Machine Learning Techniques for Detecting Hierarchical Interactions in GLM's for Insurance Premiums

José Garrido

Department of Mathematics and Statistics
Concordia University, Montreal

EAJ–2016
Lyon, September 6–8, 2016

(joint work with Sandra–Maria Nawar, Deloitte, Toronto)

Research funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Concordia University Montreal

Introduction

In this project we use machine learning techniques to try improving the predictive models of a P/C portfolio of a large Canadian insurer.

(I) The advantages of using regularization methods for actuarial models are studied.

(II) The model proposed is via a group-Lasso interaction network (Hastie & Lim, 2015, JCGS) – a method to detect linear and non-linear effects by learning pairwise hierarchical interactions; we extend it to frequency and severity claims model.

Overview:

1. 1. Preliminaries
   - 1.1 Generalized Linear Models
   - 1.2 Regularization
   - 1.3 Hierarchical Models
2. 2. Group–Lasso Interaction Network
   - 2.1 Overview
   - 2.2 Algorithm and Optimization
   - 2.3 Regularized Claims Model
   - 2.4 FISTA Algorithm
3. 3. Actuarial Application and Results
   - 3.1 Simulation Study
   - 3.2 Example 1: Singapore Insurance Data
   - 3.3 Example 2: Ontario Collision Data

1.1 Generalized Linear Models

The family of GLMs is an extension of the linear regression model (McCullagh & Nelder, 1989, C&H) that transforms the mean response by a chosen link function.

The log-link is the most popular for insurance data, where the linear predictor gets exponentiated to ensure that premiums are positive and to preserve the multiplicative structure of the variable relativities.

In GLMs, responses $y_1, \ldots, y_n$ are assumed to be independent and linearly related to the predictor variables through a non-linear link function as follows:

$$g(\mathbb{E}[Y_i|X_i = x_i]) = \sum_{j=0}^{p-1} \beta_j x_{ij}$$

and the linear predictor is given by,

$$\eta_i = \sum_{j=0}^{p-1} \beta_j x_{ij}$$

The true mean can always be retrieved by taking the inverse transformation

$$\mu_i = g^{-1}(\eta_i).$$

The Exponential Dispersion Family

The family of distributions that can be written as

$$f(\theta;\phi,y_i) = \exp\left[\frac{y_i\theta - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right]$$

yielding a likelihood function of the form

$$L(\theta;\phi,y_i) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right].$$

The functions $a(.)$, $b(.)$, and $c(.)$ vary according to the particular distributions that are member of the exponential dispersion family.

The Log Likelihood Function

The log-likelihood function for a member of the EDF becomes

$$l(\theta;\phi,y)=\sum_{i=1}^{n}\left[\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right].$$

Coefficients are then estimated by maximizing the negative log-likelihood function. The system of partial derivatives also known as gradients of the log-likelihood is called score functions, is defined as:

$$s(\theta,y)=\frac{\partial}{\partial\beta}l(\beta;\phi,y)$$

The maximum likelihood estimator (MLE) $\hat{\beta}$ is then found as the solution to the system of equations $s(\theta;y)=0$.

Estimation

The partial derivative with respect to β_j is given by:

$$\begin{aligned}
\frac{\partial l(\beta;\phi,y)}{\partial\beta_j} &= \sum_{i=1}^{n}\frac{\partial}{\partial\beta_j}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right] \\
&= \sum_{i=1}^{n}\frac{1}{a_i(\phi)}\left[y_i\frac{\partial\theta_i}{\partial\beta_j} - \frac{\partial b(\theta_i)}{\partial\beta_j}\right] \\
&= \sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ij}}{a(\phi)b''(\theta_i)g'(\mu_i)}.
\end{aligned}$$

Poisson–gamma Frequency–Severity Modeling

For example, the log-link for the Poisson distribution generates a multiplicative model as follows,

$$\log(\mu_i) = X_i^T \beta$$

transforming it to get the true mean,

$$\begin{aligned} \mu_i &= e^{X_i^T \beta} \\ &= \left(e^{\beta_1}\right)^{X_{i0}} \left(e^{\beta_2}\right)^{X_{i1}} \dots \left(e^{\beta_p}\right)^{X_{i,p-1}}. \end{aligned}$$

The mean is then found as the product of the relativities.

Generalized Linear Models and Non-Life Insurance

Assume that the frequency and severity risks are independent and modeled separately using a different GLM distribution. Aggregate losses Y_i are represented by the sums

$$Y_i = \sum_{k=1}^{N_i} Y_{ik},$$

for $N_i > 0$, otherwise $Y_i = 0$. The mean and variance are given as

$$\mathbb{E}(Y_i) = \mathbb{E}(N_i)\mathbb{E}(Y_{ik}), \qquad \mathbb{V}(Y_i) = \mathbb{E}(Y_{ik})^2\mathbb{V}(N_i)+\mathbb{V}(Y_{ik})\mathbb{E}(N_i.)$$

Generalized Linear Models and Non-Life Insurance

The pure premium is calculated as the product of the mean frequency and severity as follows

$$\begin{aligned}\text{Pure Premium} &= \text{Frequency} \times \text{Severity}\\ &= \left(\frac{\text{Number of losses}}{\text{Exposure}}\right) \times \left(\frac{\text{Amount of losses}}{\text{Number of losses}}\right)\end{aligned}$$

The Loss Ratio $= \frac{\text{Expected Losses}}{\text{Premium}}$.

1.2 Regularization

Why use regularization?

Regularization techniques are crucial for modeling big data, which means dealing with high-dimensionality, sometimes noisy data that often contains many irrelevant predictors and propose challenges in interpretable models.

The Problem of Overfitting

Models that replicate sample data well but do not generalize well for out of sample data.

Regularization

Penalized regression is a set of techniques that impose a penalty on the regression coefficients and can be used as a powerful variable selection tool.

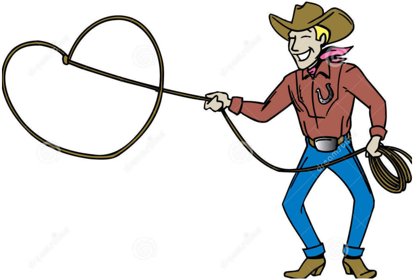
The penalty term λ controls which variables are included in the model based on how well they explain the response Y (size of coefficients).

The objective function:

$$\textit{Loss Function} + \textit{Penalty on Coefficients}$$

Lasso (Tibshirani, 1996, JRS)

The Least Absolute Shrinkage and Selection Operator (Lasso) is a regularization technique that performs feature selection and coefficient estimation by adding an ℓ_1–penalty term $||\beta||_1$ that constraints the minimum size of the estimated model coefficients.



Lasso for Linear Models

For linear models the Lasso

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2 \qquad \text{subject to} \sum_{j=1}^{p-1} \mid \beta_j \mid \leq s.$$

In Lagrange form the objective function becomes,

$$\arg\min_{\beta} \frac{1}{n} \left\| Y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 .$$

The tuning parameter λ determines how many coefficients are set to zero.

Lasso for GLMs

For GLMs, the objective function

$$\hat{\beta} = \arg\min_{\beta} \; l(Y, X; \beta) + \lambda \|\beta\|_1.$$

It is not necessary to necessarily use the log-likelihood function, any convex loss function denoted as $\mathcal{L} canbeused$.

For example the Poisson regression

$$\mathcal{L}(Y, X; \beta) = -Y(X^\top \beta) + \exp(X^\top \beta) + log(Y!),$$

which is here the negative log likelihood.

Lasso for Linear Models and GLMs

For any convex loss function, numerical optimization methods are needed for parameter updates. Small steps are taken in the opposite direction of the gradient of the negative log-likelihood function to ensure convergence.

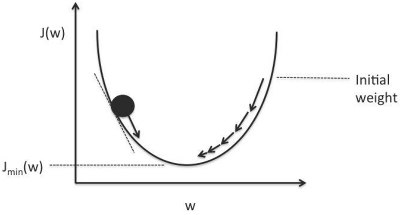


Figure: Schematic Gradient Descent

Lasso for Linear Models and GLMs (. . . continued

The figure shows how coefficients change as s changes.

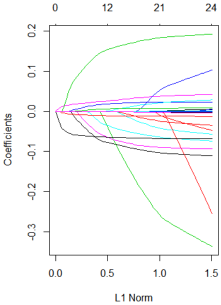


Figure: Shrinkage Coefficients for Lasso GLMs

- Bottom: The value of constraint s.
- Top: Number of variables captured out of 27.
- Each curve represents a coefficient as a function of the scaled Lasso parameter s.
- Absolute value of the coefficients tends to 0 as the value of s goes to 0.
- As s increases, λ decreases and more variables are captured.

Ridge Regression

The ridge penalty uses the ℓ_2 norm, which can shrink the size of the coefficients, but not necessarily force them to zero:

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=0}^{p-1} \beta_j x_{ij})^2 \quad \text{subject to} \quad \sum_{j=0}^{p-1} \beta_j^2 \leq s.$$

In the Lagrange form, the objective function to be minimized is:

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=0}^{p-1} \beta_j x_{ij})^2 + \lambda \sum_{j=0}^{p-1} \beta_j^2.$$

Lasso vs Ridge

Lasso has a major advantage over ridge regression, it provides a sparse solution when the problem is solved.

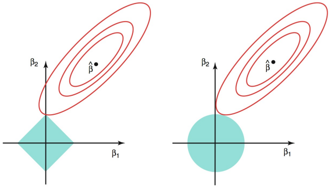


Figure: Comparing Lasso and Ridge

The figure gives a comparison of the error contours and constraint functions, for the ℓ_1 and ℓ_2 penalty.

Group–Lasso

The Group–Lasso

An extension of Lasso that performs variable selection on non-overlapping groups of variables and then sets groups of coefficients to zero.

Estimation method for models with specific sparsity patterns when the covariates are partitioned into groups, the group–Lasso leads to the selection of groups of covariates.

For high dimensional parameters a group structure is expected where the parameter space is partitioned into disjoint pieces.

Group–Lasso (. . . continued)

Sparsity at the group level can be achieved using the group–Lasso penalty for k groups of variables

$$\lambda \sum_{k=1}^{K} \gamma_k \left\| \beta_k \right\|_2 .$$

Adding a γ penalty term, yields sparsity at both the group and individual feature levels. The group–Lasso estimator can be defined by solving the following objective function in Lagrange form

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}(Y, X; \beta) + \lambda \sum_{k=1}^{K} \gamma_k \left\| \beta_k \right\|_2 ,$$

where $\mathcal{L}(Y, X; \beta)$ is the loss function for linear regression and GLMs.

Group–Lasso (. . . continued)

The “two-norm” penalty blends the Lasso ℓ_1-norm with the group–Lasso. Advantage: yields sparse solutions, both at the group and individual levels, allowing to penalize some groups more or less than others:

$$\arg\min_{\beta} \mathcal{L}(Y, X; \beta) + \lambda \sum_{k=1}^{K} \sqrt{p_k}\, \|\beta_k\|_2 \,,$$

where $\sqrt{p_k}$ accounts for the varying group sizes and $|| \cdot ||_2$ is the Euclidean norm.

Group–Lasso (. . . continued)

In the multi-variate case, the sparse group–Lasso optimization problem is given by

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}(Y, X; \beta) + \lambda_1 \sum_{k=1}^{K} \|\beta_k\|_2 + \lambda_2 \|\beta\|_1 ,$$

where β_k is the vector of coefficients for group k. When each group consists of one variable it reduces to the Lasso.

Subset Selection and Optimal Penalty Value

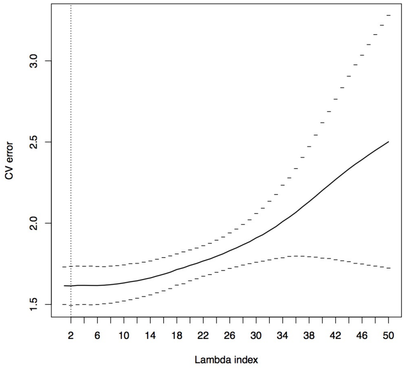


Figure: Ten Fold Cross Validation Error

1.3 Hierarchical Models

Hierarchical modeling

An extension of GLMs that gives specific model parameters their own sub-model, allowing to build models that can be grouped along a dimension containing multiple levels.

Used when data is hierarchically structured and can be grouped, outperforming classical regression in predictive accuracy.

Can be fit to multilevel and hierarchical structures by including coefficients for group indicators and adding group-level models.

Tree-Based Models

Non-parametric techniques that build ensembles of weak learners.

Generalized Boosting Models

A method that iteratively adds basis functions in a greedy way so that each additional basis function further reduces the selected loss function and the combined final model converges to a strong learner.

This is achieved by a sequential procedure of additive models $F_T(x) = \sum_{t=1}^{T} \alpha_t h_{jt}(x)$, where h_{jt} is a large pool of "weak learners".

Screening with Boosted Trees

Trees are used because of their ability to model nonlinear effects and high-order interactions.

A boosted model can be used as a screening device for interaction candidates.

2.1 Overview

The Group–Lasso Interaction Network.

Summary of the Model

- A method for learning pairwise interaction.
- Perform variable selection and dispense variables with explicitly applying constraints.
- GLM is then fit on candidate set.
- Overlapped group–Lasso penalty used to impose hierarchy.

Overview

Extension of the *glinternet* to include Poisson and gamma families to model claim frequency and severity.

The features of the *glinternet* are desirable because of its ability to do variable selection and automatic detection of interactions; satisfying the strong hierarchy property; for a non-zero estimated interaction, both its associated main effects are included in the model, and then a GLM is fitted.

The goal $\Longrightarrow$ fit insurance claims models.

Modeling Interactions

Mathematically, an interaction exists in f, between x and y, when $f(x, y)$ cannot be expressed as $g(x) + h(y)$ only, for any functions g and h.

To model an interaction, given a response Y,

$$Y = \sum_{i=0}^{p-1} X_i \theta_i + \sum_{i<j} X_{i:j} \theta_{i:j} + \xi.$$

Interactions occurs when the effect of one independent variable on the response variable changes depending on the level of another independent variable.

Modeling Interactions in Actuarial Science

The relationship between levels of one variable is not constant for all levels of another variable.

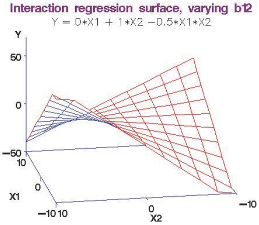


Figure: Interaction Regression Surface

Existence of interactions is evident when we see a varying interaction surface with respect to the response variable Y.

Strong Hierarchy

Satisfied when an interaction model includes those variables that have both of its main effects present.

There are 4 possible cases that satisfy a strong hierarchy:

1. $\mu_{ij} = \mu$ (no main effects, no interactions),
2. $\mu_{ij} = \mu + \theta_1^i$ (one main effect Z_1 or X_1),
3. $\mu_{ij} = \mu + \theta_1^i + \theta_2^j$ (two main effects),
4. $\mu_{ij} = \mu + \theta_1^i + \theta_2^j + \theta_{1:2}^{ij}$ (main effects and interaction).

Main effects can be viewed as deviations from the global mean, and interactions are deviations from the main effects, it does not make sense to have interactions without main effects.

2.2 Algorithm and Optimization

The model introduced is solved by dividing the solution path into two phases,

I The Screening Phase

  i Strong rules for discarding predictors in Lasso-type problems.

II Variable Selection and Model Fitting

It works by solving a group–Lasso with $p + \binom{p}{2}$ variables. .The model starts by fitting $\lambda = \lambda_{\text{max}}$, for which no variables are included, and then λ decreases to allow variables to enter the model.

Strong Hierarchy Through Overlapped Group–Lasso

As long as coefficients are unpenalized they satisfy a strong hierarchy. The problem is, when a penalty is added, results deviate from a strong hierarchy.

This property can be achieved by adding an overlapped group–Lasso penalty to the objective function in the form

$$\lambda(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \sqrt{L_2 \|\tilde{\alpha}_1\|_2^2 + L_1 \|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2}).$$

The group-Lasso has the property of "all zero" or "all nonzero" estimates,

$$\theta_{1:2} \neq 0 \Longrightarrow \theta_1 \neq 0 \Longrightarrow \theta_2 \neq 0,$$

which satisfies a strong hierarchy.

Equivalence with Unconstrained Group-Lasso Problem

If two intercepts are included in the model, one penalized and the other not, then the penalized intercept will be estimated to be zero, $\mu \leftarrow \mu + \tilde{\mu}$.

The overlapped group–Lasso reduces to a group–Lasso. Solving the constrained optimization problem with the tilde parameters is equivalent to solving the unconstrained problem.

$$\arg\min_{\mu,\beta} \frac{1}{2} \| Y - \mu 1 - X_1\beta_1 - X_2\beta_2 - X_{1:2}\beta_{1:2}) \|_2^2 \\ +\lambda\big(\|\beta_1\|_2 + \|\beta_2\|_2 + \|\beta_{1:2}\|_2 \big).$$

Screening Approach with Strong Rules

Strong rules discards large number of inactive variables that should not be added to the active set of variables.

The calculation for group–Lasso has to be conducted by computing $s_i = || X_i^T (Y - \hat{Y}) ||_2$ for every group of variables.

Then a filter is applied, where group i is discarded if the test $s_i < 2\lambda_{\text{current}} - \lambda_{\text{previous}}$ is satisfied.

Then KKT conditions are checked after the algorithm has converged to ensure that all discarded variables are actually equal to zero.

Finally, all pairwise interactions are taken for the variables that passed the screen in the combined expanded set.

Regularized Poisson Model

For count data (claim frequency), we consider a Poisson regression, using the log–link with negative log-likelihood:

$$-\sum_{i=1}^{n}\mathcal{L}(Y_i, X_i; \beta) = \sum_{i=1}^{n} -Y_i(X_i^\top \beta) + \exp(X_i^\top \beta).$$

Adding the Lasso ℓ_1 penalty the optimization problem becomes

$$\hat{\beta} = \arg\min_{\beta} \mathcal{L}(Y, X, \beta) + \lambda \left\|\beta\right\|_1.$$

The Poisson count is relative to a unit "exposure" time t_i and the expected count would be $\frac{\mu}{t}$:

$$\log(\mu(x)) = \log(t) + \sum_{i=1}^{n} X_i^\top \beta.$$

Regularized Gamma Model

For continuous responses (severity) consider a gamma GLM. The log–link is used to get a multiplicative model even though it is not the canonical link. The negative log-likelihood is then

$$-\sum_{i=1}^{n}\mathcal{L}(Y_i,X_i;\beta) = -(\alpha-1)\sum_{i=1}^{n}\log X_i + \frac{X_i}{\nu} + \alpha\log\nu + \log\Gamma(\alpha).$$

Adding the Lasso ℓ_1 penalty results in

$$\hat{\beta} = \arg\min_{\beta}\mathcal{L}(Y,X;\beta) + \lambda\left\|\beta\right\|_1.$$

The parameters α and ν can be determined by matching moments from the data as $\alpha = \frac{mean^2}{var}$ and $\nu = \frac{var}{mean}$.

FISTA (Beck and Teboulle, 2009, JIST)

Fast iterative soft thresholding algorithm (FISTA)

The FISTA is basically a generalized gradient method with a first order method of Nesterov style acceleration. It is used to solve the Lasso estimation problem. The algorithm can be applied to any objective function as long as the loss function is convex and differentiable.

At each iteration it takes a step of size s in the direction of the gradient to solve the majorization minimization scheme: $M(\beta) =$

$$\mathcal{L}(Y, X; \beta_0) + (\beta - \beta_0)^\top g(\beta_0) + \frac{1}{2s} \|\beta - \beta_0\|_2^2 + \lambda \sum_{j=1}^{p-1} \|\beta_i\|_2 .$$

3.1 Simulation Study

A simulation study was conducted to test how well Group–Lasso Interaction Network retrieves interactions.

- 10 variables which consist of 7 continuous and 3 categorical with different levels.
- Response was simulated with the 10 variables, 10 interactions and some random noise.
- Using 10-fold cross-validation the best model was chosen to minimize the cross-validation error.
- Model fit with the grid of 50 λ values and the corresponding number of variables and interactions captured.

Simulation Study

The table shows the number of coefficients captured at each λ value.

Fit	λ	ObjValue	Cat	Cot	CatCat	CotCot	CatCot
1	3.07e-03	4.310	0	0	0	0	0
2	2.80e-03	4.300	0	3	0	0	0
3	2.54e-03	4.270	0	4	0	0	0
.	..	..	..	..	...	..	...
6	1.92e-03	4.080	0	5	0	1	0
.	..	..	..	..	...	..	...
49	3.37e-05	0.646	3	7	1	4	5
50	3.07e-05	0.633	3	7	1	4	5

Table: Example of the Glinternet Output

Simulation Study

Running a 10-fold cross-validation with errors reveals that the model with the lowest λ value with all 10 variables and 10 interactions is the optimal one.

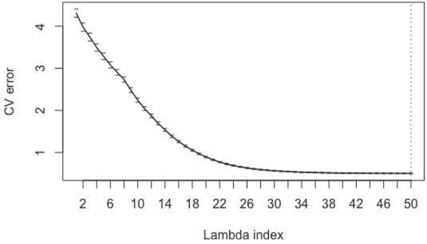


Figure: Cross-validation Error for Simulation Study

Simulation Study

The same study was repeated for the *glinternet*2 for the gamma distribution.

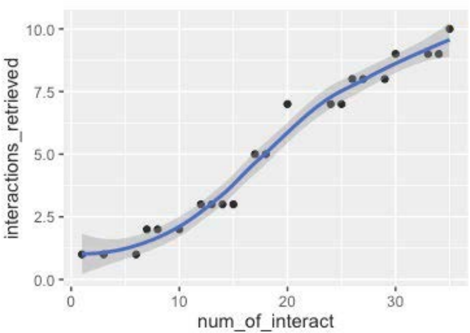


Figure: Discovery Rate in Glinternet2

3.2 Example 1: Singapore Insurance Data

The goal

Understand how driver characteristics affect the accident experience with an emphasis on variables' interactions. For pricing actuaries to understand these relationships so that they can charge the right price for the risk they cover.

Singapore Insurance Data

The model is fitted using the *glinternet*2 function in R, with the Poisson family to **model claim frequency**.

The data is split into **training** and **testing** set. The training set is used to fit the model and then the test set is used for model validation.

A **comparison** is performed on a Poisson GLM, Lasso GLM and a gradient boosting model (GBM) with Poisson loss.

Singapore Insurance Data Results - Lasso GLM

The Figure shows how the Poisson deviance changes with the number of variables included.

GLM vs. Lasso GLM

The model selected the right variables to capture the same signal with less variables.

Mean Square Error for Model Comparison

A **higher penalty** value, which would lead to **lower number** of covariates in the model, thus returns **lower** error rates.

Group-Lasso Interaction Network Results

Table shows the *glinternet*2 for a Poisson Fit with 5 Variables.

Fit	λ	ObjValue	Cat	Cot	CatCat	CotCot	CatCot
1	4.68e-04	-Inf	0	0	0	0	0
2	4.15e-04	3.84e+13	0	1	0	0	0
..	.....	......	......	......	.....	.....	....
6	2.56e-04	3.84e+13	2	1	0	1	2
7	2.26e-04	3.84e+13	2	2	0	1	3
..	.....	......	......	......	.....	.....	....
19	5.29e-05	3.84e+13	3	2	3	1	4
20	4.68e-05	3.84e+13	3	2	3	1	5

The algorithm has retrieved a maximum of 9 interactions with the order of inclusion.

Group-Lasso Interaction Network Results

Order of inclusion in the model as the penalty value decreases.

Glinternet2: Interactions Detected
1. No Claim Discount × Driver Age
2. Gender × Driver Age
3. Driver Age × Vehicle Age
4. Vehicle Type × Driver Age
5. Gender × Vehicle Type
6. Vehicle Type × No Claim Discount
7. Vehicle Type × Vehicle Age
8. Gender × No Claim Discount
9. Gender × Vehicle Age

Group-Lasso Interaction Network Results

Gradient Boosting Results

Running a gradient boosting model with 3000 tree splits, shows

- The most significant variable is Vehicle Type with relative influence of 24.83%.
- The 5 main variables were ranked highest influence.
- All 14 variables have a non-zero influence.
- Looking at the gains curve for GBM, only 4 prediction groups are recognizable.

3.3 Example 2: Ontario Collision Data

The *glinternet* was applied to a subset of the Ontario collision coverage data from a large Canadian insurer for frequency and severity modeling.

Example 2: Group-Lasso Interaction Network Frequency Results

Example 2: Group-Lasso Interaction Network Severity Results

- Same decreasing trend
- The mean predicted responses over-estimate
- The model does not differentiate much between predictions
- (1) the model was fitted to a much smaller dataset (only policies that did file a claim).
- (2) due to limitations of the algorithm in terms of memory.

Example 2: Model Comparison

Comparisons between a GLM, GLMNET and Glinernet2 for a claim frequency model is conducted, with *glinternet*2 having lowest train and test sets errors.

Conclusions

- Using regularization methods for high-dimensional data analysis is necessary to avoid overfitting and helps in variable selection without losing in model predictability.
- Results of the fitted models for out of sample predictions reveal that the model with fewer variables (Lasso vs. GLM) can capture the same signal and generalizes better.
- Linear models are not always sufficiently discerning yet the linearity characteristic is not always undesirable when trying to capture linear signals.
- Linear and generalized linear models do not capture nonlinearities and interactions among variables.

Conclusions

- Group-Lasso Interaction Network combines the detection of non-linear effects and the advantages of linear models.
- Results obtained from the group-Lasso interaction network model are able to capture linear and non-linear effects while performing variable selection and improving predictability.
- Methods based on a machine learning techniques can add value to the limitations of linear models.
- Multivariate analytical techniques focus on individual level data, so that estimates of risks are more granular.

Motivation

Thank you

Bibliography

Beck, A. and Teboulle, M. (2009) “A fast iterative shrinkage thresholding algorithm for linear inverse problems”, *Journal of Imaging Science and Technology*, **2**, 183-202.

Hastie, T. and Lim, M. (2015) “Learning interactions via hierarchical group-lasso regularization”, *Journal of Computational and Graphical Statistics*, **3**, 627-654.

McCullagh, P. and Nelder, J.A. (1989) “Generalized Linear Models”, *Chapman and Hall*.

Tibshirani, R. (1996) “Regression Shrinkage and Selection via the Lasso”, *Journal of Royal Statistical Society*, **58**, 267-288.