IA|BE Data Science Certificate

Module 1 on Foundations of machine learning in actuarial sciences
We shrunk the parameters - Lasso, friends of Lasso and the actuary

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

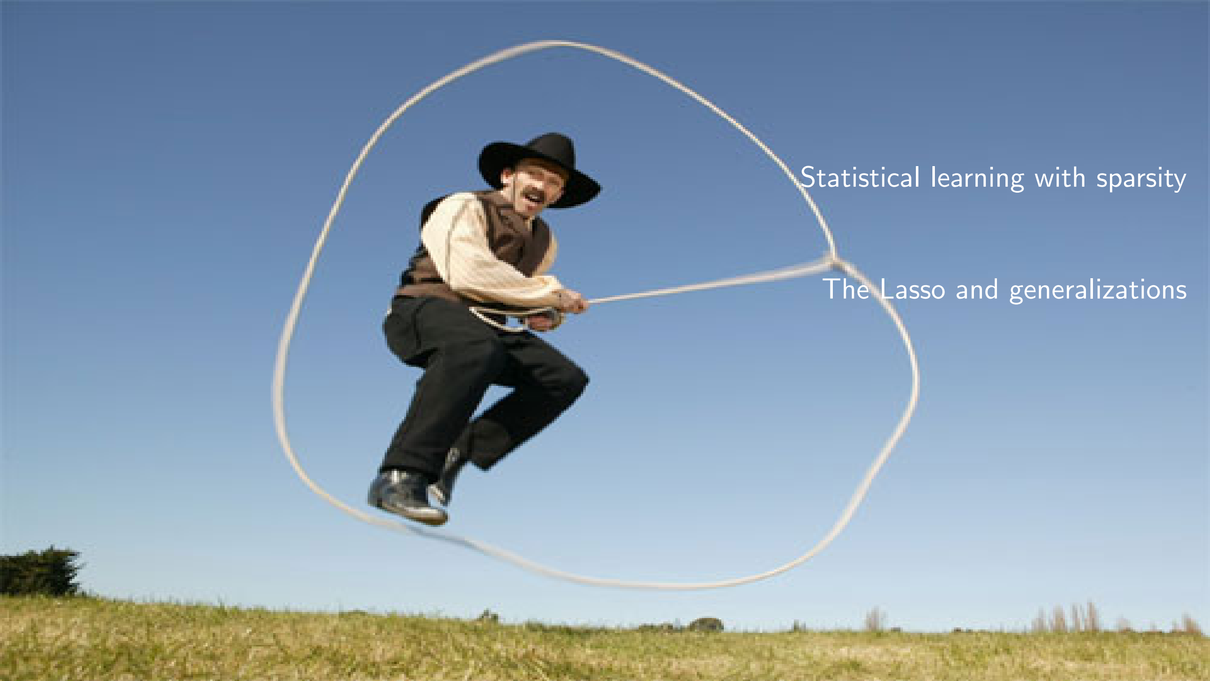
March 10, 2022

Acknowledgement

Some of the figures in this presentation are taken from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Some of the figures in this presentation are taken from *The Elements of Statistical Learning: Data mining, Inference and Prediction* (Springer, 2009) with permission from the authors: T. Hastie, R. Tibshirani and J. Friedman.

Some of the figures in this presentation are taken from *Applied Predictive Modeling* (Springer, 2013) with permission from the authors: M. Kuhn and K. Johnson.

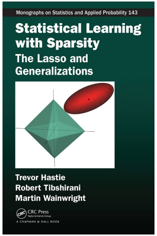


Statistical learning with sparsity

The Lasso and generalizations

Motivation

Sparsity


Monographs on Statistics and Applied Probability 143
Statistical Learning with Sparsity
The Lasso and Generalizations
Trevor Hastie
Robert Tibshirani
Martin Wainwright
CRC Press

- Crucial need to sort through the mass of information and bring it down to its bare essentials.

- One form of simplicity is sparsity.

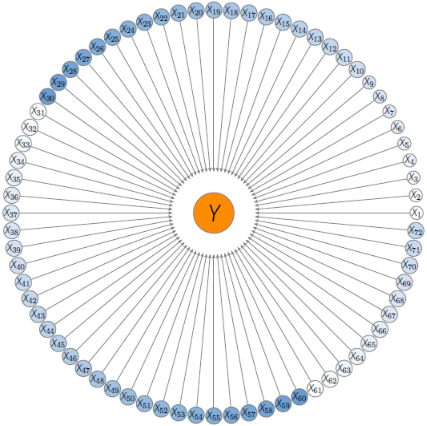
- In a sparse statistical model only a relatively small number of parameters (or predictors) play a role.

- The 'bet on sparsity' principle:

  Use a procedure that does well in sparse problems, since no procedure does well in dense problems.

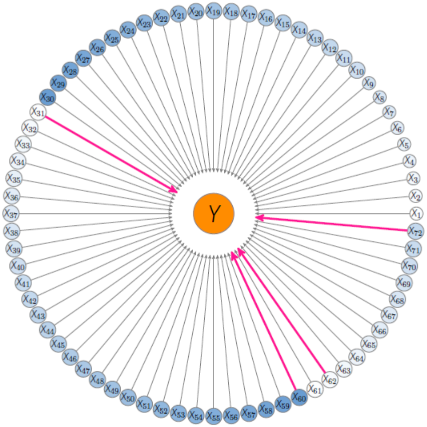
Motivation

Bet on sparsity



Motivation

Bet on sparsity



Motivation

Shrinkage methods

- Picture a predictive modelling problem with p input features $\Rightarrow$ best subset selection strategy to select relevant predictors?

- Alternative strategy:

  - fit a model with all p predictors

  - constrain or regularize the coefficient estimates $\rightsquigarrow$ shrink the coefficient estimates to zero.

- Shrinking introduces bias, but may significantly decrease the variance of the estimates. If the latter effect is larger, this would decrease the test error.

- Some types of shrinkage put some of the coefficients exactly equal to zero!

Ridge (least squares) regression

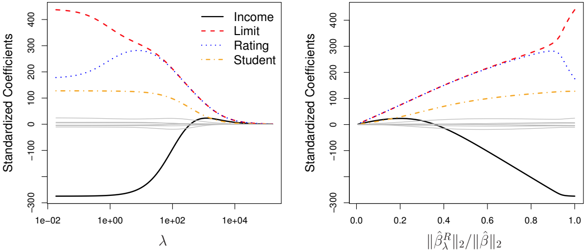
- The least-squares optimization problem

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \min_{\beta_0,\boldsymbol{\beta}} \text{RSS}$$

subject to a 'budget' t constraint

$$\sum_{j=1}^{p} \beta_j^2 \leq t \text{ or } \|\boldsymbol{\beta}\|_2^2 \leq t.$$

Ridge regression



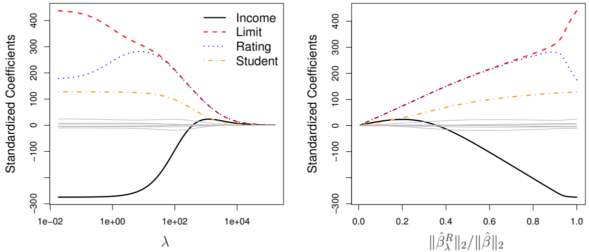
- Dual problem formulation

$$\min_{\beta_0,\boldsymbol{\beta}} \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

with

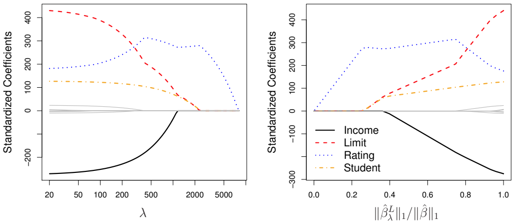
- $\lambda \geq 0$ a tuning parameter and $\lambda \sum_{j=1}^{p} \beta_j^2$ a shrinkage penalty
- with $\lambda = 0$ the least squares estimates result (all $\neq 0$!)
- with $\lambda \to \infty$ coefficients will approach zero.

Ridge regression



- Points of attention:
  - a set of coefficient estimates $\hat{\beta}^R_\lambda$ for each value of λ!
  - we do not shrink the intercept
  - standard least squares coefficients are scale invariant, not the case for ridge regression coefficients!
  - therefore, best to apply ridge regression after centering and standardizing the predictors.

Lasso



- The ridge penalty shrinks all coefficients to zero, but does not set any of them exactly to zero.

- The lasso shrinks coefficient estimates to zero, and performs variable selection

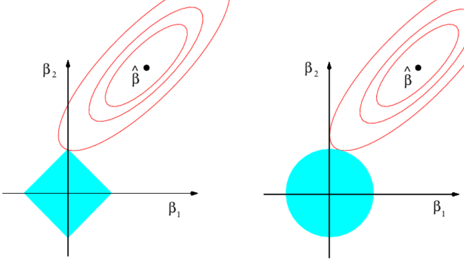
$$\min_{\beta_0,\boldsymbol{\beta}} \text{ RSS subject to } \sum_{j=1}^{p} |\beta_j| \leq t \text{ or } \min_{\beta_0,\boldsymbol{\beta}} \text{ RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

  Thus, lasso uses the ℓ_1 penalty instead of ℓ_2 penalty.

- Lasso is for Least absolute shrinkage and selection operator.

Lasso

Variable selection property



- When $p = 2$:
  - lasso coefficient estimates have smallest RSS out of all points in the diamond

    $$|\beta_1| + |\beta_2| \leq t$$

  - ridge coefficient estimates have smallest RSS out of all points in the circle

    $$\beta_1^2 + \beta_2^2 \leq t$$

  - ellipses (around least-squares $\hat{\beta}$) represent regions of constant RSS
  - since lasso has corners at each of the axes, ellipse will often intersect the constraint region at an axis.

Lasso

Variable selection property

- Recall the best subset selection problem

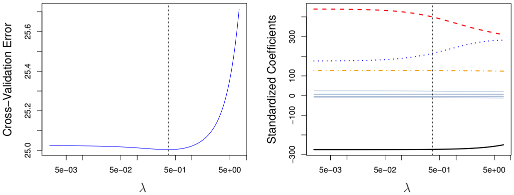
$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} I(\beta_j \neq 0) \leq t.$$

  Solving this problem is computationally infeasible when p is large!

- In general: with the ℓ_q norm of $\boldsymbol{\beta}$ as penalty
  - $q < 1$ the solution is sparse, but the problem is not convex
  - $q > 1$ the problem is convex, but the solution is not sparse.
- The value $q = 1$ is the smallest value that yields a convex problem.
- Convexity, as well as the sparsity assumption, greatly simplifies the computation.

Ridge and Lasso

Selecting the tuning parameter



- Use cross-validation to select a value for λ (or, equivalently, for the budget t).
- Choose a grid of λ values:
  - compute the cross-validation (CV) error for each value of λ
  - select the tuning parameter value for which the CV error is smallest.
- Refit the model using all available observations and the selected value of λ.

Ridge and Lasso

Generalized Linear Model setting

- Minimize

$$\min_{\beta_0,\boldsymbol{\beta}} -\frac{1}{n}\log\mathcal{L}(\beta_0,\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X})+\lambda\|\boldsymbol{\beta}\|_1.$$

  Here $\mathcal{L}$ is the log-likelihood of a GLM.

- Some examples:

Gaussian	$\frac{1}{2\sigma^2}\|\boldsymbol{y}-\beta_0 1-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$
logistic	$\sum_{i=1}^{n} y_i(\beta_0+\boldsymbol{\beta}^t x_i)-\log\left(1+e^{\beta_0+\boldsymbol{\beta}^t x_i}\right)$
Poisson	$\sum_{i=1}^{n} y_i(\beta_0+\boldsymbol{\beta}^t x_i)-e^{\beta_0+\boldsymbol{\beta}^t x_i}.$

Ridge and Lasso

The glmnet package in R

- Family members: (a.o.) gaussian, binomial, poisson.
- Penalties:

$$\lambda P_{\alpha}(\boldsymbol{\beta}) = \lambda \cdot \sum_{j=1}^{p} \left\{ \frac{(1-\alpha)}{2}\beta_j^2 + \alpha|\beta_j| \right\},$$

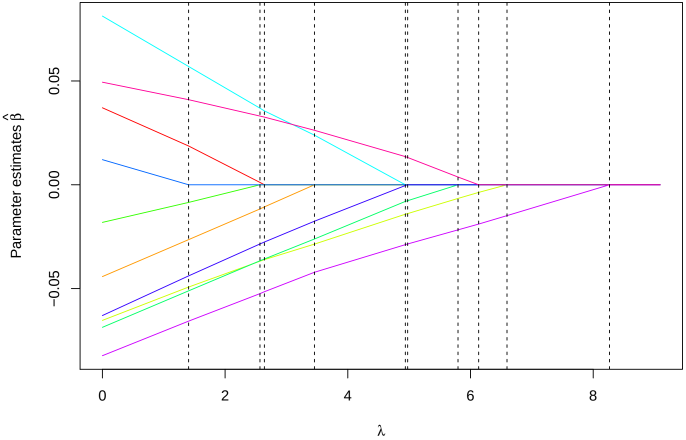
with

- $\alpha \in [0,1]$ the elastic-net parameter (to mix ridge and lasso).

Ridge and Lasso

A typical Lasso plot with `glmnet`

overfitting ⟵ λ ⟶ underfitting



Lasso and friends

- We turn to some useful variations of the basic lasso ℓ_1-penalty:
  - groups of correlated features

    $\rightsquigarrow$ lasso does not perform well, elastic net is better and selects correlated features (or not) together
  - structurally grouped features

    $\rightsquigarrow$ select or omit all within a group together via group lasso
  - neighbouring coefficients to be the same

    $\rightsquigarrow$ fused lasso.

Lasso and friends

Elastic net

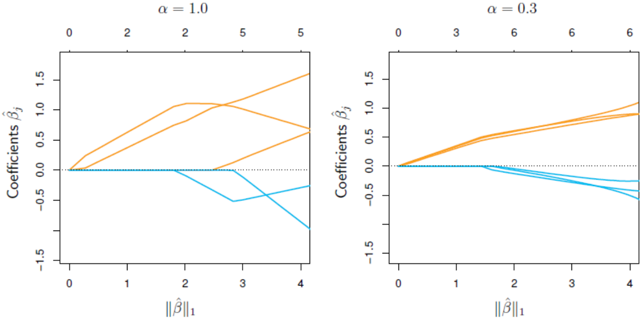


Figure 4.1 *Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the correlated groups are pulled together.*

Lasso and friends

And the actuary ...

- Adjust lasso regularization to the type of risk factor:
  - determine type (nominal / numeric $\sim$ ordinal / spatial)
  - allocate logical penalty.
- Thus, for J risk factors, each with convex regularization term $g_j(.)$, we want to optimize:

$$-\frac{1}{n}\log\mathcal{L}\left(\beta_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_J\right)+\lambda\cdot\sum_{j=1}^{J}g_j\left(\boldsymbol{\beta}_j\right).$$

A multi-type regularized predictive model!

Regularization with multi-type penalty

- Continuous or binary risk factors: lasso

$$g_{\text{Lasso}}(\boldsymbol{\beta}_j) = \sum_i w_{j,i} |\beta_{j,i}|.$$

- Ordinal risk factors: fused lasso

$$g_{\text{fLasso}}(\boldsymbol{\beta}_j) = \sum_i w_{j,i} |\beta_{j,i+1} - \beta_{j,i}| = ||\boldsymbol{D}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1.$$

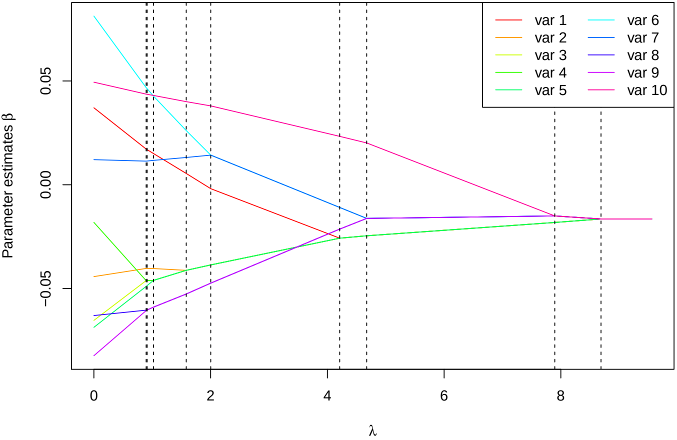
- Nominal risk factors: generalized fused lasso

$$g_{\text{gflasso}} = \sum_{(i,l)\in\mathcal{G}} w_{j,il} |\beta_{j,i} - \beta_{j,l}| = ||\boldsymbol{G}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1.$$

Lasso and friends

Fused Lasso with `genlasso`

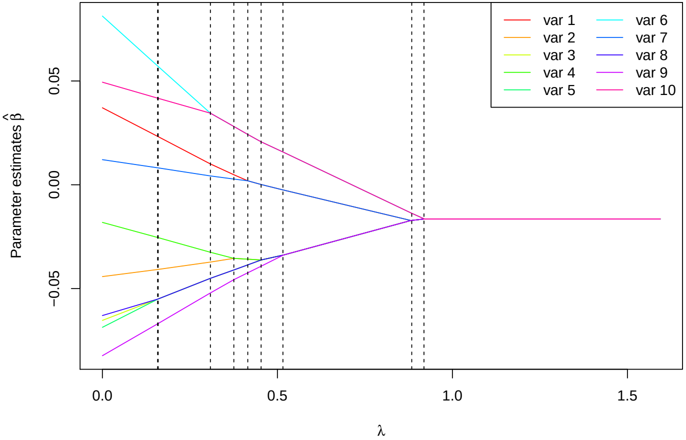
overfitting ⟵ λ ⟶ underfitting



Lasso and friends

Generalized Fused Lasso with `genlasso`

overfitting ⟵ λ ⟶ underfitting



SMuRF

Sparse Multi-type Regularized Feature modeling in Devriendt et al. (2021)

- SMuRF unifies penalty-specific (machine learning) literature with statistical (or: actuarial) literature!
- Efficient algorithm (with proximal operators).
- Scalable to large (big) data (splits into smaller sub-problems).
- Flexible regularization
  - penalty takes type of risk factor into account
  - works for all popular penalties.

MTPL data: Poisson with multi-type penalty

- Model claim frequencies with regularized Poisson GLM

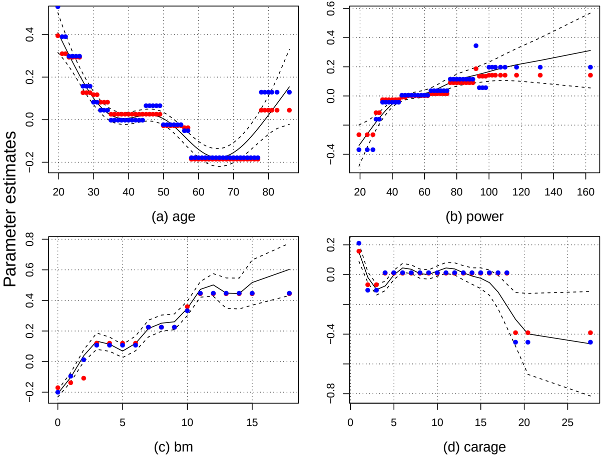
$$-\frac{1}{n}\log\mathcal{L}(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$$
$$+\lambda\left(\sum_{j\in\text{bin}}|w_j\beta_j|+\sum_{j\in\text{ord}}||\boldsymbol{D}(\boldsymbol{w}_j)\boldsymbol{\beta}_j||_1+||\boldsymbol{G}(\boldsymbol{w}_{\text{muni}})\boldsymbol{\beta}_{\text{muni}}||_1\right).$$

- Incorporate multi-type penalty, with:
  - standard Lasso for binary `use`, `fleet`, `mono`, `four`, `sports`, `sex` and `fuel`
  - fused Lasso for ordinal `payfreq`, `coverage`, `ageph`, `bm`, `power`, `agec`
  - generalized fused Lasso for spatial `muni`.

MTPL data: Poisson with multi-type penalty

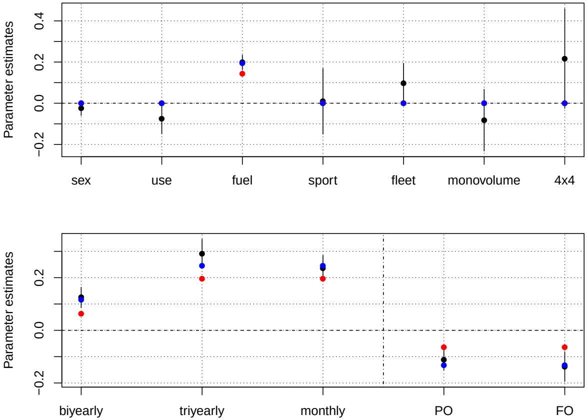
- Settings:
  - incorporate adaptive (GLM) and standardization weights for better consistency and predictive performance
  - tune λ with 10-fold stratified cross-validation where the deviance is used as error measure and the one-standard-error rule is applied
- Re-estimate the final sparse GLM with standard GLM routines (from 422 to 71 params.).

MTPL data: Poisson with multi-type penalty



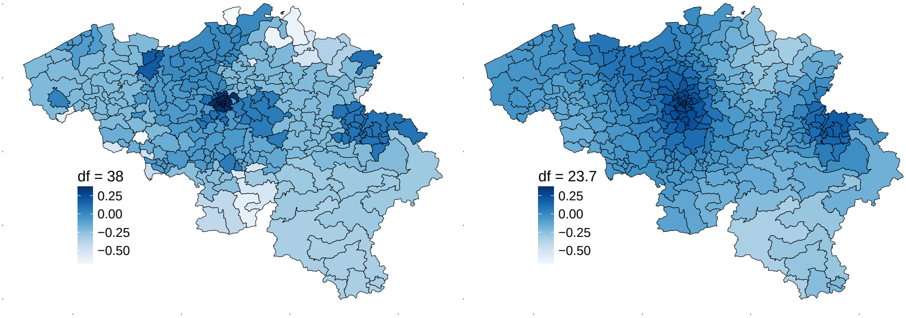
GAM fit, penalized GLM fit, GLM refit with new bins

MTPL data: Poisson with multi-type penalty



GAM fit, penalized GLM fit, GLM refit with new bins

MTPL data: Poisson with multi-type penalty



(a) SMuRF estimates

(b) GAM estimates

Wrap-up

- Flexible regularization can help predictive modeling tasks.
- `SMuRF` package, vignette and published paper (2021) available online.

References

Devriendt, S., Antonio, K., Reynkens, T. and Verbelen, R. (2021).
Sparse Regression with Multi-type Regularized Feature Modeling
Insurance: Mathematics and Economics, 96, 248-261.

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.

References

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.