Anchoring-and-Adjustment to Improve the Quality of Significant Features

Eunkyung Park[1(✉)], Raymond K. Wong[1], Junbum Kwon[2], and Victor W. Chu[3]

[1] Computer Science and Engineering, University of New South Wales, Sydney, Australia
{eunkyung.park,ray.wong}@unsw.edu.au
[2] School of Marketing, University of New South Wales, Sydney, Australia
junbum.kwon@unsw.edu.au
[3] Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
wchu@ntu.edu.sg

Abstract. There is an enormous demand for Explainable Artificial Intelligence to obtain human-understandable models. For example, advertisers are keen to understand what makes video ads successful. In our investigation, we have analysed heterogeneous visual, auditory, and textual content features from YouTube video ads. This paper proposes a two-stage anchoring-and-adjustment approach. In the first stage, we search for the optimum penalized value in the regularization path of Lasso that maximizes the number of Significant Features (SFs). After that, we improve the quality of SFs by dropping features with high Variance-Inflation-Factor (VIF) because high VIF often makes a spurious set of SFs. Experiments show that, compared to the one-stage approach without the adjustment stage, our proposed two-stage approach results in a smaller number of SFs but a higher ability to identify true features that appeal to ad viewers from human evaluation. Furthermore, our approach can identify much more SFs while maintaining similar prediction accuracy as what Lasso and Elastic-net can obtain.

Keywords: Explainable models · Content features · Anchoring-and-adjustment · Variance-inflation-factor · Significance test

1 Introduction

Recently, there is an enormous demand for Explainable Artificial Intelligence (XAI) methods [4], as decision-makers often need to fully understand what factors drive the outcomes in many practical applications.

This paper proposes a two-stage anchoring-and-adjustment approach to improve the quality of Significant Features (SFs). In the first stage, we maximize the number of the candidate of SFs by adopting SFLasso [10] and SFLasso-SI (Selective Inference) [9] that search for the optimum penalized value λ in the

W. Zhang et al. (Eds.): WISE 2021, LNCS 13080, pp. 189–197, 2021.
https://doi.org/10.1007/978-3-030-90888-1_15

regularization path of Lasso [14] that maximizes the number of SFs. Considering two opposing factors: model size (i.e., the number of active variables) and the correlations among the active variables, SFLasso-SI chooses λ, generating the biggest number of SFs via statistical significance test using SI [7,13] in training data. After that, we further improve the quality of SFs by dropping features with high Variance-Inflation-Factor (VIF), which measures the amount of correlation with other features, because high VIF can inflate either the magnitude of coefficients or its variance, and thus often makes a spurious set of SFs.

Post-hoc explainability approach typically runs complex DNN first for high prediction accuracy and then runs the post model to explain the first model's prediction by selecting input features (DeepLIFT [11], L2X [2], and ACD [12]). Given that our data have small observations, which would not be enough to train DNN, we extend Lasso, an intrinsic explainable model. This allows us to do a statistical test for selecting features using recently developed selective inference.

Most Lasso variants focus on improving prediction accuracy, such as Elastic-net[16], and Enumerate Lasso [3] or finding more or better active variables, but not finding more SFs. Recently, SFLasso [10], and SFLasso-SI [9] was developed to find the maximum number of SFs.

OLS post-Lasso [1] proposed to rerun OLS with the active variables resulted from Lasso. However, this naive approach results in many false SFs [5,7,13]. Covariance Test [8] pioneered for significance test after variable selection when signal variables are not too correlated with noise variables [5]. For more general explanatory variables, Lee et al. [7] derived closed-form p-values for selected active variables after fitting Lasso with a fixed value of hyperparameter λ. While this selective inference can exclude some false SFs, we screen such false SFs using VIF iteration to drop highly correlated features with other explanatory variables.

Experiments using YouTube video ads show that, compared to the one-stage approach without adjustment stage (SFLasso-SI), our proposed two-stage approach (SFLasso-SI+VIF) results in a smaller number of SFs but much higher accuracy in identifying true features that appeal to ad viewers from human evaluation. Furthermore, our approach can identify more SFs while maintaining similar prediction accuracy as Lasso and Elastic-net can obtain.

2 Proposed Two-Stage Anchoring-and-Adjustment Approach

Anchoring-and-Adjustment heuristic is one of the strategies to estimate unknown quantities starting with information one does know and then adjust until an acceptable value is reached [15]. To find many true features, we propose a two-stage anchoring-and-adjustment method in the framework of the Explainability maximized method.

2.1 Anchoring Stage Using Explainability Maximized Method

Lasso penalizes non-zero coefficients by adding a regularization term λ in the objective function of OLS as follows: $\min_\beta \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1$, where

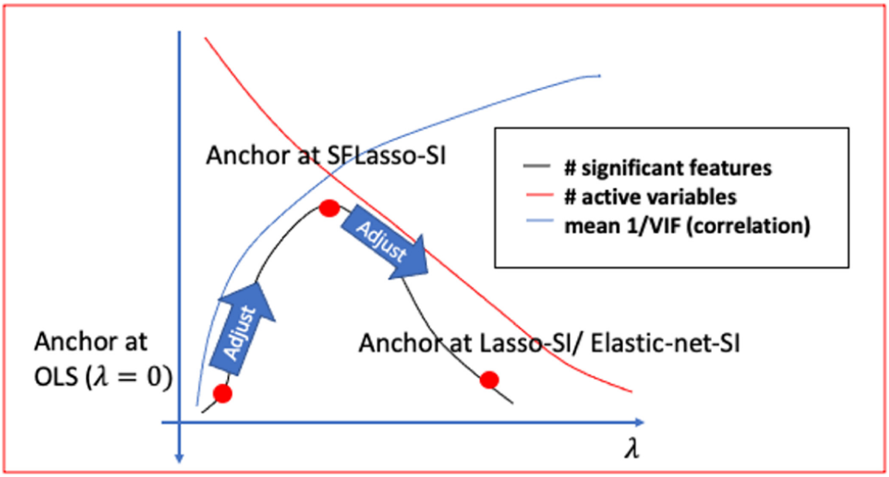

Fig. 1. Visualization of anchoring and adjustment

$X \in \mathbb{R}^{n \times p}$ (e.g., visual objects, text), $y \in \mathbb{R}^n$ (e.g., the number of likes of YouTube video ads), and $\beta \in \mathbb{R}^p$ is a vector of regression coefficient. Figure 1 shows that as λ decreases, the number of active variables increases, but the correlation among the active variables also increases (i.e., the mean of the inverse of VIF decreases). This relationship suggests that too big or too small values of λ are not good at identifying SFs. When λ is very big (e.g., Lasso, Elastic-net), only a small number of active variables is generated, so multicollinearity is not likely to happen. However, only a few active variables result in even fewer SFs. On the other hand, when λ is very small (e.g., $\lambda = 0$ for OLS), there are many active variables, so the correlation among active variables is high. This multicollinearity reduces the number of SFs.

Considering those two competing forces, one can expect that the maximum number of SFs is likely to occur at some point of λ where there is a good enough number of active variables while the correlation among x variables is not too high. Based on the above behaviors, SFLasso and SFLasso-SI search for the optimum penalized value in the regularization path of Lasso that maximizes the number of SFs as follows. $f(\lambda) = \max \sum_{i=1}^{p} I(p\text{-value}_{\beta_i} \leq 0.05)$, where $I = 1$ if $p\text{-value}_{\beta_i} \leq 0.05$ or $I = 0$ otherwise. For a given λ, the SFLasso selects active variables A with non-zero coefficients that minimize the objective function. For the testing of each selected variable's significance, while SFLasso runs OLS with the selected active variables and uses p-value, SFLasso-SI applies the selective inference [7,13]. The selective test does more conservative test by making false SFs insignificant [5,7,13]. Then, the number of SFs is counted. This process is repeated within a range of λ values. Finally, the λ that generates the maximum number of SFs is chosen.

2.2 Adjustment Stage

While SFLasso-SI finds many SFs, SFLasso-SI might include false SFs because it tends to choose features with an inflated magnitude of the coefficient to pass the significance test. James et al. [6] suggested that VIF that exceeds 10 indicates a problematic amount of multicollinearity. To adjust the solution of SFLasso-SI, we drop the feature with the highest VIF in the active variable set if the VIF is higher than 10. Note that we do not drop all the features with VIF greater than 10 at one time because the VIF of a focal variable is affected by the other variables. Usually, once a variable with very high VIF is dropped, VIFs for the remaining variables become much smaller. Then, we drop another feature with the highest VIF among the remaining x variables. We repeat this VIF iteration until the highest VIF becomes less than 10.

Figure 1 illustrates the adjustments from the three anchoring points from OLS, SFLasso-SI, and Lasso-SI. As OLS has many highly correlated variables, dropping the most correlated variables by VIF iteration reduces the variance of the coefficient, resulting in increases in the number of SFs. However, for SFLasso-SI, which is likely to generate false SFs with inflated coefficients, VIF iteration is likely to reduce the number of SFs. This reduction is the evidence to show that the correlation among active x variables of SFLasso-SI contributes to the inflation of magnitude of the coefficient in the numerator more than that of its variance in the denominator in the t-statistics, leading to some false SFs being included in the set of SFs. Lastly, Lasso-SI often has a few SFs due to the small number of active variables, making a low correlation among active x variables. Therefore, the highest VIF is often smaller than the threshold (10), resulting in no adjustment.

3 Empirical Evaluation

We conduct experiments using YouTube video ads. We divide our datasets into the training (70%), validation (15%), and test (15%) sets (Table 1). We extract visual objects, speech tones, and spoken words. We use a bag-of-words model to count word frequency. We include features with occurrences of more than 5 across video ads. The first block in Table 2 shows the result of World Vision US YouTube videos. The first 4 models run only the first anchoring stage without VIF iteration. Among these 4 models, SFLasso-SI identifies the biggest number

Table 1. Size of training, variables, and unmatched ratio

	# train (n)	# vars (p)	Unmatched ratio (%)
World Vision US	370	1,629	21.42
World Vision CA	318	1,374	25.98

(68) of SFs. The next 4 models do the second adjustment stage by doing the VIF iterations. As expected, our proposed SFLasso-SI+VIF generates a smaller number of SFs than SFLasso-SI but the biggest number of SFs among all other competing models that go through the VIF iteration. Furthermore, SFLasso-SI+VIF shows a similar prediction accuracy (RMSE) level with other predictability maximized methods such as Lasso+VIF. Table 3 shows that our SFLasso-SI+VIF outperforms all the other models in the F1 score. In particular, compared to SFLasso-SI, adding VIF iteration improve precision substantially from 26 to 69 for the US and from 28 to 63 in Canada, while recall increases only slightly in Canada. The recall is low across all models, which means that identifying truly appealing features among video content features is a challenging problem. Nevertheless, the VIF iteration helps improve the quality of identified SFs (i.e., higher precision).

Table 2. Results from models - World Vision US and Canada

		λ	α	RMSE (train)	RMSE (test)	VIF Mean	VIF Max	#act	#sig
US	OLS	0.0		0.57	2840.62	1374.35	106505.60	368	0
	Lasso-SI	37.9		44.40	202.95	1.27	1.66	4	2
	Elastic-SI	285.8	0.1	44.04	204.13	1.69	4.00	6	1
	SFLasso-SI	0.6		20.77	209.29	3.78	30.15	151	68
	OLS+VIF			23.65	219.56	5.50	9.79	237	16
	Lasso-SI+VIF			44.40	202.95	1.27	1.66	4	2
	Ela-SI+VIF			44.04	204.13	1.69	4.00	6	1
	SF-SI+VIF			20.99	207.47	3.35	9.57	144	26
CA	OLS	0.0		2.20	3622.45	664.79	139282.80	316	0
	Lasso-SI	140.7		148.82	277.41	1.13	1.17	3	1
	Elastic-SI	1406.6	0.1	148.82	277.41	1.13	1.17	3	0
	SFLasso-SI	2.7		87.85	302.01	2.96	18.28	107	58
	OLS+VIF			58.29	468.40	5.66	9.58	218	15
	Lasso-SI+VIF			148.82	277.41	1.13	1.17	3	1
	Ela-SI+VIF			148.82	277.41	1.13	1.17	3	0
	SF-SI+VIF			65.85	295.85	2.70	9.64	102	27

OLS identifies zero SF. Through the VIF iteration (threshold 10), many highly correlated variables are dropped from 368 to 237. As a result, 16 SFs are identified. This VIF process confirms that multicollinearity hides SFs. On the contrary, Lasso-SI and Elastic-SI drop many variables due to the mismatch between train and validation data. As a result, only 4 (Lasso-SI) and 6 (Elastic-SI) features are active, and only 2 (Lasso-SI) and 1 (Elastic-SI) features have significance. Since VIF is already low enough (i.e., smaller than threshold 10), there are no additional variables to be dropped with VIF criteria. Therefore, Lasso-SI+VIF and Elastic-SI+VIF do not gain additional SFs. As discussed earlier, SFLasso-SI finds the biggest number (68) of SFs by trading off the reduction in the amount of correlation against the size of active features. Specifically, SFLasso-SI still keeps 151 active variables after dropping many correlated vari-

Table 3. Results from human evaluation

	US			Canada		
	Precision	Recall	F1	Precision	Recall	F1
OLS	0	0	0	0	0	0
Lasso-SI	50.00	0.19	0.37	100.00	0.20	0.39
Elastic-SI	0.00	0.00	0.00	0.00	0.00	0.00
SFLasso-SI	26.47	3.36	5.96	27.59	3.16	5.66
OLS+VIF	43.75	1.31	2.54	53.33	1.58	3.07
Lasso-SI+VIF	50.00	0.19	0.37	100.00	0.20	0.39
Ela-SI+VIF	0.00	0.00	0.00	0.00	0.00	0.00
SFLasso-SI+VIF	69.23	3.36	6.41	62.96	3.35	6.37

ables. As a result, its max VIF is about 30, which substantially reduces OLS (106,505), while it keeps much more active variables than Lasso-SI (151 vs. 4).

Table 4. Significant features - World Vision US

OLS (0/0)	
Lasso-SI (1/2)	⟨visual objects⟩mammal ⟨spoken words⟩**water**
Elastic-SI (0/1)	⟨visual objects⟩mammal
SFLasso-SI (18/68)	⟨visual objects⟩indoor man sky **tree grass boy** woman girl people **smiling** field road mountain beach plant athleticGame bed house orange **eating** red room snow suit militaryUniform old box chicken tattoo ⟨speech tones⟩joy ⟨spoken words⟩**action** address **area believe** bone call **continue** doctor Donna famine **gift happen** heart information involve **kid leader** leave letter life local **make** people poverty problem program reach really **right** sense sponsor **stand stop** teacher **vital** water wonderful worker
OLS+VIF (7/16)	⟨visual objects⟩**tree animal little** ⟨spoken words⟩**hope leave end** challenge man teacher sit cool plan follow ready malaria **development**
Lasso-SI+VIF (1/2)	⟨visual objects ⟩ mammal ⟨spoken words⟩ **water**
Elastic-SI+VIF (0/1)	⟨visual objects⟩mammal
SFLasso-SI+VIF (18/26)	⟨visual objects⟩**ground tree rollingcredits boy little** window **smiling** playing **eating** ⟨spoken words⟩ **action day god happen** information **kid leader** main **make** plan problem **right** Sabina **stand stop vital** wonderful

More importantly, the max VIF of SFLasso-SI is still relatively high compared to the suggested threshold ($30 > 10$). After several VIF iterations until its threshold, our proposed SFLasso-SI+VIF obtains only 26 SFs among 144 active variables. Surprisingly, more than half of the initial SFs are gone. This is because SFLasso-SI favours the inflated magnitudes of the coefficient to maximize the number of SFs. This VIF iteration substantially improves the identified SFs' quality by cutting off the false SFs with such inflated coefficients.

Our SFLasso-SI+VIF identifies much more SFs than OLS+VIF (26 vs. 16), although SFLasso-SI+VIF uses a smaller number of active variables than OLS+VIF (144 vs. 237). Note that the max VIFs in both are lower than 10. This surprising result can be explained by the difference in the initial anchoring set of active variables. While OLS+VIF starts with all the 368 variables, SFLasso-SI+VIF does with 144 active variables from SFLasso-SI. The correlation structure among variables is complex. VIF measures the total redundancy of a focal x variable with the rest of the active variables. Depending on the set of active variables, VIF iteration can journey very different paths and result in very different final sets of active variables and SFs. This result emphasizes (1) the important role of the first anchoring stage to identify many candidates of SFs, and (2) the second adjustment stage to drop false SFs via VIF iteration. Table 4 shows the SFs from each model. Although SFLasso-SI+VIF identifies smaller SFs than SFLasso-SI (26 vs. 68), both models include the same number (18) of true features. In other words, SFLasso-SI+VIF has a higher quality (i.e., precision) of SFs (69 vs. 26).

Table 5. Deleted false SFs and added true SFs by VIF iteration - World Vision US

False significant features (48)	True significant features (5)
⟨visual objects⟩indoor man sky woman girl people field road mountain beach plant athleticGame bed house orange red room snow suit militaryUniform old box chicken tattoo ⟨speech tones⟩joy ⟨spoken words⟩address bone call doctor Donna famine heart information involve leave letter life local people poverty program reach really sense sponsor teacher water worker	⟨visual objects⟩ground rollingcredits little ⟨spoken words⟩day god

The next question is how SFLasso-SI+VIF can achieve higher accuracy in hitting truly appealing features to ad viewers than SFLasso-SI, although SFLasso-SI+VIF identifies a smaller number of SFs than SFLasso-SI. As discussed above, SFLasso-SI tends to have false SFs with inflated coefficients. Through the VIF iteration, 48 false SFs are excluded, as seen in Table 5. The visual objects 'indoor', 'man', and 'sky' are examples. As a result, precision increases. Furthermore, recall that high VIF can increase the coefficient variance, leading to the insignificance of a focal feature. In other words, high VIF can hide true features. Hence, the VIF process help reveal true features. The visual objects 'ground', 'little', and 'rolling credits' and the spoken words 'day' and 'god' are those features. These additions improve recall as well as precision. In summary, the second adjustment stage via the VIF iterations increases identified SFs by excluding false SFs and adding new true features. The results on World Vision Canada are similar to those on World Vision US.

4 Conclusion

In this paper, we propose a two-stage anchoring-and-adjustment approach. In the first stage, we adopt the recently developed SFLasso-SI to find many candidates of SFs. After then, through the VIF iterations, we adjust SFs by dropping false SFs and adding true SFs. Human evaluations show that our proposed two-stage approach achieves higher accuracy in identifying true features than SFLasso-SI without the VIF iterations. Furthermore, our approach maintains similar prediction accuracy as what Lasso and Elastic-net can obtain.

Acknowledgements. This research is supported by the Australian Government Research Training Program Scholarship.

References

1. Belloni, A., Chernozhukov, V.: Least squares after model selection in high-dimensional sparse models. In: Bernoulli, vol. 19, pp. 521–547 (2013)
2. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: Learning to explain: an information-theoretic perspective on model interpretation. In: ICML, vol. 80, pp. 882–891 (2018)
3. Hara, S., Maehara, T.: Enumerate lasso solutions for feature selection. In: AAAI, pp. 1985–1991 (2017)
4. Harder, F., Bauer, M., Park, M.: Interpretable and differentially private predictions. In: AAAI, pp. 4083–4090 (2020)
5. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations (2015)
6. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: With Applications in R. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-7138-7
7. Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E.: Exact post-selection inference, with application to the lasso. Ann. Stat. **44**(3), 907–927 (2016)
8. Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R.: A significance test for the lasso. Ann. Stat. **42**(2), 413–468 (2014)
9. Park, E., Wong, R.K., Kwon, J., Chu, V.W.: Maximizing explainability with sf-lasso and selective inference for video and picture ads. In: Advances in Knowledge Discovery and Data Mining, pp. 566–577 (2021)
10. Park, E., Wong, R.K., Kwon, J., Chu, V.W., Rutz., O.J.: Video ads content analysis using significant features lasso. In: The 43rd ISMS Marketing Science Conference (2021)
11. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: ICML, vol. 70, pp. 3145–3153 (2017)
12. Singh, C., Murdoch, W.J., Yu, B.: Hierarchical interpretations for neural network predictions. In: ICLR (2019)
13. Taylor, J., Tibshirani, R.: Post-selection inference for $l1$-penalized likelihood models. Can. J. Stat. **46**(1), 41–61 (2018)
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)

15. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. Science **185**(4157), 1124–1131 (1974)
16. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B **67**, 301–320 (2005)