SFI Centre for Research Training in Foundations of Data Science

Centre for Research Training

Science Foundation Ireland For what's next

Combining a smooth information criterion with neural networks

Andrew McInerney,

University of Limerick

LMU, 07 July 2023

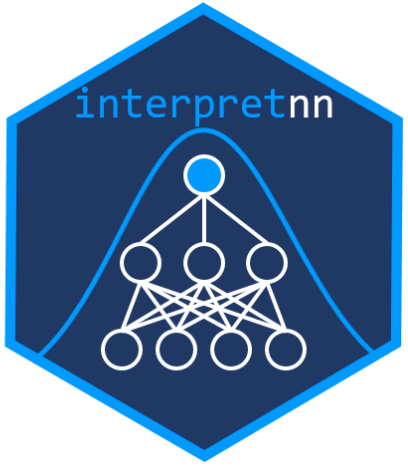
UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

Introduction

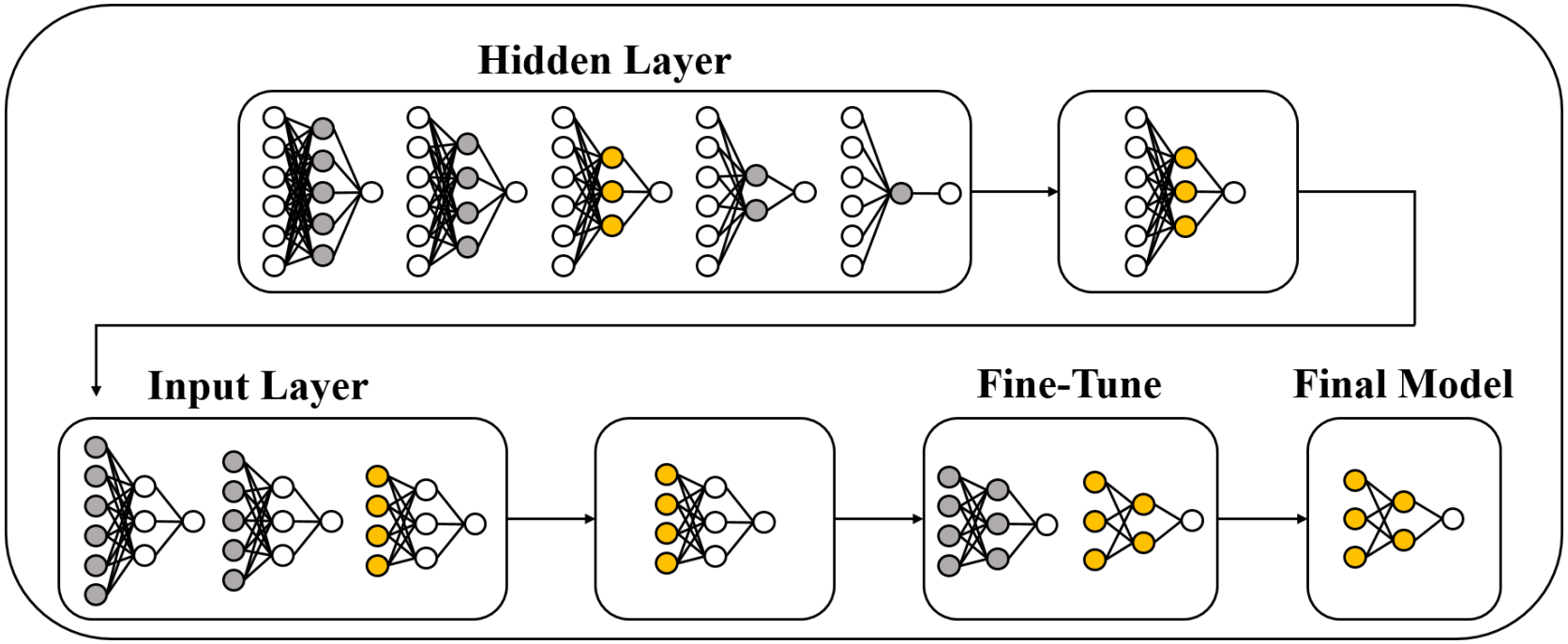
Background

SFI Centre for Research Training in Foundations of Data Science

- Research: Neural networks from a statistical-modelling perspective



Model Selection



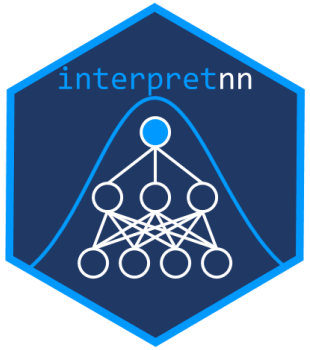
A Statistically-Based Approach to Feedforward Neural Network Model Selection (arXiv:2207.04248)

Insurance: Model Selection


```
library(selectnn)
nn <- selectnn(charges ~ ., data = insurance, Q = 8,
               n_init = 5)
summary(nn)
```

```
## [...]
## Number of input nodes: 4
## Number of hidden nodes: 2
##
## Value: 1218.738
##          Covariate Selected Delta.BIC
##        smoker.yes      Yes  2474.478
##                bmi      Yes   919.500
##                age      Yes   689.396
##           children      Yes    13.702
## [...]
```

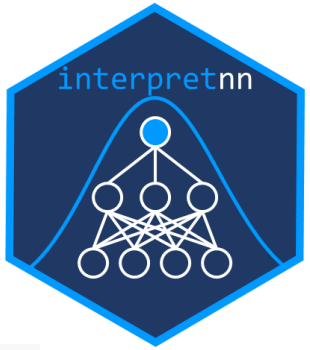
Interpreting FNNs



Extend packages: **nnet**, **neuralnet**, **keras**, **torch**

- Significance testing
- Covariate-effect plots

Insurance: Model Summary

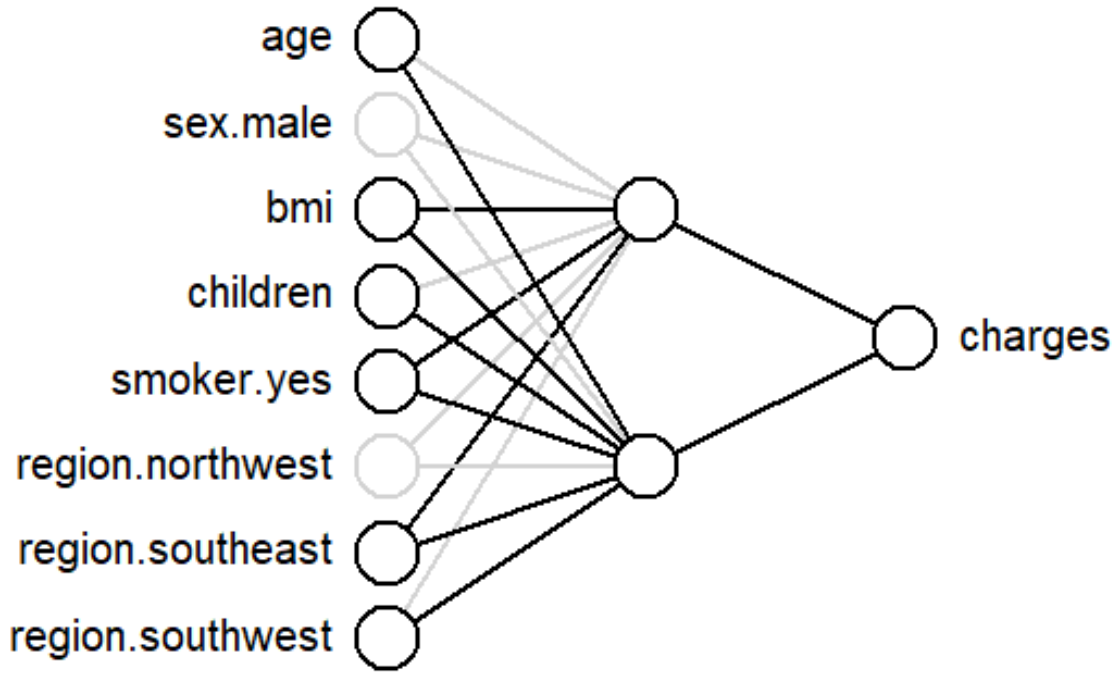



```
intnn <- interpretnn(nn)
summary(intnn)
```

```
## Coefficients:
##                                    Weights |      X^2    Pr(> X^2
##                age      (0.19, -0.41***) | 24.1009 5.84e-06 ***
##           sex.male        (-0.25, 0.05.) |  3.6364 1.62e-01
##                bmi   (-26.11***, -0.03*) | 14.7542 6.25e-04 ***
##           children      (0.16, -0.07***) | 13.1946 1.36e-03 **
##         smoker.yes (63.64***, -2.83***) | 62.8237 2.28e-14 ***
##  region.northwest        (-3.65., 0.03) |  3.4725 1.76e-01
##  region.southeast       (-1.95*, 0.08*) |  7.8144 2.01e-02 *
##  region.southwest       (-1.27, 0.12**) |  9.1267 1.04e-02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

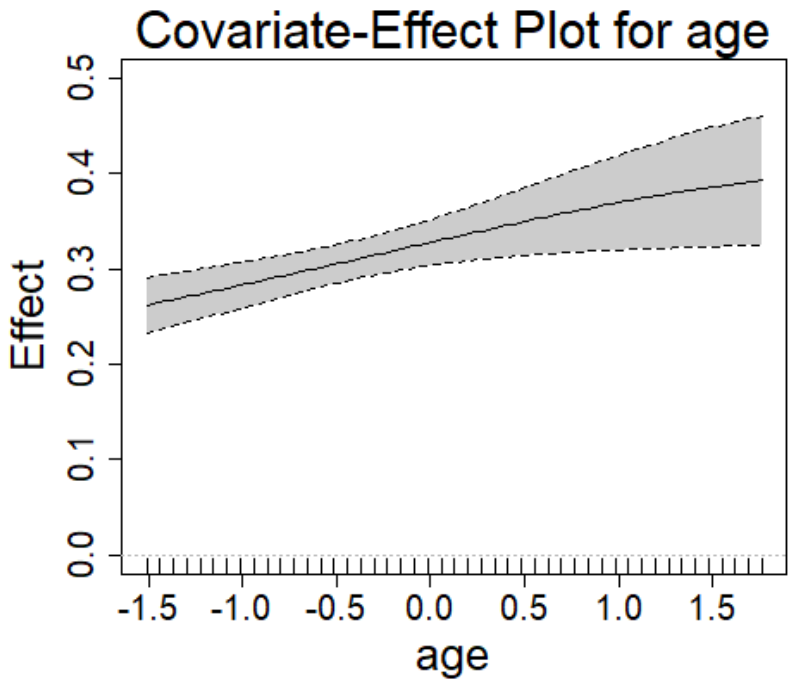
Insurance: Model Summary

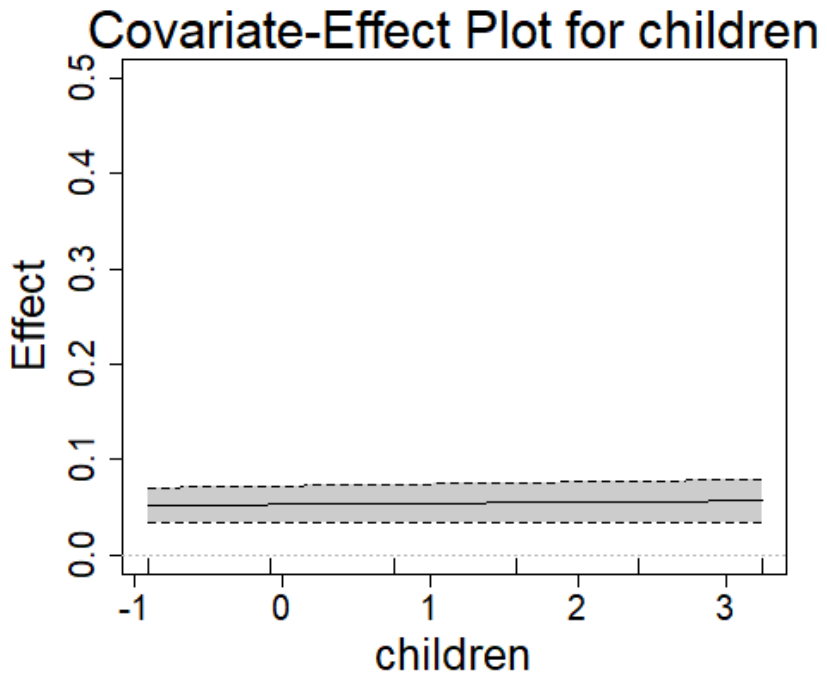
```
plotnn(intnn)
```



Insurance: Covariate Effects

```
plot(intnn, conf_int = TRUE, which = c(1, 4))
```





Current Work

Statistics and Computing (2023) 33:71
https://doi.org/10.1007/s11222-023-10204-8

ORIGINAL PAPER

Check for updates

Variable selection using a smooth information criterion for distributional regression models

Meadhbh O'Neill[1] · Kevin Burke[1]

Received: 7 March 2022 / Accepted: 3 January 2023 / Published online: 21 April 2023

Smooth Information Criterion

$$\mathrm{IC} = -2\ell(\theta) + \lambda[\|\tilde{\theta}\|_0 + 1]$$

where $\lambda = \log(n)$ for the BIC.

Rearrange as an IC-based penalized likelihood:

$$\ell^{\mathrm{IC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2}[\|\tilde{\theta}\|_0 + 1]$$

Smooth Information Criterion

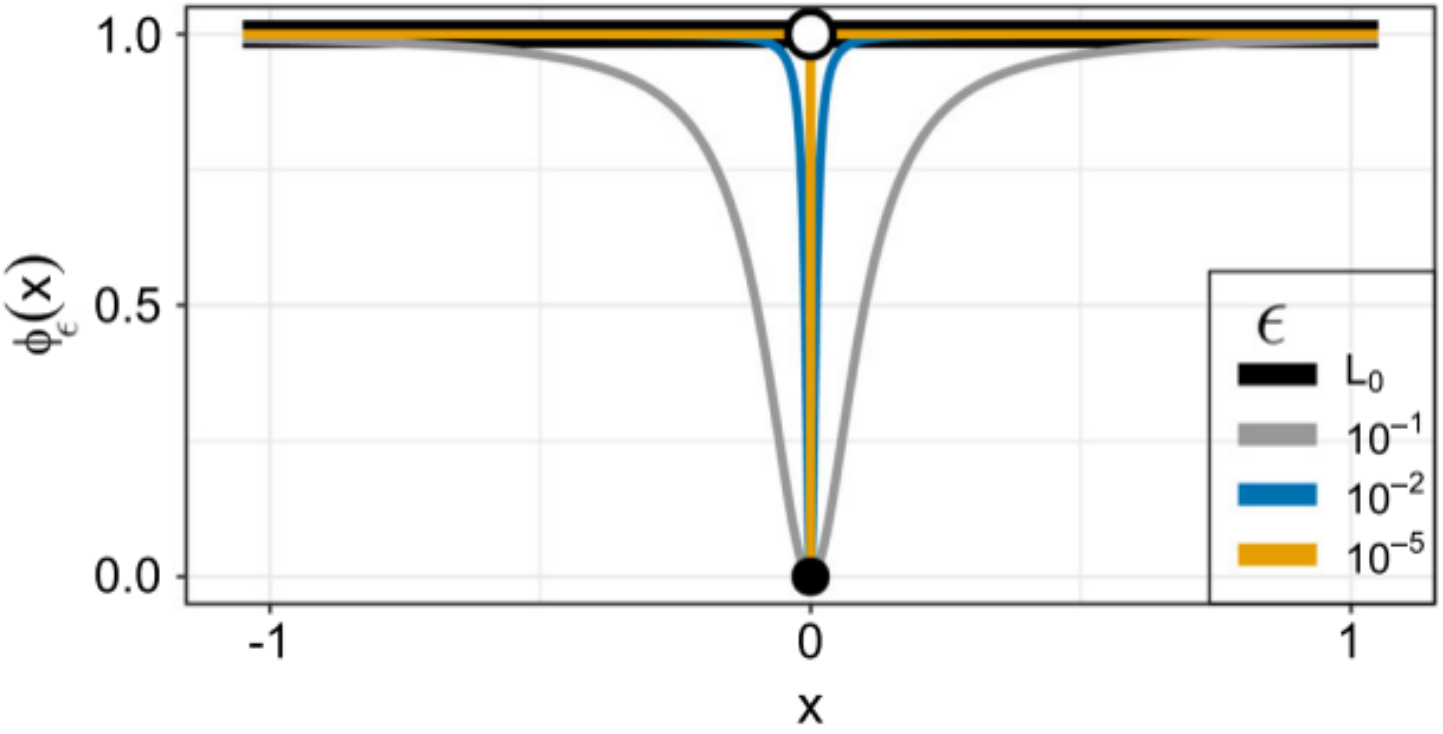
Introduce "smooth L_0 norm":

$$\|\theta\|_{0,\epsilon} = \sum_{j=1}^{p} \phi_\epsilon(\theta_j)$$

where

$$\phi_\epsilon(\theta_j) = \frac{\theta_j^2}{\theta_j^2 + \epsilon^2}$$

Smooth Information Criterion



Motivation

- Tuning parameter automatically selected in one step

- Computationally advantageous

ϵ-telescoping

- Optimal ϵ is zero
- Smaller $\epsilon \implies$ less numerically stable
- Start with larger ϵ, and "telescope" through a decreasing sequence of ϵ values using warm starts

Algorithm

Algorithm 1 Implementation of the MPR-SIC ϵ-telescope Method

1. **Initialization:** Set $\theta^{(0)} = (\beta^{(0)T}, \alpha^{(0)T})^T$, where $\beta^{(0)}$ and $\alpha^{(0)}$ are the initial values for the location and dispersion parameters respectively (see Sect. 2.3).
2. **Telescoping:** Go through the exponentially decaying sequence of telescope values of length T from ϵ_1 to ϵ_T, where $\epsilon_t = \epsilon_1 r^{t-1}$ for step $t = 1, \ldots, T$ and $r \in (0, 1)$ is the rate of decay (see Sect. 2.4).
   - **For $t = 1, \ldots, T$:**
     Optimization: Maximize $\ell^{\text{SIC}}(\theta)$ in (6) by iteratively re-solving the system of equations in (10) with initial values $\theta_{\epsilon_t}^{(0)}$, to obtain $\hat{\theta}_{\epsilon_t}$. Convergence is achieved when $|\theta_{\epsilon_t}^{(m+1)} - \theta_{\epsilon_t}^{(m)}| \leq \omega$ for some small tolerance, e.g., $\omega = 10^{-8}$. For warm starts, set $\theta_{\epsilon_{t+1}}^{(0)} = \hat{\theta}_{\epsilon_t}$ so that the obtained estimates are used as initial values for the next step in the telescope. Note that we set $\hat{\theta}_{\epsilon_0} = \theta^{(0)}$.
3. **Output:** At $t = T$, the final estimates $\hat{\theta}_{\epsilon_T}$ are obtained and any estimates that are very close to zero (below 10^{-8} for example) can be treated as being zero. The corresponding standard errors are computed by evaluating (11) at $\hat{\theta}_{\epsilon_T}$. Note that because we are applying penalized variable selection, the predictors are scaled to have unit variance. However, the final estimates are converted back to their original scale.

Results

Table 4 Simulation results: model selection metrics

		MPR-SIC				BAMLSS			
	n	C(6)	IC(0)	PT	MSE	C(6)	IC(0)	PT	MSE
β	100	5.25	0.15	0.44	0.14	5.55	0.24	0.52	0.17
	500	5.88	0.00	0.88	0.01	5.67	0.00	0.73	0.02
	1000	5.95	0.00	0.95	0.00	5.70	0.00	0.74	0.01
α	100	5.52	0.80	0.30	0.62	5.60	1.08	0.20	0.46
	500	5.92	0.00	0.93	0.04	5.35	0.00	0.67	0.08
	1000	5.95	0.00	0.95	0.02	5.08	0.00	0.63	0.06
		SPR-SIC				ALASSO-IC			
	n	C(6)	IC(0)	PT	MSE	C(6)	IC(0)	PT	MSE
β	100	5.59	3.20	0.00	2.37	5.43	2.99	0.01	2.02
	500	5.84	1.57	0.11	0.64	5.58	1.13	0.22	0.61
	1000	5.88	0.70	0.37	0.27	5.70	0.45	0.45	0.27
α	100	6.00	6.00	0.00	6.43	6.00	6.00	0.00	6.62
	500	6.00	6.00	0.00	7.12	6.00	6.00	0.00	7.16
	1000	6.00	6.00	0.00	7.15	6.00	6.00	0.00	7.17

C, average correct zeros; IC, average incorrect zeros; PT, the probability of choosing the true model; MSE, the average mean squared error

R Package

smoothic

For more information, check out the `smoothic` website.

Implementation of the SIC epsilon-telescope method, either using single or multi-parameter regression. Includes classical regression with normally distributed errors and robust regression, where the errors are from the Laplace distribution. The "smooth generalized normal distribution" is used, where the estimation of an additional shape parameter allows the user to move smoothly between both types of regression. See O'Neill and Burke (2022) "Robust Distributional Regression with Automatic Variable Selection" for more details on arXiv. This package also contains the data analyses from O'Neill and Burke (2023). "Variable selection using a smooth information criterion for distributional regression models" in Statistics & Computing.

Installation

CRAN

You can install the released version of smoothic from CRAN with:

```
install.packages("smoothic")
```

Extending to Neural Networks

$$\mathbb{E}(y) = \mathrm{NN}(X, \theta)$$

where

$$\mathrm{NN}(X, \theta) = \phi_o \left[\gamma_0 + \sum_{k=1}^{q} \gamma_k \phi_h \left(\sum_{j=0}^{p} \omega_{jk} x_j \right) \right]$$

Extending to Neural Networks

We can then formulate a **smooth** BIC-based penalized likelihood:

$$\ell^{\mathrm{SIC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2}[\|\tilde{\theta}\|_{0,\epsilon} + q + 1],$$

where

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathrm{NN}(x_i))^2$$

Extending to Group Sparsity

The smooth approximation of the L_0 norm can be written for groups as

$$\phi_\epsilon(\theta^{(g)}) = |\theta^{(g)}| \frac{\|\theta^{(g)}\|_2^2}{\|\theta^{(g)}\|_2^2 + \epsilon^2}.$$

Group Sparisty

Input-node penalization

$$\ell^{\text{IN-SIC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2}\left[\sum_{j=1}^{p} \|\omega_j\|_{0,\epsilon} + \|\tilde{\gamma}\|_{0,\epsilon} + q + 1\right],$$

where $\omega_j = (\omega_{j1}, \omega_{j2}, \ldots, \omega_{jq})^T$

Group Sparisty

Hidden-node penalization

$$\ell^{\text{HN-SIC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2}\left[\sum_{k=1}^{q}\|\theta^{(k)}\|_{0,\epsilon} + q + 1\right],$$

where $\theta^{(k)} = (\omega_{1k}, \omega_{2k}, \ldots, \omega_{pk}, \gamma_k)^T$

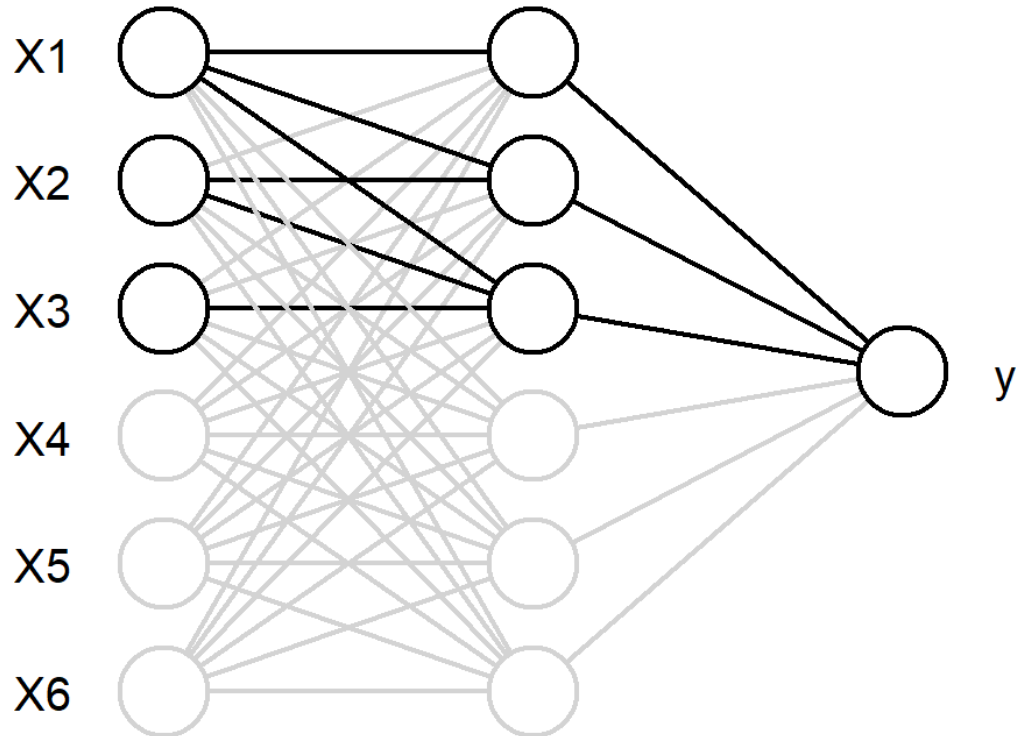
Combined Penalty

- Implement a group penalty and the single-parameter penalty in one optimization procedure
- Start with group penalization and telescope through the ϵ values until some predefined change point, τ
- Switch to single-parameter penalization for the remainder of the ϵ values

Approaches

- Single-parameter penalization
- Input-node penalization
- Hidden-node penalization
- Combined approaches (perform group penalization initially and then switch to single-parameter penalization)

Preliminary Simulation



Preliminary Results

References

- O'Neill, M. and Burke, K. (2023). Variable selection using a smooth information criterion for distributional regression models. *Statistics and Computing*, *33(3)*, *p.71*.

R Packages

```
devtools::install_github(c("andrew-mcinerney/selectnn",
                           "andrew-mcinerney/interpretnn"))
```


andrew-mcinerney @amcinerney_ andrew.mcinerney@ul.ie