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

Where is Limerick?

Limerick

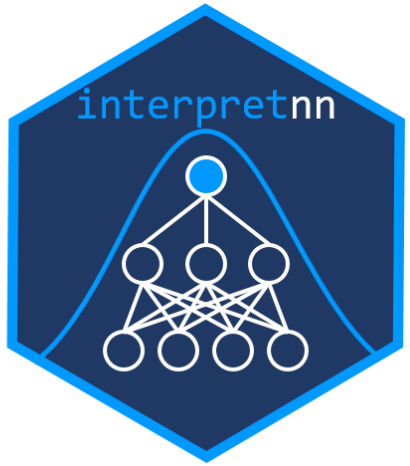
MÜNCHEN COLMCILLES
G.A.A.

Univeristy of Limerick

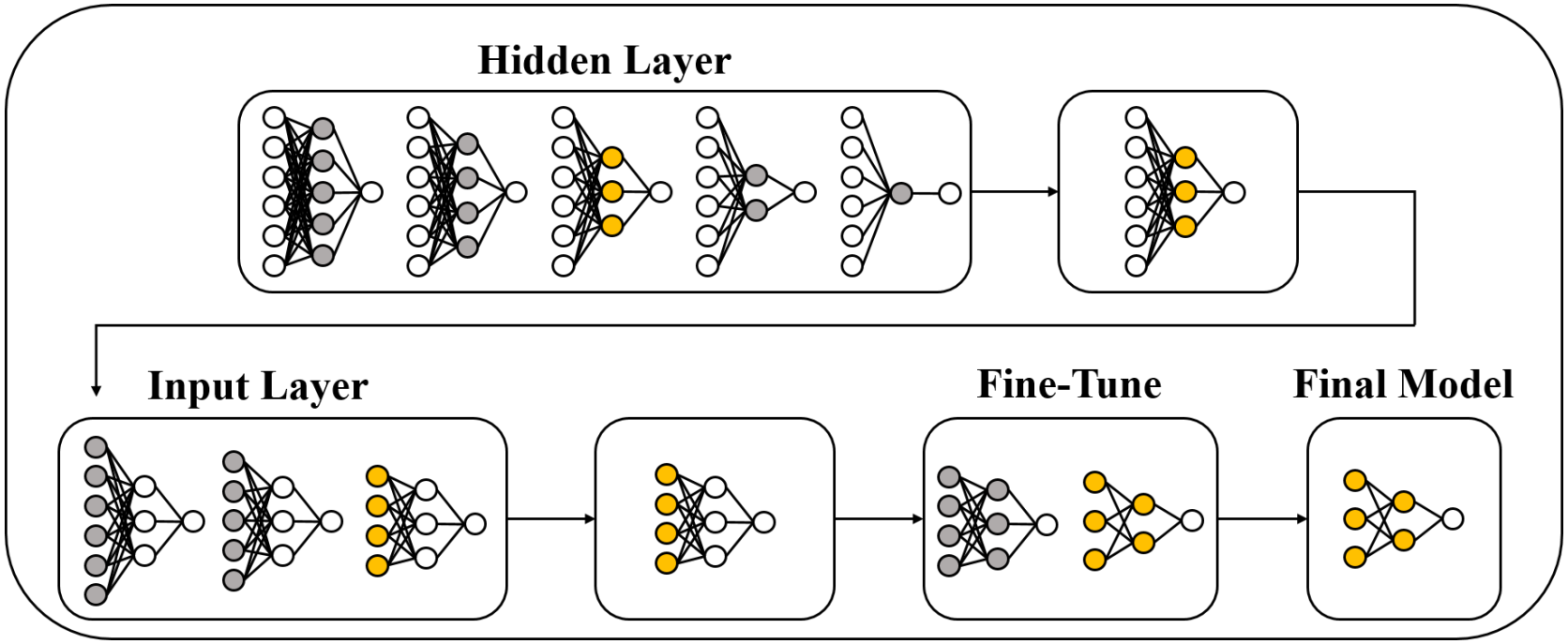
Background

SFI Centre for Research Training in Foundations of Data Science

- Research: Neural networks from a statistical-modelling perspective



Model Selection



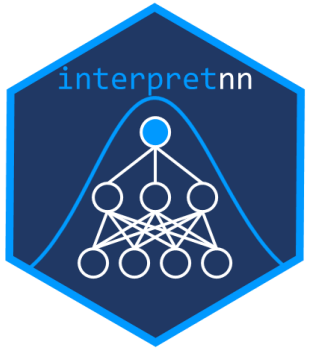
A Statistically-Based Approach to Feedforward Neural Network Model Selection (arXiv:2207.04248)

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
## [...]
```

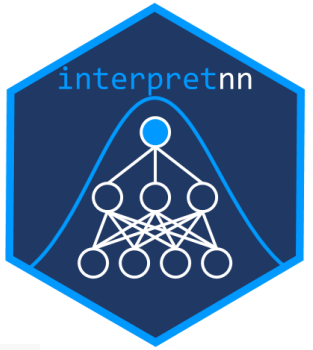
Interpreting FNNs



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

- Significance testing
- Covariate-effect plots

Insurance: Model Summary

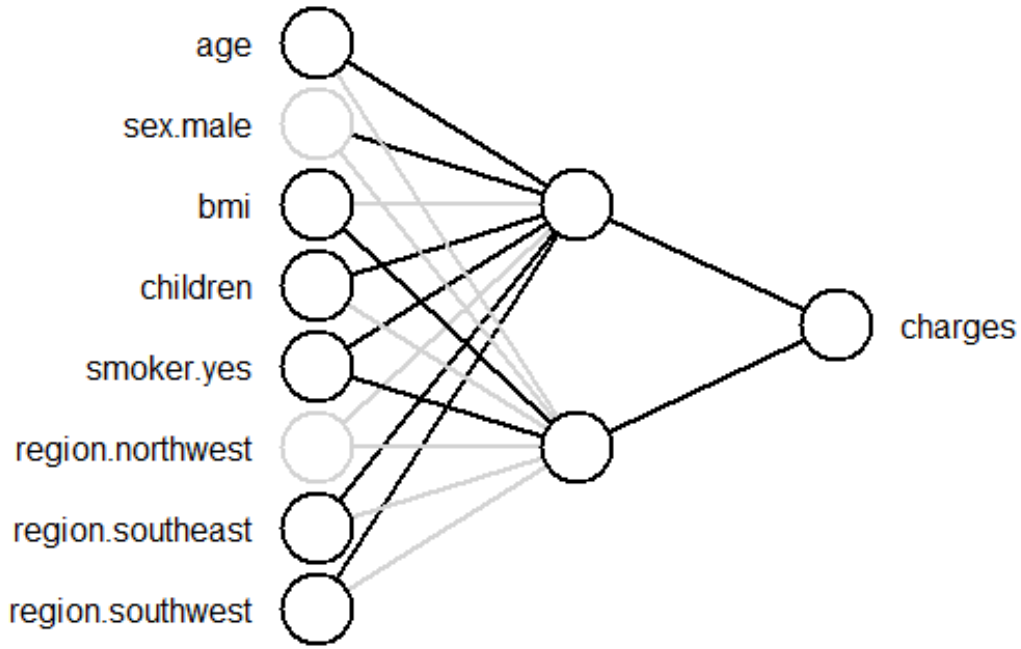



```
intnn <- interpretnn(nn)
summary(intnn)
```

```
## Coefficients:
##                                    Weights |       X^2    Pr(> X^
##                age      (-0.43***, 0.04) |   41.4363 1.01e-09 *
##           sex.male         (0.08*, 0.13) |    5.5055 6.38e-02 .
##                bmi       (0.03, 2.19***) | 105.6106 0.00e+00 *
##           children   (-0.08***, -0.11.) |   19.0146 7.43e-05 *
##         smoker.yes (-3.16***, -6.19***) | 250.6393 0.00e+00 *
##  region.northwest         (0.07., 0.15) |    2.8437 2.41e-01
##  region.southeast         (0.11*, 0.12) |    6.2560 4.38e-02 *
##  region.southwest        (0.15**, 0.14) |   10.8218 4.47e-03 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

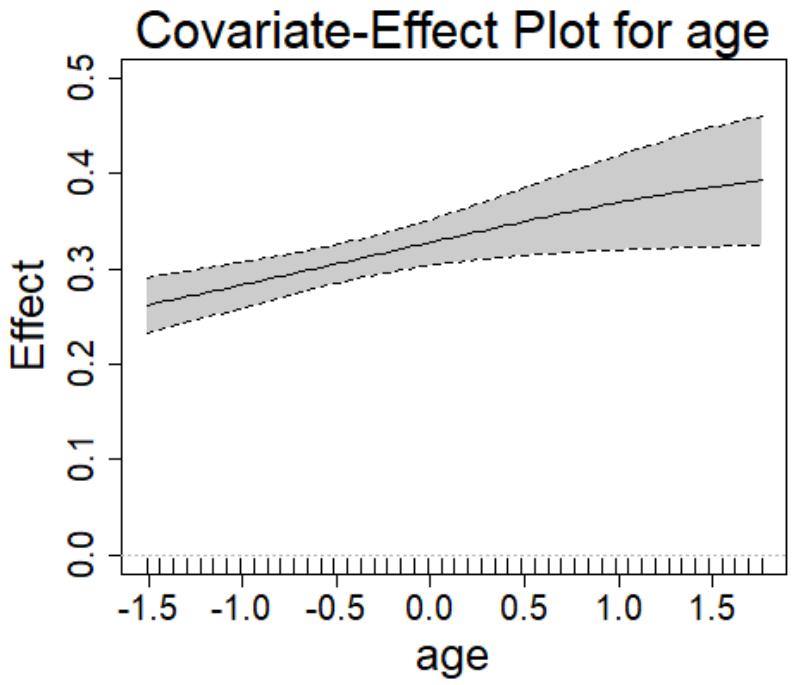
Insurance: Model Summary

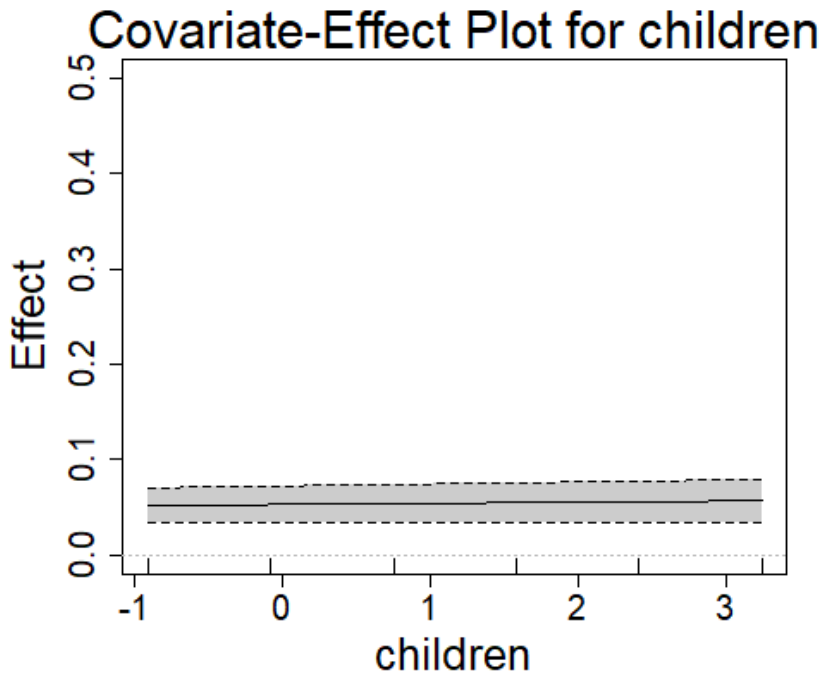
```
plotnn(intnn)
```



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
© The Author(s) 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

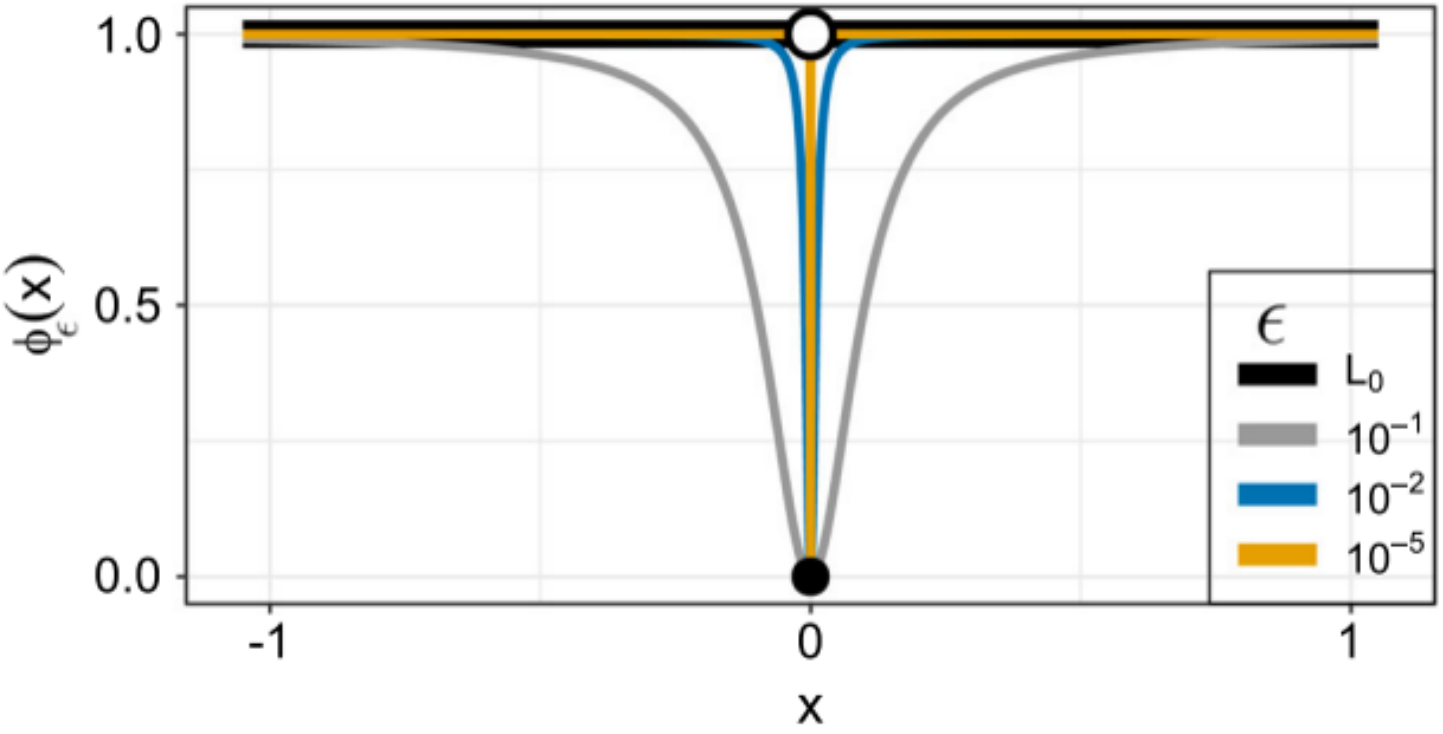
Introduce "smooth L_0 norm":

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

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

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

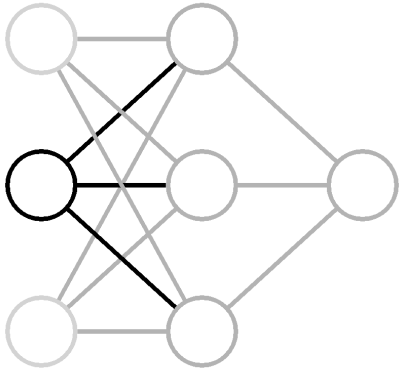
$$\ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathrm{NN}(x_i))^2$$

Extending to Group Sparsity

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

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

Group Sparisty

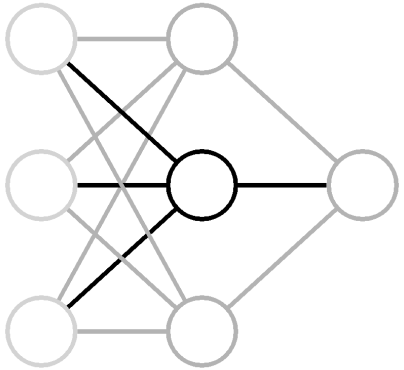


Input-neuron penalization

$$\ell^{\text{IN-SIC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2}\left[q \times \sum_{j=1}^{p} \left\| \|\omega_j\|_2^2 \right\|_{0,\epsilon} + \|\tilde{\gamma}\|_{0,\epsilon} + q + 1\right]$$

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

Group Sparisty



Hidden-neuron penalization

$$\ell^{\text{HN-SIC}}(\theta) = \ell(\theta) - \frac{\log(n)}{2}\left[(p+1) \times \sum_{k=1}^{q} \left\| \|\theta^{(k)}\|_2^2 \right\|_{0,\epsilon} + q + 1\right]$$

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

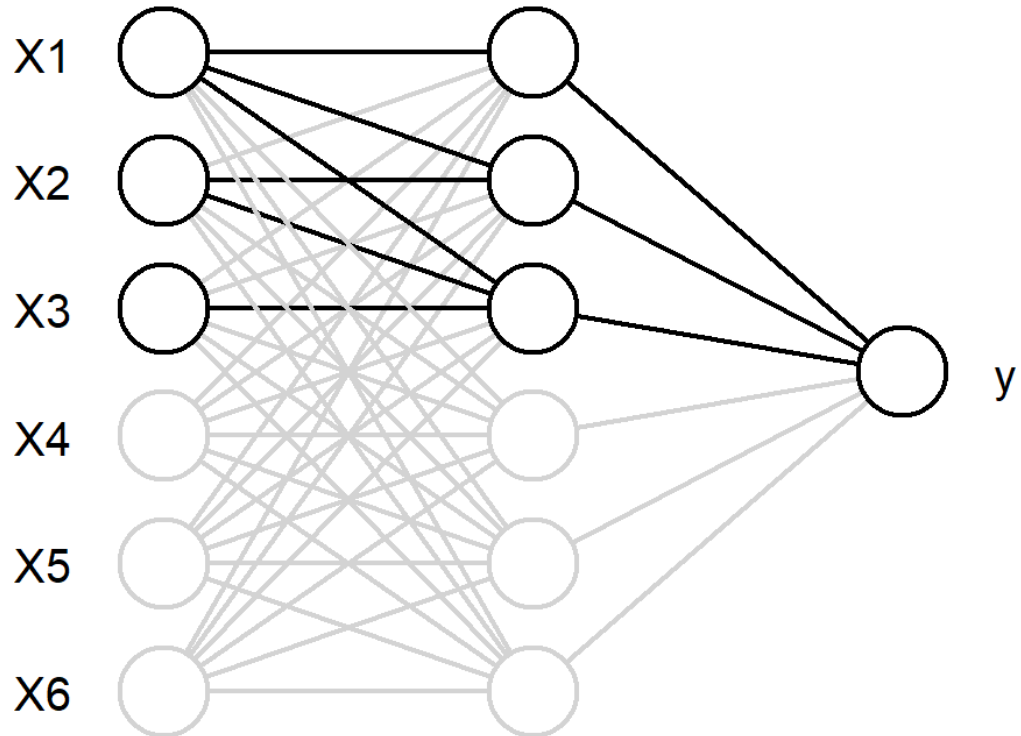
Combined Penalty

- 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 value, τ
- Switch to single-parameter penalization for the remainder of the ϵ values

Approaches

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

Preliminary Simulation



Preliminary Results

Table 1: Preliminary simulation results

Approach	Weights (33)		Inputs (3)		Hidden (3)	
	TPR	FPR	TPR	FPR	TPR	FPR
Single	0.67	0.05	0.32	0.00	0.24	0.04
Group: Input	0.56	0.00	0.99	0.00	0.21	0.00
Group: Hidden	0.63	0.04	0.00	0.00	0.94	0.06
Combined: Input-Single	0.67	0.10	0.94	0.00	0.20	0.03
Combined: Hidden-Single	0.83	0.04	0.36	0.00	0.94	0.06
Combined: Hidden-Single *	0.96	0.07	0.92	0.00	0.93	0.08

References

- McInerney, A., & Burke, K. (2022). A statistically-based approach to feedforward neural network model selection. *arXiv preprint arXiv:2207.04248*.
- McInerney, A., & Burke, K. (2023). Interpreting feedforward neural networks as statistical models. *To appear on arXiv*.
- 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