# Ratemaking application of Bayesian LASSO with conjugate hyperprior

Himchan Jeong and Emiliano A. Valdez

University of Connecticut

Actuarial Science Seminar
Department of Mathematics
University of Illinois at Urbana-Champaign

26 October 2018

# Outline of talk

# Regularization or least squares penalty

- $L_q$ penalty function:

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ ||Y - X\beta||^2 + \lambda ||\beta||_q \right\},$$

  where $\lambda$ is the regularization or penalty parameter and $||\beta||_q = \sum_{j=1}^{p} |\beta_j|^q$.
- Special cases include:
  - LASSO (Least Absolute Shrinkage and Selection Operator): $q = 1$
  - Ridge regression: $q = 2$
- Interpretation is to penalize unreasonable values of $\beta$.
- LASSO optimization problem:

$$\min_{\beta} \left\{ ||Y - X\beta||^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| = ||\beta||_1 \leq t$$

- See Tibshirani (1996)

# A motivation for regularization: correlated predictors

- Let $y$ be a response variable with potential predictors $x_1$, $x_2$, and $x_3$ and consider the case when predictors are highly correlated.

```
> x1 <- rnorm(50); x2 <- rnorm(50,mean=x1,sd=0.05); x3 <- rnorm(50,mean=-x1,sd=0.02)
> y <- rnorm(50,mean=-2+x1+x2-2*x3); x <- data.frame(x1,x2,x3); x <- as.matrix(x)
> # correlation matrix
> upper
        x1      x2 x3
x1       1
x2  0.9984       1
x3 -0.9997 -0.9982  1
```

- Fitting the least squares regression:

```
> coef(lm(y~x1+x2+x3))
(Intercept)          x1          x2          x3
 -2.3347410 -16.5839237   0.2353327 -19.9617757
```

- Fitting ridge regression and lasso:

```
> library(glmnet)

> lm.ridge <- glmnet(x,y,alpha=0,lambda=0.1,standardize=FALSE); t(coef(lm.ridge))
1 x 4 sparse Matrix of class "dgCMatrix"
   (Intercept)       x1       x2        x3
s0   -2.359547 1.114166 1.104729 -1.356508

> lm.lasso <- glmnet(x,y,alpha=1,lambda=0.1,standardize=FALSE); t(coef(lm.lasso))
1 x 4 sparse Matrix of class "dgCMatrix"
   (Intercept) x1 x2        x3
s0   -2.381575  .  . -3.496807
```

# Bayesian interpretation of LASSO (Naive)

Park and Casella (2008) demonstrated that we may interpret LASSO in a Bayesian framework as follows:

$$Y|\beta \sim N(X\beta, \sigma^2 I_n), \quad \beta_i|\lambda \sim \text{Laplace}(0, 2/\lambda)$$

so that $p(\beta_i|\lambda) = \frac{\lambda}{4} e^{-\lambda|\beta_i|}$.

According to this specification, we may write out the likelihood for $\beta$ as

$$L(\beta|Y, X, \lambda) \ \propto \ \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - X_i\beta)^2\right] - \lambda||\beta||_1\right)$$

and the log-likelihood as

$$\ell(\beta|Y, X, \lambda) = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - X_i\beta)^2\right] - \lambda||\beta||_1 + \text{Constant.}$$

# Bayesian LASSO with conjugate hyperprior

- Choice of the optimal $\lambda$ is critical in penalized regression.
- Here, let us assume that

$$Y|\beta \sim N(X\beta, \sigma^2 I_n),$$

$$\beta_j|\lambda_j \sim \text{Laplace}(0, 2/\lambda_j), \quad \lambda_j|r \overset{i.i.d.}{\sim} \text{Gamma}(r/\sigma^2 - 1, 1).$$

- In other words, the 'hyperprior' of $\lambda$ follows a gamma distribution so that $p(\lambda|r) = \lambda^{(r/\sigma^2)-p-1} e^{-\lambda}/\Gamma(r/\sigma^2 - p)$, then we have

$$L(\beta, \lambda_1, \ldots, \lambda_p|Y, X, r) \;\propto\; \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - X_i\beta)^2\right]\right) \times \prod_{j=1}^{p} \exp\left(-\lambda_j\left[|\beta_j| + 1\right]\right) \lambda_j^{r/\sigma^2 - 1}.$$

# Log adjusted absolute deviation (LAAD) penalty

Integrating out the $\lambda$ and taking the log of the likelihood, we get

$$\ell(\beta|Y, X, r) = -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - X_i\beta)^2 + 2r\sum_{j=1}^{p}\log(1+|\beta_j|)\right) + \text{Const.}$$

Therefore, we have a new formulation for our penalized least squares problem. This gives rise to what we call LAAD penalty function:

$$||\beta||_L = \sum_{j=1}^{p}\log(1+|\beta_j|)$$

so that

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\, ||y - X\beta||^2 + 2r||\beta||_L.$$

# Analytic solution for the univariate case

To understand the characteristics of the new penalty, consider the simple example when $X'X = I$, in other words, design matrix is orthonormal so that it is enough to solve the following:

$$\widehat{\theta}_j = \underset{\theta_j}{\operatorname{argmin}} \frac{1}{2}(z_j - \theta_j)^2 + r\log(1 + |\theta_j|).$$

By setting $\ell(\theta|r,z) = \frac{1}{2}(z-\theta)^2 + r\log(1+|\theta|)$, then we can show that minimizer will be given as $\widehat{\theta} = \theta^* \mathbb{1}_{\{|z| \geq z^*(r) \vee r)\}}$ where $z^*(r)$ is the unique solution of

$$\Delta(z|r) = \frac{1}{2}(\theta^*)^2 - \theta^* z + r\log(1+|\theta^*|) = 0,$$

$$\theta^* = \frac{1}{2}(z + \mathsf{sgn}(z)\left[\sqrt{(|z|-1)^2 + 4|z| - 4r} - 1\right].$$

Note that $\widehat{\theta}$ converges to $z$ as $|z|$ tends to $\infty$.

# Sketch of the proof

We have $\widehat{\theta} \times z \geq 0$ so we start from the case that $z$ is nonnegative number and we have the following;

$$\ell'(\theta|r,z) = (\theta - z) + \frac{r}{1+\theta},\ \ell''(\theta|r,z) = 1 - \frac{r}{(1+\theta)^2},$$
$$\ell'(\theta^*) = 0 \Leftrightarrow \theta^* = \frac{z-1}{2} + \frac{\sqrt{(z-1)^2 + 4z - 4r}}{2}$$

Case (1) $z \geq r \Rightarrow \ell''(\theta^*|r,z) > 0$ so that $\theta^*$ is the local minimum. Moreover, $\ell'(0|r,z) \leq 0$ implies $\theta^*$ is the global minimum.

Case (2) $z < r, z < 1 \Rightarrow \theta^* < 0$ so that $\ell'(\theta|r,z) > 0\ \forall\ \theta \geq 0$. Therefore, $\ell(\theta|r,z)$ strictly increasing and $\widehat{\theta} = 0$.

Case (3) $r \geq (\frac{z+1}{2})^2 \Rightarrow$ in this case, $\theta^* \notin \mathbb{R}$. Moreover, $(\frac{z+1}{2})^2 \geq z$, $\ell'(0|r,z) = r - z \geq 0$ and $\ell'(\theta|r,z) > 0\ \forall\ \theta > 0$. Therefore, $\widehat{\theta} = 0$.

UCONN

# Contour map of $\widehat{\theta}$

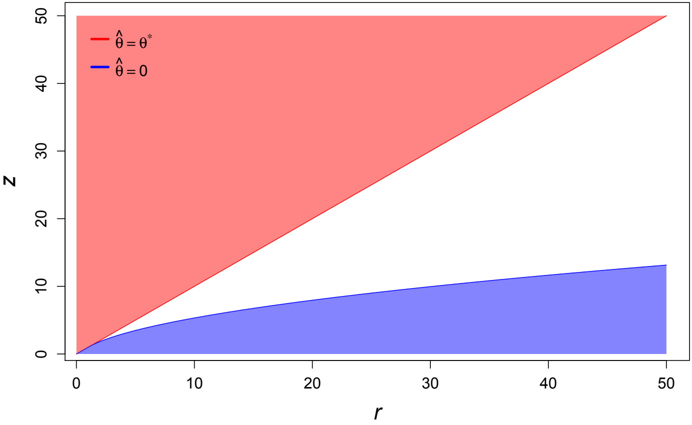


Figure 1: Distribution of the optimizer for the three cases

# - continued

Case (4) $1 \leq z < r < (\frac{z+1}{2})^2 \Rightarrow$ First, we show that $\ell''(\theta^*|r,z) > 0$ so that $\theta^*$ is a local minimum of $\ell(\theta|r,z)$ and $\widehat{\theta}$ would be either $\theta^*$ or $0$. In this case, we compute $\Delta(z|r) = \ell(\theta^*|r,z) - \ell(0|r,z)$ and

$$\widehat{\theta} = \begin{cases} \theta^* \ , & \text{if } \Delta(z|r) < 0 \\ 0 & \text{if } \Delta(z|r) > 0 \end{cases} ,$$

$$\Delta'(z|r) = \left(\theta^* - z + \frac{r}{1+\theta^*}\right)\frac{\partial \theta^*}{\partial z} - \theta^* = -\theta^* < 0.$$

Thus, $\Delta(z|r)$ is strictly decreasing w.r.t. $z$ and $\Delta(z|r) = 0$ has a unique solution because

$$\Delta(z|r) < 0 \Leftrightarrow \widehat{\theta} = \theta^*, \text{if } z = r$$

and

$$\Delta(z|r) > 0 \Leftrightarrow \widehat{\theta} = 0, \text{if } z = 2\sqrt{r} - 1.$$

# - continued

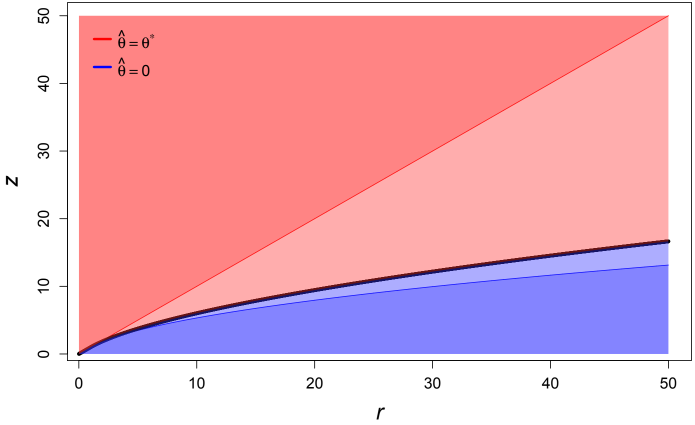


Figure 2: Distribution of the optimizer for all the cases

# Estimate behavior

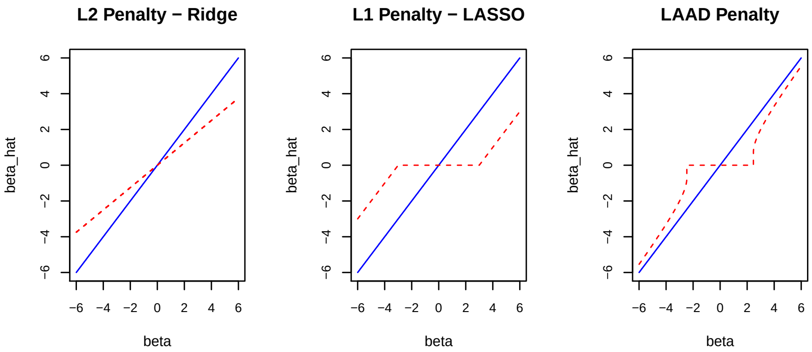


Figure 3: Estimate behavior for different penalties

# Penalty regions

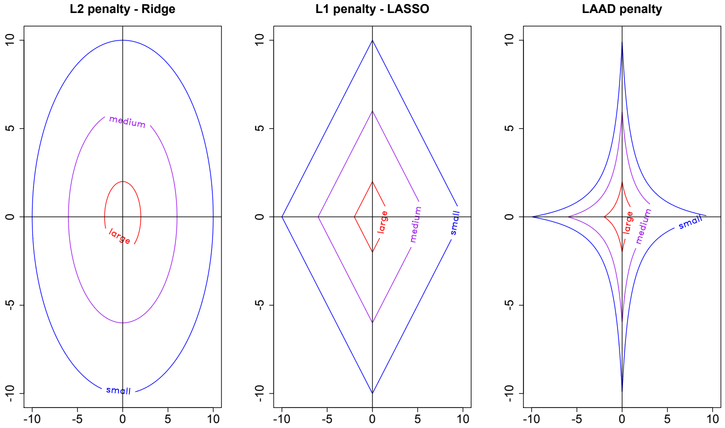


Figure 4: Penalty regions for different penalties

# Coordinate descent algorithm

- Model estimation is an optimization problem
- Coordinate descent algorithm: Luo and Tseng (1992), Wu and Lange (2008)
  - start with an initial estimate and then successively optimize along each coordinate or blocks of coordinates

Do Loop

$$y_{(1)} = y - \sum_{j=2}^{p} X_j \beta_j^{[old]}$$
$$\beta_{(1)}^{[new]} = \mathbf{1}' y_{(1)} / n$$

for ($j$ in $2:p$) {

$$y_{(j)} = y_{(j-1)} - X_{j-1} \beta_{j-1}^{[new]} + X_j \beta_j^{[old]}$$
$$z_{(j)} = X_j' y_{(j)}$$
$$\beta_{(j)}^{[new]} = \text{argmin}[0, \theta^*(z_{(j)}, r)]$$

}

Until $\frac{||\beta^{[new]} - \beta^{[old]}||}{||\beta^{[new]}||} < \epsilon$

# The frequency-severity two-part model

- For ratemaking, e.g., in auto insurance, we have to predict the aggregate claims $S = \sum_{k=1}^{n} C_k$.
- Traditional approach is

$$\text{Cost of Claims} = \text{Frequency} \times \text{Average Severity}$$

- The joint density of the number of claims and the average claim size can be decomposed as

$$\begin{aligned} f(N, \overline{C}|\mathbf{x}) &= f(N|\mathbf{x}) \times f(\overline{C}|N, \mathbf{x}) \\ \text{joint} &= \text{frequency} \times \text{conditional severity.} \end{aligned}$$

- This natural decomposition allows us to investigate/model each component separately and it does not preclude us from assuming $N$ and $\overline{C}$ are independent.

# The two-part model specifications

- For the frequency component:
  - $N$ is assumed to follow a Poisson distribution so that $\mathbb{E}\left[N|\mathbf{x}\right] = e^{\mathbf{x}\alpha}$.
  - typically used in practice
  - penalized log-likelihood for estimation
- For the average severity component $\overline{C}|N$:
  - We use lognormal distribution so that $\mathbb{E}\left[\log \overline{C}|N, \mathbf{x}\right] = \mathbf{x}\beta$ and $\mathbb{V}ar\left(\log \overline{C}|N, \mathbf{x}\right) = \sigma^2$.
  - penalized least squares for estimation
- For both components: the log-adjusted absolute deviation (LAAD) penalty is used:

$$||\beta||_L = \sum_{j=1}^{p} \log(1 + |\beta_j|)$$

# Penalized estimation for the two-part model

- For the frequency part, $\widehat{\alpha}$ from the penalized likelihood is given as follows:

$$\widehat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \left( -\sum_{i=1}^{n} (n_{it} X_{it}\alpha - e^{X_{it}\alpha}) \right) + r||\alpha||_L .$$

- For the average severity part, $\widehat{\beta}$ from the penalized likelihood is given as follows:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} || \log \overline{C} - X\beta||^2 + r||\beta||_L$$

# Observable policy characteristics used as covariates

| Categorical variables | Description | | Proportions | |
|---|---|---|---|---|
| VehType | Type of insured vehicle: | Car | 97.75% | |
| | | MotorBike | 1.64% | |
| | | Others | 0.6% | |
| Gender | Insured's sex: | Male = 1 | 82.78% | |
| | | Female = 0 | 17.22% | |
| Cover Code | Type of insurance cover: | Comprehensive = 1 | 74.57% | |
| | | Others = 0 | 25.43% | |
| Continuous variables | | Minimum | Mean | Maximum |
| VehCapa | Insured vehicle's capacity in cc | 10.00 | 1560.91 | 9990.00 |
| VehAge | Age of vehicle in years | -1.00 | 7.84 | 46.00 |
| Age | The policyholder's issue age | 17.00 | 39.98 | 99.00 |
| NCD | No Claim Discount in % | 0.00 | 23.88 | 50.00 |

- Singapore insurance data (1993–2000: Training set, 2001: Test set)
- 208,107 of aggregated total number of observations observed on training set.

# Covariates for frequency estimation

- Original: **VTypeCar, VTypeMBike, logVehCapa**, VehAge, **SexM, Comp, NCD, Age**, Age2, Age3
- Interactions: **MlogVehCapa, MVehAge, MAge, MAge2, MAge3**

Even after adding the interaction terms, almost every covariate is significant for frequency estimation.

# Estimation results: frequency component

| | Reduced model | Full model | Naive LASSO | Bayesian LASSO |
|---|---|---|---|---|
| (Intercept) | -0.740957 | -3.258836 | -1.792429 | -1.791314 |
| VTypeCar | -0.585375 | -0.566404 | -0.000077 | -0.000254 |
| VTypeMBike | -2.085336 | -2.102879 | -0.000873 | -0.000102 |
| logVehCapa | 0.214138 | 0.334423 | 0.000039 | 0.000001 |
| VehAge | -0.009061 | -0.000031 | -0.000020 | -0.000004 |
| SexM | 0.105565 | 3.166574 | 0.000531 | 0.000341 |
| Comp | 0.910381 | 0.909633 | 0.000517 | 0.000377 |
| Age | -0.150428 | -0.055286 | 0.000005 | 0.000005 |
| Age2 | 0.002705 | 0.000936 | 0.000000 | 0.000000 |
| Age3 | -0.000015 | -0.000005 | 0.000000 | 0.000000 |
| NCD | -0.009976 | -0.009943 | -0.000004 | 0.000000 |
| MlogVehCapa | | -0.140558 | 0.000100 | 0.000082 |
| MVehAge | | -0.010818 | 0.000041 | 0.000018 |
| MAge | | -0.119687 | 0.000007 | 0.000003 |
| MAge2 | | 0.002232 | 0.000000 | 0.000000 |
| MAge3 | | -0.000013 | 0.000000 | 0.000000 |
| Loglikelihood | -54811.696563 | -54796.659753 | -55542.756271 | -55547.849702 |
| AIC | 109645.393127 | 109625.319506 | 111117.512543 | 111127.699405 |
| BIC | 109758.097011 | 109789.252428 | 111281.445465 | 111291.632327 |

# Tuning the frequency penalty parameter

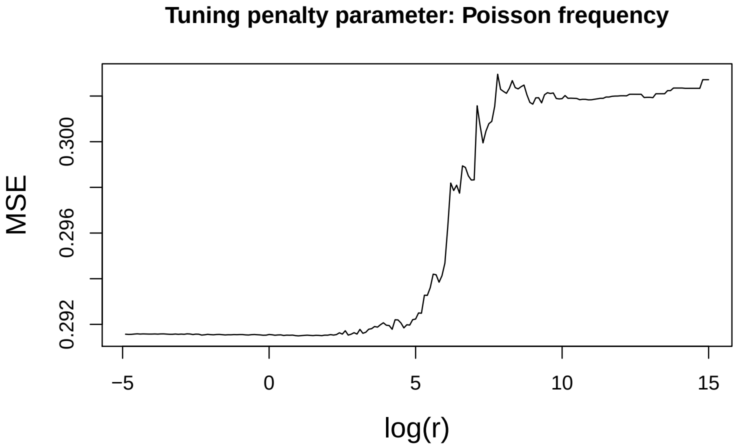


Figure 5: Tuning the penalty parameter: frequency component

# Validation results: Poisson frequency

Comparing the MAE and MSE for the various models

| | Reduced model | Full model | Naive LASSO | Bayesian LASSO |
|---|---|---|---|---|
| MAE | 0.13343 | 0.13344 | 0.13883 | 0.13890 |
| MSE | 0.27873 | 0.27876 | 0.28043 | 0.28044 |

# Frequency validation results - Gini index

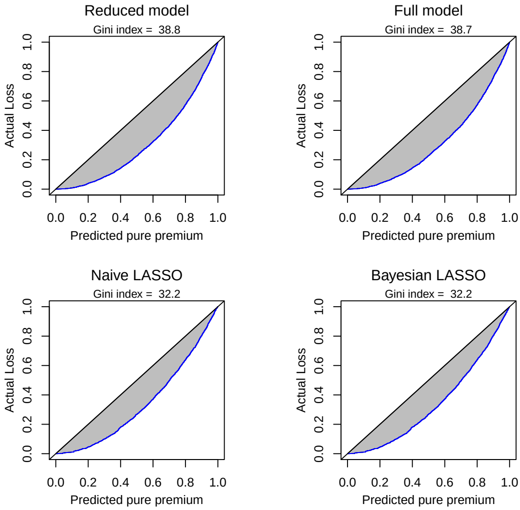


Figure 6: Gini indices for the Poisson frequency models

# Covariates for average severity estimation

- Original: **VTypeCar, VTypeMBike, logVehCapa, VehAge, Comp, NCD, Age, Age2, Age3, Count**, SexM
- Interactions: **FintNCD, FintVehAge, FintComp**, FintVTypeCar, FintlogVehCapa, FintSexM, FintAge, FintAge2, FintAge3

After adding the interaction terms, only some covariates are significant for the average severity estimation.

# Estimation results: average severity component

| | Reduced model | Full model | Naive LASSO | Bayesian LASSO |
|---|---|---|---|---|
| (Intercept) | 7.153653 | 7.444373 | 7.673586 | 7.533879 |
| VTypeCar | -0.613385 | -0.718301 | | -0.579297 |
| VTypeMBike | -0.699988 | -0.804159 | -0.336990 | -0.680890 |
| logVehCapa | 0.226679 | 0.238242 | | 0.221583 |
| VehAge | -0.010845 | -0.012367 | -0.014867 | -0.011665 |
| SexM | -0.022395 | -0.034773 | | -0.028442 |
| Comp | 0.321747 | 0.279826 | 0.285615 | 0.292984 |
| Age | -0.072443 | -0.068476 | | -0.077812 |
| Age2 | 0.001406 | 0.001330 | | 0.001538 |
| Age3 | -0.000008 | -0.000008 | 0.000000 | -0.000009 |
| NCD | -0.002662 | -0.002899 | -0.003135 | -0.002876 |
| Count | 0.725876 | 0.453060 | 0.208421 | 0.451539 |
| Fint_VTypeCar | | 1.144692 | 0.009752 | |
| Fint_logVehCapa | | -0.151809 | | |
| Fint_VehAge | | 0.019121 | 0.012669 | 0.010154 |
| Fint_SexM | | 0.115636 | 0.046500 | 0.074754 |
| Fint_Comp | | 0.642902 | 0.605135 | 0.481527 |
| Fint_Age | | -0.037246 | | -0.010455 |
| Fint_Age2 | | 0.000723 | | |
| Fint_Age3 | | -0.000004 | 0.000000 | 0.000001 |
| Fint_NCD | | 0.003337 | 0.002222 | 0.003331 |
| Loglikelihood | -21589.565106 | -21569.896261 | -22832.531841 | -22049.633473 |
| AIC | 43205.130212 | 43183.792522 | 45685.063682 | 44141.266945 |
| BIC | 43303.550718 | 43350.350300 | 45760.771763 | 44300.253916 |

# Tuning the average severity penalty parameter

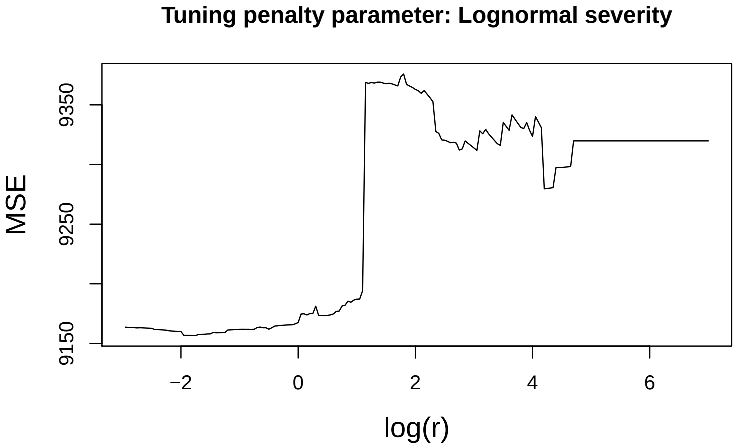


Figure 7: Tuning the penalty parameter

# Validation results: Lognormal average severity

Comparing the MAE and MSE for the various models

| | Reduced model | Full model | Naive LASSO | Bayesian LASSO |
|---|---|---|---|---|
| MAE | 3002.512 | 2995.511 | 3112.826 | 2993.567 |
| MSE | 4985.503 | 4970.821 | 5396.835 | 4967.892 |

# Severity validation results - Gini index

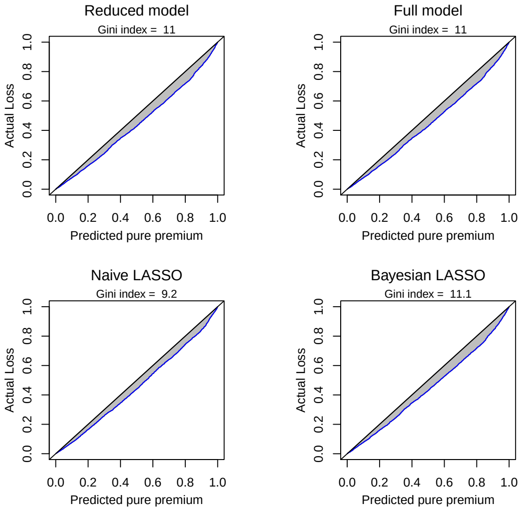


Figure 8: Gini indices for the Lognormal average severity models

# Concluding remarks

- We suggest a model using a hyperprior for the $\lambda$ in the Bayesian LASSO, which yielded a new penalty function with good properties such as variable selection as well as reversion to the true regression coefficients.
- While our proposed LASSO model did not perform well for the frequency component, it was the optimal choice for the average severity component. Note that we could not have enough degree of sparsity from fitting the frequency, but moderate degree of sparsity for fitting the average severity component.
- Compared to Naive LASSO model which uses $L_1$ penalty for regularization, our proposed LASSO model showed better performance with respect to all of the validation measures, such as MSE, MAE, and Gini index, which support the assertion that our proposed model enables variable selection with less bias on the regression coefficient estimate.

# Acknowledgment

- We thank the financial support of the Society of Actuaries through our CAE grant on data mining.
- Thank you to all present here.