A novel differentiable unification of least absolute deviations and least squares

Kevin Burke | University of Limerick

Science Foundation Ireland sfi For what's next

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least *squares*

$$\min_{\beta} \sum_i \left(y_i - x_i^T \beta \right)^2$$

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least *squares*

$$\min_{\beta} \sum_i \left(y_i - x_i^T \beta\right)^2$$

Least *absolute deviations*

$$\min_{\beta} \sum_i \left|y_i - x_i^T \beta\right|$$

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least *squares*

$$\min_{\beta} \sum_{i} \left(y_i - x_i^T \beta\right)^2$$

Least *absolute deviations*

$$\min_{\beta} \sum_{i} \left|y_i - x_i^T \beta\right|$$

Model

$$y_i = x_i^T \beta + \sigma \varepsilon_i$$

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least ***squares***

$$\min_{\beta} \sum_i \left(y_i - x_i^T \beta\right)^2$$

Gaussian

Least ***absolute deviations***

$$\min_{\beta} \sum_i \left|y_i - x_i^T \beta\right|$$

Model

$$y_i = x_i^T \beta + \sigma \boxed{\varepsilon_i}$$

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least *squares*

$$\min_{\beta} \sum_i \left(y_i - x_i^T \beta\right)^2$$

Gaussian

Least *absolute deviations*

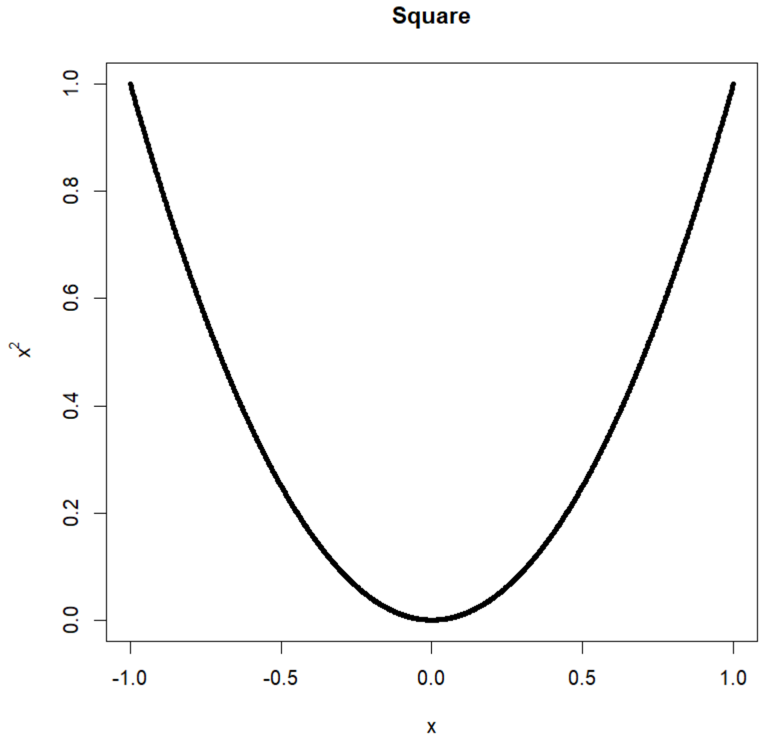
$$\min_{\beta} \sum_i \left|y_i - x_i^T \beta\right|$$

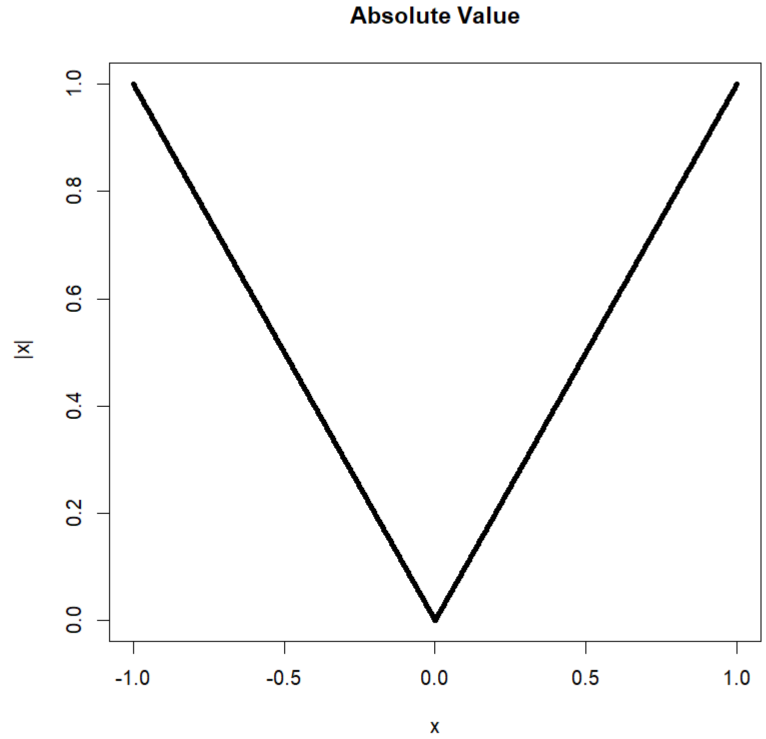
Laplace

Model

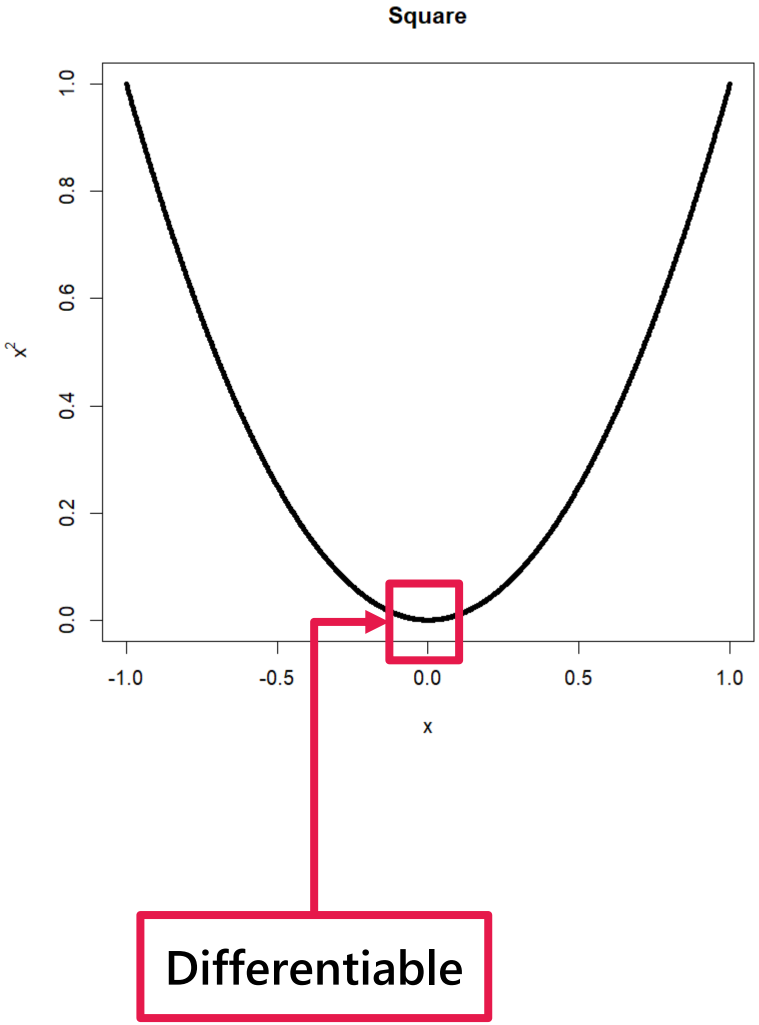
$$y_i = x_i^T \beta + \sigma \boxed{\varepsilon_i}$$

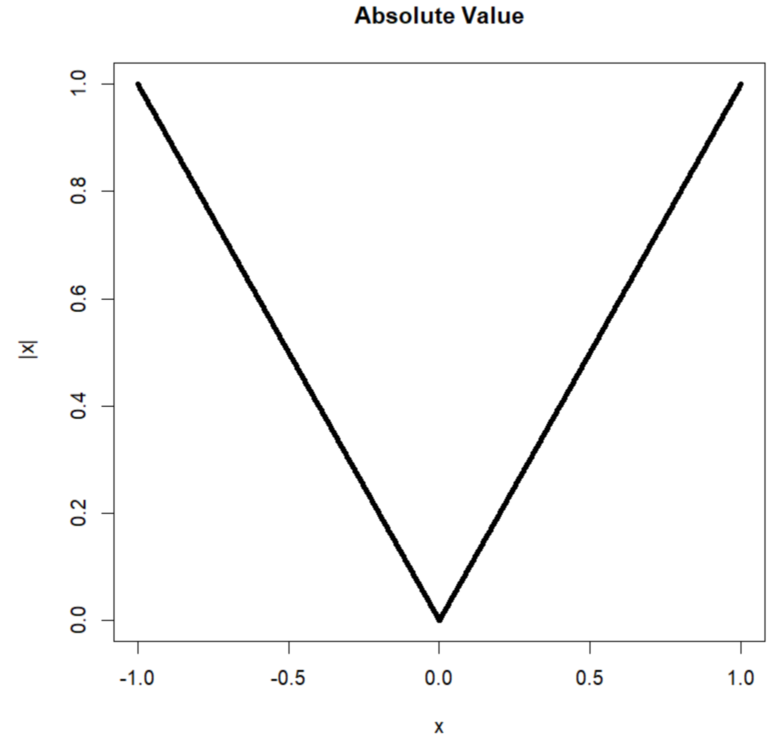
SQUARE VS ABSOLUTE VALUE FUNCTION



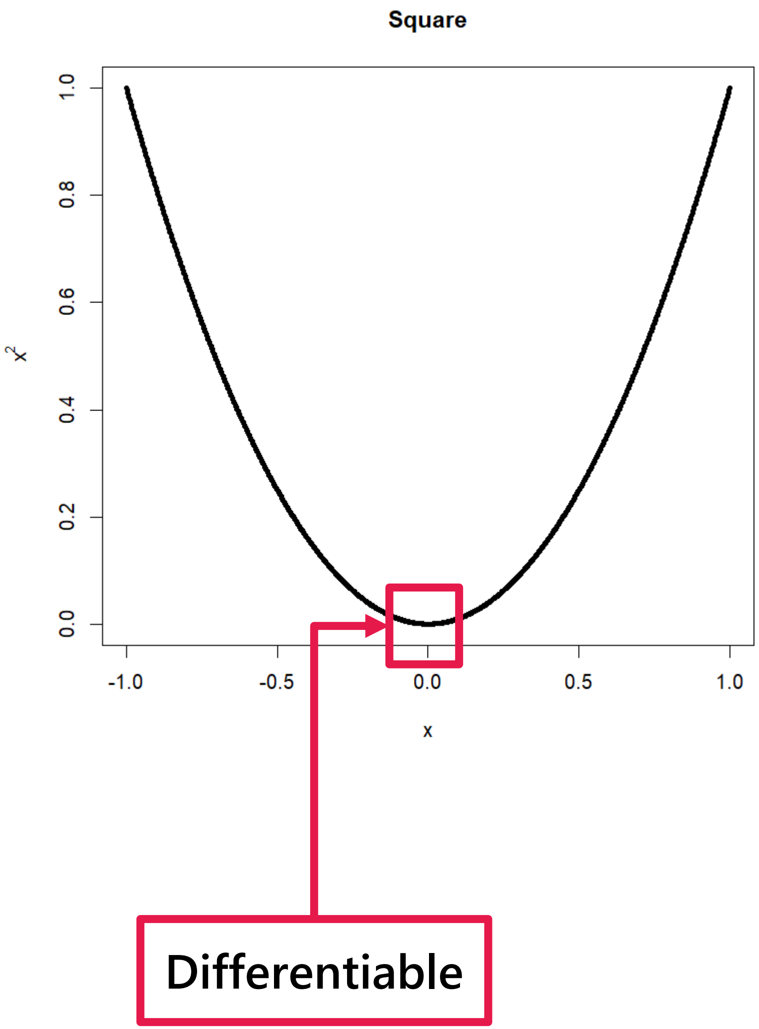


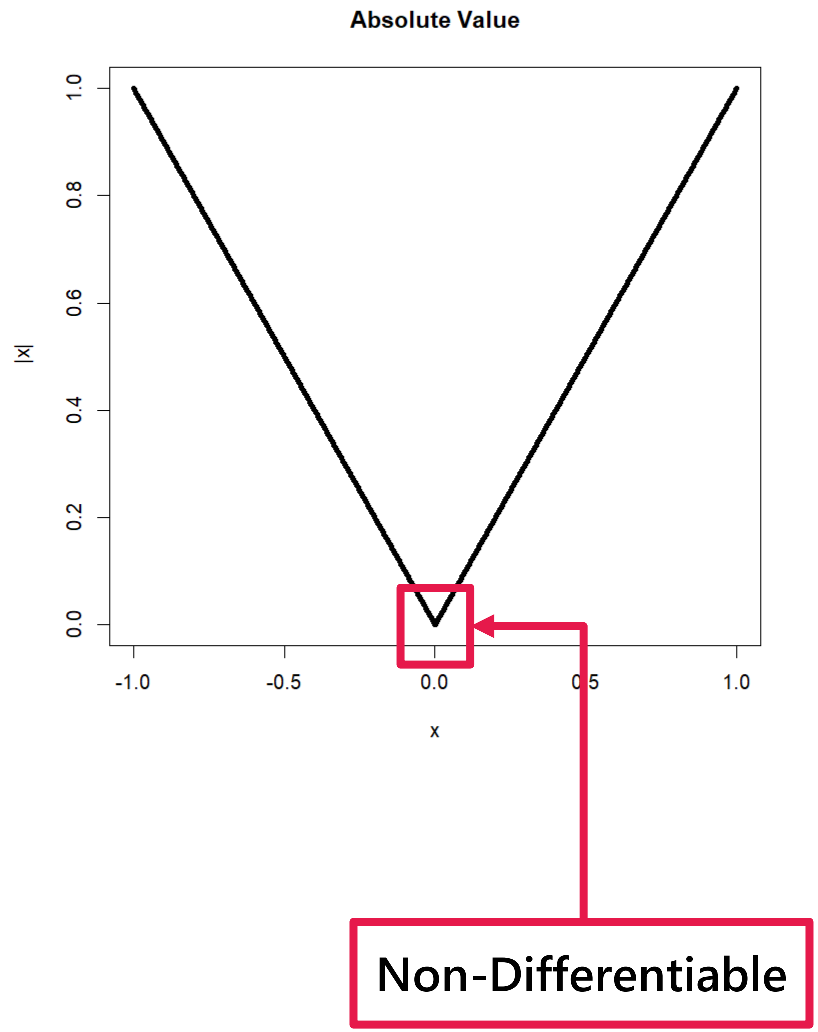
SQUARE VS ABSOLUTE VALUE FUNCTION





SQUARE VS ABSOLUTE VALUE FUNCTION





DIFFERENTIABLE APPROXIMATION

$$a_\tau(x) = \sqrt{x^2 + \tau^2} - \tau$$

Jaouimaa, Ha, & Burke (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

DIFFERENTIABLE APPROXIMATION

$$\sqrt{x^2} = |x|$$

$$a_\tau(x) = \overbrace{\sqrt{x^2 + \tau^2}} - \tau$$

Jaouimaa, Ha, & Burke (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

DIFFERENTIABLE APPROXIMATION

$$\sqrt{x^2} = |x|$$

$$a_\tau(x) = \underbrace{\sqrt{x^2 \boxed{+\,\tau^2}}} - \tau$$

$$a'_\tau(x) = \frac{x}{\sqrt{x^2 + \tau^2}}$$

(differentiable)

Jaouimaa, Ha, & Burke (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

DIFFERENTIABLE APPROXIMATION

$$\sqrt{x^2} = |x|$$

$$a_\tau(0) = 0 \quad (\text{since } |0| = 0)$$

$$a_\tau(x) = \underbrace{\sqrt{x^2 \boxed{+ \tau^2}}}_{} \boxed{- \tau}$$

$$a'_\tau(x) = \frac{x}{\sqrt{x^2 + \tau^2}}$$

(differentiable)

Jaouimaa, Ha, & Burke (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

DIFFERENTIABLE APPROXIMATION

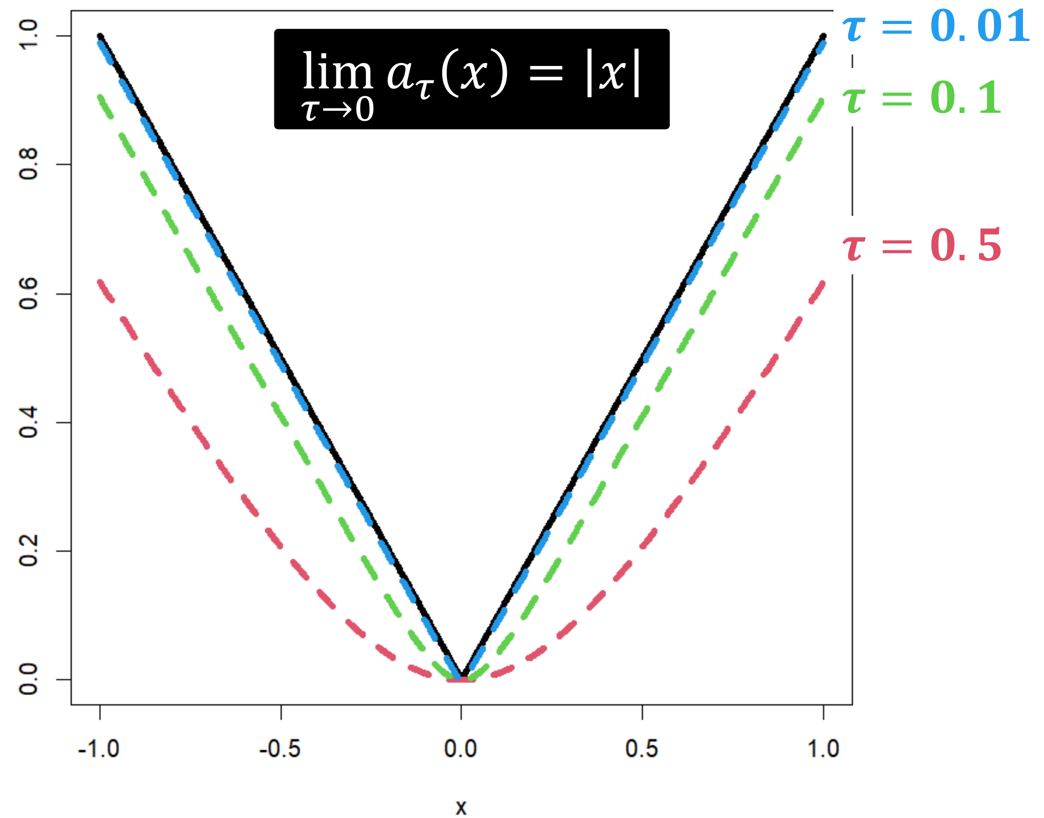
$$\sqrt{x^2} = |x|$$

$$a_\tau(0) = 0 \quad (\text{since } |0| = 0)$$

$$a_\tau(x) = \overbrace{\sqrt{x^2 \boxed{+\tau^2}}}^{} \boxed{-\tau}$$

$$a_\tau'(x) = \frac{x}{\sqrt{x^2+\tau^2}}$$

(differentiable)



Jaouimaa, Ha, & Burke (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

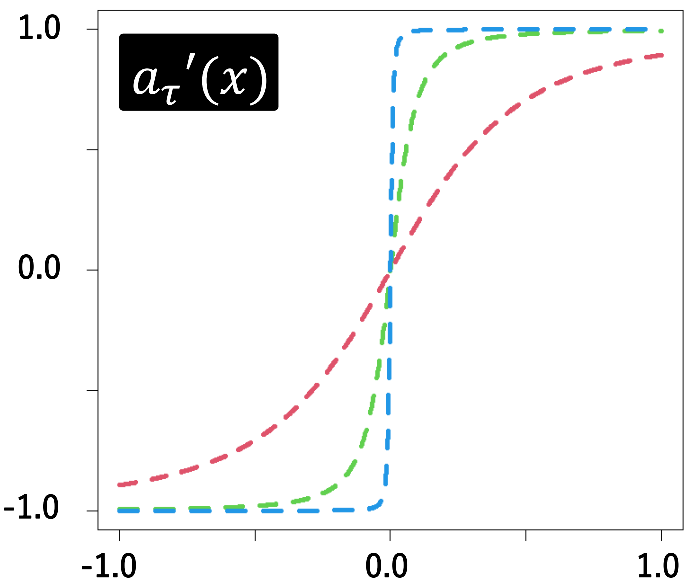
DIFFERENTIABLE APPROXIMATION

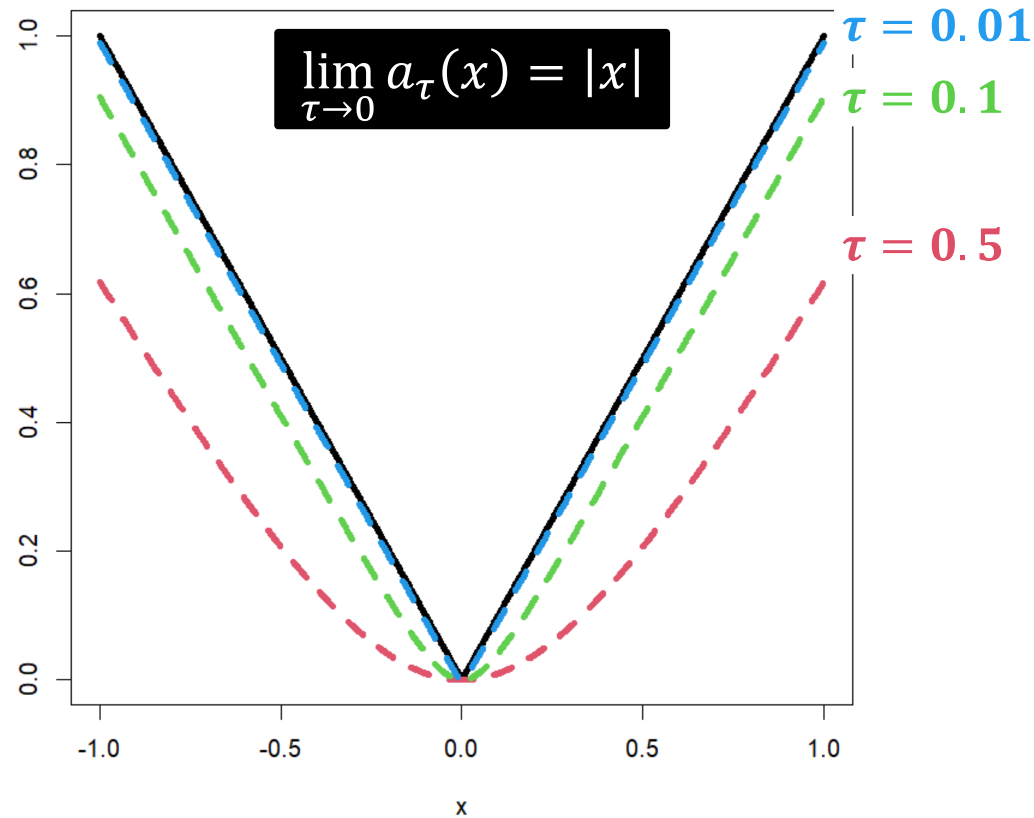
$a_\tau(0) = 0 \quad (\text{since } |0| = 0)$

$$a_\tau(x) = \overbrace{\sqrt{x^2 \boxed{+\,\tau^2}}}^{\sqrt{x^2} = |x|} \boxed{-\,\tau}$$

$$a_\tau'(x) = \frac{x}{\sqrt{x^2 + \tau^2}}$$

(differentiable)





Jaouimaa, Ha, & Burke (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
Burke & Patilea (2021). A likelihood-based approach for cure regression models. TEST.

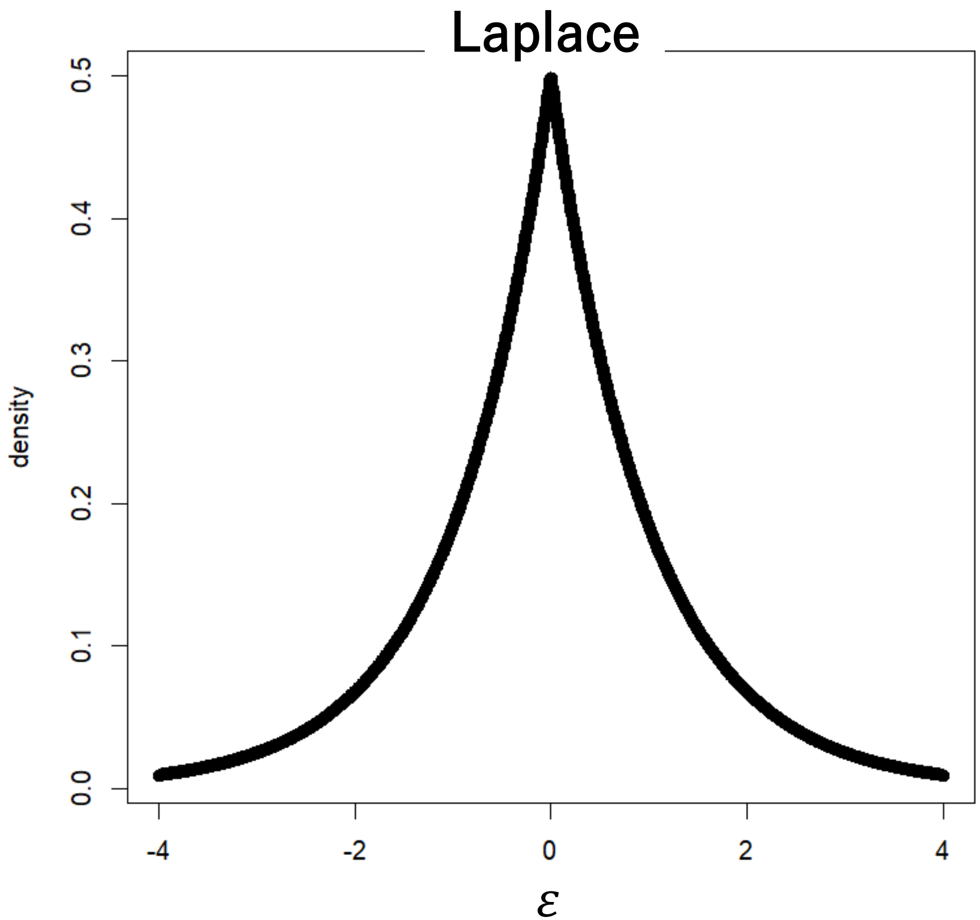
SMOOTH LAPLACE DISTRIBUTION

$$y = x^T \beta + \sigma \varepsilon$$

SMOOTH LAPLACE DISTRIBUTION

$$y = x^T\beta + \sigma\varepsilon$$

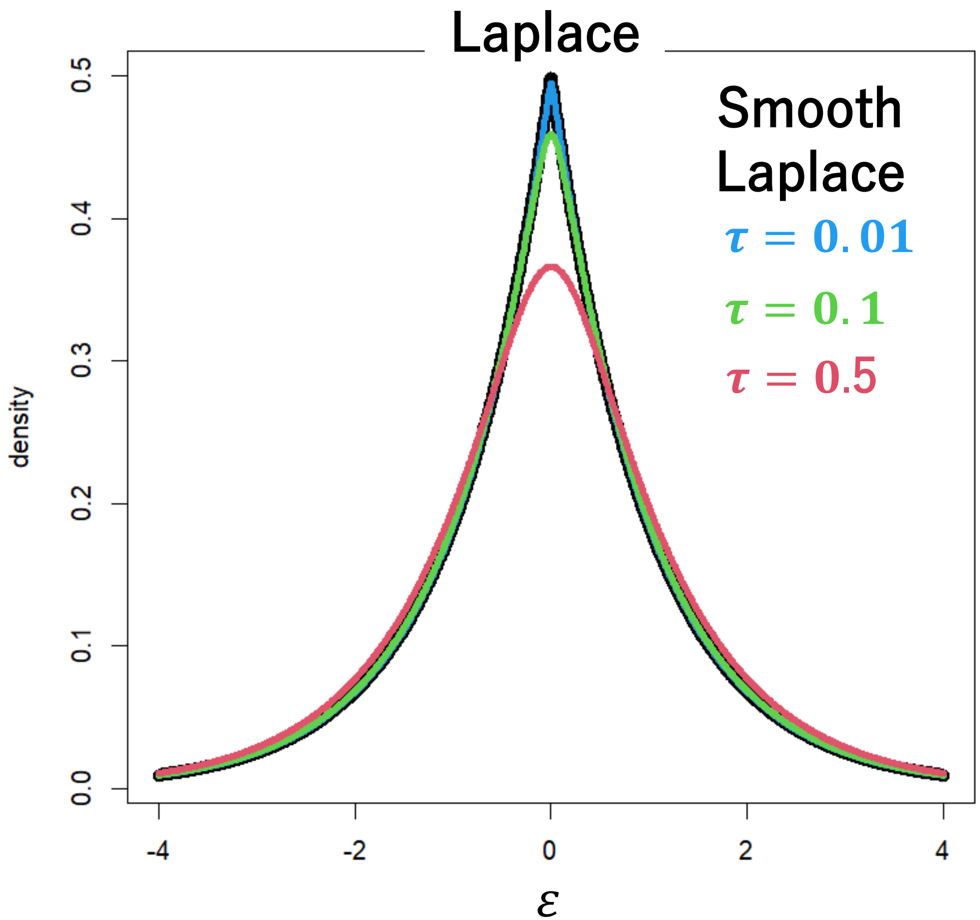
$$f(\varepsilon) = \frac{1}{2}e^{-|\varepsilon|}$$



SMOOTH LAPLACE DISTRIBUTION

$$y = x^T\beta + \sigma\varepsilon$$

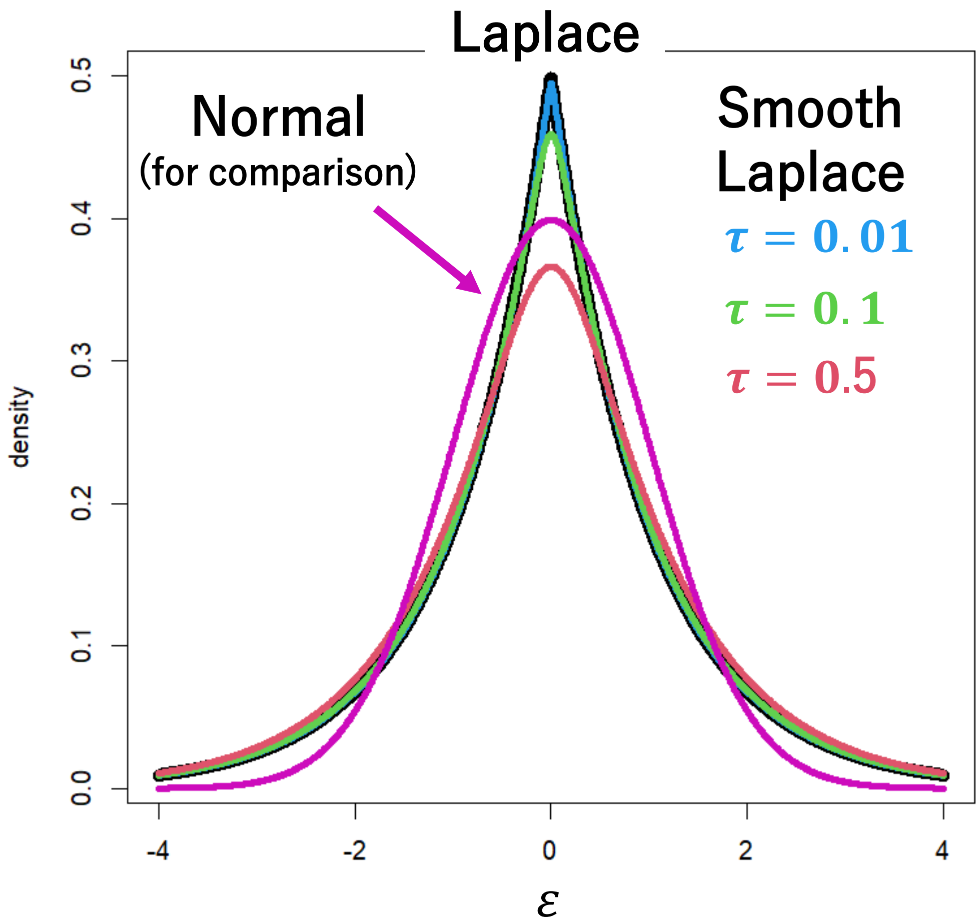
$$f(\varepsilon) = \tfrac{1}{2}e^{-|\varepsilon|} \qquad f_\tau(\varepsilon) = \boldsymbol{c_\tau} e^{-\boldsymbol{a_\tau(\varepsilon)}}$$



SMOOTH LAPLACE DISTRIBUTION

$$y = x^T\beta + \sigma\varepsilon$$

$$f(\varepsilon) = \frac{1}{2}e^{-|\varepsilon|} \qquad f_\tau(\varepsilon) = \boldsymbol{c_\tau} e^{-\boldsymbol{a_\tau(\varepsilon)}}$$



LIKELIHOOD ESTIMATION

- Log-likelihood function

$$\ell(\beta,\sigma) = n\log c_\tau - n\log\sigma - \sum_i a_\tau\left(\frac{y_i - x_i^T\beta}{\sigma}\right)$$

LIKELIHOOD ESTIMATION

- Log-likelihood function

$$\ell(\beta, \sigma) = n \log c_\tau - n \log \sigma - \sum_i a_\tau\left(\frac{y_i - x_i^T \beta}{\sigma}\right)$$

- Differentiable in β and σ

LIKELIHOOD ESTIMATION

- Log-likelihood function

$$\ell(\beta, \sigma) = n \log c_\tau - n \log \sigma - \sum_i a_\tau\left(\frac{y_i - x_i^T \beta}{\sigma}\right)$$

- Differentiable in β and σ
- Standard, gradient-based optimisation can proceed, e.g., **`nlm`**

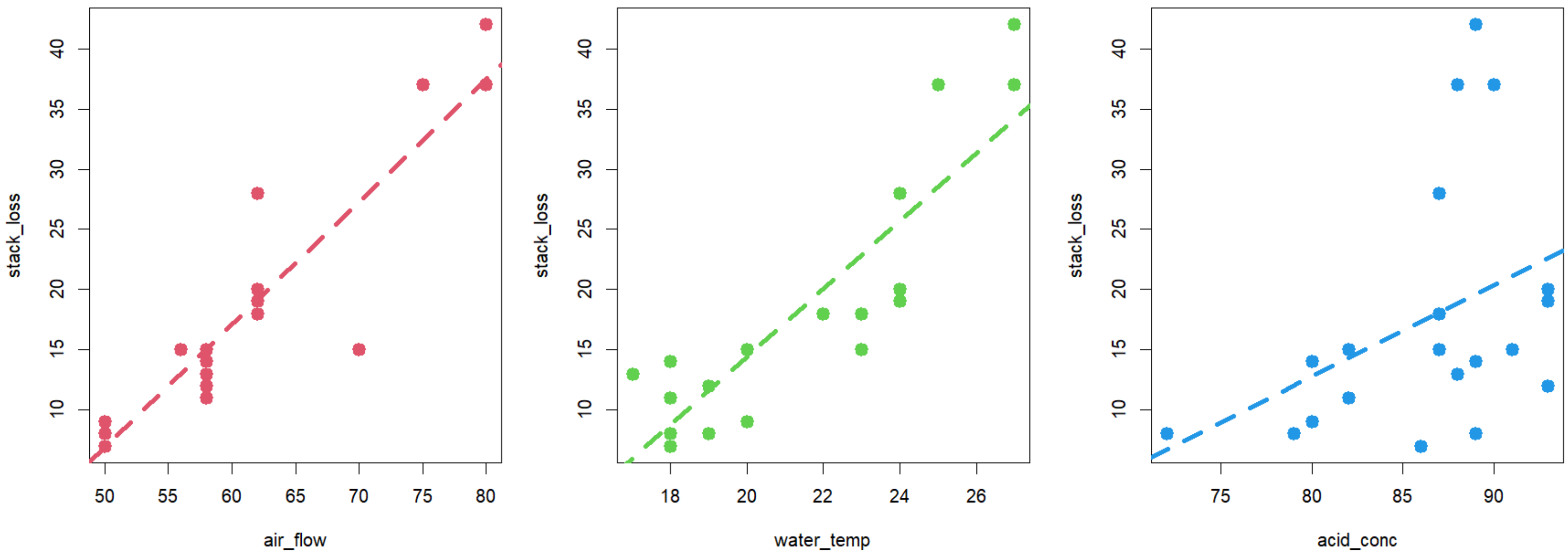
STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- **Response:** `stack_loss` (inefficiency)
- **Inputs:** `air_flow`, `water_temp`, `acid_conc`

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- **Response:** `stack_loss` (inefficiency)
- **Inputs:** `air_flow`, `water_temp`, `acid_conc`



Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- **Response:** `stack_loss` (inefficiency)
- **Inputs:** `air_flow`, `water_temp`, `acid_conc`

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- **Response:** `stack_loss` (inefficiency)
- **Inputs:** `air_flow`, `water_temp`, `acid_conc`

	LAD	0.01
`air_flow`	0.83	0.83
`water_temp`	0.57	0.57
`acid_conc`	-0.06	-0.06

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- Response: `stack_loss` (inefficiency)
- Inputs: `air_flow`, `water_temp`, `acid_conc`

`L1pack::lad`

	LAD	0.01
`air_flow`	0.83	0.83
`water_temp`	0.57	0.57
`acid_conc`	-0.06	-0.06

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- Response: `stack_loss` (inefficiency)
- Inputs: `air_flow`, `water_temp`, `acid_conc`

`L1pack::lad` τ

	LAD	0.01
`air_flow`	0.83	0.83
`water_temp`	0.57	0.57
`acid_conc`	-0.06	-0.06

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- **Response:** `stack_loss` (inefficiency)
- **Inputs:** `air_flow`, `water_temp`, `acid_conc`

`L1pack::lad` τ larger τ

	LAD	0.01	0.1	0.5
`air_flow`	0.83	0.83	0.83	0.83
`water_temp`	0.57	0.57	0.59	0.69
`acid_conc`	-0.06	-0.06	-0.07	-0.10

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- Response: `stack_loss` (inefficiency)
- Inputs: `air_flow`, `water_temp`, `acid_conc`

`L1pack::lad` | τ | larger τ | even larger τ

	LAD	0.01	0.1	0.5	2	5	10
`air_flow`	0.83	0.83	0.83	0.83	0.81	0.77	0.75
`water_temp`	0.57	0.57	0.59	0.69	0.92	1.09	1.18
`acid_conc`	-0.06	-0.06	-0.07	-0.10	-0.12	-0.14	-0.14

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- Response: `stack_loss` (inefficiency)
- Inputs: `air_flow`, `water_temp`, `acid_conc`

`L1pack::lad` τ larger τ even larger τ

	LAD	0.01	0.1	0.5	2	5	10	lm
`air_flow`	0.83	0.83	0.83	0.83	0.81	0.77	0.75	0.72
`water_temp`	0.57	0.57	0.59	0.69	0.92	1.09	1.18	1.30
`acid_conc`	-0.06	-0.06	-0.07	-0.10	-0.12	-0.14	-0.14	-0.15

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

STACK LOSS DATA FIT

- "`stackloss`": data on industrial process for oxidising ammonia to nitric acid
- **Response:** `stack_loss` (inefficiency)
- **Inputs:** `air_flow`, `water_temp`, `acid_conc`

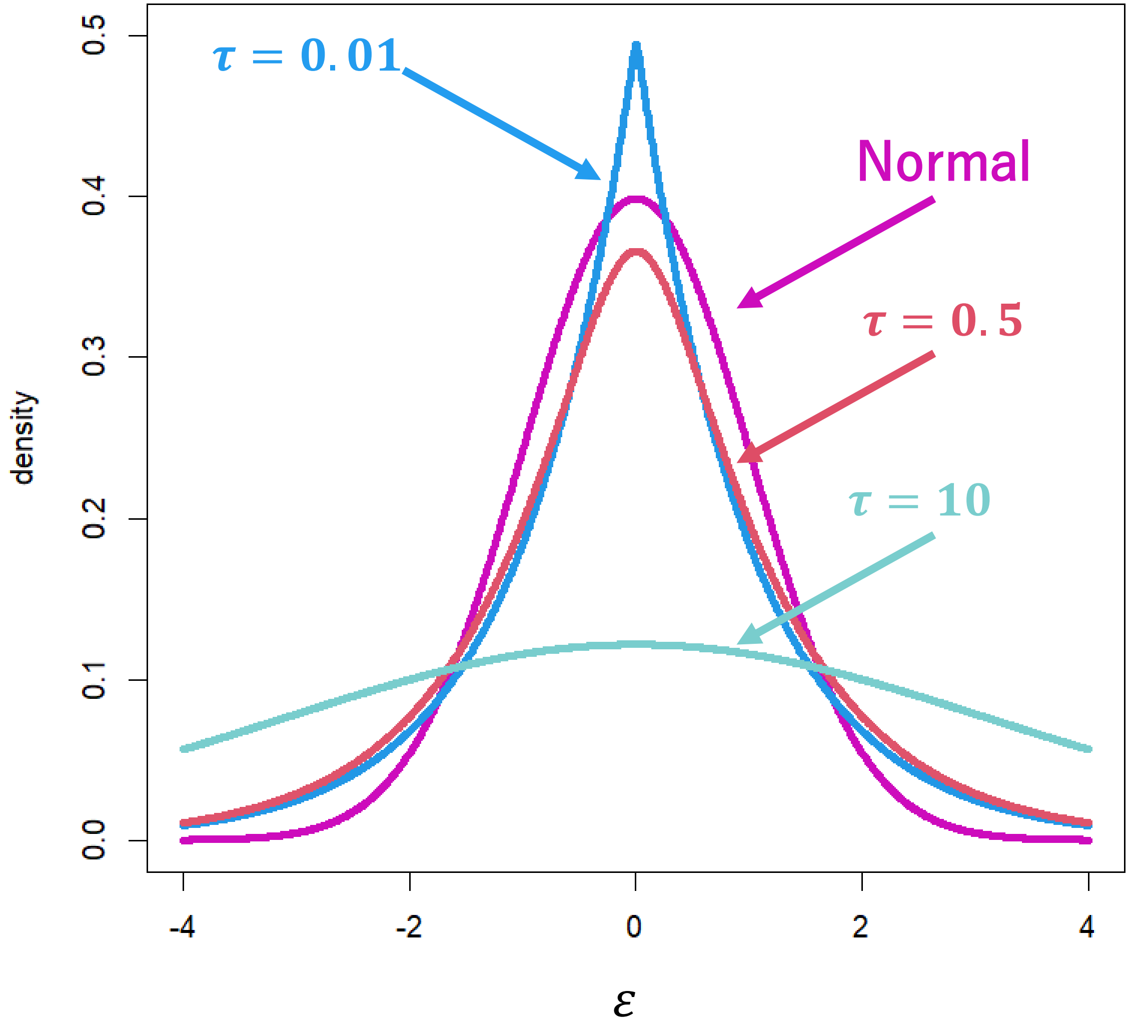
`L1pack::lad` τ larger τ even larger τ

	LAD	0.01	0.1	0.5	2	5	10	lm
`air_flow`	0.83	0.83	0.83	0.83	0.81	0.77	0.75	0.72
`water_temp`	0.57	0.57	0.59	0.69	0.92	1.09	1.18	1.30
`acid_conc`	-0.06	-0.06	-0.07	-0.10	-0.12	-0.14	-0.14	-0.15

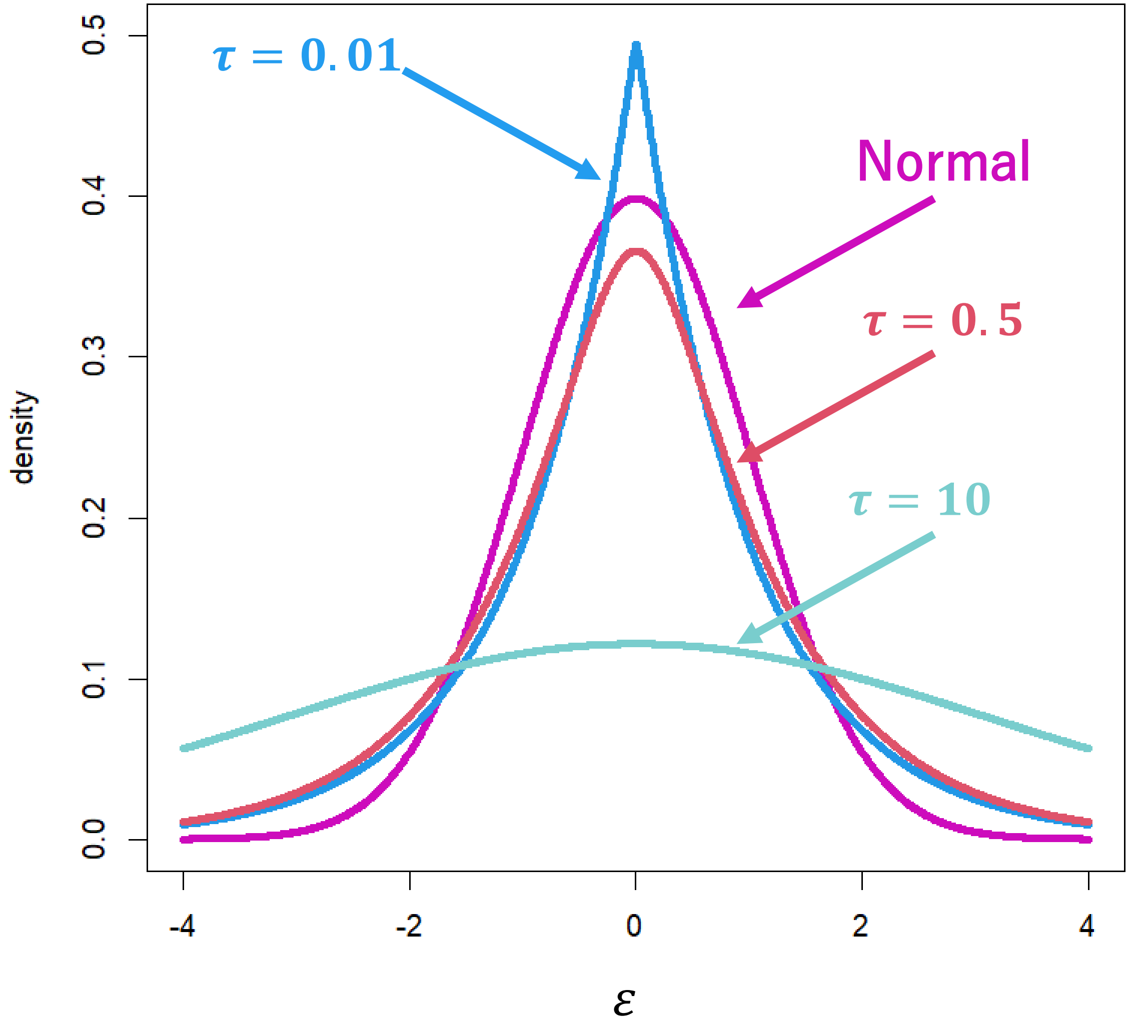
..tending to least squares?

Brownlee (1960) Statistical Theory and Methodology in Science and Engineering

INCREASING τ

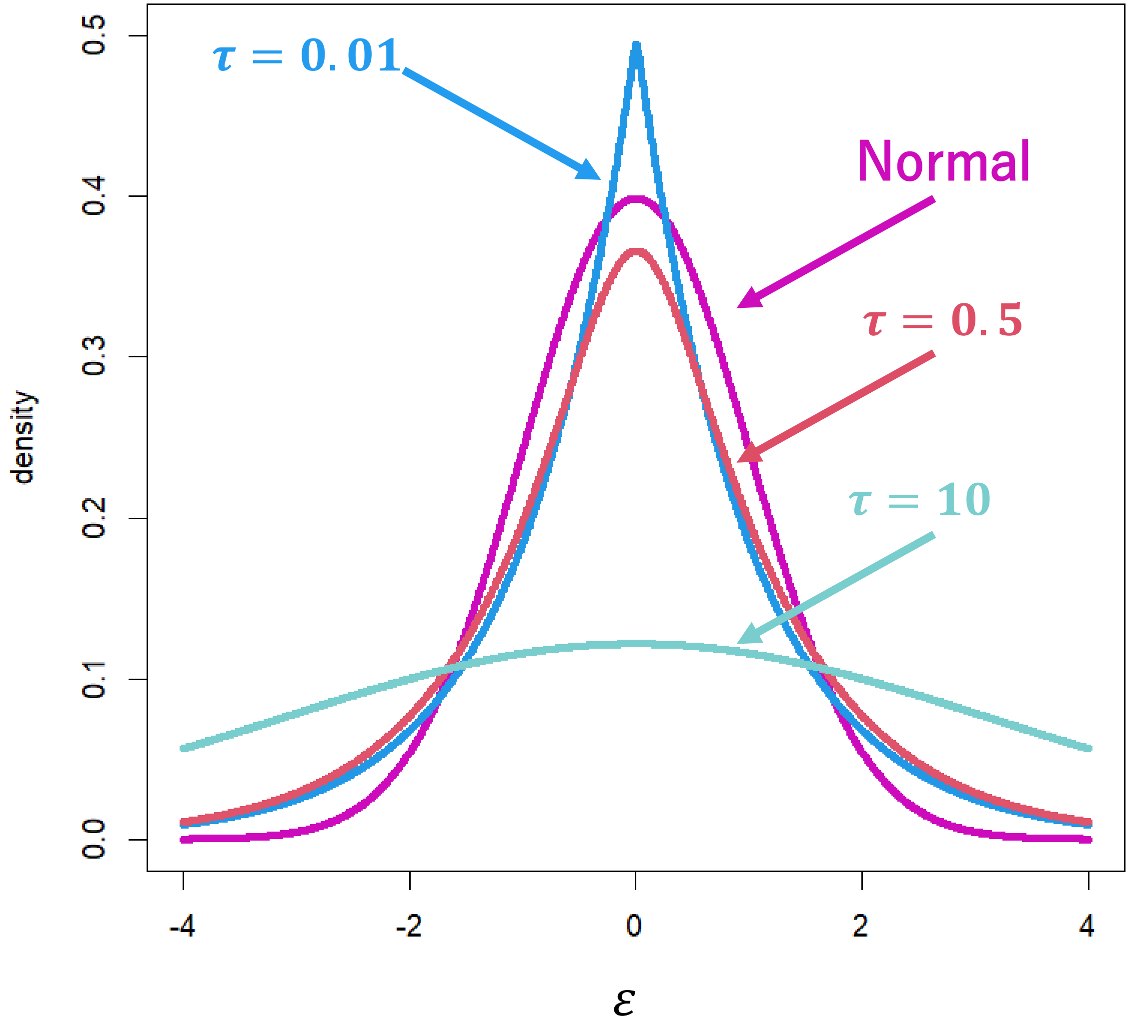


INCREASING τ



$$f_\tau(\varepsilon) = c_\tau e^{-a_\tau(\varepsilon)}$$

INCREASING τ

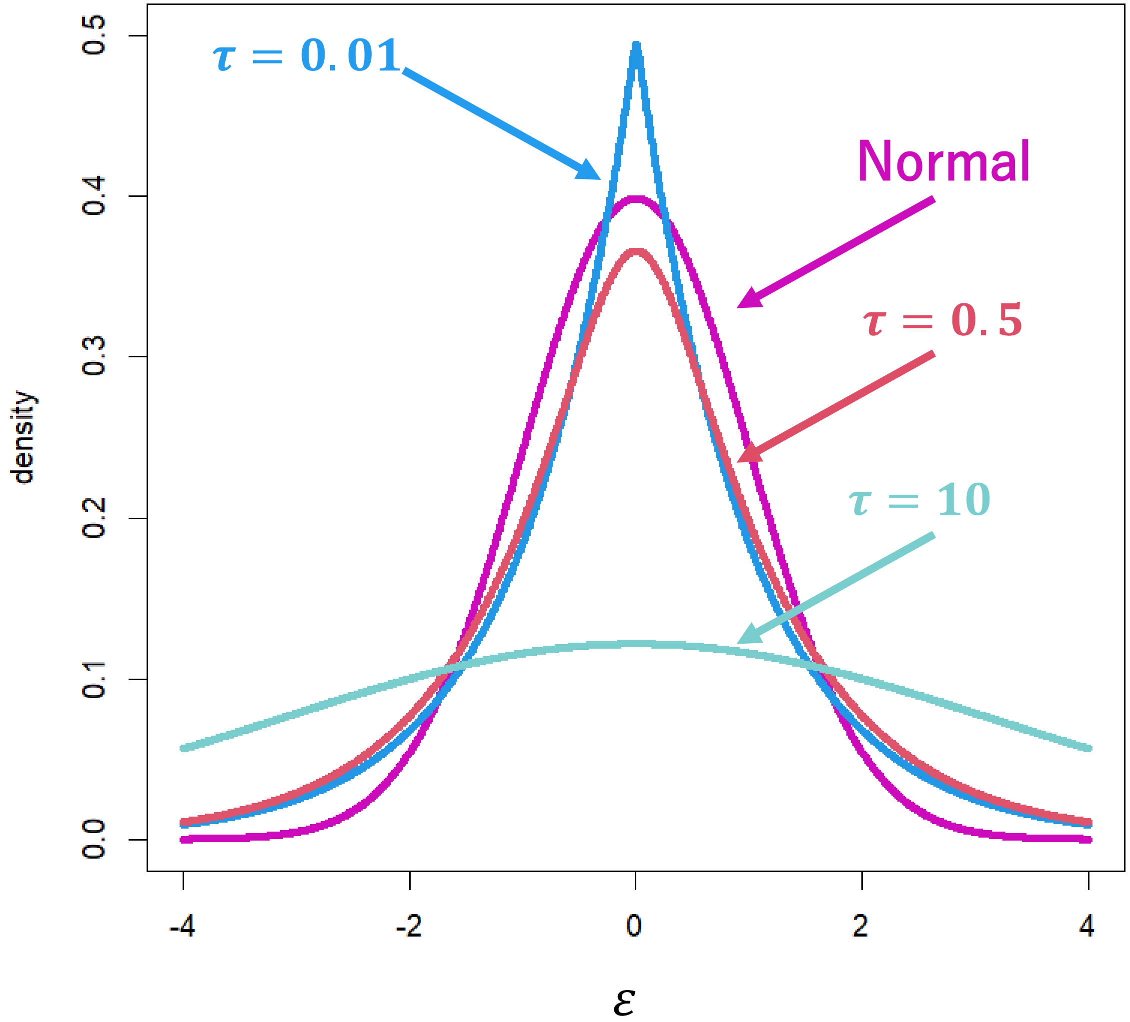


$$f_\tau(\varepsilon) = c_\tau e^{-a_\tau(\varepsilon)}$$

Small τ

$a_\tau(\varepsilon) \sim |\varepsilon|$

INCREASING τ



$$f_\tau(\varepsilon) = c_\tau e^{-a_\tau(\varepsilon)}$$

Small τ

$a_\tau(\varepsilon) \sim |\varepsilon|$

Large τ

$a_\tau(\varepsilon) \sim \varepsilon^2$?

BEHAVIOUR FOR LARGE τ

BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$

BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$
- Suggest using: $\tilde{a}_\tau(\varepsilon) = (\tau + 1)\, a_\tau(\varepsilon)$

BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$
- Suggest using: $\tilde{a}_\tau(\varepsilon) = (\tau + 1)\, a_\tau(\varepsilon)$
  - $\lim_{\tau \to 0} \tilde{a}_\tau(\varepsilon) = |\varepsilon|$

BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$
- Suggest using: $\tilde{a}_\tau(\varepsilon) = (\tau + 1)\, a_\tau(\varepsilon)$
  - $\lim_{\tau\to 0} \tilde{a}_\tau(\varepsilon) = |\varepsilon|$
  - $\lim_{\tau\to\infty} \tilde{a}_\tau(\varepsilon) = \varepsilon^2/2$

BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$
- Suggest using: $\tilde{a}_\tau(\varepsilon) = (\tau + 1)\, a_\tau(\varepsilon)$
  - $\lim_{\tau \to 0} \tilde{a}_\tau(\varepsilon) = |\varepsilon|$
  - $\lim_{\tau \to \infty} \tilde{a}_\tau(\varepsilon) = \varepsilon^2/2$
- $\tilde{f}_\tau(\varepsilon) = \tilde{c}_\tau e^{-\tilde{a}_\tau(\varepsilon)} = \tilde{c}_\tau e^{-(\tau+1)(\sqrt{\varepsilon^2+\tau^2}-\tau)}$

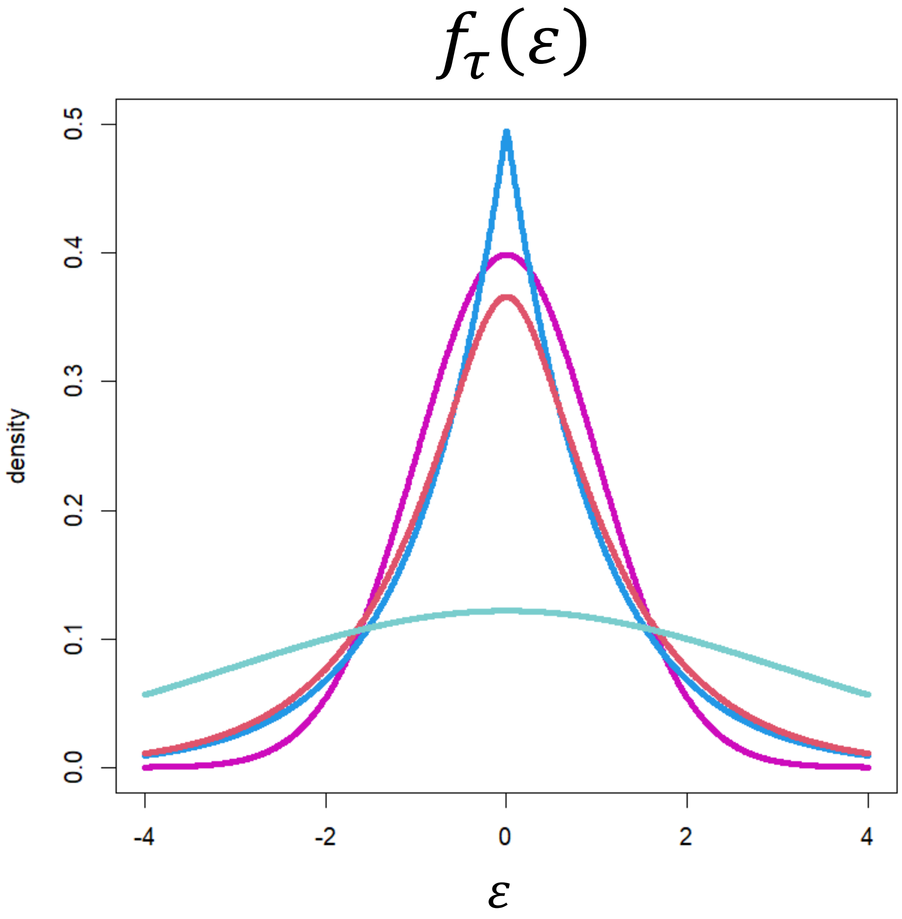
BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$
- Suggest using: $\tilde{a}_\tau(\varepsilon) = (\tau + 1)\, a_\tau(\varepsilon)$
  - $\lim_{\tau \to 0} \tilde{a}_\tau(\varepsilon) = |\varepsilon|$
  - $\lim_{\tau \to \infty} \tilde{a}_\tau(\varepsilon) = \varepsilon^2/2$
- $\tilde{f}_\tau(\varepsilon) = \tilde{c}_\tau e^{-\tilde{a}_\tau(\varepsilon)} = \tilde{c}_\tau e^{-(\tau+1)(\sqrt{\varepsilon^2+\tau^2}-\tau)}$
  - $\lim_{\tau \to 0} \tilde{f}_\tau(\varepsilon) =$ Laplace

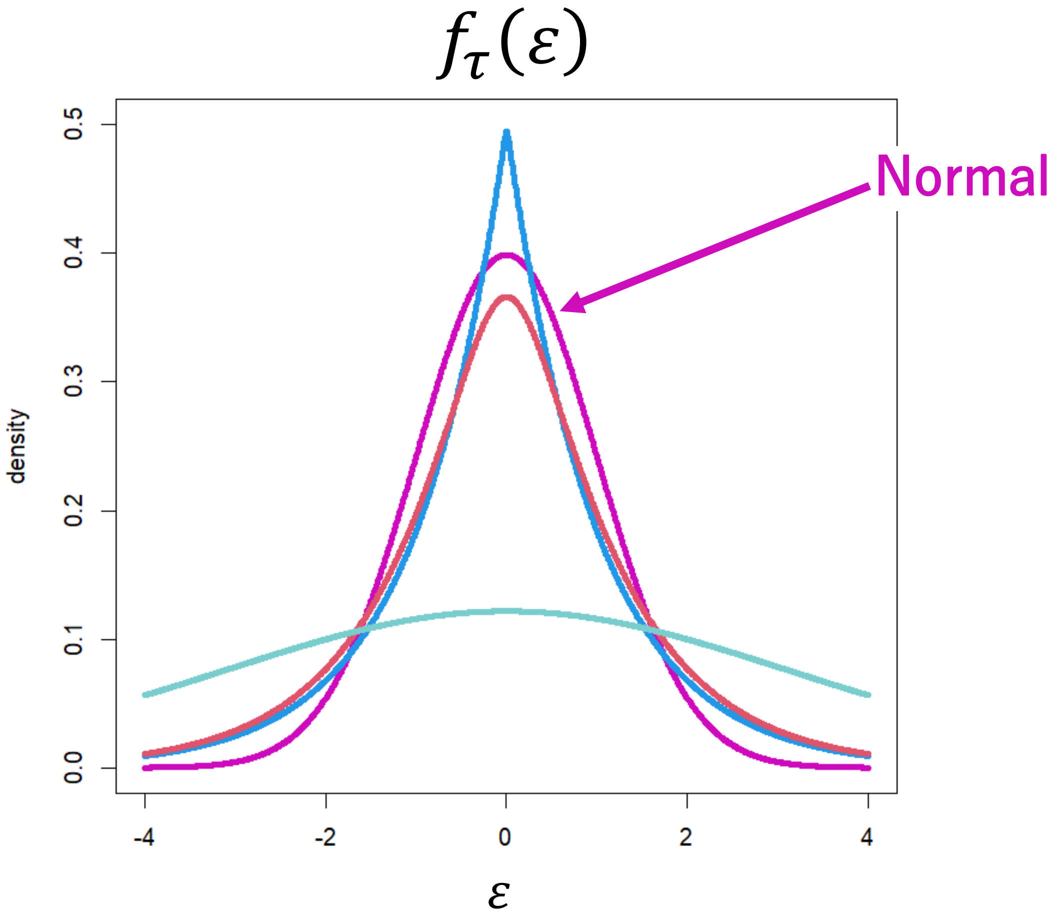
BEHAVIOUR FOR LARGE τ

- Expand at $\tau = \infty$: $a_\tau(\varepsilon) \approx \varepsilon^2/(2\tau)$
- Suggest using: $\tilde{a}_\tau(\varepsilon) = (\tau + 1)\, a_\tau(\varepsilon)$
  - $\lim\limits_{\tau \to 0} \tilde{a}_\tau(\varepsilon) = |\varepsilon|$
  - $\lim\limits_{\tau \to \infty} \tilde{a}_\tau(\varepsilon) = \varepsilon^2/2$
- $\tilde{f}_\tau(\varepsilon) = \tilde{c}_\tau e^{-\tilde{a}_\tau(\varepsilon)} = \tilde{c}_\tau e^{-(\tau+1)(\sqrt{\varepsilon^2+\tau^2}-\tau)}$
  - $\lim\limits_{\tau \to 0} \tilde{f}_\tau(\varepsilon) =$ Laplace
  - $\lim\limits_{\tau \to \infty} \tilde{f}_\tau(\varepsilon) =$ Gaussian

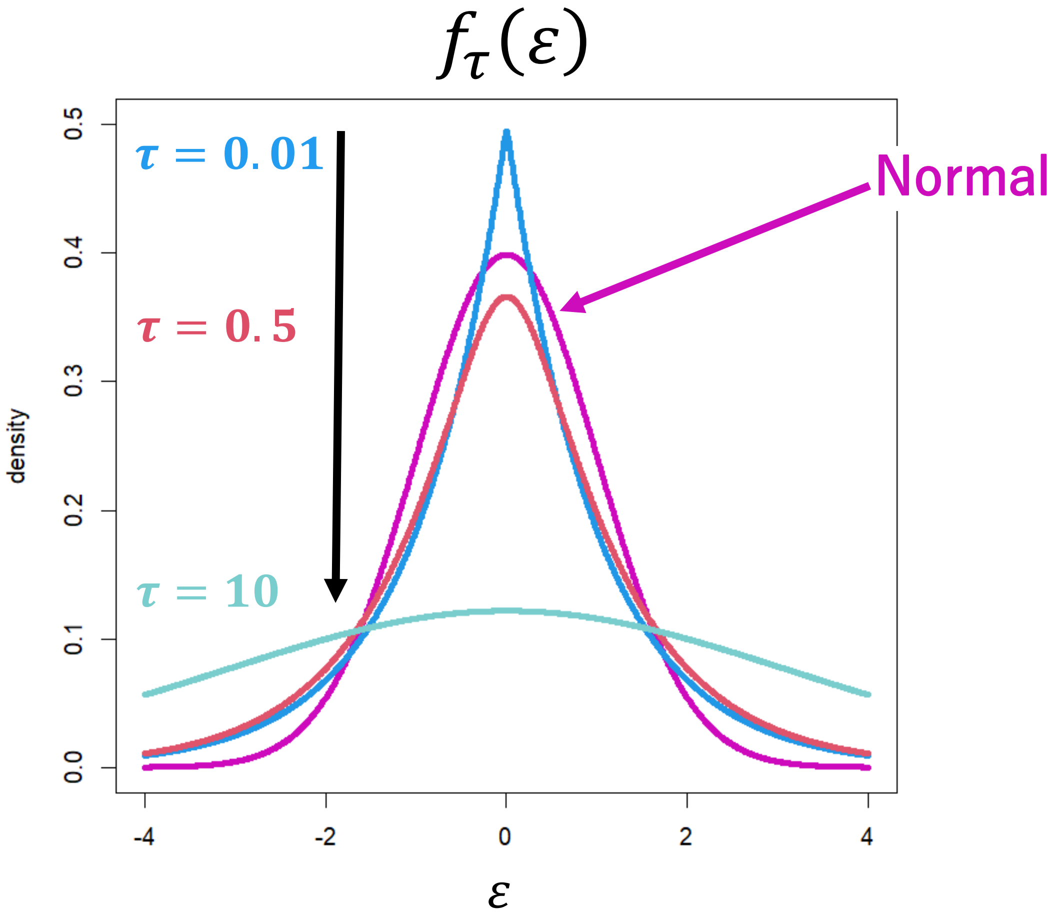
NEW PARAMETERISATION



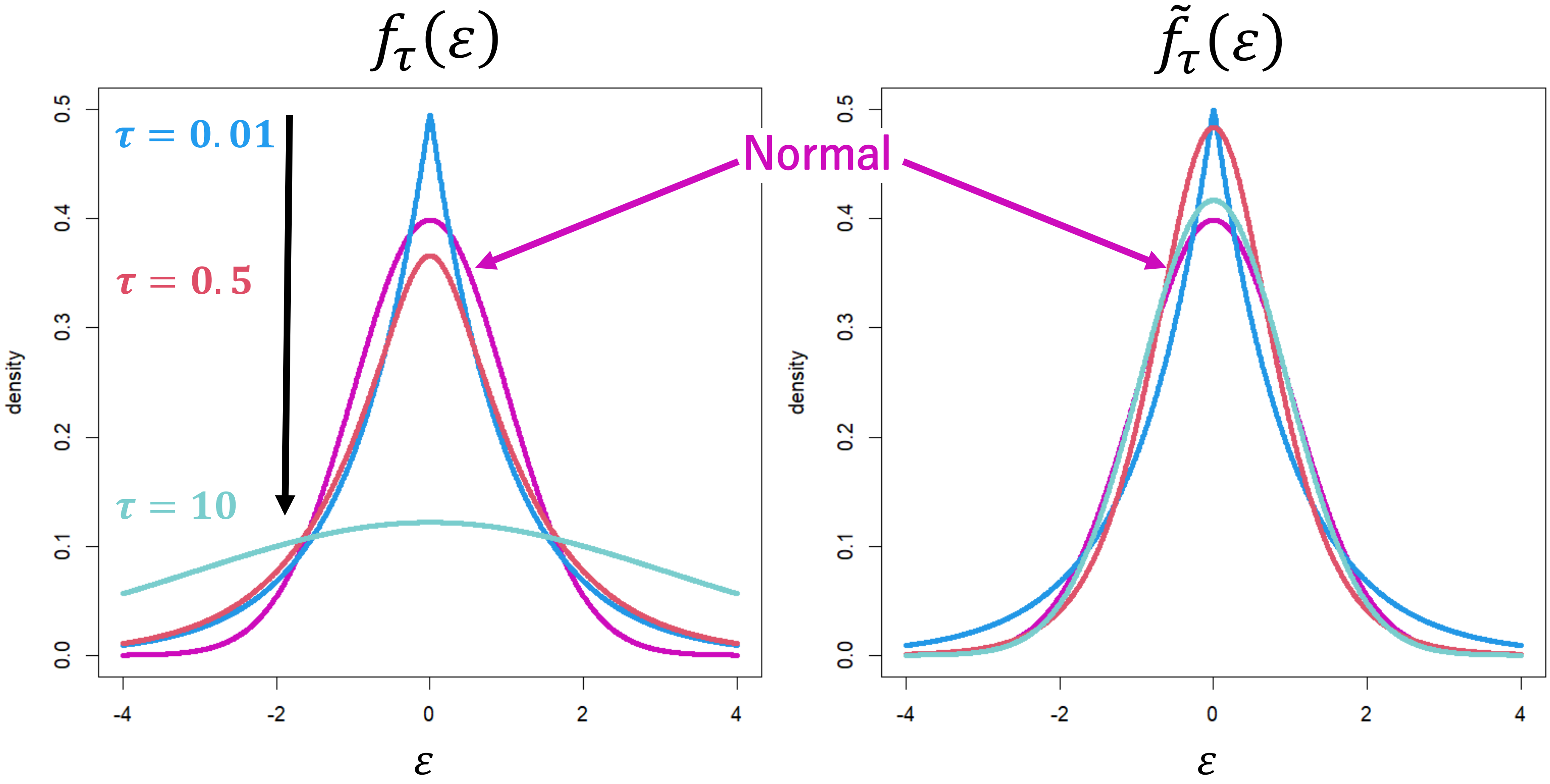
NEW PARAMETERISATION



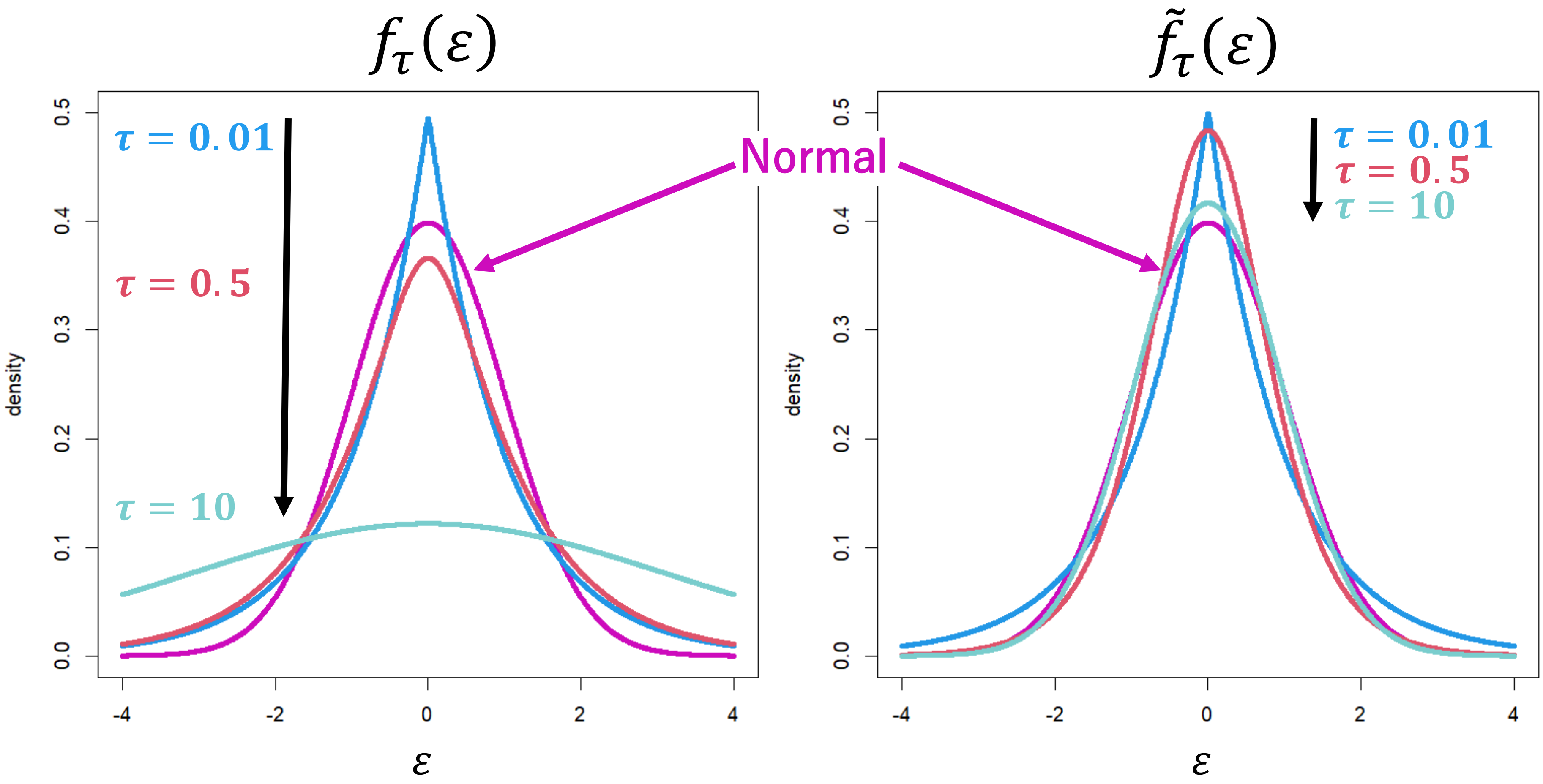
NEW PARAMETERISATION



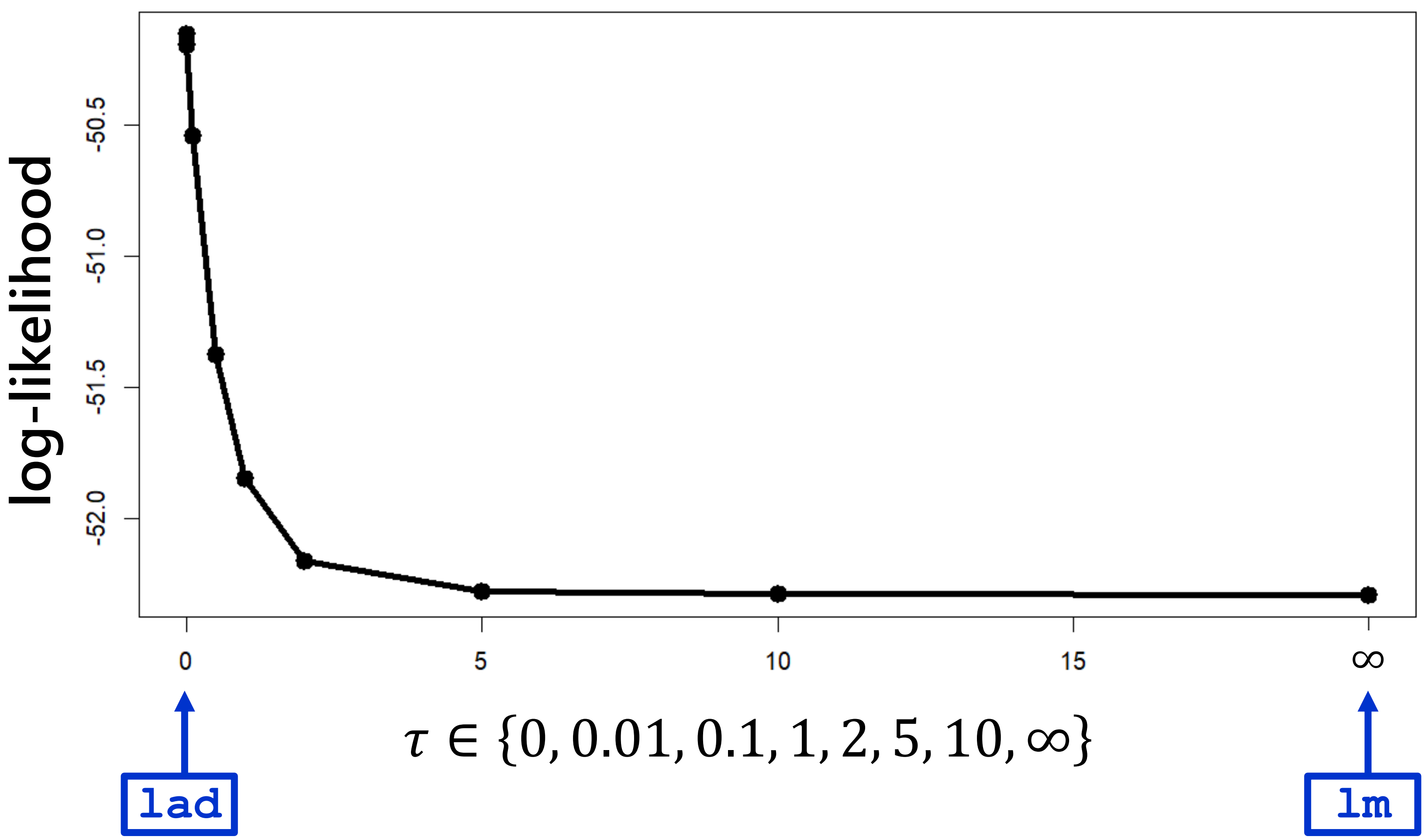
NEW PARAMETERISATION



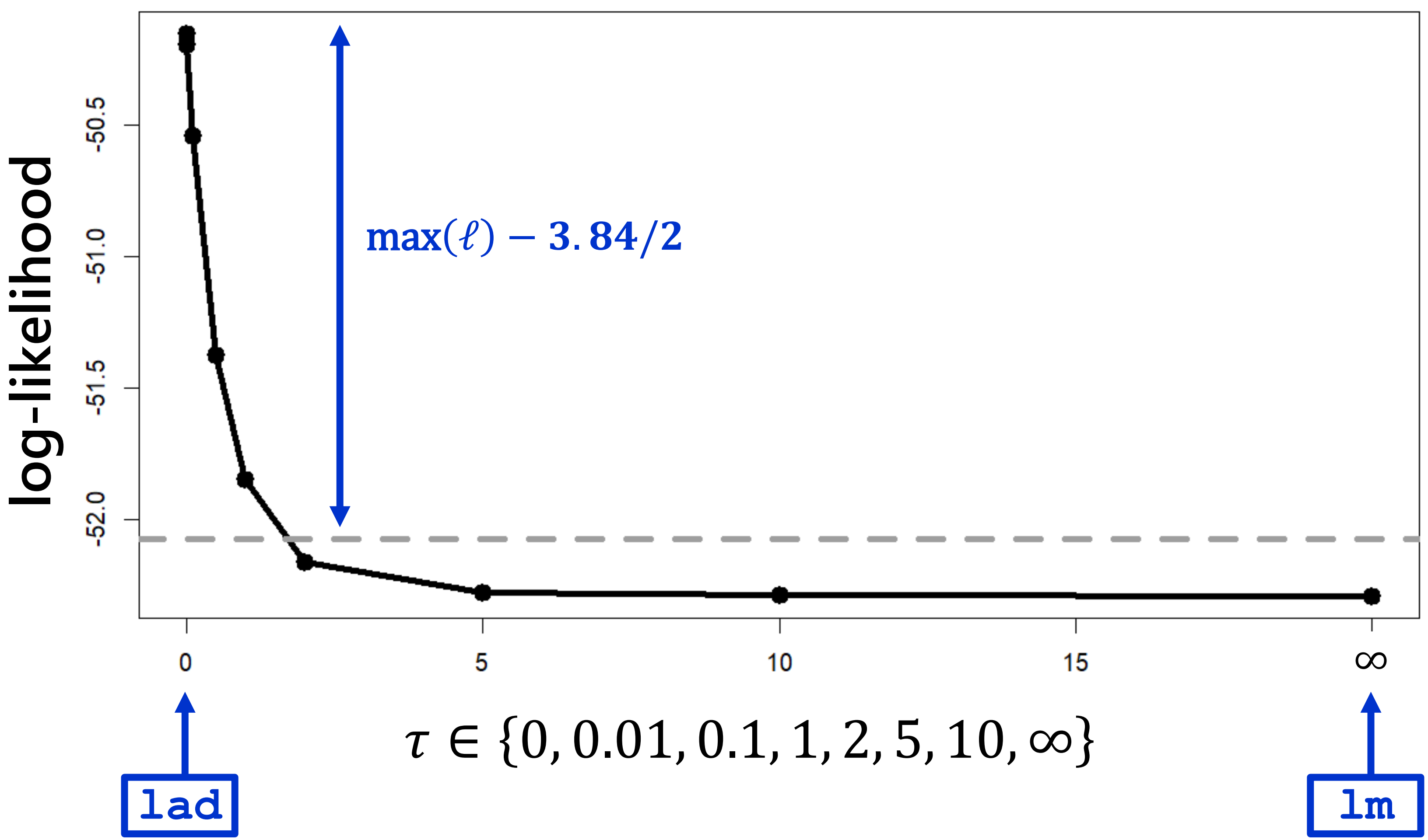
NEW PARAMETERISATION



STACKLOSS: LOG-LIKELIHOOD

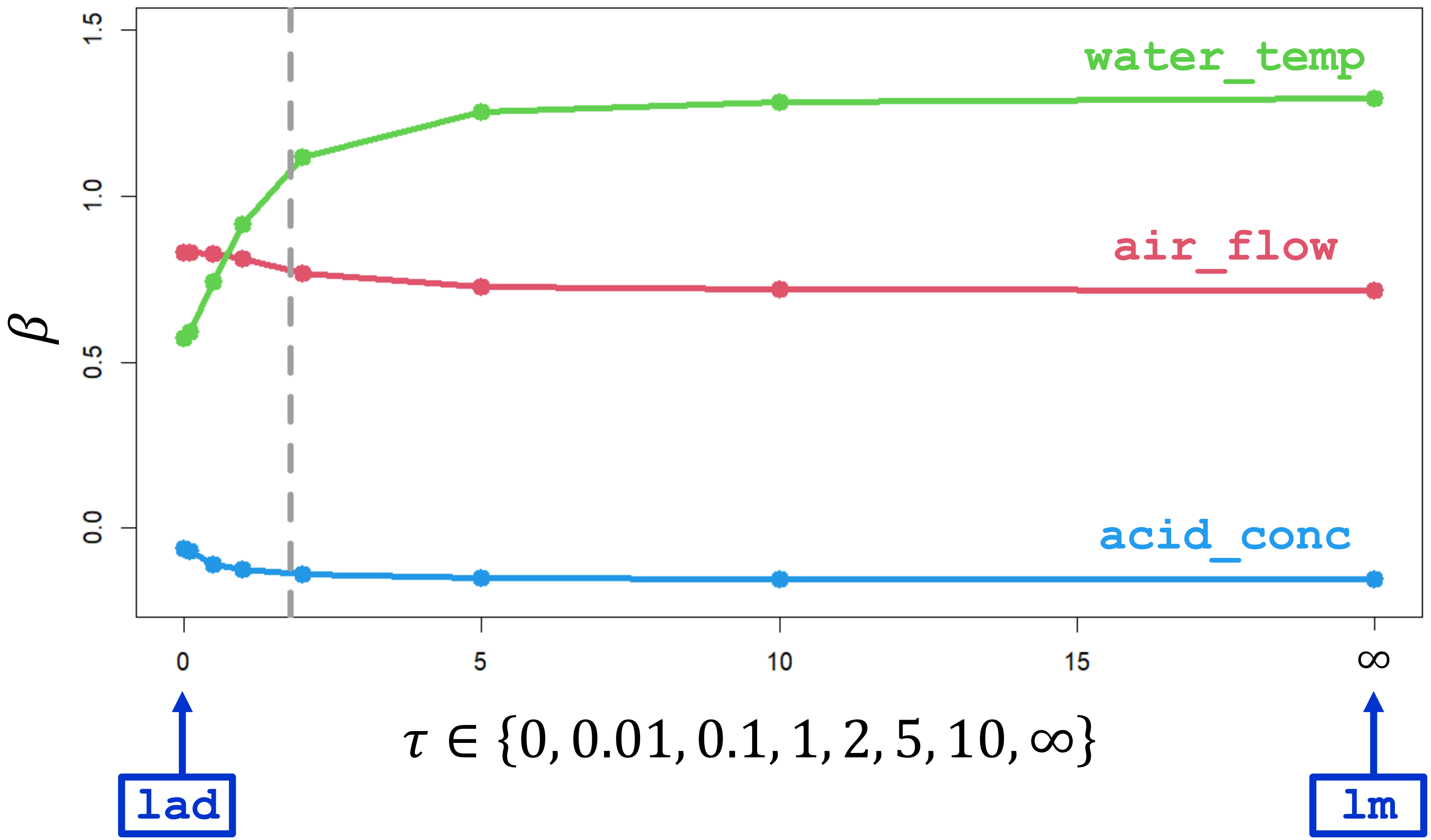


STACKLOSS: LOG-LIKELIHOOD

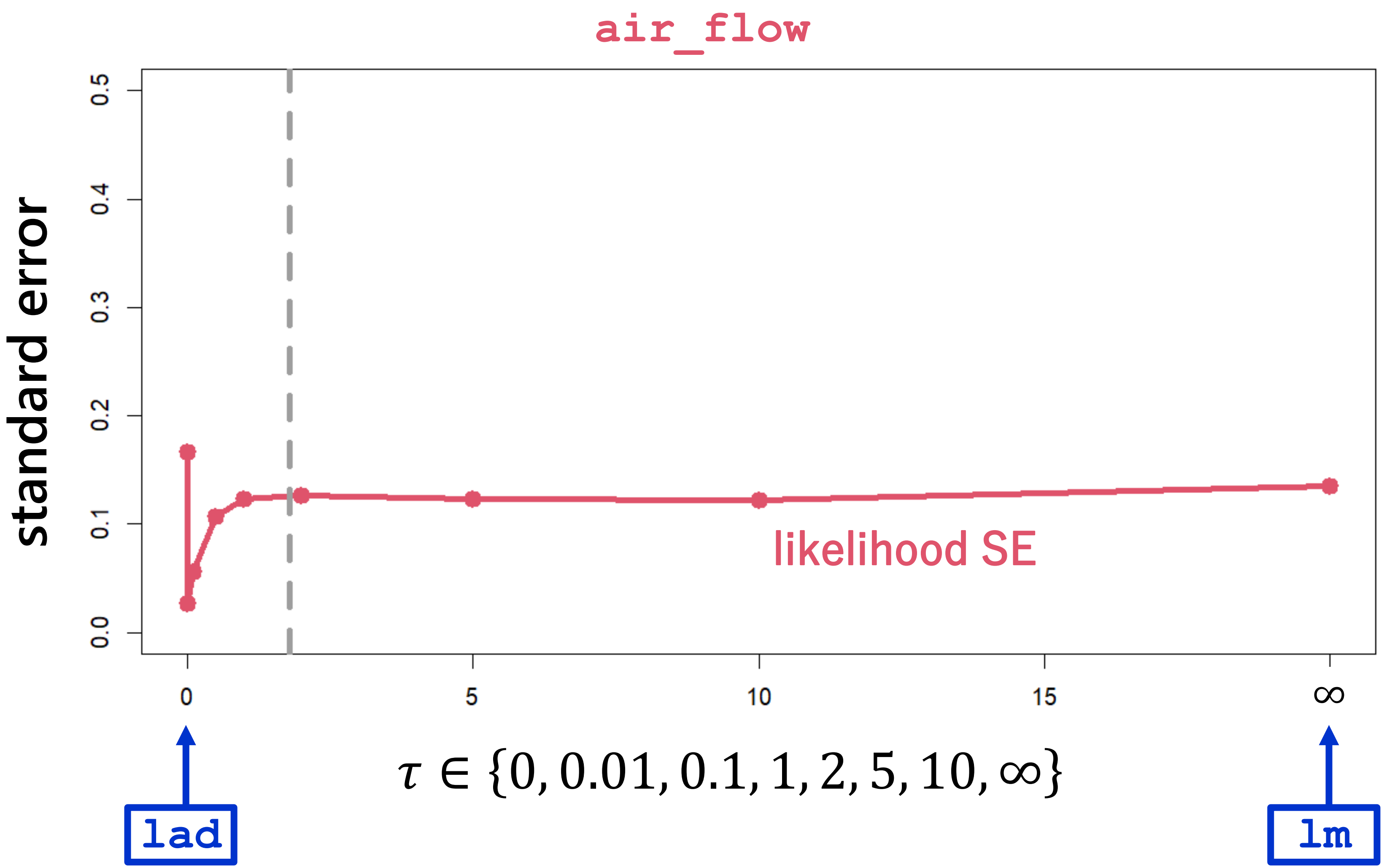


$\tau \in \{0, 0.01, 0.1, 1, 2, 5, 10, \infty\}$

STACKLOSS: BETA COEFFICIENTS

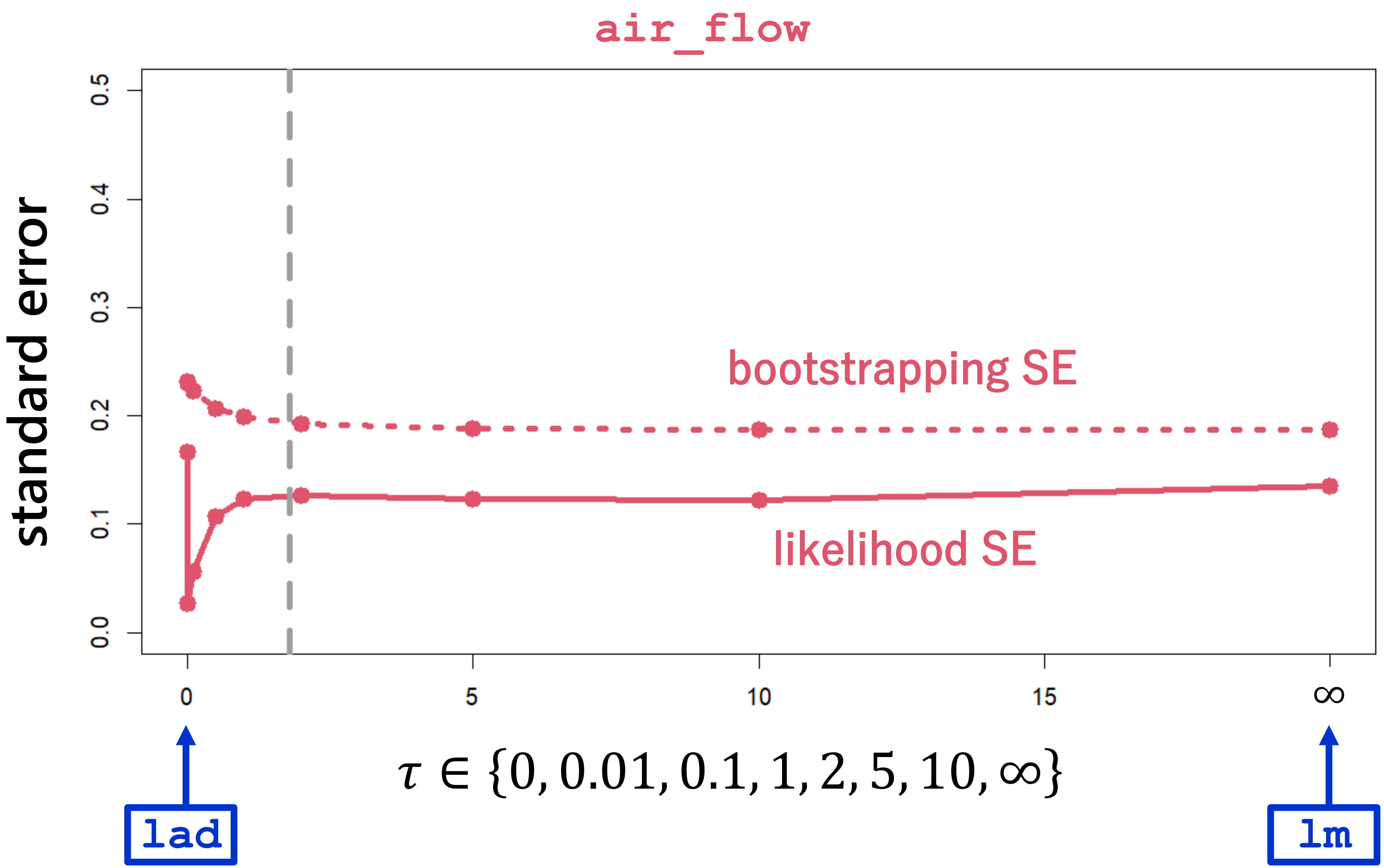


STACKLOSS: STANDARD ERRORS



$\tau \in \{0, 0.01, 0.1, 1, 2, 5, 10, \infty\}$

STACKLOSS: STANDARD ERRORS



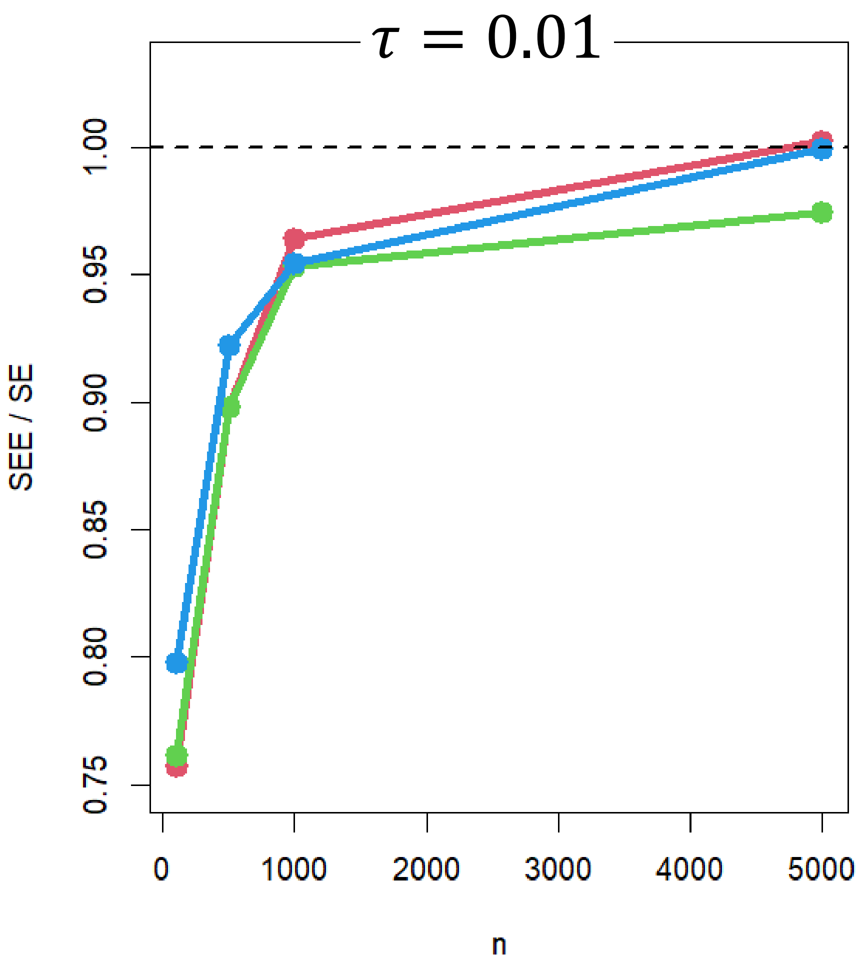
$$\tau \in \{0, 0.01, 0.1, 1, 2, 5, 10, \infty\}$$

SIMULATION: SE ESTIMATION

- $y = \beta_0 + \mathbf{1}x_1 + \mathbf{0.5}x_2 + \mathbf{0}x_3 + \sigma\varepsilon_\tau,$
- $\boldsymbol{\tau} \in \{\mathbf{0.01}, \mathbf{0.1}\}, \quad x_j \sim N(0,1), \quad n \in \{100, 500, 1000, 5000\}$

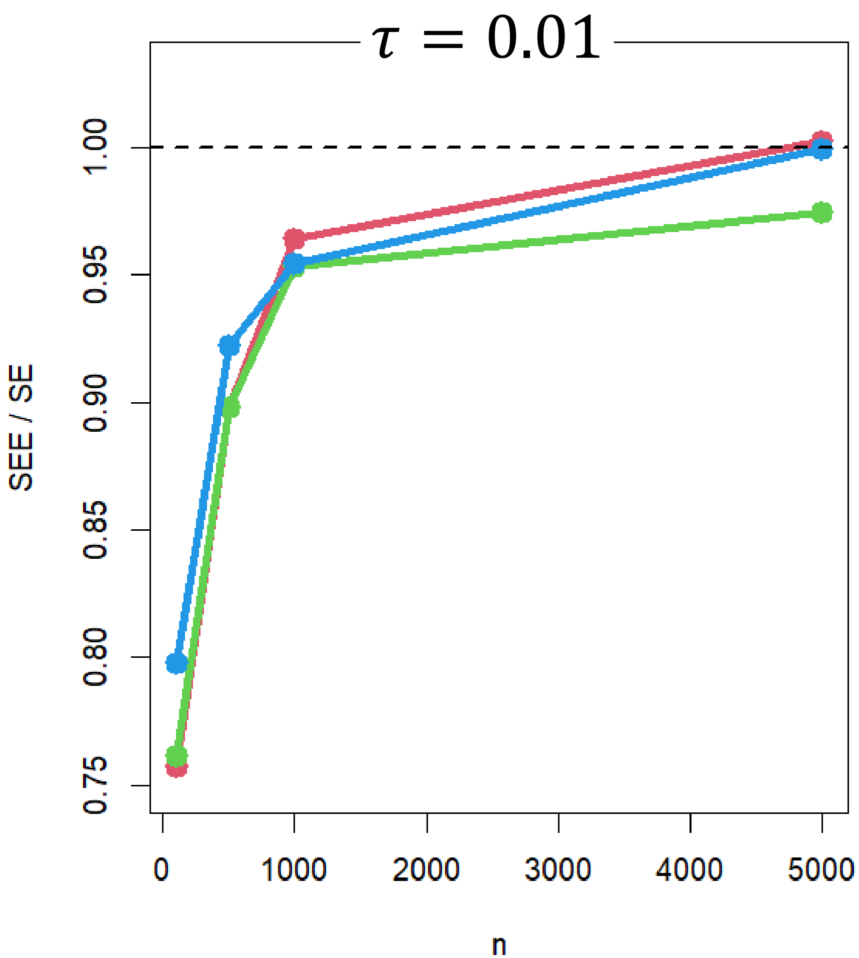
SIMULATION: SE ESTIMATION

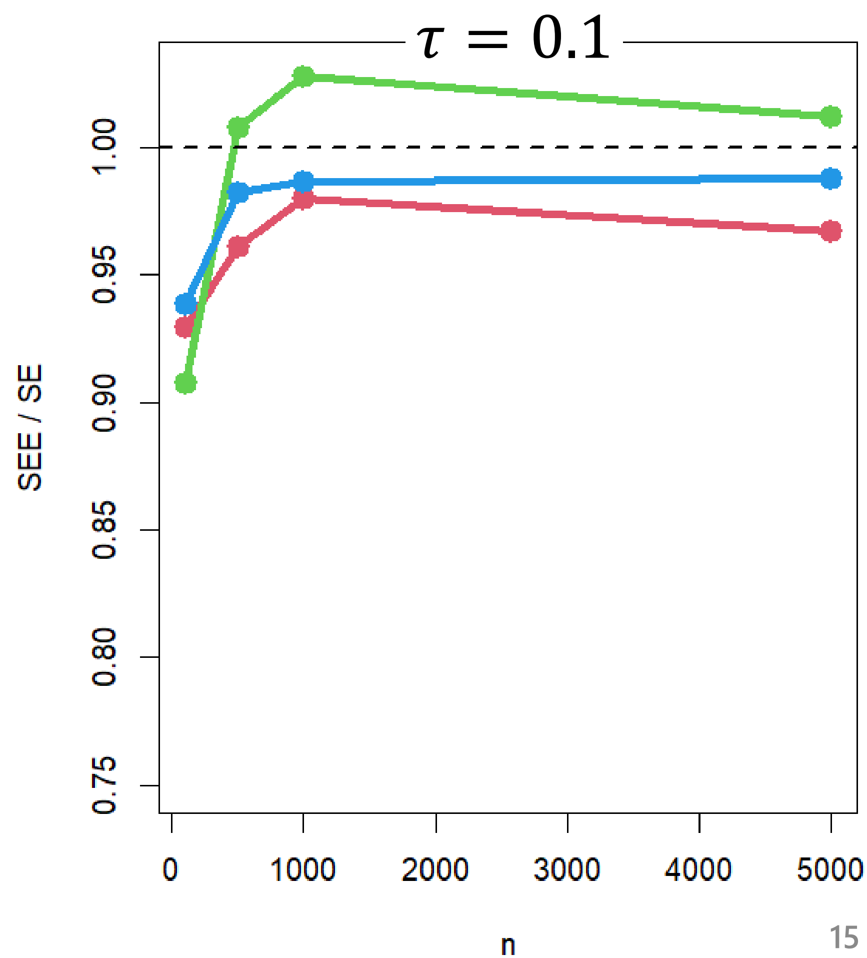
- $y = \beta_0 + \mathbf{1}x_1 + \mathbf{0.5}x_2 + \mathbf{0}x_3 + \sigma\varepsilon_\tau$,
- $\boldsymbol{\tau} \in \{\mathbf{0.01}, \mathbf{0.1}\}$, $\quad x_j \sim N(0,1)$, $\quad n \in \{100, 500, 1000, 5000\}$



SIMULATION: SE ESTIMATION

- $y = \beta_0 + \mathbf{1}x_1 + \mathbf{0.5}x_2 + \mathbf{0}x_3 + \sigma\varepsilon_\tau$,
- $\boldsymbol{\tau} \in \{\mathbf{0.01}, \mathbf{0.1}\}$, $x_j \sim N(0,1)$, $n \in \{100, 500, 1000, 5000\}$



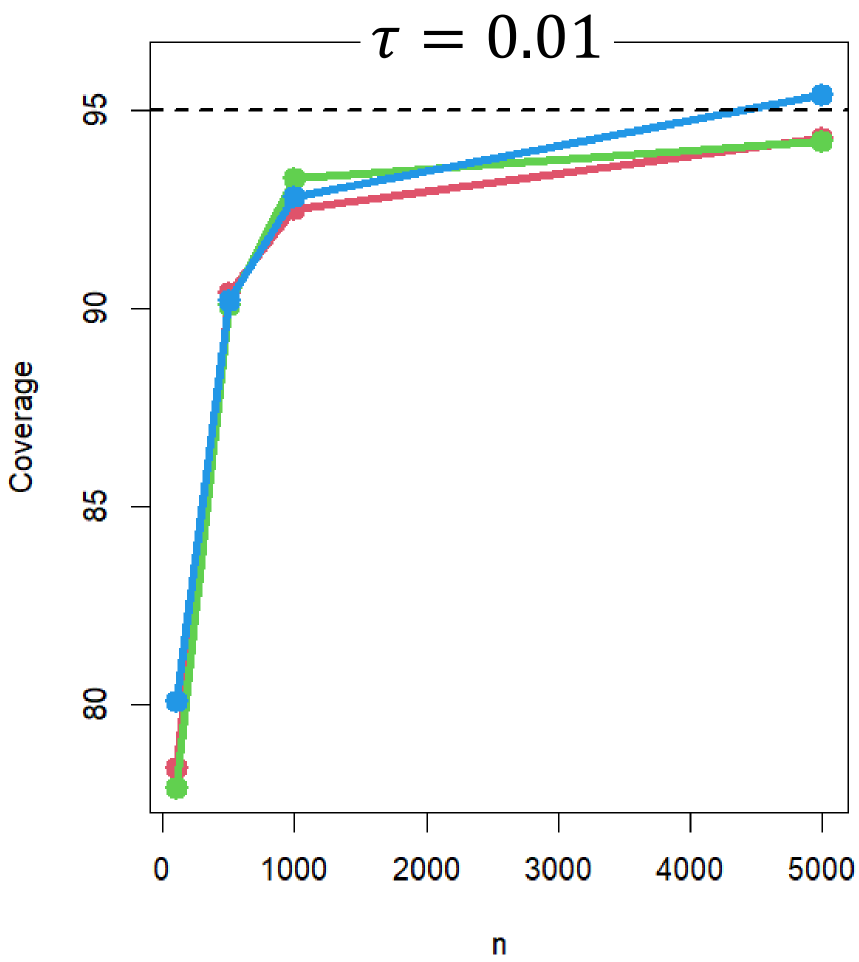


SIMULATION: 95% CI COVERAGE

- $y = \beta_0 + \mathbf{1}x_1 + \mathbf{0.5}x_2 + \mathbf{0}x_3 + \sigma\varepsilon_\tau,$
- $\boldsymbol{\tau} \in \{\mathbf{0.01}, \mathbf{0.1}\}, \quad x_j \sim N(0,1), \quad n \in \{100, 500, 1000, 5000\}$

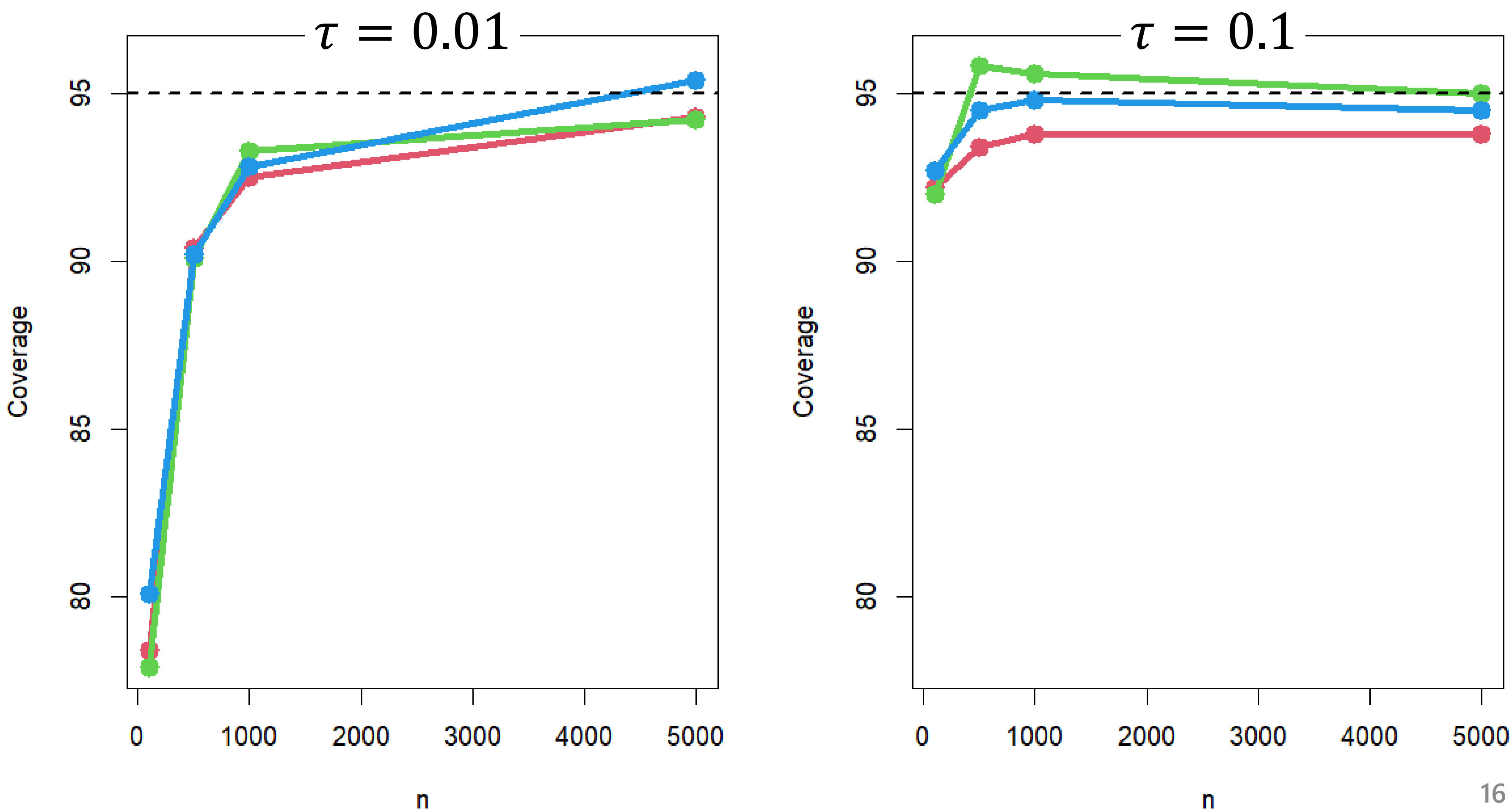
SIMULATION: 95% CI COVERAGE

- $y = \beta_0 + \mathbf{1}x_1 + \mathbf{0.5}x_2 + \mathbf{0}x_3 + \sigma\varepsilon_\tau,$
- $\boldsymbol{\tau} \in \{\mathbf{0.01}, \mathbf{0.1}\}, \quad x_j \sim N(0,1), \quad n \in \{100, 500, 1000, 5000\}$



SIMULATION: 95% CI COVERAGE

- $y = \beta_0 + \mathbf{1}x_1 + \mathbf{0.5}x_2 + \mathbf{0}x_3 + \sigma\varepsilon_\tau,$
- $\boldsymbol{\tau} \in \{\mathbf{0.01}, \mathbf{0.1}\}, \quad x_j \sim N(0,1), \quad n \in \{100, 500, 1000, 5000\}$



SUMMARY

- New differentiable approximation to L1 regression / Laplace
- Extension includes L2 regression / Gaussian
- Smoothly joins two common regression approaches
- SEs can be improved
- References
  - O'Neill & **Burke** (2022). Robust Distributional Regression with Automatic Variable Selection. arXiv.
  - O'Neill & **Burke** (2021) Variable Selection Using a Smooth Information Criterion for Multi-Parameter Regression Models. arXiv.
  - **Burke** & Patilea (2021). A likelihood-based approach for cure regression models. TEST.
  - Jaouimaa, Ha, & **Burke** (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.
  - **Also see: kevinburke.ie and arxiv.org/a/burke_k_1**

Session EC814 ***Room:* S-1.04**
Variable selection *Sunday 18.12.2022* 08:15 - 09:55
Chair: Asaf Weinstein *Organizer:* CMStatistics
B1717: M. ONeill, K. Burke
Distributional regression models with automatic variable selection