparts, Monte Carlo simulation studies of a sparse multinomial model with three classes with 250 repetitions were conducted to pit the performance of LASSO and the+-lastic Net against Logistical Regression, referred to henceforth as Logit regression. Moreover, for each repetition, differing sizes of samples will be regressed to evaluate how well these regression methods perform at low to high sample sizes. Sample sizes tested will be in order, 500, 1000, 2000 and 4000. The process of generating the simulation data will be as follows:

3.1 Data Generation

Ten explanatory variables will be randomly generated for this Monte Carlo study numbered $x1$ through $x10$, using the means and standard deviations shown in the table below.

Variables	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Means	0	0	0	0	0	0	0	0	0	0
Std Dev	2	2.5	1.5	2	1.25	3	5	2.75	4	1

Figure 1. Means and Standard deviations of Explanatory Variables

Using these explanatory variables, a dependent variable y with values of 1, 2 or 3 which will dictate its class will be generated with the following binary regressions as a base:

$$Pr(Y_i = K) = \frac{1}{1+\sum_{k=1}^{K-1} e^{\beta_k * X_i}} \quad (1)$$

$$Pr(Y_i = k) = \frac{e^{\beta_k * X_i}}{1+\sum_{k=1}^{K-1} e^{\beta_k * X_i}}, k < K \quad (2)$$

Wherein K, the number of classes present, equals 3, $Pr(Y_i = 1) = Pr(Y_i = 3)$ = 0.2 and $Pr(Y_i = 2)$ = 0.6, or in plain terms, on average, there will be a 20% chance of y being class 1 or 3 and a 60% chance of it being class 2. The coefficients for the explanatory variables are as follows:

True Parameters	β1	β2	β3	β4	β5	β6	β7	β8	β9	β10
Class 1	0.2	-0.2	0.005	0	0	0	0	0	0	0
Class 2	-0.2	0.005	0.2	0	0	0	0	0	0	0
Class 3 (Base Class)	0	0	0	0	0	0	0	0	0	0

Figure 2. Coefficient Values with respect to their explanatory variables and their classes.

Note that in order to create a sparse model environment, a majority of the variables are unrelated to determining the determinant variables class, and for the purposes of regression, Class 3 will be deemed the 'base' class, with coefficient values of 0 across the board. Moreover, in order to evaluate the regression method's sensitivity, differing coefficient values will be tested. As can be seen in Figure 2, 4 of the relevant variables will have a strong effect, with a coefficient value of 0.2 and 2 of the relevant variables will have a 'weak' effect, with coefficient values of 0.005

3.2 Evaluation

Two metrics will be used to compare the ability of the regression methods. The first is selection consistency, or the percentage number of times that a method correctly marked a relevant variable as significant or an

irrelevant one as insignificant. Selection consistency will be demonstrated as a %, out of the 250 repetitions that a regression method marked a variable's coefficient as significant. In regards to shrinkage based methods, selection will be defined by the coefficient being assigned a non-zero value. For Logit, selection will be determined through a p-value less than 0.05, worked out through a Wald test. The second metric used will be coefficient variance or how consistent the values of the coefficients are assigned by the regression method - for both shrinkage and Logit, this will be shown by a histogram showing the spread of predicted coefficient values over the 250 repetitions.

4 Results of Monte Carlo Simulation

4.1 Selection Consistency

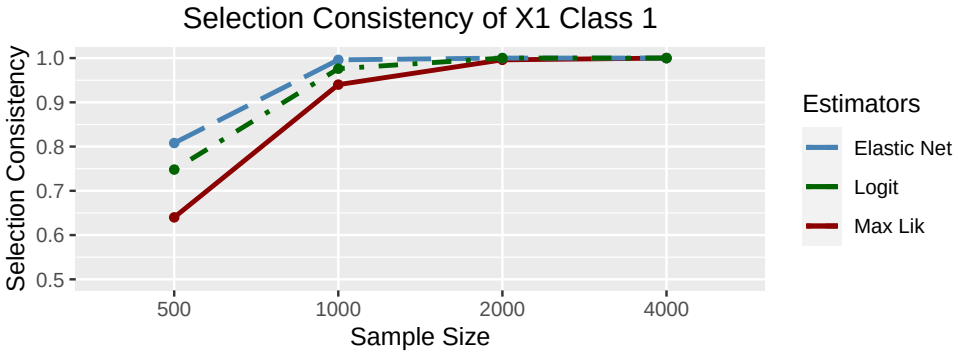

Figure 3. Selection consistency for X1 in relation to Class 1

As can be seen above, Figure 3 demonstrates the method's selection consistency for one of the "strong" coefficients, X1 in relation to class 1. At the smallest sample size of 500, the Elastic Net performs the best, followed by the Logit and the LASSO (marked on this Figure and the following Figures as Maximum Likelihood) Accordingly, all methods improve with increasing sample sizes, and once the second largest sample size of 2000 is reached, all methods accurately assign relevance to X01 100% of the time.

Figure 4. Selection consistency for X1 in relation to Class 2

Following up, Figure 4 displays the selection consistency results for another X1 in relation to class 2 instead, another one of the "strong" coefficients. Unlike the results of Figure 3, LASSO outperforms Logit at the smallest size - demonstrating that Logit's superiority to the LASSO in detecting theser variables are not guarenteed, however, Elastic Net again, like in Figure 3 outperforms both at all sample sizes.

Overall, these results demonstrate that, in terms of detecting "strong" coefficients Elastic Net consistently outperforms the other two methods, especially at small sample sizes, while LASSO and Logit's ability may vary.

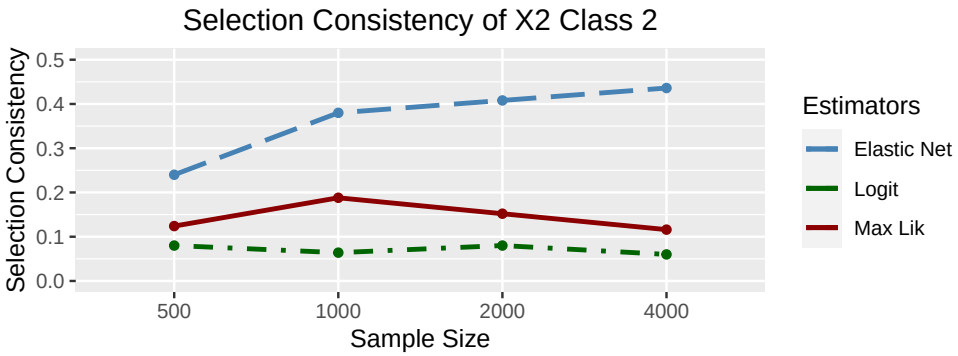

Figure 5. Selection consistency for X2 in relation to Class 2

Comparatively, Figure 5 displays selection consistency results for one of the 'weak' coefficients, X02 in relation to class 2. Like in the previous Figures, the Elastic Net demonstrates the best performance, outperforming the other methods when marking X03 as relevant at all sample sizes and additionally improving with higher sample sizes. Comparatively, while LASSO does also outperform Logit, its selection consistency actually peaks with the sample size of 1000 and begins to decrease in accuracy, demonstrating some irregularity in behaviour, which is also reflected in Logit, which also seems to decrease in efficacy with the higher sample sizes.

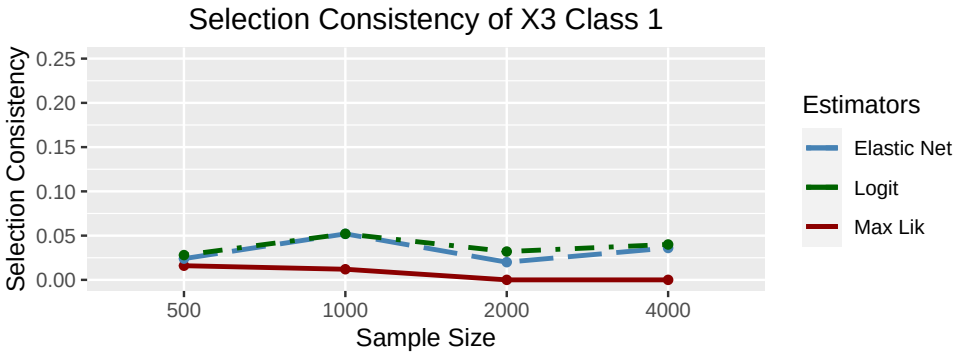

Figure 6. Selection consistency for X3 in relation to Class 1

In contrast to the result of Figure 5, LASSO performs the worst out of the three methods, hardly even marking X3 in relation to class 1 as relevant, while the performance of the Elastic Net and Logit varies by sample size, neither being able to clearly be called superior in this case.

Over, unlike with the strong coefficients, none of the tested regression methods can be clearly marked as 'definitively' better compared to the others when selecting weakly relevant factors. That being said, LASSO performs the worst out of the three methods in this field consistently.

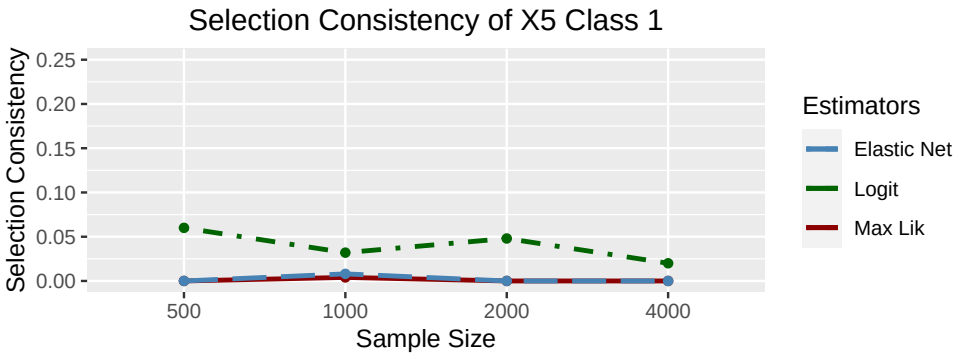

Figure 7. Selection consistency for X5 in relation to Class 1

Finally, Figure 7 demonstrates the selection consistency of an irrelevant variable - in this case X05 in regards to class 1 - meaning the less it is selected by the relevant methods the better.

In this regard, both shrinkage based methods outperform Logit at all sample sizes, consistently marking it as irrelevant. Despite initial poor performance Logit gradually improves as the sample size increases however it never matches the ability of LASSO or Elastic Net, demonstrating shrinkage's superior ability in this regard.

4.2 Estimator Consistency

Figure 8. Estimator Consistency for X1 in relation to Class 1

Figure 8 displays the estimator variance of the three methods of x1 in relation to Class 1. There appears to not much difference in the variance in predicted values for the coefficient value, with the main difference being, as expected, the shrinkage-based methods having a bias in value compared to the true value of 0.2 due to the nature of shrinkage.

5 CONCLUSION

With the aim of improving the understanding of how machine learning techniques, this paper performed a Monte Carlo simulation of a sparse multinomial environment to compare the performance of shrinkage based techniques such as the Least Absolute Shrinkage and Selection Operator or LASSO, and the Elastic Net against the more traditional Logit Regression.

Overall, the results of this paper show that when dealing with sparse multinomial environments, of the methods tested, the Elastic Net provides the clearest advantage in selection consistency, both being able to mark strongly relevant variables as strongly relevant more often than LASSO and Logit, but also managing to avoid marking irrelevant variables as relevant. Moreover, when compared to its peers Elastic Net holds the greatest advantage when dealing with datasets with low samples. However, this advantage is not universal, as it holds a natural bias when assigning values to coefficients and its ability to pick out weakly relevant variables is comparable to that of Logit.

REFERENCES

Athey, S. and G. W. Imbens (2017, may). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Athey, S. and G. W. Imbens (2019). Machine Learning Methods That Economists Should Know about. *Annual Review of Economics 11*, 685–725.

Breiman, L. (2001). Random Forests. *Machine learning 45*, 5–32.

Caner, M., E. Maasoumi, and J. A. Riquelme (2016). Moment and IV Selection Approaches: A Comparative Simulation Study. *Econometric Reviews 35*(8-10), 1562–1581.

Chen, J. M. (2021). An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, 1–16.

Gentzkow, M., B. Kelly, and M. Taddy (2017, mar). Text as Data. Technical report, National Bureau of Economic Research, Cambridge, MA.

Hoerl, A. E. and R. W. Kennard (1970, feb). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics 12*(1), 55.

Kuan, C. M. and H. White (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews 13*(1), 1–91.

McCulloch, W. S. and W. Pitts (1943, dec). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics 5*(4), 115–133.

Mullainathan, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Shi, Z. (2016). Estimation of Sparse Structural Parameters with Many Endogenous Variables. *Econometric Reviews 35*(8-10), 1582–1608.

Tibshirani, R. (1996, jan). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Varian, H. R. (2014, may). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Wilkinson, L. and G. E. Dallal (1981, nov). Tests of Significance in Forward Selection Regression With an F -to-Enter Stopping Rule. *Technometrics 23*(4), 377–380.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 67*(2), 301–320.