UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

Automating variable selection in distributional regression

Kevin Burke | University of Limerick

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

Science Foundation Ireland sfi For what's next

Smooth information criterion (SIC)

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

ORIGINAL PAPER

Variable selection using a smooth information criterion for distributional regression models

Meadhbh O'Neill[1] · Kevin Burke[1]

Scientific question

Does X cause Y?

Scientific question

Does X ~~cause~~ Y?
relate to

Scientific question

Does X ~~cause~~ Y?

relate to

linearly

Scientific question

the mean of

Does X ~~cause~~ Y?

relate to

linearly

Mean regression (obsession?)

Mean regression (obsession?)

- $Y = X^\top \beta + \sigma\varepsilon, \quad \varepsilon \sim N(0,1)$

- $\Rightarrow Y \sim N(X^\top \beta, \boldsymbol{\sigma^2})$

- Why should σ^2 be constant?

Mean regression (obsession?)

- $Y = X^\top \beta + \sigma\varepsilon, \quad \varepsilon \sim N(0,1)$

- $\Rightarrow Y \sim N(X^\top \beta, \boldsymbol{\sigma^2})$

- Why should σ^2 be constant?

Kneib et al (2021). Rage against the mean–a review of distributional regression approaches.

Distributional regression

- Any/all distributional parameters could depend on X

Burke, Mackenzie (2017). Multi-parameter regression survival modelling.

Distributional regression

- Any/all distributional parameters could depend on X

- $Y = \mu + \sigma\varepsilon, \qquad \varepsilon \sim N(0,1)$

- $\mu = X^\top \beta, \;\; \boldsymbol{\log \sigma^2 = X^\top \alpha}$

- $\Rightarrow Y \sim N(X^\top \beta, \boldsymbol{\exp(X^\top \alpha)})$

Burke, Mackenzie (2017). Multi-parameter regression survival modelling.

Distributional regression

- Any/all distributional parameters could depend on X
- $Y = \mu + \sigma\varepsilon, \qquad \varepsilon \sim N(0,1)$
- $\mu = X^\top \beta, \;\; \boldsymbol{\log \sigma^2 = X^\top \alpha}$
- $\Rightarrow Y \sim N(X^\top \beta, \boldsymbol{\exp(X^\top \alpha)})$
- Also known as **"multi-parameter regression"**

Burke, Mackenzie (2017). Multi-parameter regression survival modelling.

Model space

- Consider $\mathrm{X} = (X_1, \mathrm{X}_2, \mathrm{X}_3)$

$\mu(X)$
$\emptyset$
X_1
X_2
X_3
X_1, X_2
X_1, X_3
X_2, X_3
X_1, X_2, X_3

Model space

- Consider $X = (X_1, X_2, X_3)$

$\boldsymbol{\mu(X)}$
$\emptyset$
$\boldsymbol{X_1}$
$\boldsymbol{X_2}$
$\boldsymbol{X_3}$
$\boldsymbol{X_1, X_2}$
$\boldsymbol{X_1, X_3}$
$\boldsymbol{X_2, X_3}$
$\boldsymbol{X_1, X_2, X_3}$

Mean regression

Model space

- Consider $X = (X_1, X_2, X_3)$

$\mu(X)$	$\sigma(X)$
$\emptyset$	$\emptyset$
X_1	X_1
X_2	X_2
X_3	X_3
X_1, X_2	X_1, X_2
X_1, X_3	X_1, X_3
X_2, X_3	X_2, X_3
X_1, X_2, X_3	X_1, X_2, X_3

Mean regression

Distributional regression (mean & variance)

Model space

- Consider $X = (X_1, X_2, X_3)$

$\boldsymbol{\mu(X)}$	$\boldsymbol{\sigma(X)}$
$\emptyset$	$\emptyset$
$\boldsymbol{X_1}$	$\boldsymbol{X_1}$
$\boldsymbol{X_2}$	$\boldsymbol{X_2}$
$\boldsymbol{X_3}$	$\boldsymbol{X_3}$
$\boldsymbol{X_1, X_2}$	$\boldsymbol{X_1, X_2}$
$\boldsymbol{X_1, X_3}$	$\boldsymbol{X_1, X_3}$
$\boldsymbol{X_2, X_3}$	$\boldsymbol{X_2, X_3}$
$\boldsymbol{X_1, X_2, X_3}$	$\boldsymbol{X_1, X_2, X_3}$

Mean regression

Distributional regression (mean & variance)

- Models
  - $2^3 \times 2^3 = 2^{2\times 3}$
- In general
  - $2^{d\times p}$
  - d distributional parameters
  - p covariates

How to select variables?

How to select variables?

- $\mathrm{BIC} = -2\ell + (\log n)(\|\beta\|_0 + \|\alpha\|_0)$

How to select variables?

- $\mathrm{BIC} = -2\ell + (\log n)(\|\beta\|_0 + \|\alpha\|_0)$
- $\ell_{\mathrm{BIC}} = -\mathrm{BIC}/2 = \ell - \{(\log n)/2\}(\|\beta\|_0 + \|\alpha\|_0)$

How to select variables?

- $\mathrm{BIC} = -2\ell + (\log n)(\|\beta\|_0 + \|\alpha\|_0)$

- $\ell_{\mathrm{BIC}} = -\mathrm{BIC}/2 = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

- $\left(\hat{\beta}, \hat{\alpha}\right) = \max_{\beta,\alpha} \ell_{\mathrm{BIC}}$

How to select variables?

- $\mathrm{BIC} = -2\ell + (\log n)(\|\beta\|_0 + \|\alpha\|_0)$

- $\ell_{\mathrm{BIC}} = -\mathrm{BIC}/2 = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

- $(\hat{\beta}, \hat{\alpha}) = \max_{\beta,\alpha} \ell_{\mathrm{BIC}}$

- Discrete optimisation: fit all models

How to select variables?

- $\mathrm{BIC} = -2\ell + (\log n)(\|\beta\|_0 + \|\alpha\|_0)$

- $\ell_{\mathrm{BIC}} = -\mathrm{BIC}/2 = \ell - \{(\log n)/2\}(\|\beta\|_0 + \|\alpha\|_0)$

- $(\hat{\beta}, \hat{\alpha}) = \max_{\beta,\alpha} \ell_{\mathrm{BIC}}$

- Discrete optimisation: fit all models

- Heuristic: stepwise search

Related problem

Jaouimaa, Ha, Burke (2023) Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

Related problem

- Problem: $\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

Jaouimaa, Ha, Burke (2023) Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

Related problem

- Problem: $\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

- Related: $\ell_{\mathrm{LASSO}} = \ell - \lambda(\|\beta\|_1 + \|\alpha\|_1)$

Jaouimaa, Ha, Burke (2023) Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.

Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

Related problem

- Problem: $\ell_{\text{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$
- Related: $\ell_{\text{LASSO}} = \ell - \lambda(\|\beta\|_1 + \|\alpha\|_1)$
- What about λ?
  - Given λ, obtain $\left(\hat{\beta}_{\text{LASSO}}, \hat{\alpha}_{\text{LASSO}}\right) = \max \ell_{\text{LASSO}}$
  - Select λ using ℓ_{BIC}
  - Thus, $\max \ell_{\text{BIC}}$ subject to solutions of the form $\left(\hat{\beta}_{\text{LASSO}}, \hat{\alpha}_{\text{LASSO}}\right)$

Jaouimaa, Ha, Burke (2023) Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

Related problem

- Problem: $\ell_{\text{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

- Related: $\ell_{\text{LASSO}} = \ell - \lambda(\|\beta\|_1 + \|\alpha\|_1)$

- What about λ?
  - Given λ, obtain $\left(\hat{\beta}_{\text{LASSO}}, \hat{\alpha}_{\text{LASSO}}\right) = \max \ell_{\text{LASSO}}$
  - Select λ using ℓ_{BIC}
  - Thus, $\max \ell_{\text{BIC}}$ subject to solutions of the form $\left(\hat{\beta}_{\text{LASSO}}, \hat{\alpha}_{\text{LASSO}}\right)$

- **Note that $\|x\|_1$ is not differentiable**

Jaouimaa, Ha, Burke (2023) Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

Recall the original problem

Recall the original problem

- Maximise $\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

Recall the original problem

- Maximise $\ell_{\text{BIC}} = \ell - \{(\log n)/2\} (\|\beta\|_0 + \|\alpha\|_0)$

- Discrete optimisation challenging

Recall the original problem

- Maximise $\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

- Discrete optimisation challenging

- Solve "related" LASSO-type problem

Recall the original problem

- Maximise $\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\} (\|\beta\|_0 + \|\alpha\|_0)$
- Discrete optimisation challenging
- Solve "related" LASSO-type problem
- Not fully satisfying …
  - λ tuning required
  - $(\hat{\beta}, \hat{\alpha})$ **not** solution to original problem

Idea: smooth information criterion (SIC)

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- "Brilliant"

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- “Brilliant”
- “Ingenious”

O’Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- “Brilliant”
- “Ingenious”
- “Remarkable”

O’Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- “Brilliant”
- “Ingenious”
- “Remarkable”

Robitzsch (2023)

a. L0 and Lp Loss Functions in Model-Robust Estimation of Structural Equation Models

b. Implementation aspects in regularized structural equation models

c. Implementation Aspects in Invariance

O’Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- “Brilliant”
- “Ingenious”
- “Remarkable”

Robitzsch (2023)

a. L0 and Lp Loss Functions in Model-Robust Estimation of Structural Equation Models

b. Implementation aspects in regularized structural equation models

c. Implementation Aspects in Invariance

- $\ell_{\text{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$

O’Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- "Brilliant"
- "Ingenious"
- "Remarkable"

Robitzsch (2023)

a. L0 and Lp Loss Functions in Model-Robust Estimation of Structural Equation Models

b. Implementation aspects in regularized structural equation models

c. Implementation Aspects in Invariance

- $$\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\}\,(\|\beta\|_0 + \|\alpha\|_0)$$
$$= \ell - \{(\log n)/2\}\left(\sum_j \left|\boldsymbol{\beta}_j\right|^{\mathbf{0}} + \sum_j \left|\boldsymbol{\alpha}_j\right|^{\mathbf{0}}\right)$$

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Idea: smooth information criterion (SIC)

- "Brilliant"
- "Ingenious"
- "Remarkable"

Robitzsch (2023)

a. L0 and Lp Loss Functions in Model-Robust Estimation of Structural Equation Models

b. Implementation aspects in regularized structural equation models

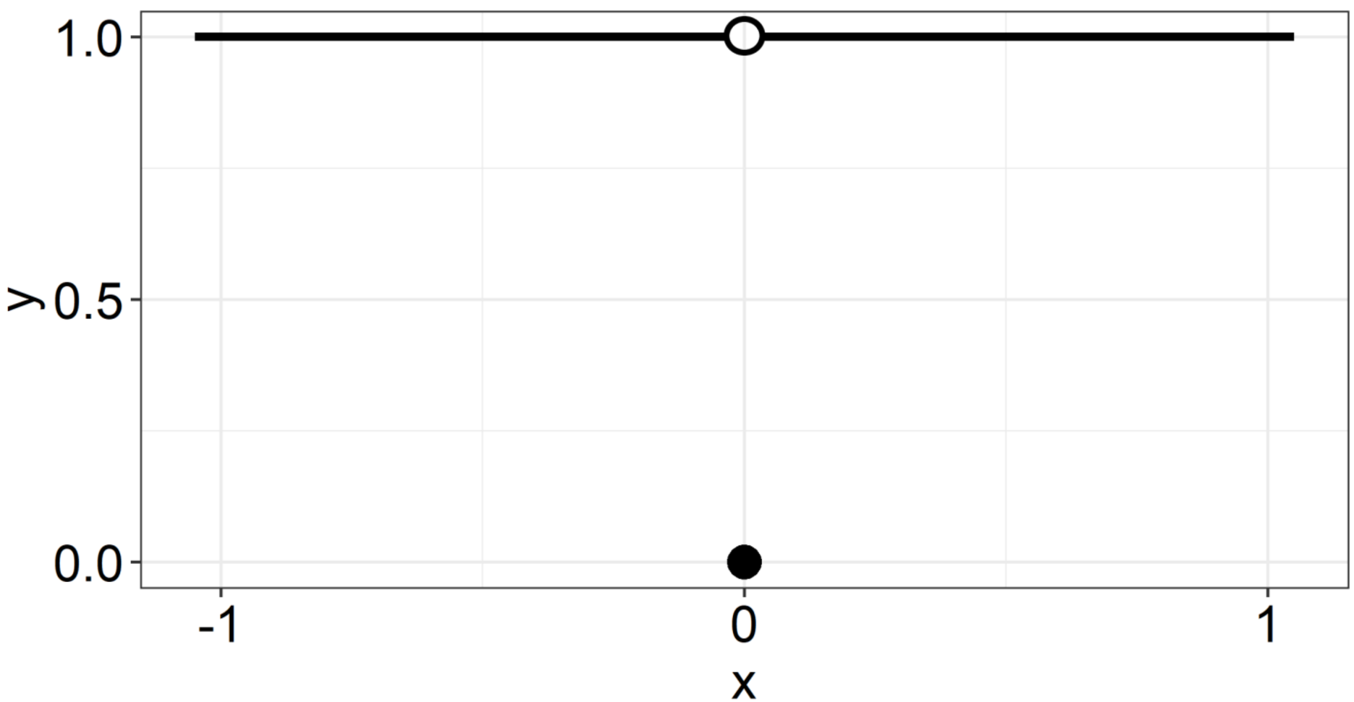
c. Implementation Aspects in Invariance

- $$\ell_{\text{BIC}} = \ell - \{(\log n)/2\}\left(\|\beta\|_0 + \|\alpha\|_0\right)$$
$$= \ell - \{(\log n)/2\}\left(\sum_j |\boldsymbol{\beta}_j|^{\mathbf{0}} + \sum_j |\boldsymbol{\alpha}_j|^{\mathbf{0}}\right)$$

- $$\ell_{\text{SIC}} = \ell - \{(\log n)/2\}\left(\sum_j \boldsymbol{\phi}_{\epsilon}(\boldsymbol{\beta}_j) + \sum_j \boldsymbol{\phi}_{\epsilon}(\boldsymbol{\alpha}_j)\right)$$

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Smooth L_0 norm

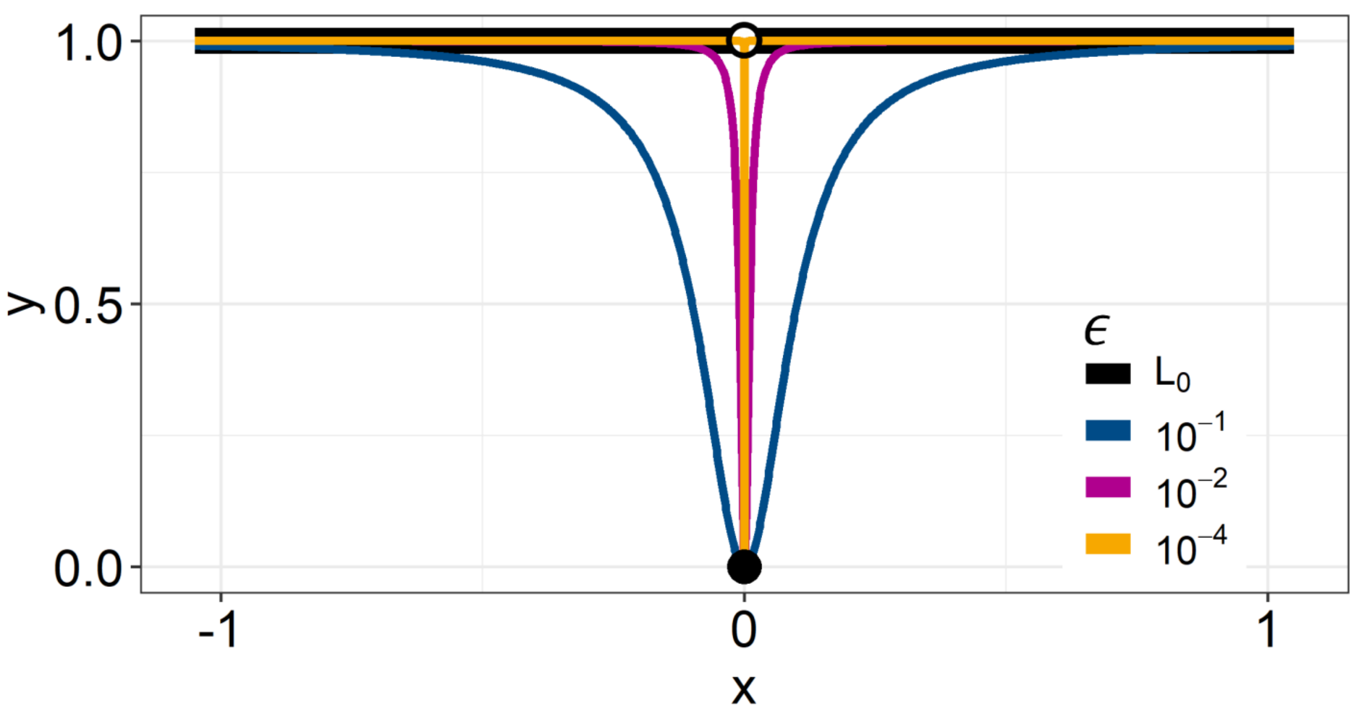


O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Smooth L_0 norm

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

- Differentiable for $\epsilon > 0$
- $\lim_{\epsilon \to 0} \phi_\epsilon(x) = |x|^0$



O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

ϵ-telescoping

- $\lim_{\epsilon \to 0} \ell_{\text{SIC}} = \ell_{\text{BIC}}$
- Maximising ℓ_{SIC} equivalent to ℓ_{BIC} for $\epsilon \approx 0$

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

ϵ-telescoping

- $\lim_{\epsilon \to 0} \ell_{\text{SIC}} = \ell_{\text{BIC}}$
- Maximising ℓ_{SIC} equivalent to ℓ_{BIC} for $\epsilon \approx 0$

Warm Starts

Optimization more stable

Closer to the discontinuity

ϵ
10^{-1}
10^{-2}
10^{-3}
10^{-4}

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

ϵ-telescoping

- $\lim_{\epsilon \to 0} \ell_{\text{SIC}} = \ell_{\text{BIC}}$
- Maximising ℓ_{SIC} equivalent to ℓ_{BIC} for $\epsilon \approx 0$

Warm Starts

Optimization more stable

Closer to the discontinuity

ϵ	true zero $\beta_1 = 0$	true nonzero $\beta_2 = 1$
10^{-1}	-0.0249883798	1.008
10^{-2}	-0.0005789613	1.009
10^{-3}	-0.0000058339	1.009
10^{-4}	**-0.0000000006**	1.009

O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

smoothic package

O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

smoothic package

Prostate data example

- Level of prostate-specific antigen (PSA)
- $p = 8$, various clinical measures
- $n = 97$

O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

`smoothic` package

Prostate data example

- Level of prostate-specific antigen (PSA)
- $p = 8$, various clinical measures
- $n = 97$

```
# Smooth Information Criterion --------------------
library(smoothic)
fit <- smoothic(formula = lpsa ~ .,
                data = pcancer,
                family = "normal",
                model = "mpr")    # or model = "spr"
```

O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

`smoothic` package

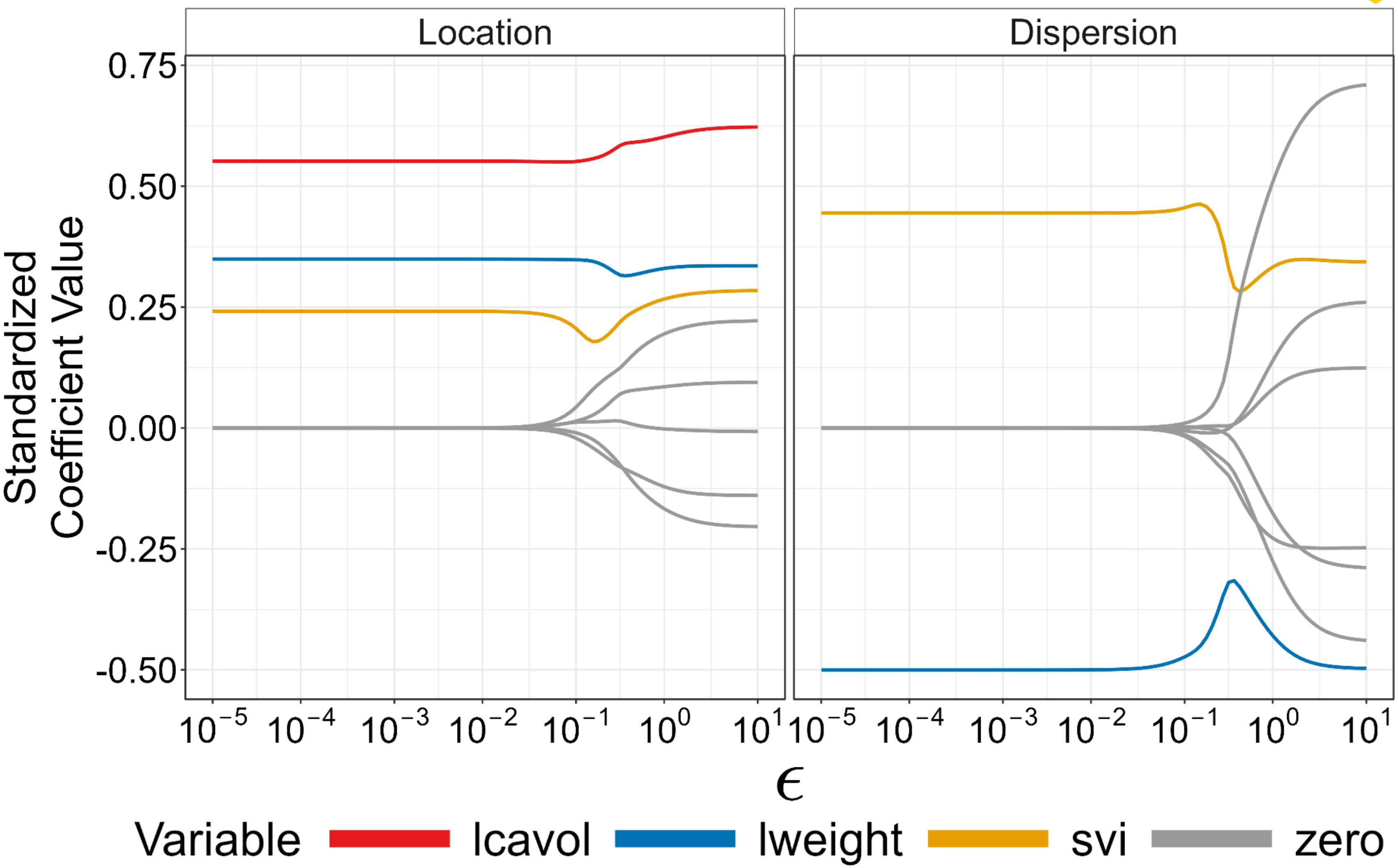
ϵ-telescope

O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

smoothic package

plot_paths()

ϵ-telescope



O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

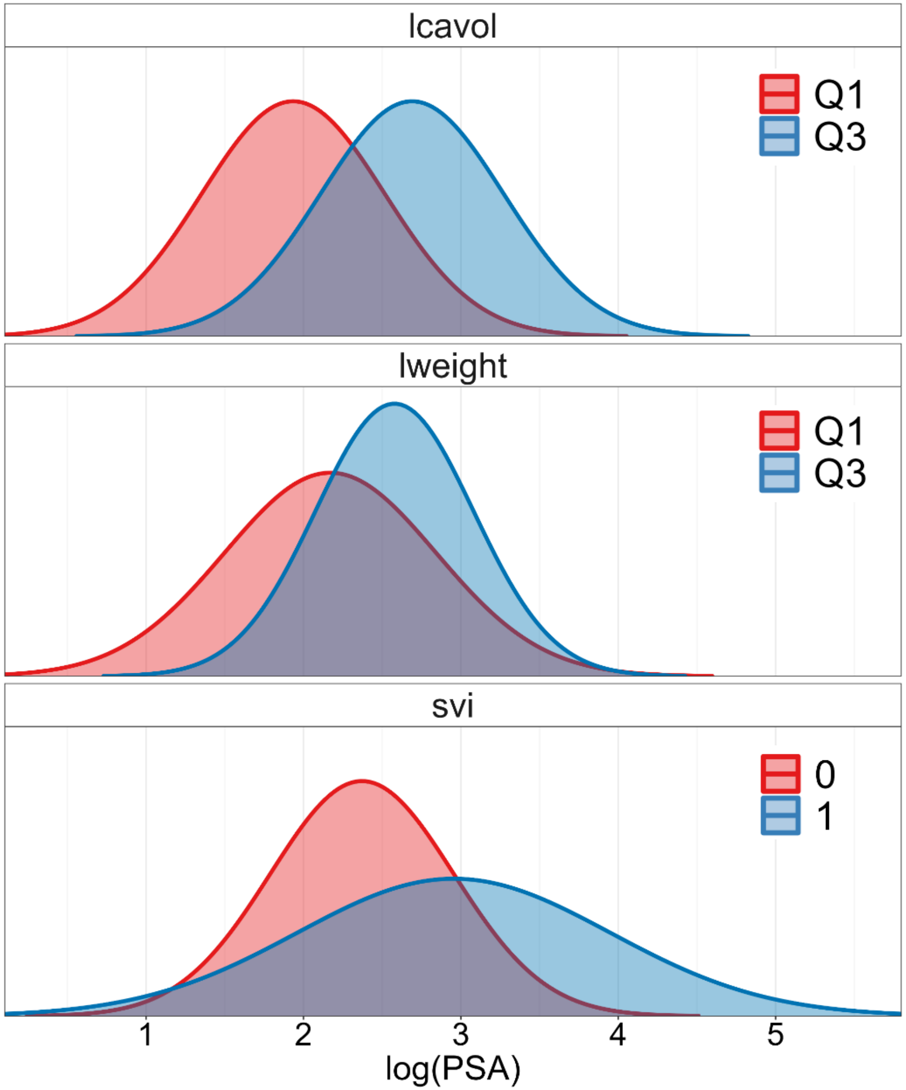
smoothic package

Conditional Density Curves

O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

smoothic package

plot_effects() Conditional Density Curves

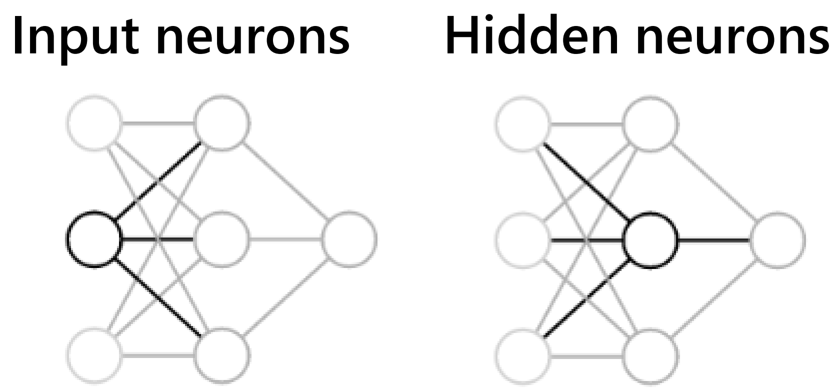


- **lcavol** | log(cancer volume)
- **lweight** | log(prostate weight)
- **svi** | presence of seminal vesicle invasion

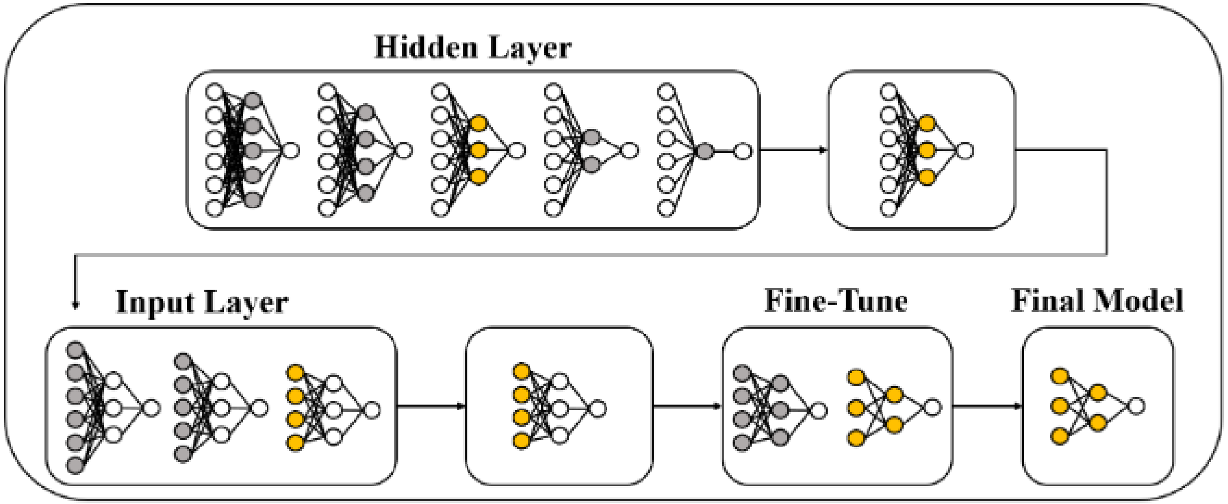
O'Neill & Burke (2023). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R Package.

Neural network selection

- Developing ℓ_{SIC} in neural networks
- Penalty types



- Traditional stepwise procedures



McInerney & Burke (2023). A Statistically-Based Approach to Feedforward Neural Network Model Selection.

selectnn package

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

```
Number of input nodes: 3
Number of hidden nodes: 1

Value: 235.005
Inputs:
 Covariate Selected Delta.BIC
    lcavol      Yes    33.685
   lweight      Yes    10.325
       svi      Yes     6.794
       age       No
      lbph       No
       lcp       No
   gleason       No
     pgg45       No
```

... also see `interpretnn`

McInerney & Burke (2023). A Statistically-Based Approach to Feedforward Neural Network Model Selection.

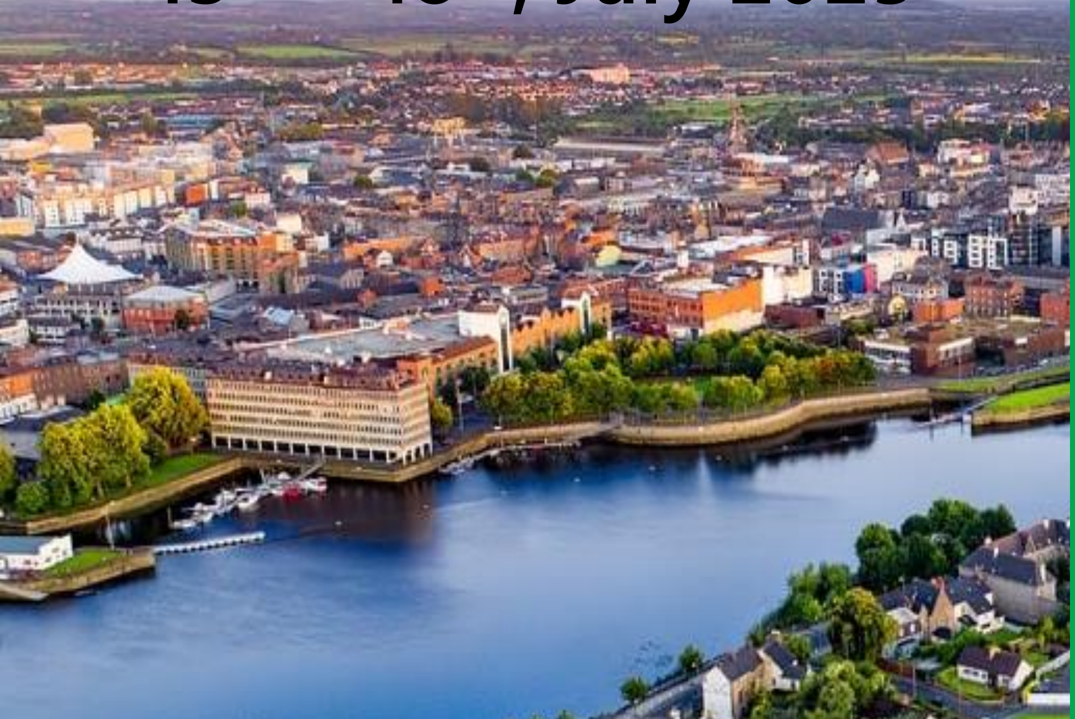
Summary

- Distributional regression more flexible than standard mean regression
- Variable selection more challenging but the smooth information criterion is efficient
- Developing extensions to neural networks
- O'Neill & Burke (2023) Variable selection using a smooth information criterion for distributional regression models. Stats & Computing.
- **Also see: kevinburke.ie and arxiv.org/a/burke_k_1**

39th IWSM
Limerick, Ireland
13th – 18th, July 2025

MACSI
MATHEMATICS APPLICATIONS CONSORTIUM FOR SCIENCE & INDUSTRY

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH



Brendan Murphy | Ireland

Ruth King | UK

Sonja Greven | Germany

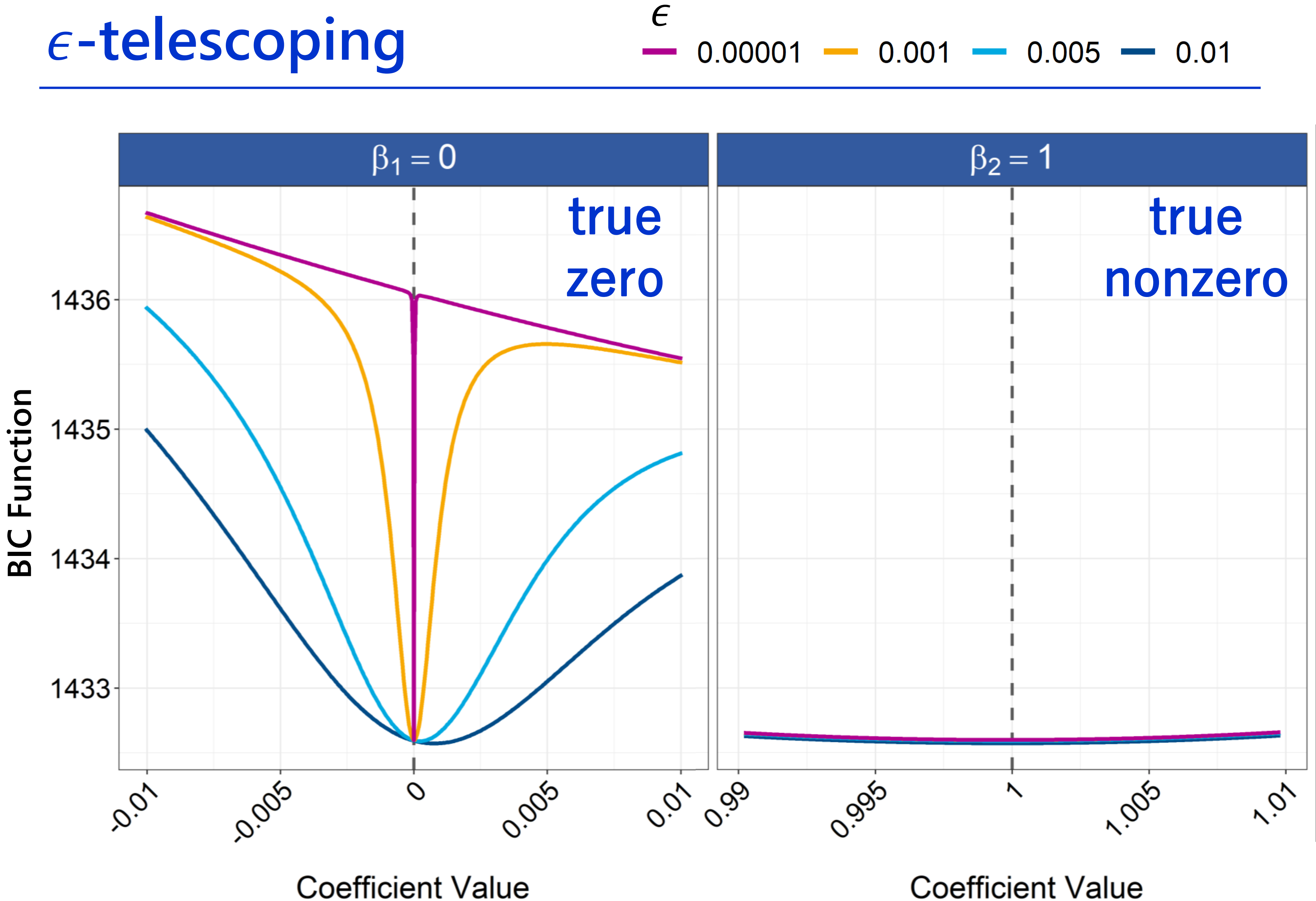
Daniele Durante | Italy

Cynthia Rudin | US

SMS
Statistical Modelling Society

IWSM
International Workshop on Statistical Modelling

ϵ-telescoping



O'Neill & Burke (2023). Variable selection using a smooth information criterion for distributional regression models.

Further details

- Separate tuning parameters

$$\ell_{\mathrm{SAL}} = \ell - \left(\boldsymbol{\lambda_{\beta}} \sum_j w_{\beta_j} a_{\epsilon}(\beta_j) + \boldsymbol{\lambda_{\alpha}} \sum_j w_{\alpha_j} a_{\epsilon}(\alpha_j) \right)$$

- Standard errors

$$\operatorname{cov}(\hat{\theta}) \\ = \left\{-\nabla_\theta \nabla_\theta^\top \boldsymbol{\ell_{SAL}(\widehat{\theta})}\right\}^{-1} \left\{-\nabla_\theta \nabla_\theta^\top \boldsymbol{\ell(\widehat{\theta})}\right\} \left\{-\nabla_\theta \nabla_\theta^\top \boldsymbol{\ell_{SAL}(\widehat{\theta})}\right\}^{-1}$$

- Effective degrees of freedom

$$\ell_{\mathrm{BIC}} = \ell - \{(\log n)/2\}\,(\boldsymbol{edf})$$

$$\boldsymbol{edf} = \operatorname{trace}\left[\left\{-\nabla_\theta \nabla_\theta^\top \boldsymbol{\ell_{SAL}(\widehat{\theta})}\right\}^{-1} \left\{-\nabla_\theta \nabla_\theta^\top \boldsymbol{\ell(\widehat{\theta})}\right\}\right]$$

Jaouimaa, Ha, Burke (2023) Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.