MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Elastic net constrained stereotype logit model for ordered categorical data

Abstract

Gene expression studies are of growing importance in the field of medicine. In fact, sub-types within the same disease have been shown to have differing gene expression profiles. Often, researchers are interested in differentiating a disease by a categorical classification indicative of disease progression. For example, it may be of interest to identify genes that are associated with progression and to accurately predict the state of progression using gene expression data. One challenge when modeling microarray gene expression data is that there are more genes (variables) than there are observations. In addition, the genes usually demonstrate a complex variance-covariance structure. Therefore, modeling a categorical variable reflecting disease progression using gene expression data presents the need for methods capable of handling an ordinal outcome in the presence of a high dimensional covariate space. We present a method that combines the stereotype regression model with an elastic net penalty as a method capable of modeling an ordinal outcome for high-throughput genomic data sets. Results from the application of the proposed method to gene expression data are reported and the effectiveness of the proposed method is discussed.

Keywords: stereotype logit, high dimensional, affymetrix, elastic net

Volume 2 Issue 7 - 2015

André AA Williams,[1] Kellie J Archer[2]
[1]College of Public Health, Temple University, USA
[2]Department of Biostatistics, Virginia Commonwealth University, USA

Correspondence: André AA Williams, College of Public Health, Temple University, 1301 Cecil B. Moore Avenue, Ritter Annex, Philadelphia, PA 19122, USA, Phone: 215-204-5534, Fax: 215-204-1854; Email andre.williams@temple.edu

Received: July 12, 2015 | **Published:** October 20, 2015

Introduction

In biomedical studies the outcome of interest may be an ordinal, rather than a dichotomous, class label such as progression of disease. An example of an ordinal variable is drug toxicity levels evaluated as mild, moderate or severe. Another example is the Breast Imaging Reporting and Data System (BI-RADS)[1] classification system. After a mammogram is read, a subjective score is assigned based on the condition of the breast tissue. These categories are: Category 0 - Incomplete; Category 1- Negative; Category 2 - Benign; Category 3 -Probably Benign; Category 4 - Suspicious Abnormality; Category 5 - Highly Suspicious of Malignancy; and Category 6 - Known Biopsy Proven Malignancy. The ordinality of these categories is evident. As another example, when cancer treatments are applied there is usually an interest in how patients respond. A typical way to measure this response is called the Revised Response Evaluation Criteria in Solid Tumors (RECIST).[2] Based on a wide variety of tools, as well as defined rules for classification, Revised RECIST defines the responses as: Complete Response, Partial Response, Stable Disease, and Progressive Disease. The types of models used to model ordinal data include the multinomial, adjacent category logit, continuation ratio logit, proportional odds logit, stereotype logit, and cumulative link models.[3] These models have the assumption, among others, that there are considerably more observations than variables. However, there are many types of data for which there are more variables than observations. When using microarray based, or other high throughout technologies, due to the expense of obtaining samples, there may be few observations but thousands of variables. The afore-mentioned models are not estimable by traditional means or without additional assumptions. Although there are data dimensionality reduction techniques, such as principle component analysis, due to the severely unbalanced nature of the data it may still be impossible to satisfactorily reduce the subset of variables to be less than the number of observational units without a significant loss of information in the data. This paper is concerned with the development of an ordinal classification model using the Least Absolute Shrinkage and Selection Operator (LASSO) and ridge penalizations to accommodate the case where there are considerably more variables than observations. This described procedure uses the stereotype logit model[4] with the applied penalty in an attempt to overcome said problems. The proposed method is applied to simulated data. An algorithm is presented in which the above penalized likelihood is utilized to model high dimensional data with an ordinal outcome; the algorithm is applied to an actual data set with promising results.

Motivating example

The motivating example came from a study titled "Genes Involved in Viral Carcinogenesis and Tumor Initiationin Hepatitis C Virus-Induced Hepato cellular Carcinoma".[5] The primary aim of this National Institutes of Health (NIH) grant funded project was to find genes related to: Hepato-cellular Carcinoma (HCC), a malignancy of the liver; and cirrhosis of the liver. Although the incidence of HCC is relatively low in developed countries, in the past few decades there has been an increase in the number of reported cases in countries such as Japan, the United Kingdom and France.[6] The number of reported cases is al soon the rise in the USA and is due, in part, to the prevalence of the Hepatitis C Virus (HCV). It is estimated that 4,000,000 persons are HCV sero-positive and it is known that one of the main causes of HCC is HCV infection.[7] In this study the issue of cirrhosis is also covered. Cirrhosis is a condition in the liver where the tissue is replaced by fibrosis, scartissue, and re-generative nodules. Like HCC, some of the causes of cirrhosis are HCV, Hepatitis B, and alcohol abuse.[8] It is believed that almost every carcinogenic path way is altered in the development of the HCC.[7] In the diagnosis of HCC the following guidelines are provided by:[9] Once HCC is suspected, a computed tomography (CT) scan of the liver and thorax is a common detection mechanism. In addition, a magnetic resonance imaging (MRI), followed by a CT scan, may offer a more effective means of detecting lesions on the liver.

Once HCC has been successfully confirmed, transplantation is one of two successful methods proved at resolving this disease, along

with hepatic resection.[9] In the case where surgery is not possible, non-surgical techniques, such as percutaneous ethanol injection (PEI), may provide some benefit.[9] In the case the cancer cannot be resolved; the disease usually results in death to the patient in approximately 3-6months, although it has been common for some people to survive longer. Although there are ways and methods aimed at detecting this disease, HCC is usually caught in the later stages and there is still no set standard of care.[9] In therefore mentioned study there were various diseased states of the liver.

The tissue types are:

a. Normal liver tissue (normal)

b. Cirrhotic liver tissue (pre-malignant)

c. HCC liver tissue (malignant)

As stated by Thomas & Zhu[7] Because of the heterogeneity of the under lying etiologies, it is a challenge to provide a clear and consistent portrait of the principal molecular abnormalities in this disease."Due to the complex nature of this cancer, some people believe that estimating HCC for HCV patients cannot be done with a given level of accuracy.[11] The issue of diagnosis for HCC provides an excellent opportunity to evaluate our model framework as the proposed method will not only find genes related to HCC, but also to the progression of the disease based on the ordinal out come. There has been recent work on fitting a penalized model to a subset of the data from this study.[12]

The stereotype logit model

The stereo type logit model was initially proposed by Anderson [4] and is based on the multinomial distribution. For a given observation, J, denote the outcome vector, of length J, $\mathbf{y}_i$ as $(y_{i1}, y_{i2}, ...y_{iJ})$ where $y_{ij} = 1$ if for that observation, the outcome is in the J category; the other entries in the vector are 0. There are J possible outcomes. In addition, the J^{th} level is defined as the reference level against which all other levels are compared. There is also a covariate vector $(y_{i1}, y_{i2}, ...y_{iJ})$ consisting of p covariates possibly related to the outcome. For the one dimensional stereo type logit the model has the log likelihood of

$$\sum_{i=1}^{n}\left[\sum_{j=1}^{J-1} y_{ij}\theta_{ij} + \log \pi_J(\mathbf{x}_i)\right] \tag{1}$$

where

$$\theta_{ij} = \log\frac{\pi_j(\mathbf{x}_i)}{\pi_J(\mathbf{x}_i)}, \tag{2}$$

and

$$\pi_j(\mathbf{x}_i) = \frac{e^{\theta_{ij}}}{\sum_{j=1}^{J} e^{\theta_{ij}}} \tag{3}$$

In addition θ_{ij} is represented as $\alpha_j + \phi_j\{\mathbf{x}_i'\beta\}$, though $\pi_j(\mathbf{x}_i)$ is now modeled as

$$\frac{\exp(\alpha_j + \phi_j \mathbf{x}_i'\beta)}{1 + \sum_{j'=1}^{J-1}\exp(\alpha_{j'} + \phi_{j'}\mathbf{x}_i'\beta)} \tag{4}$$

The baseline category, against which all others are compared, is the J^{th} level. For each ordered level the effect of the independent variables is equal to an overall effect multiplied by a value ϕ_j. This is applicable when modeling disease progression where it is believed that a group of covariates are related to the disease and that this relationship is proportional for all levels. The intensity parameter, ϕ_j, is used to determine the ordinal structure. That is, for $j \neq j'$ if $\phi_j \geq \phi_{j'}$ then outcome level j is higher than level j'. We can also determine whether there is a statistically significant difference between two levels based on their magnitude parameters. In a medical setting there may be case of a given diseased tissue, say lung tissue, having varied states. It may not be clear whether a given state is better or worse than another. In other words, there is no ordering among the states. The stereotype model has a direct appeal to this setting; as a preliminary analysis, researchers can gain insight into the ordering of the tissue states as well as whether there is a significant difference between two states previously considered to differ. As this is an ordinal model we are concerned with modeling the logits θ_{ij}. The log likelihood is now represented as

$$L(\beta,\alpha,\phi \mid y,x) = \sum_{i=1}^{n}\sum_{j=1}^{J-1} y_{ij}\left(\alpha_j + \phi_j\{x_i'\beta\}\right) - \log 1 + \sum_{j=1}^{J-1} e^{\alpha_j + \phi_j\{x_i'\beta\}} \tag{5}$$

Elastic net constrained Stereo type logit model

This model takes a multinomial likelihood, with a stereotype logit, and adds a penalty to it in an attempt to model high dimensional data. Currently, there is not a set standard method to analyze the discussed data structure employing a likelihood approach. The applied penalty is known as the elastic net penalty.[13] A constraint on the sum of the absolute value of the parameters of interest is imposed. In addition, to ensure stability of the estimates, an additional smaller penalty on the sum of the squared values of the parameters is also enforced.[13] For a set of parameters, represented in a p length vector $\hat{\mathbf{a}}$, the elastic net penalty is defined as follows

$$\lambda\sum_{k=1}^{p}\left(\psi\beta_k^2 + (1-\psi)|\beta_k|\right) \tag{6}$$

where $0 < \lambda < \infty$ is allowed to vary, and ψ is a tuning parameter which ranges from 0 to 1.[14] The value of ψ represents the proportion of the penalty attributed to the LASSO and $1-\psi$ is the proportion of the penalty attributed to the ridge. The second term places a penalty on the sum of the absolute value of the parameters and is known as the LASSO penalty. This penalty has the effect of restricting the selected number of parameters to be less than the number of observations.[15] In addition, this penalty leads to a theoretical unique solution when the objective function is convex.[16] If this case does not hold, the first term of the penalty as defined on the sum of the squared values of the parameters[17] becomes relevant. This term is known as the ridge penalty. Applying the ridge penalty alone yields an estimate for each covariate; the LASSO yields a parsimonious model forcing many parameters to 0. As such, the goal is that the LASSO penalty will have more effect than the ridge penalty. For the stereo type logit representation, there is a common underlying effect for each level of the ordinal outcome; it is the intensity of this effect that differs from level to level. Based on the multinomial distribution, we are concerned with finding a set of estimates for our parameters, $\hat{\beta}$, such that

$$(\hat{\beta},\hat{\alpha},\hat{\phi}) = \underset{\beta,\alpha,\phi}{\arg\max}\, L(\beta,\alpha,\phi \mid y,x) \tag{7}$$

where $\boldsymbol{\alpha}$ denotes the vector of length J-1 containing the intercepts for the J-1 logits, and ϕ denotes the vector on length J-1 containing the intensity parameters. In addition, the intensity parameters, ϕ_j, are bounded such that $0 \leq \phi_j \leq 1$ for $\forall_j$.[4] We note that minimizing the

Citation: Williams AAA, Archer KJ. Elastic net constrained stereotype logit model for ordered categorical data. *Biom Biostat Int J.* 2015;2(7):191–198. DOI: 10.15406/bbij.2015.02.00049

negative log likelihood is equivalent to maximizing the log likelihood. Therefore, after imposing the elastic net constraints, we are concerned with finding parameter estimates such that:

$$\left(\hat{\beta},\hat{\alpha},\hat{\phi}\right)=\underset{\beta,\alpha,\phi}{\arg\min}\left\{-L(\beta,\alpha,\phi\mid \mathbf{y},x)+\lambda\sum_{k=1}^{p}\left(\psi\beta_k^2+(1-\psi)|\beta_k|\right)\right\} \quad (8)$$

The goal is to emphasize the LASSO penalty over the ridge. For our research, as λ is allowed to vary, there is a desire to trace the corresponding solutions of our nonlinear objective function through this parameter. In addition to specifying the starting value of λ, the maximum penalty, λ , needs to be specified. Additionally, the step length is defined as the distance between adjacent values of our varying penalty parameter λ, or $\delta_k=\lambda_k-\lambda_{k+1}$. Termination criteria must also be specified. This concept of tracing the solution through the given parameter is formally referred to as the λ trace.[15] An approach aimed at modeling our nonlinear objective function with the elastic net constraint is now presented and is based on an approach employed by Park and Hastie.[15] An implemented algorithm based on the modeling approach is subsequently presented.

Implemented algorithm

The implemented algorithm attempts to model equation (5) over the λ trace. This approach utilizes nonlinear programming to find optimal solutions for a given value of λ. This algorithm gives the user control over the range of λ values. The user is allowed to select $\lambda_{\min}$, $\lambda_{\max}$, and δ_k. The parameter $\forall k$ is set at a fixed parameter over $\forall k$ and is now denoted δ. The general algorithm describes the application of modeling procedure over the λ trace and is called *lambda trace*; leading to numerous models being fitted. For a given value of λ, a sub-algorithm is invoked; we call this procedure *model estimation*. This sub-algorithm uses an optimization algorithm developed by Ye [18] in model fitting. For the *model estimation*, the entry order of the variables into the model needs to be specified, the procedure *variable entry into the model* performs this step.The *lambda trace* algorithm is now presented.

Lambda trace algorithm

1. Determine the smallest value of λ, $\lambda_{\min}$. This is selected by the user.
2. Determine the largest value of λ, $\lambda_{\max}$. This is selected by the user.
3. Determine δ, the step length. This is selected by the user
4. Calculate the number of models to be fit. This is calculated as $\left\lceil(\lambda_{\max}-\lambda_{\min})/\delta\right\rceil$ and is denoted as $k_{\max}$
5. Set $k=1$
6. For λ_k invoke the procedure *model fit* to find the solution. Denote the solution at *k*. $(\hat{\beta},\hat{\alpha},\hat{\phi})'_k$
7. If $k=k+1$, terminate, else set $k=k+1$ and repeat step 6.

The above procedure is responsible for providing and storing the parameter estimates for all models along the λ trace. The benefit of having the user selects $\lambda_{\min}$, $\lambda_{\max}$ and δ is that the solution can be tailored for specific circumstances. If there is a desire to highly penalize the model so that only a few covariates will be included, or to place a lower range of penalty on the model so that a larger amount of covariates will enter the model, or evaluate a larger range of models, $\lambda_{\max}$, $\lambda_{\max}$ can be selected accordingly. Also if δ is smaller the fitted models will be closer with regards to the parameter estimates; a larger version of δ will result in models where the parameter estimates could be more varied. This approach allows more flexibility with respect to the execution of the algorithm. In addition, define the active set *A*, as the set of covariates that are included in the solution at λ_k .The algorithm *model estimation*, which is used to provide a solution at a set value of λ, is now presented.

Model estimation algorithm

1. Determine the order of entry of the variables into the model using the model entry procedure. Denote this list of entry as the vector of length p , ν . Include the first 5 important variables in the preliminary model. Set $t=1$.
2. Use nonlinear programming to find the solution to equation (5), for a given value of λ. Denote the set of parameter estimates at this stage $\hat{\beta}_t$, *t*=1,2,… .
3. For the parameter estimates $\hat{\beta}_t$ if $|\hat{\beta}_{tk}|\le\varepsilon$, then it is removed from the model.
4. If the length of $\hat{\beta}_t$ is $n-(2\times J)+3$ or if all variables in ν have been considered then go to step 5, else include the next important variable, as determined by ν, set $t=t+1$, and repeat step 2.
5. Among the candidate models choose the one which has the best classification performance (the one with the highest percent correctly classified). For the given value λ_k denote this model as $\left(\hat{\beta}',\hat{\alpha}',\hat{\phi}\right)_2'$
6. The head variable entry into the model should not be a main heading; it is just the title of an algorithm.

This algorithm attempts to model the nonlinear objective function subject to the elastic net penalty at a given value of λ. Before the variables are passed to this algorithm they are scaled and centered. In addition, as the whole aim of this paper is to correctly model and classify ordinal outcome data with a high and complex covariate space, the sub model that correctly determines the highest proportions of correctly classified outcomes is selected. As the multinomial log likelihood model, with the stereotype logit representation, is a generalized nonlinear model, adequate starting values must be determined. Once the order of entry has been determined by the *model entry* procedure, the first five entries are input and the model is fit. Each literation includes the next important variable in $^{\nu}$. For a given iteration t, a parameter estimate is removed if $\left|\hat{\beta}_{tk}\right|\le\varepsilon$ where ε is a user adjusted parameter with the default of. It is desirable that parameter estimates with values close to 0 be removed from the model; therefore ε should be set to a small value. This process continues until the maximum number of variables is $n-(2\times J)+3$, or until all variables have been considered.

The *variable entry* into the model procedure, used to create is **v** now explained.

Variable entry into the model algorithm

1. Discretize all continuous variables.

Citation: Williams AAA, Archer KJ. Elastic net constrained stereotype logit model for ordered categorical data. *Biom Biostat Int J.* 2015;2(7):191–198. DOI: 10.15406/bbij.2015.02.00049

a. For a variable, compute its minimum (min) and maximum (max) values.

b. For a specified number of bins, calculate interval width as follows (max-min)/(number of bins). By default they are 4 bins.

c. Place the variables into corresponding bin and return the bin number (1, 2, 3, or 4)

2.For the newly created ordinal variables, calculate the gamma statistic[19] using the ordinal outcome variable. Rank the variables in order of importance based on the absolute value of the statistic.

Applied bootstrap resampling procedure

For our model we need to estimate the standard errors of our parameter estimates. For the purposes of this paper, the bootstrapping pairs design is used.[20] Denote B as the number of bootstrap re-samples. The size of B is set to 200; this is based on the fact that B ranging from 50 to 200 is sufficient.[21] For a covariate matrix $\mathbf{y}$ and an ordinal outcome vector $\mathbf{y}$, which are viewed as the population, define the tuple $(y_i, \mathbf{x}_i)$ which denotes the i^{th} entry and row respectively, $i = 1, 2, ..., n$. For each bootstrap resample we take a sample of n tuples, with replacement from the original data giving rise to a new data set $\mathbf{x}_b$ and $\mathbf{y}_b$, $b = 1, 2, ..., B$. Once we have the B re-samples, the corresponding model is fit to each data set. For the original data, once the model is selected based on the elastic net penalty, the corresponding value of λ is used in model fitting for all bootstrap re-samples. This value is fixed as allowing it to vary may introduce additional variation into our model[22] and it is desirable that the variances be correctly attributed to the parameters and their interactions with each other. Once the B models are fit the corresponding parameter estimates are obtained.

Denote the b^{th} bootstrap parameter estimates as $(\hat{\alpha}, \hat{\beta}, \hat{\phi})_b$. Having these B parameter estimates allow us to gain insight into the distributions of the parameter estimates as well as construct confidence intervals. The potential of using a bootstrap re-sampling procedure to obtain the distribution for a given parameter is that it no longer has to conform to a known form; it is no longer bounded. In addition, the corresponding confidence intervals can be developed; they need not be symmetric. The resulting B estimates for a given parameter can also be used to assess significance; by the proportion of them that are non-zero. In addition, a covariance matrix will also be calculated from the bootstrap re-samples. In the construction of the confidence intervals the bootstrap-t confidence interval method is used. In short, the bootstrap-t confidence intervals are of the from

$$\left[\hat{\beta}_p - \hat{t}^{(1-\alpha)} \times s\hat{e}, \hat{\beta}_p + \hat{t}^{(1-\alpha)} \times s\hat{e}\right] \tag{9}$$

where

$$s\hat{e}(\hat{\beta}_j^*) = \sqrt{\hat{V}(\beta_j)} \,/\, B \tag{10}$$

with $\hat{V}(\beta_j)$ being defined as

$$V(\beta_j) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}(.)_j - \hat{\beta}_{jb}^*\right)^2 \tag{11}$$

where $j = 1, 2, ..., p$ and

$$\hat{\beta}(.)_j = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_{j,b}^* \tag{12}$$

In addition $\hat{t}^{(\alpha)}$ is chosen from the standard normal distribution such that

$$\#\{Z^*(b) \le \hat{t}^{(\alpha)}\} \Big/ B = \alpha \tag{13}$$

where $Z^*(b)$ is defined as

$$Z^*(b) = \frac{\hat{\beta}(.)_j - \hat{\beta}_j}{se(\hat{\beta}(.)_j)} \tag{14}$$

Model selection criteria

As the proposed modeling procedure leads to a group of models along the λ trace, there is a need to select one model using objective criteria; different criteria may lead to different candidate models. The goal of any modeling procedure is to approximate, or make an educated guess at, the truth. The goal is to select models that will shed some light on the true phenomenon in the data. In this section three model selection procedures are employed; model selection based on Akaikie information criterion (AIC) and Bayesian information criterion (BIC) and cross validation using a data set. This may lead to the selection of up to three models for further consideration. For model selection, using AIC and BIC, once the corresponding models along the λ trace are found the model yielding the lowest AIC and BIC value are selected. In applying these criteria the log likelihood in equation (5) is used; the penalty is not included. As these model selection procedures are calculated using non-penalized likelihoods, not pseudo-log likelihood like that in equation (8), it is best to adhere with convention.[23] In practice AIC has a tendency to select models that include a larger number of variables, some of which are truly unimportant.[24] Model selection, using BIC, leans towards models that select fewer covariates as compared with AIC.[25]

The third approach selects the ideal model based on its predictive capabilities. The modeling procedure is used to obtain various models for values of λ along the λ trace; their performance is evaluated on a given data set. For the ordinal outcome data, a simple definition of classification performance is employed and is defined as follows: divide the number of correctly classified observations by the total number of observations. The model that is able to correctly classify the most observations is selected. For this paper, validation was performed using the data to which the model was fit. Once our three candidate models have been selected, they can be compared using various objective measures.

Data simulation

For assessing the proposed model the multivariate normal distribution was used to simulate datasets consisting of a large number of covariates and a smaller number of samples, such as is the case with microarray gene expression data. For each simulation, N=80 observations and P=400 covariates were generated. The correlation matrix is of dimension 400×400. The correlation matrix has a compound symmetric structure where the off diagonal entries, ρ, equal 0.01. This structure was selected to test the strength of the proposed modeling scheme against different relationships among the covariates. After the full correlation matrix was created, to make its form acceptable for further processing the correlation matrix was converted to its positive definite form. Thereafter, ten covariates were selected to be the truly important predictors by setting their corresponding coefficient values to either 0.5 or -0.5. The remaining 390 covariates were not related to the response and hence were unimportant to the predictive structures; their coefficient values were set to 0. After this, the correlation matrix

Citation: Williams AAA, Archer KJ. Elastic net constrained stereotype logit model for ordered categorical data. *Biom Biostat Int J.* 2015;2(7):191–198. DOI: 10.15406/bbij.2015.02.00049

was converted to a variance matrix $\sum$.

Using the Cholesky's decomposition the covariance matrix can be decomposed as follows:

$$\sum = \mathbf{A}\mathbf{A}' \quad (15)$$

Then, a matrix of dimension 400×80, denoted $\mathbf{X}^*$, was generated so that the columns of the matrix demonstrate an independent multivariate normal distribution. An intermediate covariate matrix $\mathbf{X}^*$ with the desired covariate structure was created by the following transformation:

$$\mathbf{X}^* = \mathbf{A}\mathbf{X} \quad (16)$$

Following this, a random normal vector, of length 390, with parameters N (0, 1) was generated to represent the means of the unimportant variables; the means of the 10 important variables correspond to a random normal vector whose entries are generated from N (6, 1). These combined vectors were used to represent the mean of the 400 covariates. The mean vector, $\mathbf{i}$, was then combined with the covariate matrix to create the final covariate matrix $\mathbf{X}^{\kappa}$ as follows

$$\mathbf{X}^{\kappa} = \mathbf{M} + \mathbf{X}, \quad (17)$$

Where $\mathbf{M}$ is 400×80 matrix with each column equaling $\boldsymbol{\mu}$ For the desired outcome there are four levels. After the 10 truly predictive covariates were selected, the matrix $\mathbf{Z}$ was generated. $\mathbf{Z}$ is a 80×40 matrix containing the output row vectors, of length 4, for the 80 observations. Let i = 1,2,…,80 index the observations and j = 1,2,3,4 index the outcome levels. The probabilities $\pi_j(\mathbf{x}_i)$ were generated using the stereotype logit model.[4]

The simulated parameters for β contain {0.5, 0.5, 0.5, 0.5, 0.5, -0.5, -0.5, -0.5, -0.5, -0.5}. The baseline level was set at j=4. The simulated α parameters are {-0.7, -0.1, 0.1, 0.0}. The ϕ parameters values are {1.00, 0.67, 0.33, 0}. For a given observation, i, denote the true outcome, h_i, as $y_i = \arg\max_h \pi_h(\mathbf{x}_i)$. Denote the vector containing these values as y^*. The proposed modeling scheme was applied to y^* and $\mathbf{X}^{\kappa}$. Bootstrapping re-sampling techniques were used to estimate the 95% confidence intervals of the model parameters. The variable selection capabilities are acceptable as all 10 variables are selected in the final model. However, a large portion of the non-significant variables were included in the final model, 152 to be exact. Based on Figure 1 the signs of the estimates are correct. The first five should have a positive average value and the last five should be negative; this is preserved. For the confidence intervals of the parameter estimates, the intervals are somewhat small. This is the anticipated result as the proposed method is supposed to yield parameter estimates with low variability. This is seen in Table 1. The results indicate that the modeling framework is adept at variable selection; the classification capabilities require finer tuning.

Application to liver data

In the application of the method the only independent variables used were gene expression values. The cRNA samples were hybridized to the HG-U133A and the HG-U133A2.0. Chips. These chips are able to measure approximately 17000 and 14500 unique genes respectively (Affymetrix, CA) [26]. The raw data, CEL files were downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14323. The expression summaries were obtained using the robust median average (RMA). In an attempt to filter the genes the MAS 5Present/Absent calls were used. The only genes included were those for which there was a present call in all the samples.

Table 1 Parameter estimates, along with 95% confidence interval, for truly important variables included in the final model of the compound symmetric correlated data. The final model was chosen based on the highest percent correctly classified

Truly Important Variable	Parameter Estimate	95% Confidence Interval
V1	1.40	(1.04,1.75)
V2	1.77	(1.38, 2.16)
V3	1.16	(0.78,1.54)
V4	0.97	(0.55,1.39)
V5	1.06	(0.65,1.47)
V6	-3.22	(-3.81, -2.63)
V7	-0.97	(-1.33, -0.62)
V8	-1.65	(-2.15, -1.16)
V9	-1.19	(-1.54, -0.84)
V10	-1.66	(-2.08, -1.23)

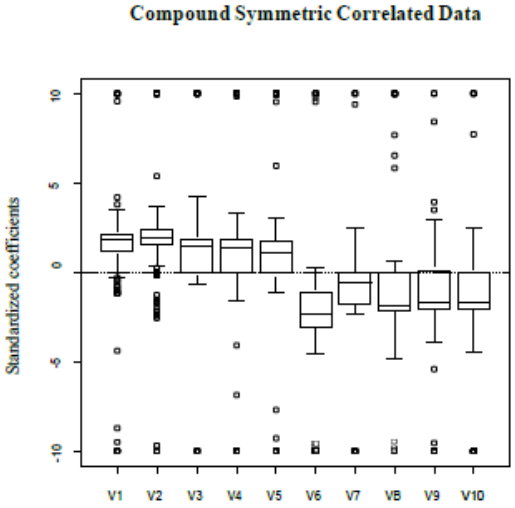

Figure 1 Boxplot of parameter estimates over the B=200 bootstrap resamples for the ten truly important covariates in the compound symmetric simulation. The final model was chosen based on the highest percent correctly classified.

The proposed model was applied to the gene expression data set, with associated ordinal outcome, in an attempt to determine genes associated with disease progression of liver tissue classified by normal-cirrhosis-HCC. The algorithm *lambda trace* was formally applied to the data. The gene expression values were the only variables used. The variables were standardized (centered and scaled) prior to model fitting. Among the 98 samples, 19 had normal liver tissue, 41 had cirrhosis of the liver and 38had HCC. The genes used to fit the model were selected if the MAS 5calls were declared as present in all samples. MAS 5calls are used to remove the genes that cannot be detected with a given degree of reliability. If, for a given gene, the MAS 5 called is present then the readings for this gene are reliable. This reduced the number of candidate genes to 4406; this set was passed to the model building process. The results are shown for the model corresponding to λ=0.001.

Citation: Williams AAA, Archer KJ. Elastic net constrained stereotype logit model for ordered categorical data. *Biom Biostat Int J.* 2015;2(7):191–198. DOI: 10.15406/bbij.2015.02.00049

Results

Boot strap re-sampling was used to provide estimates of the standard error which were used in the construction of the confidence intervals. B=200 re-samples were used. In each re-sample, a fixed sample size of 98 patients was randomly drawn with replacement from the original sample. When the method was applied to these re-samples, the value of λ was fixed at 0.001. To assess the significance of the parameter estimates, the bootstrap-t confidence intervals were used. The aim was to as certain if the parameters estimates were significantly different from 0. Figure 2 shows the genes (and their distributions) selected by the final model. The boot strap procedure, previously described, was used to generate the 95% confidence intervals presented in Table 2. The model that resulted in the lowest misclassification rates was selected. Genes whose coefficient values are greater than 0 are over expressed; coefficient values less than 0 imply the gene is under expressed. Table 2 shows the selected genes, along with their confidence intervals. The corresponding gene names and definitions are also presented. Similar to the results from the simulation the 95% confidence intervals are very small; the percentage correctly classified is somewhat low, below 50%. Figure 2 presents the underlying gene expression profile of HCC; it is assumed that this particular expression profile is present at all stages of the disease. The intensity of this profile increases as the disease progresses to severer stages. The analysis was performed in R.[27]

Table 2 Final Variable produced by the Stereotype Logit model when applied to the liver data. Parameter estimates, along with 95% confidence interval, for the variables included in the final model fitted to the liver expression data. The final model was chosen based on the highest percent correctly classified

Gene name	Definition	Parameter Estimate	95% Confidence Interval
NCOR1	Nuclear receptor corepressor1	-9.46	(-9.62, -9.31)
CALD1	caldesmon1	9.95	(9.87,10.04)
OSBP	Oxysterol binding protein	0.02	(-0.59,0.63)
ATMIN	ATM interactor	6.17	(5.68,6.66)
TJP2	Tight junction protein2	-4.66	(-5.08, -4.24)
RABAC1	Rab acceptor 1	-6.5	(-6.85, -6.14)
MTRR	5-methyltetrahydrofolate- homocysteine methyltransferase reductase protein tyrosine phosphatase	0	(-0.35,0.34)
PTPN3		-1.51	(-2.02, -1.00)
EEF2	non-receptor type 3 eukaryotic elongation factor-2 Kinas	4.11	(3.46,4.77)
PLGLB1	Kinas plasminogen-like B1	2.72	(2.36,3.09)
PDIA3	Protein disulfide isomerase family A, member 3 valosin-containing protein	-2.93	(-3.31, -2.55)
VCP		-0.66	(-1.57, -0.24)
STAT3	signal transducer and activator of transcription 3 transmembrane protein41B	4.73	(4.27,5.20)
TMEM41B		6	(5.51,6.48)
ACVR1B	Activin A receptor, type IB	-1.33	(-1.89,-0.78)
KRAS	v-Ki-ras2Kirstenrat sarcomaviral oncogenehomologserine	-4.65	(-5.02,-4.28)
SHMT2		-4.25	(-4.78, -3.71)
TMEM93	Hydroxyl methyl transferase 2 transmembrane protein93	1.35	(0.82,1.88)
SPCS3	Signal peptidase complex subunit 3 homolog aquaporin 3	3.84	(3.19,4.49)
AQP3		-4.66	(-5.06, -4.26)
INTS5	Integrator complex subunit 5	-1.27	(-1.93, -0.60)
THADA	Thyroid adenoma associated	-3.98	(-4.01, -3.95)

Citation: Williams AAA, Archer KJ. Elastic net constrained stereotype logit model for ordered categorical data. *Biom Biostat Int J.* 2015;2(7):191–198. DOI: 10.15406/bbij.2015.02.00049

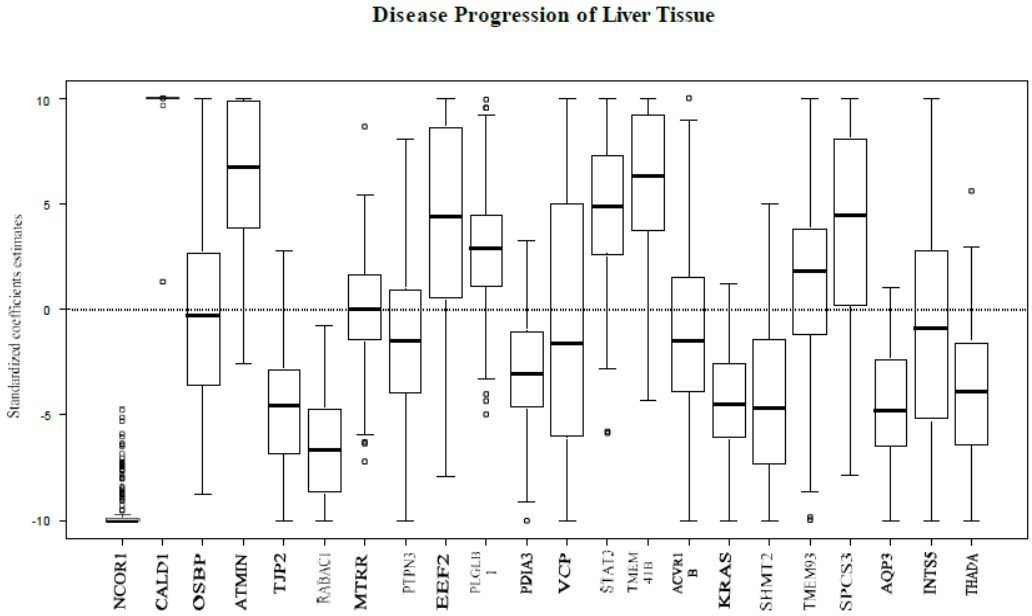

Figure 2 Boxplots based on the bootstrap resampling procedure for the genes selected from the application of the proposed stereotype logit model.

Twenty two genes were selected to be in the final model. An α level of 0.05 was used in determining significance. Figure 2 also provides insight into the variability of the coefficients for the genes. Some genes such as NCOR1 have low variability from the bootstrap re-sampling procedure. Some genes, such as VCP, have a higher variability. The entire gene scan be further assessed for significance by using a publicly available database, such as GO[28] or Entrez,[29] performing pathway analysis, seeking input from a scientific investigator, or from prior knowledge. Based on the KEGG[30] database, the KRAS gene has been linked to HCC.AccordingtoEntrezGene,NCOR1 has been implicated in prostate, bladder, colorectal, and breast cancer; STAT3 functions in many cellular processes including cell growth and apoptosis; and PTPN3 is associated with Oncogenic human papilloma virus E6 proteins.

For the gene expression data sets, a list of genes was presented in Table 2. These genes were declared significant in relation to the ordinal outcome, progression of disease. Path way analysis maybe performed on these genes; a corresponding database search can also be conducted to as certain if the gene had been previously linked to the disease progression. A clinical investigator, with prior knowledge of the disease domain can also assess the significance of these genes to provide further context.

Conclusion

This paper developed a penalized modeling procedure with an ordinal outcome. The penalized stereotype logit model is suited to the case where there are more covariates than observations; a typical scenario with genomic data. The proposed method performs automatic variable selection and model estimation by penalizing predictors with the elastic net constraint. We developed the model by adding an elastic net penalty to the stereotype logit model and provided the model fitting algorithms *Lambda trace algorithm*, *Model estimation algorithm*, and *Variable entry into the model algorithm*. We presented a simulation study demonstrating the performance of the proposed statistical methodology. The applied method was able to select all the important (nonzero) covariates in the simulated data. The method was then applied to liver tissue gene expression data set. HCV infection is one of the main causes of HCC. As such, there is a concern to determine a risk estimate for this cancer in those with HCV infection. There are those who believe that a risk estimate cannot be provided with a given degree of accuracy.[11] As such this is an ideal scenario for the application of the proposed method. The results were presented. For the selected genes additional information was presented in Table 2 and Figure 2. Some of the genes have been linked to HCC or cirrhosis; the KRAS gene has been linked to HCC based on the KEGG[30] database. Based on the bootstrap re-sampling procedure the variability and standard deviation of the coefficient estimates were presented. There were several misclassified samples. As we used publicly available data for this study we were constrained by the sample size. A larger sample size may yield more significant results and lead to lower misclassification rates. In addition we did not have relevant demographic information. If these were available, and included in the modeling procedure, we would expect lower misclassification rates.

One limitation of this study is the computational complexity of the algorithm. Because an exhaustive model search is performed, this is not unexpected. A potential way to address this is to enhance the model-fitting algorithm by passing more information to the optimization procedure previously presented. The score (first derivative) and the Hessian (second derivative) could be included the model fit algorithm. The more information that is passed to the optimization function the clearer the search path to the solution becomes, which may reduce the execution time to reach this solution. We would need to verify whether the Hessian is positive definite and, if not, invoke an appropriate function to make it positive definite. An R function make positive definite[31] could be used. However, the nonlinear programming function solnp[31] used for this study does not allow one to supply the score and Hessian matrix functions. As such the appropriate function

Citation: Williams AAA, Archer KJ. Elastic net constrained stereotype logit model for ordered categorical data. *Biom Biostat Int J.* 2015;2(7):191–198. DOI: 10.15406/bbij.2015.02.00049

would need to be utilized. This will be explored in a future manuscript. In addition, coding the functions in C++ code that is then callable by R[27] could also help to reduce execution time. Development and implementation, of statistical methods that discern genes (or single nucleotide polymorphisms (SNPs), or methylation (CpG) sites) related to disease progression will provide vital information that can be used for disease control and prevention. This information can be employed when designing treatments for diseases with a genetic contribution. Hopefully the penalized stereotype logit model, and associated algorithms, will be employed to determine genetic factors related to the particular stage in the progression of disease.

Acknowledgement

None.

Conflict of interest

None.

References

1. Liberman L, Abramson AF, Squires FB, et al. The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. *AJR Am J Roentgenol*. 1998;171(1):35–40.
2. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247.
3. Agresti A. Categorical data analysis. John Wiley & Sons; 2014.
4. Anderson JA. Regression and ordered categorical variables. *Journal of the Royal Statistical Society*; 1984:1–30.
5. Mas VR, Maluf DG, Archer KJ, et al. Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus–induced hepatocellular carcinoma. *Mol Med*. 2009;15(3-4):85–94.
6. El-Serag HB, Mason AC. Rising incidence of hepatocellular carcinoma in the United States. *N Engl J Med*. 1999;340(10):745–750.
7. Thomas MB, Zhu AX. Hepatocellular carcinoma: the need for progress. *J Clin Oncol*. 2005;23(13):2892–2899.
8. Mayo Clinic. *Liver cancer: Causes*. 2010.
9. Ryder SD. Guidelines for the diagnosis and treatment of hepatocellular carcinoma (HCC) in adults. *Gut*. 2003;52(suppl 3):1–8.
10. Llovet JM, Real MI, Montaña X, et al. Arterial embolisation or chemoembolisation versus symptomatic treatment in patients with unresectable hepatocellular carcinoma: a randomised controlled trial. *Lancet*. 2002;359(9319):1734–1739.
11. Di Bisceglie AM. Hepatitis C and hepatocellular carcinoma. *Hepatology*. 1997;26(S3):34S–38S.
12. Archer KJ, Mas VR, David K, et al. Identifying genes for establishing a multigenic test for hepatocellular carcinoma surveillance in hepatitis C virus-positive cirrhotic patients. *Cancer Epidemiol Biomarkers Prev*. 2009;18(11):2929–2932.
13. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;7(2):301–320.
14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
15. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007;69(4):659–677.
16. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*. 1970;12(1):55–67.
17. Rosset S, Zhu J. Discussion of "least angle regression" by Efron et al. *Annals of Statistics*. 2004;32(2):469–475.
18. Ye Y. Interior algorithm for linear, quadratic, and linearly constrained non-linear programming. Stanford University; 1987.
19. Goodman LA, Kruskal WH. Measures of association for cross classifications*. *Journal of the American Statistical Association*. 1954;49(268):732–764.
20. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical science*. 1996;189–212.
21. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*. 1986:54–75.
22. Osborne MR, Presnell B, Turlach BA. On the lasso and its dual. *Journal of Computational and Graphical statistics*. 2000;9(2):319–337.
23. Burnham KP, Anderson DR. Multi-model inference understanding AIC and BIC in model selection. *Sociological methods & research*. 2004;33(2):261–304.
24. George EI. The variable selection problem. *Journal of the American Statistical Association*. 2000;95(452):1304–1308.
25. Kadane JB, Lazar NA. Methods and criteria for model selection. *Journal of the American Statistical Association*. 2004;99(465):79–290.
26. Liu G, Loraine AE, Shigeta R, et al. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*. 2003;31(1):82–86.
27. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014.
28. Carbon S, Ireland A, Mungall CJ, et al. Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25(2):288–289.
29. Maglott D, Ostell J, Pruitt KD, et al. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*. 2005;33:D54–D58.
30. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28:27–30.
31. Wuertz D. Utilities: Function utilities. 2010.