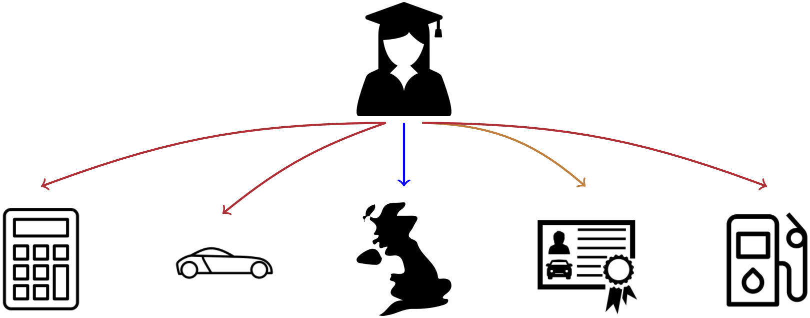
Sparsity with multi-type Lasso penalties

Tom Reynkens
Joint work with S. Devriendt, K. Antonio, E. W. Frees, R. Verbelen
Insurance Data Science, London, 16 July 2018

Motivation



Claim frequency and claim severity

as function of

binary / numeric $\sim$ ordinal / nominal / spatial

features

Research questions

- Generalised Linear Models (GLMs) for frequency ($\sim$ Poisson) and severity ($\sim$ Gamma).
- How to:
  (1) select variables or features?
  (2) cluster (or bin or fuse) levels within a variable?
  age groups / postal code clusters / clusters of car models
- Procedure should be data driven, scalable to large (big) data.
- End product is interpretable, within actuarial comfort zone.

- Generalised Linear Models (GLMs) for frequency ($\sim$ Poisson) and severity ($\sim$ Gamma).
- How to:
  (1) avoid overfitting with too many variables or levels?
  (2) avoid underfitting with a priori binning/selection?

Sparsity with multi-type Lasso regularised GLMs

Devriendt, Antonio, Frees, Reynkens, Verbelen, 2018 (in progress)

Standard GLM

Fit data as good as possible,
no constraint on parameters.

↕

Regularised GLM

Trade-off between fit and interpretability/sparsity/stability,
constraint on parameters.

- **Less is more** (Hastie, Tibshirani & Wainwright, 2015):

  Sparse model is easier to estimate and interpret than a dense model.

- Regularise:

$$\min_{\beta_0,\boldsymbol{\beta}} \{-\mathcal{L}(\beta_0, \boldsymbol{\beta})\} \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t,$$

or equivalently

$$\min_{\beta_0,\boldsymbol{\beta}} \left\{-\mathcal{L}(\beta_0, \boldsymbol{\beta}) + \lambda \cdot \sum_{j=1}^{p} |\beta_j| \right\}.$$

Shrinks coefficients and even sets some to zero.

- Adjust Lasso regularisation to the type of variable:
  - Determine type (binary / numeric $\sim$ ordinal / nominal / spatial);
  - Allocate suitable penalty.
- For J variables, each with regularisation term $P_j(.)$, we want to optimise:

$$-\mathcal{L}(\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J) + \lambda \cdot \sum_{j=1}^{J} P_j(\boldsymbol{\beta}_j).$$

- Different variable type $\rightarrow$ different penalty budget.

- Binary: *Lasso*

$$w_i|\beta_i|$$

- Ordinal: *Fused Lasso*

$$\sum_i w_i|\beta_{i+1} - \beta_i|.$$

- Nominal: *Generalised Fused Lasso*

$$\sum_{i>k} w_{i,k}|\beta_i - \beta_k|.$$

- Spatial: *Graph-Guided Fused Lasso*

$$\sum_{(i,k)\in\mathcal{G}} w_{i,k}|\beta_i - \beta_k|.$$

Unified GLM framework with multiple penalty types

- Gertheiss & Tutz (2010) and Oelker & Gertheiss (2017):
  - GLMs with various penalties.
  - R package available: `gvcm.cat` (not maintained).
- Local quadratic approximations of penalties and PIRLS:
  - non-exact selection or fusion;
  - computationally intensive.

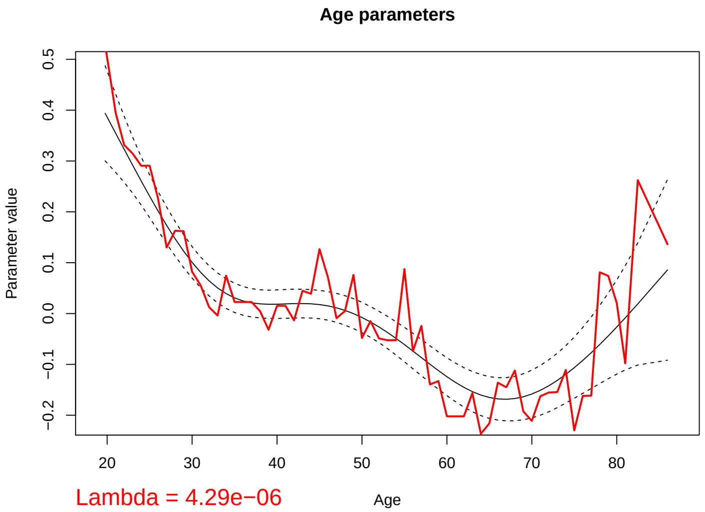
Unified GLM framework with multiple penalty types

- Our contribution:
  - Efficient algorithm (with **proximal operators**);
    - code bottleneck in `C++` (`Rcpp`)
    - efficient linear algebra (`RcppArmadillo`)
    - parallel computations (`parallel`)
  - Scalable to big data (splits into smaller sub-problems);
  - Flexible regularisation
    - penalty takes type of variable into account;
    - works for all popular penalties;
  - No approximations

⇒ `R` **package** under construction.

- Frequency (and severity) information for $n = 163,234$ policyholders.
- 14 variables: binary, ordinal and spatial.
- Exposure modelled as offset.
- Fit Poisson GLM for frequency data with different penalties.
  - $N_i \sim \text{Poisson}(\mu_i)$
  - $\log(\mu_i) = \log(\text{exposure}_i) + \beta_0 + \sum_{j=1}^{14} X_j \boldsymbol{\beta}_j$

$$\mathcal{O}(\beta; \boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n}\mathcal{L}(\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{14}) + \lambda \left(\sum_{j \in \texttt{bin}} |w_j \beta_j| + \sum_{j \in \texttt{ord}} ||\boldsymbol{D}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1 + ||\boldsymbol{G}(\boldsymbol{w}_{\texttt{muni}})\boldsymbol{\beta}_{\texttt{muni}}||_1 \right)$$



Age parameters
Parameter value
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
20
30
40
50
60
70
80
Lambda = 4.29e−06
Age

- Settings:
  - Adaptive (GLM) and standardisation penalty weights $\boldsymbol{w}_i$ for better consistency and predictive performance.
  - Tune λ with 10-fold stratified cross-validation with one standard error rule and deviance as measure.
- Re-estimate the final sparse GLM with standard GLM routines (using 94 parameters instead of 422).

MTPL: ordinal features

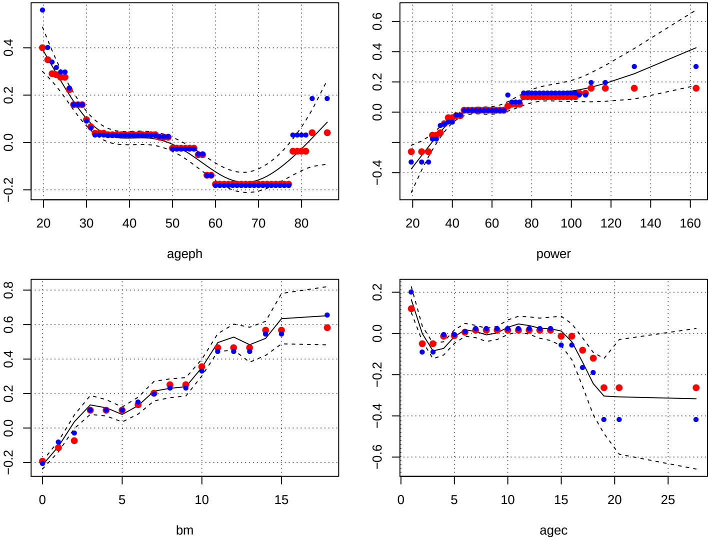


Figure: GAM fit, penalised GLM fit, GLM refit with new clusters.

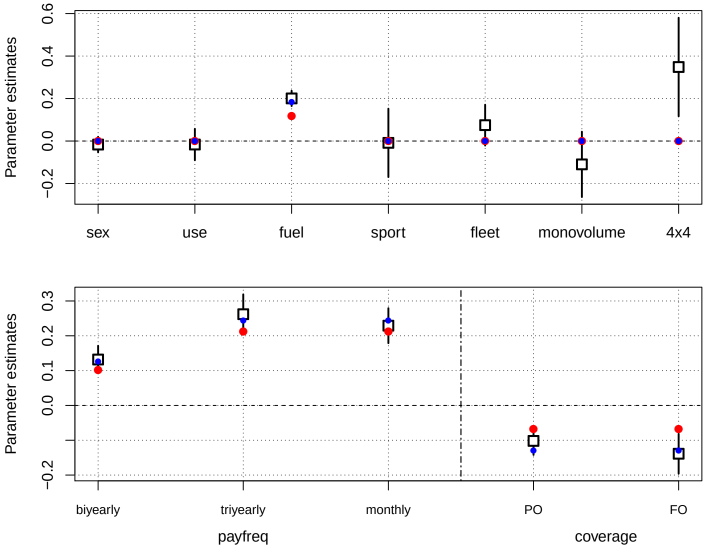
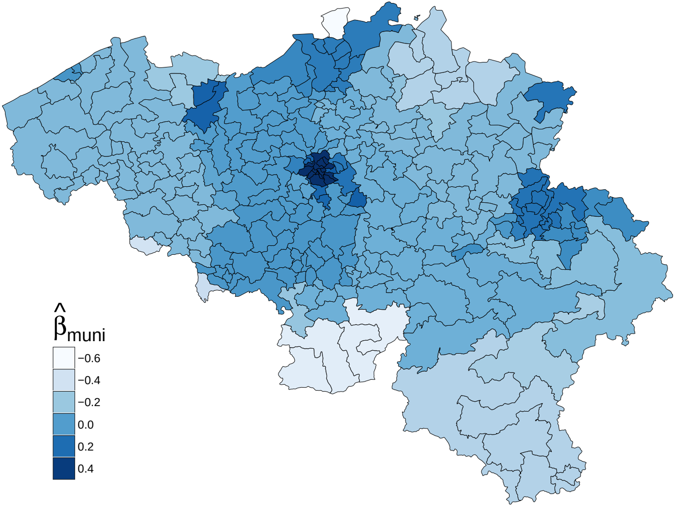


Figure: GAM fit, penalised GLM fit, GLM refit with new clusters.



$\hat{\beta}_{muni}$
−0.6
−0.4
−0.2
0.0
0.2
0.4

- Less is more.
- Flexible regularisation can help predictive modelling.
- General framework with multiple penalty types.
- Efficient algorithm using proximal operators.
- R package and working paper to be finalised.

Ageas Continental Europe

+ Sander Devriendt and colleagues

Contact: tom.reynkens@kuleuven.be
sander.devriendt@kuleuven.be

Henckaerts, R., Antonio, K., Clijsters, M. and Verbelen, R. (2018)
A data driven strategy for the construction of insurance tariff classes.
Scandinavian Actuarial Journal, published online.

Wood, S. (2006)
Generalized additive models: an introduction with R.
Chapman and Hall/CRC Press.

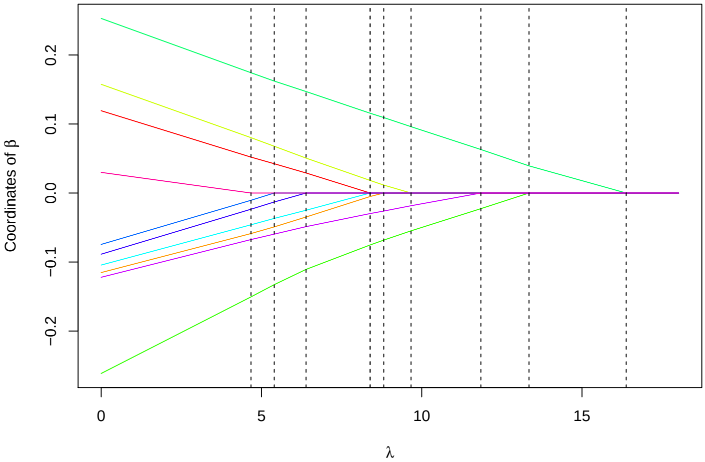
Gertheiss, J. and Tutz, G. (2010).
Sparse modeling of categorial explanatory variables.
The Annals of Applied Statistics, 4(4), 2150-2180.

Oelker, M. and Gertheiss, J. (2017).
A uniform framework for the combination of penalties in generalized structured models.
Advances in Data Analysis and Classification, 11(1),97-120.

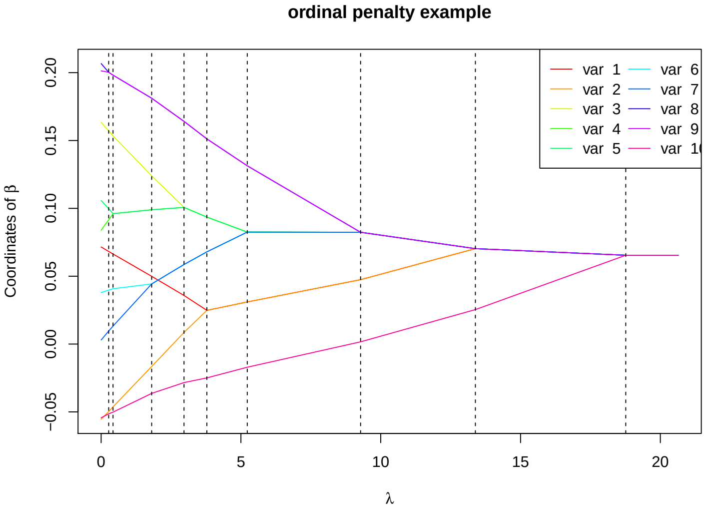
Parikh, N. and Boyd, S. (2013).
Proximal algorithms.
Foundations and Trends in Optimization, 1(3):123-231.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015)
Statistical learning with sparsity: the Lasso and generalizations.
Chapman and Hall/CRC Press.

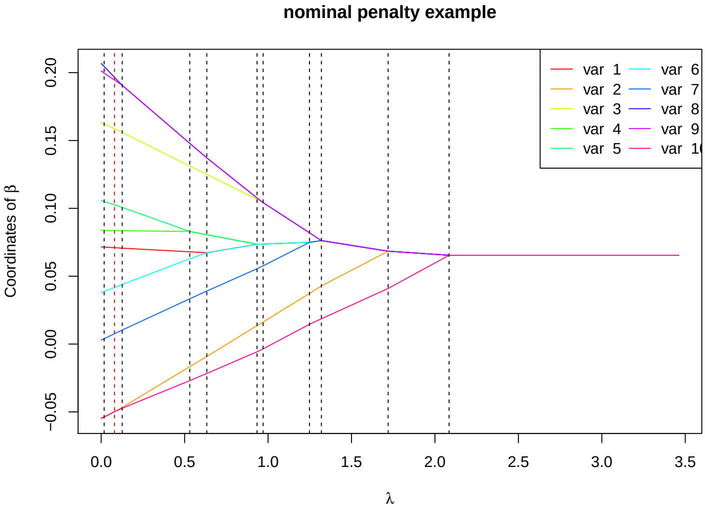
overfitting $\longleftarrow$ λ $\longrightarrow$ underfitting



overfitting ⟵ λ ⟶ underfitting



overfitting $\longleftarrow$ λ $\longrightarrow$ underfitting



$$\hat{\beta} = \min_{\beta} -\mathcal{L}(\beta) + \lambda \cdot \sum_{j=1}^{J} P_j(\beta_j)$$

- Our algorithm:
  - iteratively solve proximal operator (PO)

$$\beta^{k+1} = \min_{\boldsymbol{z}} \frac{1}{2} \left\| \left(\beta^k + s \cdot \nabla \mathcal{L}(\beta^k) \right) - \boldsymbol{z} \right\|_2^2 + \lambda \cdot \sum_{j=1}^{J} P_j(\boldsymbol{z}_j)$$

  - PO splits into $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_J$;
  - solve J smaller POs with specialised algorithms per $P_j(\cdot)$;
  - Improve algorithm with warm starts, acceleration, adaptive restarts, backtracking step size.