Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n2p236

Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty
By Algamal, Lee

Published: 14 October 2015

Electronic Journal of Applied Statistical Analysis
Vol. 08, Issue 02, October 2015, 236-245
DOI: 10.1285/i20705948v8n2p236

Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty

Zakariya Yahya Algamal[a] and Muhammad Hisyam Lee*[a]

[a]*Department of Mathematical Sciences, Universiti Teknologi Malaysia , 81310 Skudai, Johor, Malaysia*

Published: 14 October 2015

Variable selection in count data using penalized Poisson regression is one of the challenges in applying Poisson regression model when the explanatory variables are correlated. To tackle both estimate the coefficients and perform variable selection simultaneously, elastic net penalty was successfully applied in Poisson regression. However, elastic net has two major limitations. First it does not encouraging grouping effects when there is no large correlation. Second, it is not consistent in variable selection. To address these issues, a modification of the elastic net (AEN) and its adaptive modified elastic net (AAEM), are proposed to take into account the weak and mild correlation between explanatory variables and to provide the consistency of the variable selection simultaneously. Our simulation and real data results show that AEN and AAEN have advantage with weak, mild, and extremely correlated variables in terms of both prediction and variable selection consistency comparing with other existing penalized methods.

keywords: high dimensional, penalization, Poisson regression, LASSO, elastic net.

1 Introduction

With the advancement of technologies, massive amount of data with increasing dimensions have been generated in many areas such as genetics, medical, economic and social

*Corresponding author: mhl@utm.my

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

sciences. The expansion of the data is in two dimensions: the number of variables and the number of observations. High dimensional data refer to the situation where the number of variables measured is greater as the number of observations in the data. This differs from traditional datasets for statistical analysis where we have many observations on a few variables. Such high dimensional data has posed new challenges to statistical analysis, because a lot of classically statistical methods do not automatically apply into these datasets, for example, the curse of dimensionality makes many classical regression models, such as Poisson regression, ineffective, because statistical issues associated with modeling high dimensional data include model overfitting, estimation instability, computational difficulty (Pourahmadi, 2013).

How to reduce the dimensionality has been an important research question in statistical application. One way to handle the high dimensional data is to perform data reduction. To do this, various penalized methods have been proposed begin by ridge penalty (Hoerl and Kennard, 1970). It estimates the regression coefficients through L1-norm penalty. It is well-known that ridge regression shrinks the coefficients of correlated predictor variables toward each other, allowing them to borrow strength from each other (Friedman et al., 2010). The least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996) to estimate the regression coefficients through L1-norm penalty. Zou and Hastie (2005) proposed the elastic net penalty which is based on a combined penalty of LASSO and ridge regression penalties in order to overcome the drawbacks of using the LASSO and ridge regression on their own.

Usually, in high dimensional data the explanatory variables are correlated. If there is a group of highly correlated variables, the LASSO will randomly select only one variable from this group and drop the rest whereas elastic net will select the whole group of the highly correlated explanatory variables (Zou and Hastie, 2005; Zhou, 2013). Analogously, Bondell and Reich (2008) proposed a penalty called OSCAR to encourage selection of a group of highly correlated explanatory variables. Elastic net often performs better than LASSO in terms of prediction error when there is correlation among variables, also OSCAR has a comparable performance similar to elastic net (Zeng and Xie, 2011). Tutz and Ulbricht (2009) proposed correlation-based penalty to deal with grouping effects. This penalty just makes variable shrinkage rather than variable selection. Elastic net penalty lacks consistent variable selection (oracle property) even though it outperforms LASSO. Zou and Zhang (2009) proposed adaptive elastic net to handle grouping effects and enjoying oracle property simultaneously. El Anbari and Mkhadri (2014) explained though experimental studies that elastic net seems to be slightly less reliable if the correlation between explanatory variables is not so extreme (i.e. $|\rho| \leq 0.95$).

In this paper, an adjusted of the elastic net (AEN) and its adaptive adjusted elastic net (AAEM), are proposed to take into account the small and medium correlation between explanatory variables and to provide the consistency of the variable selection simultaneously. The remainder of this paper organizes as follows. Section 2 covers the penalized Poisson regression methods. Description of the AEN and AAEM is explained in section 3. Sections 4 and 5 are devoted to simulation study and results. While section 6 covered the real data analysis. We end this paper with a conclusion in section 7.

2 Penalized Poisson Regression Model

Poisson regression models have received much attention in econometrics and medicine literature as model for describing count data that assume integer values corresponding to the number of events occurring in a given interval. The Poisson regression model is the most basic model, where the mean of the distribution is a function of the explanatory variables. This model has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance (Algamal, 2012; Algamal and Lee, 2015). A procedure called penalization, which is always used in variables selection in high dimensional data, attaches a penalty term $P_\lambda(\beta)$ to the log-likelihood function to get a better estimate of the prediction error by avoid overfitting. Recently, there is growing interest in applying the penalization method in the Poisson regression models. Friedman et al. (2010) developed an efficient algorithm for the estimation of a generalized linear model including Poisson regression with a convex penalty. Hossain and Ahmed (2012) proposed Stein-type shrinkage estimator for the parameters of Poisson regression model. Wang et al. (2014) proposed a combination of minimax concave and ridge penalties and a combination of smoothly clipped absolute deviation and ridge penalties.

In Poisson regression model, the number of events y_i has a Poisson distribution with a conditional mean that depends on individual characteristics according to the structural model.

$$f(y_i) = \frac{e^{-\theta_i}{\theta_i}^{y_i}}{y_i!}, \quad y_i = 0, 1, ...; \; i = 1, 2, ..., n \tag{1}$$

and the conditional mean parameter

$$\theta_i = \exp({x_i}'\beta) \tag{2}$$

Under the assumption of independent observations, the log-likelihood function is given by

$$\ell(\beta) = \sum_{i=1}^{n} \left\{ y_i {x_i}'\beta - \exp({x_i}'\beta) - \ln y_i! \right\} \tag{3}$$

The penalized Poisson regression (PPR) is defined as

$$PPR = \ell(\beta) + \lambda P(\beta) \tag{4}$$

where λ is defined as a tuning parameter ($\lambda \geq 0$). It controls the strength of shrinkage the explanatory variables, when λ takes larger value, more weight will be given to the penalty term. Since the value of λ is depends on the data, it can be computed using cross-validation method (Fan and Tang, 2013; James et al., 2013). Before solving the PPR, it is worth to make standardization to x_j, so that, $\frac{1}{n}\sum_{i=1}^{n} X_{ij} = 0$, and $\sum_{i=1}^{n} X_{ij}^2 = 1$ for $j = 1, 2, ..., k$. This is to make the intercept (β_0) equals zero.

The LASSO for the Poisson regression model was originally proposed by Park and Hastie (2007). This technique is in some sense similar to ridge regression but it can

shrink some coefficients to zero, and thus can implement variable selection. The LASSO method estimates the coefficients by minimizing the negative log-likelihood with the constraint that the sum of the absolute values of the model coefficients is bounded above by some positive number. The LASSO estimator is

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left(-\ell(\beta) + \lambda \sum_{j=1}^{k} |\beta_j| \right) \tag{5}$$

where $\lambda \geq 0$ is the tuning parameter. For large values of λ, Eq. (5) produces shrunken estimates of the β and sets some variables to equal zero.

The elastic net estimator which proposed by Zou and Hastie (2005) is a combination between the ridge and the lasso penalty. The second term (ridge penalty) encourages highly correlated variables to be averaged, while the first term (the LASSO penalty) encourages a sparse solution in the coefficients of these average variables. The elastic net estimator for Poisson regression model is

$$\hat{\beta}_{Elastic} = \arg\min_{\beta} \left(-\ell(\beta) + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} |\beta_j|^2 \right) \tag{6}$$

As we observe from Eq. (6), elastic net estimator is depended on non-negative two tuning parameters λ_1, λ_2 and leads to penalized Poisson regression solution. However, elastic net performs well when the pairwise correlations between variables are very high. El Anbari and Mkhadri (2014) stated that if the absolute correlation between genes is less than 0.95, elastic net may be slightly less reliable. Moreover, elastic net does not take into account the correlation structure among variables (Zhou, 2013). Additionally, it was pointed out by Zou and Zhang (2009) that the elastic net fails in terms of achieving oracle property, although the grouping effect problem for elastic net remains. As a result, adaptive elastic net was introduced by Zou and Zhang (2009) and Ghosh (2011), which it combines the L2-norm penalization with the adaptive LASSO.

3 Modified Elastic Net Penalty

In this section, we present our proposed modified method, AEN and AAEN, in Poisson regression model. The main idea behind AEN is to take into account the information about the empirical correlation of the data matrix in the L2-norm term because elastic net does not. Suppose without loss of generality that the explanatory variables are scaled, we define the AEN estimator as

$$\hat{\beta}_{AEN} = \arg\min_{\beta} \left\{ -\ell(\beta) + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \left[\sum_{j=1}^{k-1} \sum_{j+1>j} (\beta_j - r_{j,j+1}\beta_{j+1})^2 + {\beta_k}^2 \right] \right\} \tag{7}$$

where λ_1 and λ_2 are non-negative tuning parameters. $r_{j,j+1}$ is the correlation between j and p explanatory variables where $p > j$. The quantity $(\beta_j - r_{j,j+1}\beta_{j+1})^2$ is

helpful to make AEN reliable if the correlation between explanatory variables is not so extreme. The last term from Eq.(7) is greater than zero for any vector β. Therefore, $(r_{j,j+1})'(r_{j,j+1})$ represents a Choleskys decomposition. After suitable data argumentation, Eq. (7) is equivalent to a LASSO. The AEN was solved using coordinate descent optimization (Friedman et al., 2010) which a computationally efficient method for solving this type of convex optimization problem. The optimal AEN model was found by a grid search over the parameters λ_1 and λ_2.

Furthermore, the adaptive version of AEN, AAEN, is defined by

$$\hat{\beta}_{AAEN} = \arg\min_{\beta} \left\{ -\ell(\beta) + \lambda_1 \sum_{j=1}^{k} w_j \left|\beta_j\right| + \lambda_2 \left[\sum_{j=1}^{k-1} \sum_{j+1>j} \left(\beta_j - r_{j,j+1}\beta_{j+1}\right)^2 + {\beta_k}^2 \right] \right\} \tag{8}$$

where

$$w_j = (1/|\hat{\beta}_{j(AEN)}|)^{\gamma}, \quad j = 1, 2, ..., p \tag{9}$$

where $\gamma > 0$. For simplicity, $\gamma = 1$ was used for both simulation study and real data application.

4 Simulation Study

In this section, simulation studies are used to investigate the performance of the proposed AEN and AAEN. Furthermore, we compare AEN and AAEN with elastic net. In all simulations the response variable was generated from Poisson distribution with conditional mean θ_i. All simulation cases are replicate 50 times. For every simulation case and in each replication we generate training, validation, and testing data. The training data were used for model fitting. The validation data were used to determine the tuning parameters. The testing data were used to evaluate the penalization methods. For each case, the observation numbers of the corresponding data sets are denoted by training/validation/testing. Based on the simulated data, we used three metrics to evaluate all penalization methods which were studied in this paper, mean-squared errors for the test data ($\mathrm{MSE_t}$), *hits* which stands for the number of correctly identified true variables, and false positive (FP) which denotes to the number of zero variables which are wrongly considered as true variables.

Since we investigate a penalization method with both variable selection and grouping property, we use simulation scenario with different values of the correlation and different numbers of training, validation, and testing observations. Simulation Scenario: In this setting, we generate data sets with sample sizes 200/200/400 and 100 explanatory variables. Four cases are studied. The grouping effects were generated as follows

$$Group1 : x_i = w_1 + \varepsilon_i, w_1 \sim N(0,1), \quad i = 1, 2, 3$$

$$Group2 : x_i = w_2 + \varepsilon_i, w_2 \sim N(0,1), \quad i = 4, 5, 6$$

$$Group3 : x_i = w_3 + \varepsilon_i, w_3 \sim N(0,1), \quad i = 7,8,9$$

Furthermore, the noisy explanatory variables were generated as $x_i \sim N(0,1), \quad i = 10,11,....,100$.

Case A: In this case we set $\varepsilon_i \sim N(0,0.01), \quad i = 1,2,...,9$. The correlations among variables within each group are 0.98. The true variables parameters were $\beta = (\underbrace{0.3,....,0.3}_{9},\underbrace{0,....,0}_{91})$.

Case B: This simulation is like case A except that $\varepsilon_i \sim N(0,0.6), \quad i = 1,2,...,9$. Thus, there are correlations within each group around 0.7.

Case C: Similar to case A, we set $\varepsilon_i \sim N(0,0.8), \quad i = 1,2,...,9$ in order to get correlations within each group equal 0.5.

Case D: Similar to previous cases, in order to get correlations within each group equal 0.3. We assume that $\varepsilon_i \sim N(0,1.5), \quad i = 1,2,...,9$.

5 Simulation Results

To examine the performance of the AEN and AAEN penalties we compare it with elastic net. For the tuning parameters of elastic net, AEN, and AAEN, a prior value of is required to transform the original training data set to the new augmented training data set. A sequence of values for λ_2 is given, where $0 \leq \lambda_2 \leq 100$. The mean-squared error for the training data ($\text{MSE}_{\text{train}}$) is computed as the criterion of evaluation. Figure 1 displays the corresponding boxplots of the $\text{MSE}_{\text{train}}$ for the three used methods for the four cases. It is clearly seen that AEN and AAEN has less variability comparing with elastic net. Also, it can be seen that AEN and AAEN are slightly similar.

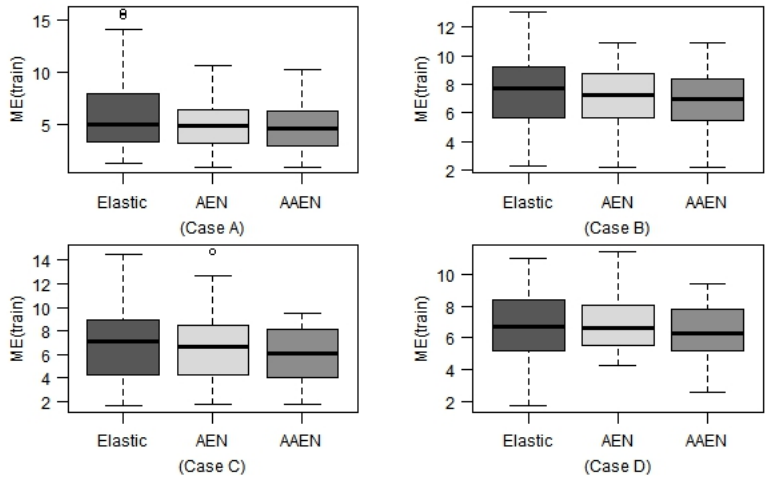

Figure 1: Comparison of median $\text{MSE}_{\text{train}}$ for three methods

Table 1 summarizes the median MSE_{t} and the standard deviation (Std. Dev.) of the median MSE_{t} which estimated by using bootstrap with $B = 100$ resampling on 50

MSE_t values. In addition, the median number of *hits* and FP are reported too. In each case, the bold font indicates the best method on MSE_t, Std. Dev., *hits*, and FP. Table 1 reveals that the AAEN method produces considerably smaller median MSE_t and standard deviation among all methods in all cases. For example, in case B the median MSE_t of AAEN is 6.967 with standard deviation equals to 2.521 which is smaller than 7.299 (2.615) and 7.509 (2.745) for AEN and elastic net methods respectively. Furthermore, the reduction of MSE_t is usually substantial compared to elastic net. For example, the reduction of AEN in case A, case B, case C, and case D is 0.67%, 1.60%, 5.01%, and 6.60% respectively. Moreover, in case A, there is high pairwise correlation among variables. Elastic net is supposed to have the best performance then AEN because elastic net deals with extremely highly correlations. In addition, our method performs well in terms of MSE_t when the correlation is small and medium. Besides, from the simulation results we can observe that elastic net came the last method.

Table 1: Comparison among methods over 50 replications for the four cases

	MSE_t(Std. Dev.)	Hits	False Positive	Length
$\rho = 0.98$				
Elastic net	4.931 (1.378)	4.5	14	18.5
AEN	4.887 (1.369)	9	13	22
AAEN	**4.811 (1.294)**	**9**	**13**	**22**
$\rho = 0.7$				
Elastic net	7.509 (2.745)	7	13	21
AEN	7.299 (2.615)	8	13	20
AAEN	**6.967 (2.521)**	**8**	**12**	**20**
$\rho = 0.5$				
Elastic net	6.874 (2.414)	6.5	12.5	19
AEN	6.787 (1.889)	7	11	18
AAEN	**6.483 (1.527)**	**7.5**	**10**	**17.5**
$\rho = 0.30$				
Elastic net	6.742 (2.637)	7	16	23
AEN	6.663 (2.152)	7	15	22
AAEN	**6.271 (2.113)**	**7.5**	**12.5**	**20**

For variable selection accuracy, the penalization methods should include all important variables (non-zero variables), *hits* and FP were used to measure the performance of AAEN, AEN, and elastic net in term of selecting the non-zero variables. From Table 1 both AAEN and AEN succeed in selecting the true non-zero variables in most of the

cases in term of *hits*. For example, AEN selects the all nine non-zero variables in case A. Moreover, when the correlation coefficient varies from small, medium, to extremely high correlation elastic net selects less non-zero variables comparing to AAEN and AEN. We can expect such a result because elastic has its limitation in biased selection. In term of FP, AAEN and AEN method usually selects less ineffective variables than elastic net in most cases. To this end, it is obvious from our simulation results that the AAEN and AEN methods perform better in term of MSE_t by obtaining smaller values, *hits*, and FP followed by elastic net for small, medium, and extremely high correlation and has greater advantage of variable selection with grouping effects in Poisson regression model.

6 Real Data Results

To evaluate our proposed method in the field of count data model, The real dataset which belong to the study of the distribution of freshwater mussels was taken from Sepkoski and Rex (1974). The study aims at the estimation of the numbers of species of mussels in 41 rivers in US by various explanatory variables. The nine explanatory variables are: area, number of stepping stones (intermediate rivers) to 4 major species-source river systems (Alabama-Coosa (AC), Apalachicola (AP), St. Lawrence (SL), and Savannah (SV)), nitrate concentration, hydronium concentration, and solid residue.

In order to enable a fair comparison, typically, the dataset was randomly partitioned into a training dataset, which comprised 70% of the samples, and a test dataset, which consisted of 30% of the samples. The partition repeated 50 times. In order to get the best value of the pair (λ_1, λ_2), the 10-fold CV was employed using the training dataset. All the applications were conducted in R using the *glmnet* package. Table 2 shows the median number of explanatory variables selected by each of the AAEN, AEN, and elastic net in the test data set, and the corresponding median MSE_t. It can be seen that AAEN performs best in term of prediction error where the MSE_t of the AAEN is approximately 0.66% lower than AEN and 5.37% lower than elastic net. Moreover, AAEN selects less explanatory variables than the other two methods.

Table 2: Comparison among methods for the real dataset

Methods	MSE_t	No. of selected variables
Elastic net	9.817	6
AEN	9.351	6
AAEN	**9.289**	**5**

7 Conclusion

A study of adjusted elastic net was proposed by applying on Poisson regression model. AAEN and AEN with elastic net were compared by using simulation studies and real data application. Both the simulation and real data results show that the AAEN and AEN are outperforming the elastic net in term of $\mathrm{MSE_t}$ of test data and variable selection accuracy. We can conclude that AAEN and AEN more reliable for grouping effects when there are broader ranges of correlation between variables in applying penalized Poisson regression model.

References

Algamal, Z. Y. (2012). Diagnostic in poisson regression models. *Electronic Journal of Applied Statistical Analysis*, 5(2):178–186.

Algamal, Z. Y. and Lee, M. H. (2015). Adjusted adaptive lasso in high-dimensional poisson regression model. *Modern Applied Science*, 9(4):170–177.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123.

El Anbari, M. and Mkhadri, A. (2014). Penalized regression combining the l1 norm and a correlation based penalty. *Sankhya B*, 76(1):82–102. Sankhya B.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22.

Ghosh, S. (2011). On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing*, 21(3):451–462.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hossain, S. and Ahmed, E. (2012). Shrinkage and penalty estimators of a poisson regression model. *Australian & New Zealand Journal of Statistics*, 54(3):359–373.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer, New York.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons, Hoboken, New Jersey.

Sepkoski, J. J. and Rex, M. A. (1974). Distribution of freshwater mussels: coastal rivers as biogeographic islands. *Systematic Biology*, 23(2):165–188.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, 19(3):239–253.

Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C.-Y., and Devarajan, P. (2014). Penalized count data regression with application to hospital stay after pediatric cardiac surgery. *Statistical Methods in Medical Research*.

Zeng, L. and Xie, J. (2011). Group variable selection for data with dependent structures. *Journal of Statistical Computation and Simulation*, 82(1):95–106.

Zhou, D. X. (2013). On grouping effect of elastic net. *Statistics & Probability Letters*, 83(9):2108–2112.

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

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4):1733–1751.