

A novel differentiable unification of least absolute deviations and least squares

Kevin Burke | University of Limerick

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least
squares

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

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least
squares

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

Least
*absolute
deviations*

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

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least
squares

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

Least
absolute
deviations

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

Model

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

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least
squares

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

Gaussian

Least
*absolute
deviations*

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

Model

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

LEAST *SQUARES* VS *ABSOLUTE DEVIATIONS*

Least
squares

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

Gaussian

Least
*absolute
deviations*

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

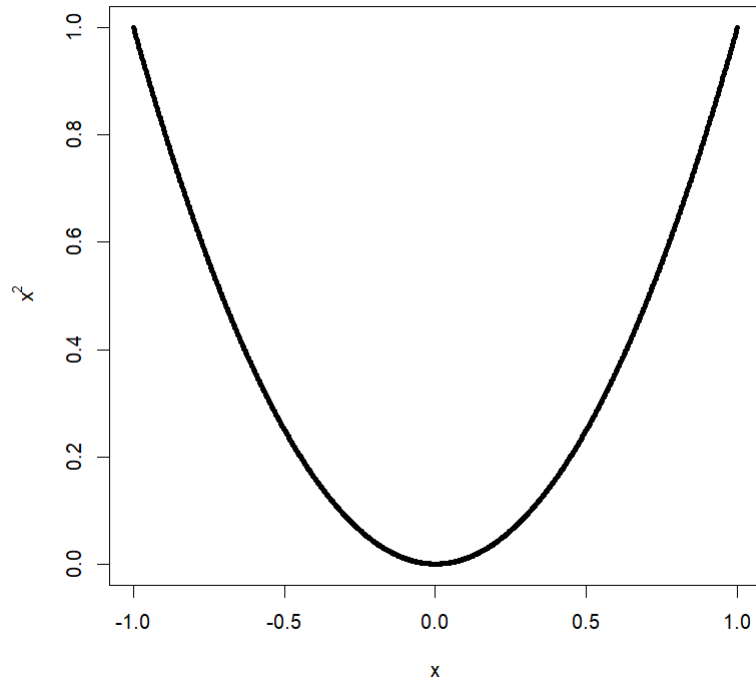
Laplace

Model

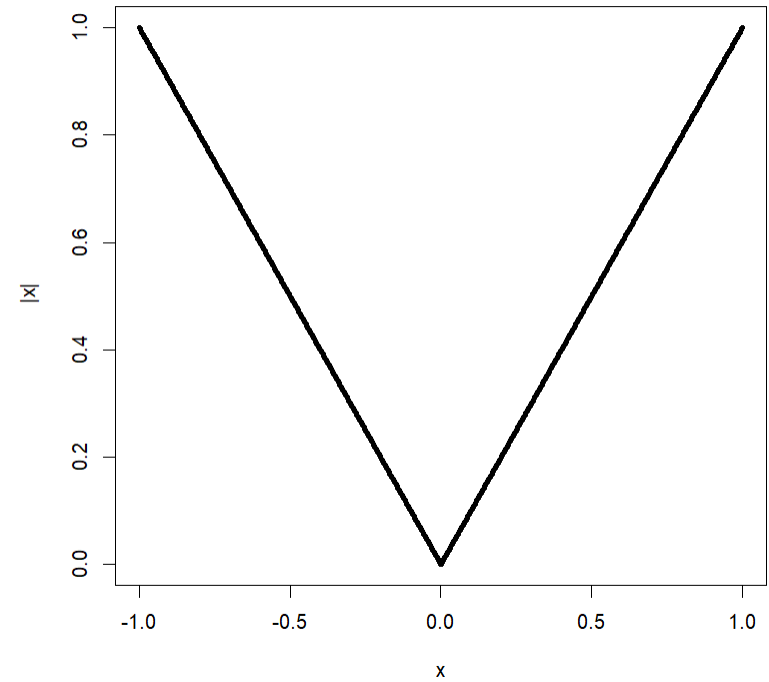
$$y_i = x_i^T \beta + \sigma \boxed{\varepsilon_i}$$

SQUARE VS ABSOLUTE VALUE FUNCTION

Square

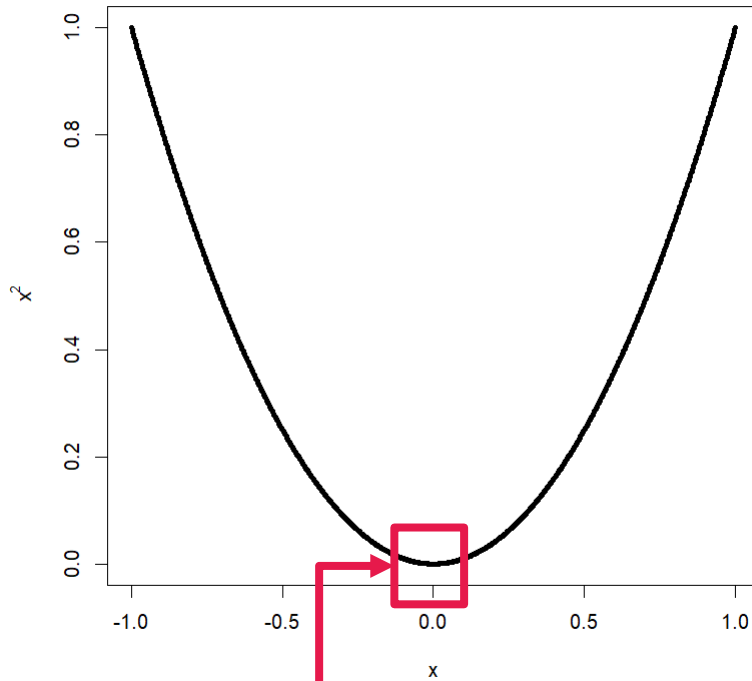


Absolute Value



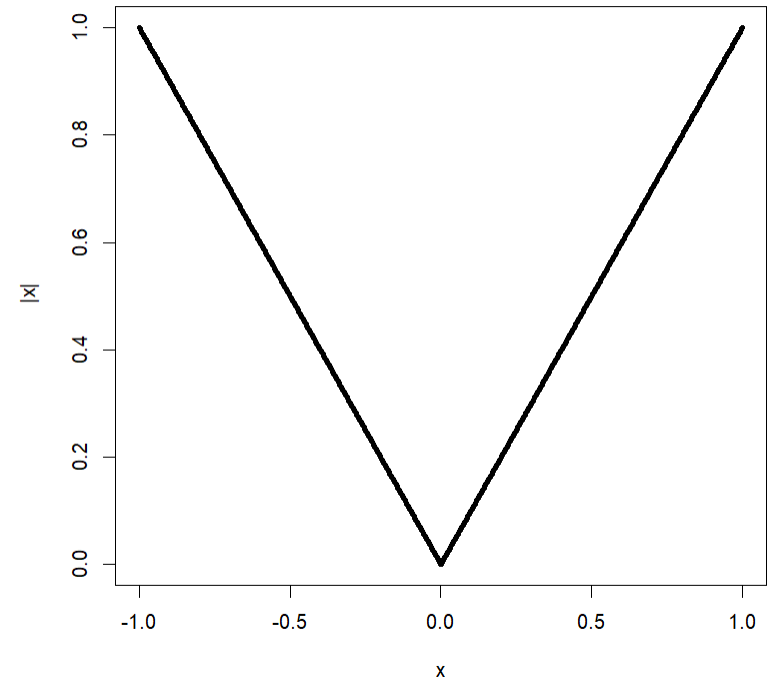
SQUARE VS ABSOLUTE VALUE FUNCTION

Square



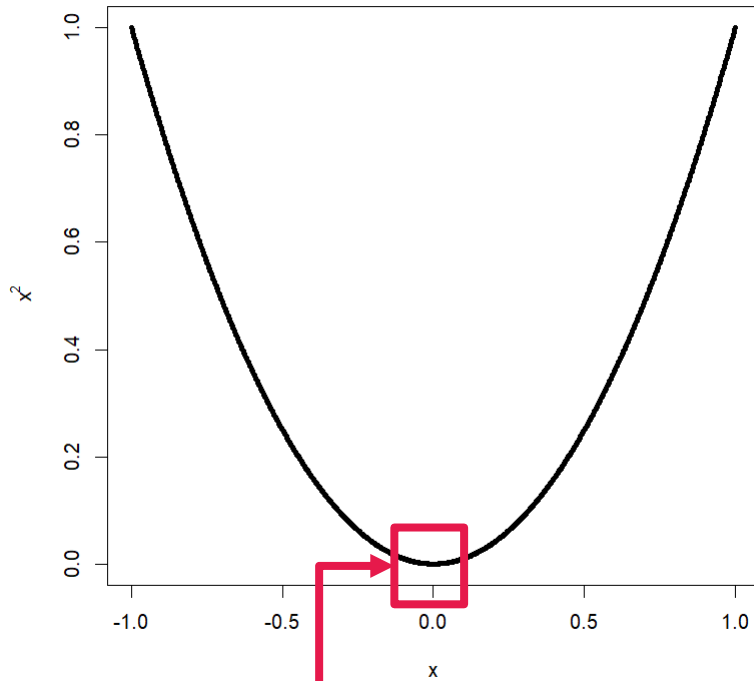
Differentiable

Absolute Value



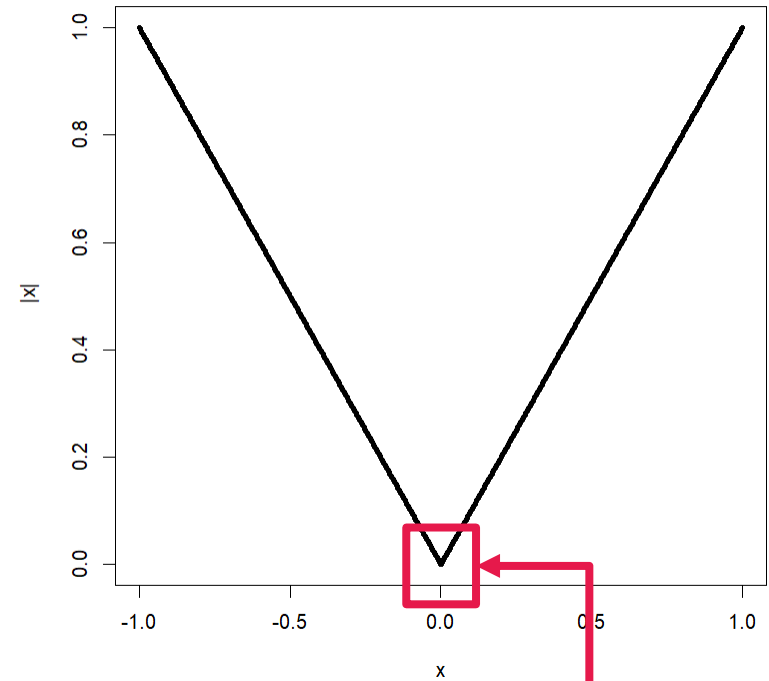
SQUARE VS ABSOLUTE VALUE FUNCTION

Square



Differentiable

Absolute Value



Non-Differentiable

DIFFERENTIABLE APPROXIMATION

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

DIFFERENTIABLE APPROXIMATION

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

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

DIFFERENTIABLE APPROXIMATION

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

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

$$\begin{array}{c} \text{(differentiable)} \quad a'_\tau(x) = \frac{x}{\sqrt{x^2 + \tau^2}} \end{array}$$

DIFFERENTIABLE APPROXIMATION

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

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

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

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

(differentiable)

DIFFERENTIABLE APPROXIMATION

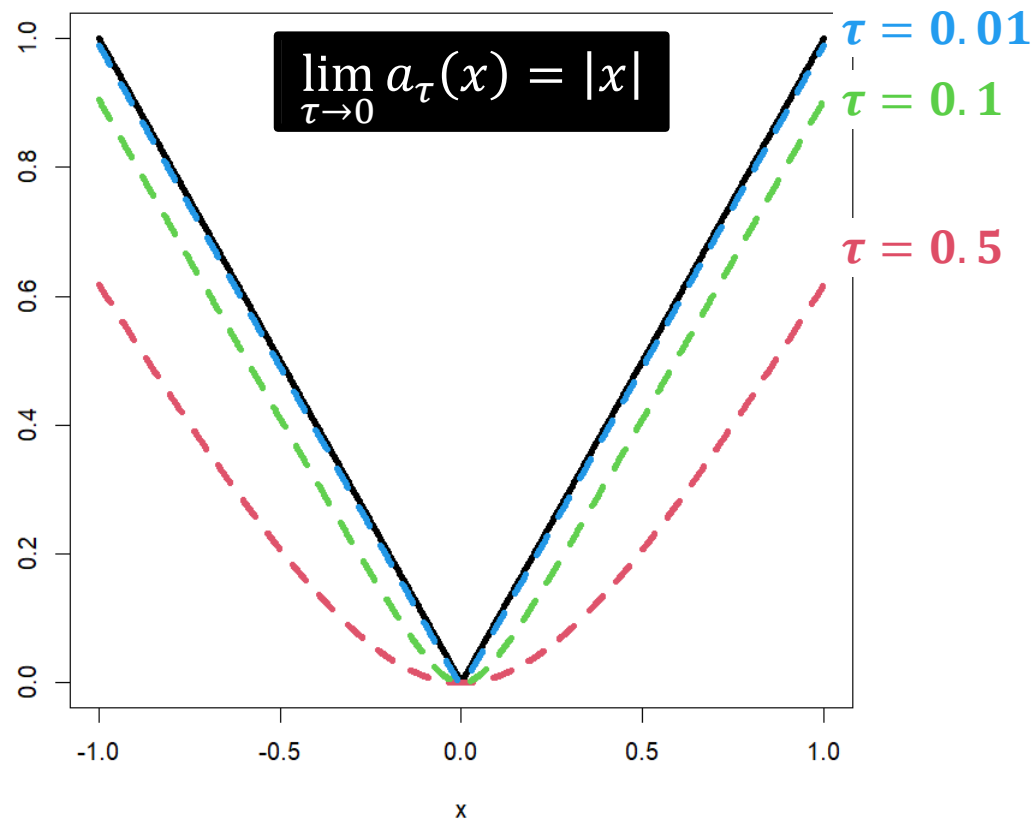
$$\sqrt{x^2} = |x|$$

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

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

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

(differentiable)



DIFFERENTIABLE APPROXIMATION

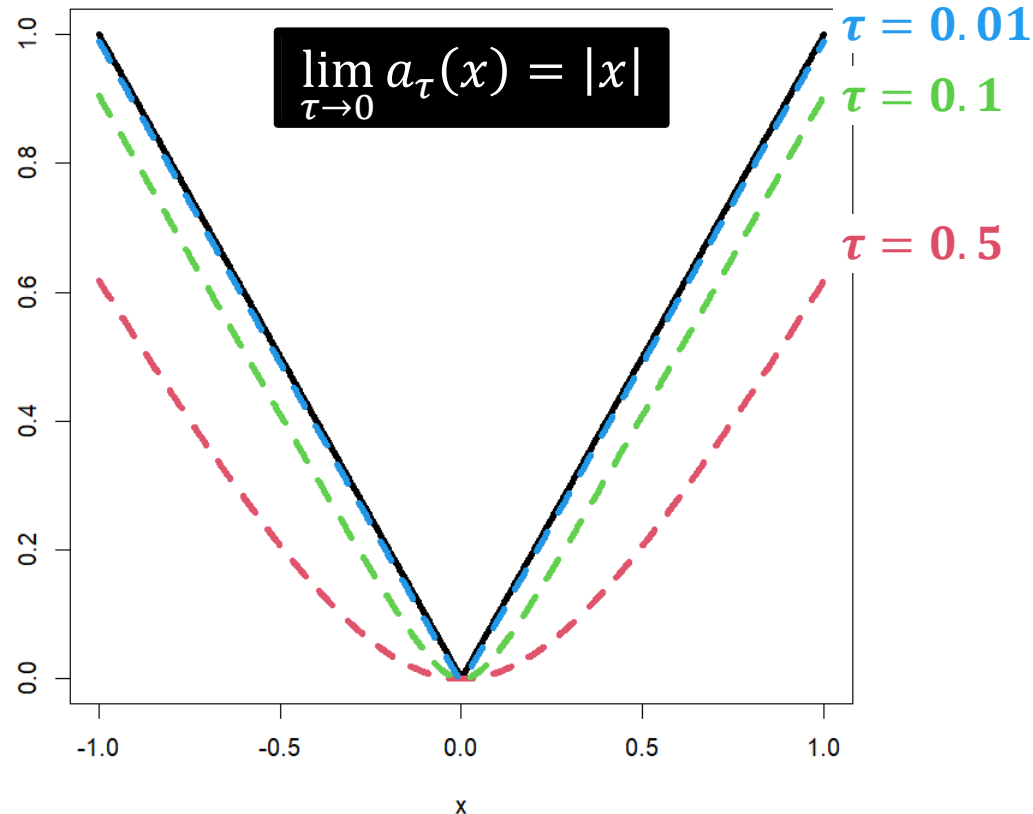
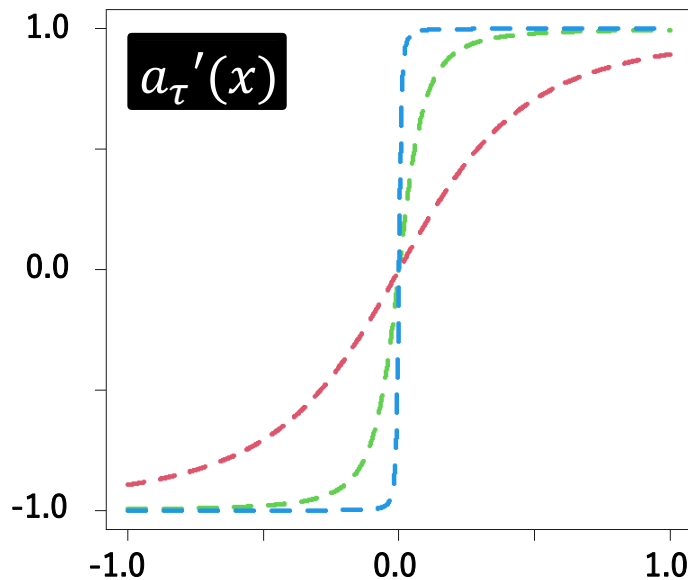
$$\sqrt{x^2} = |x|$$

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

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

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

(differentiable)



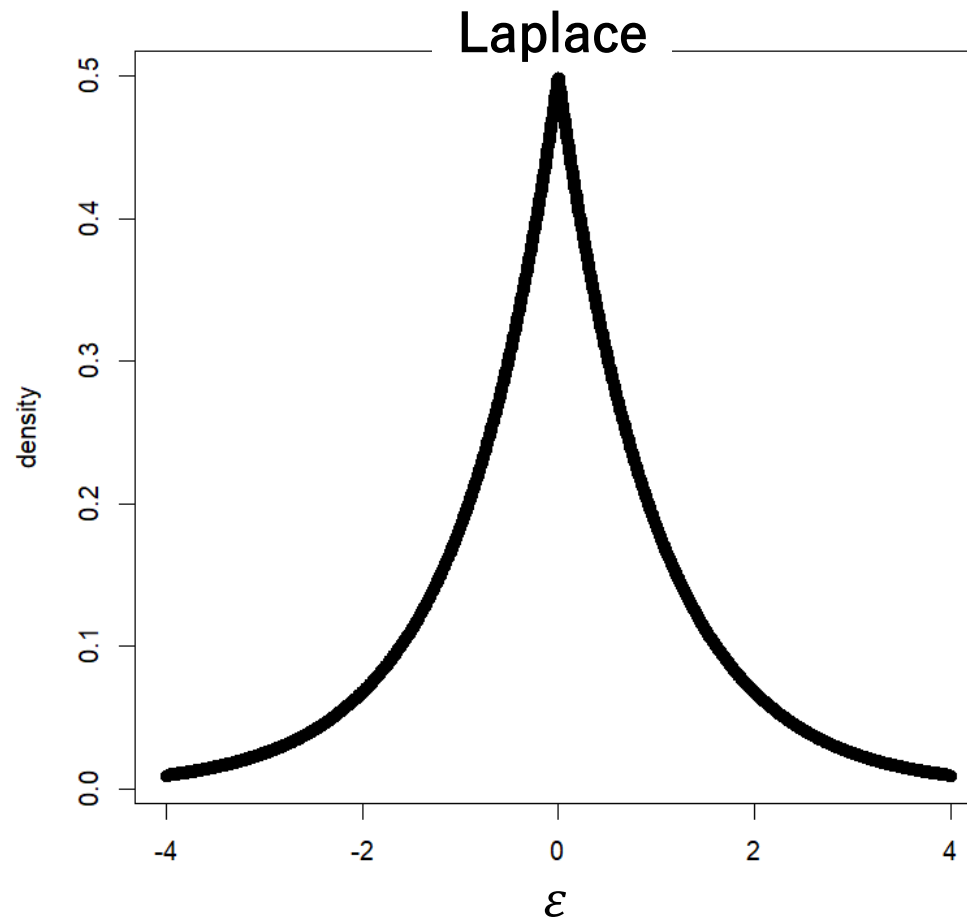
SMOOTH LAPLACE DISTRIBUTION

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

SMOOTH LAPLACE DISTRIBUTION

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

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

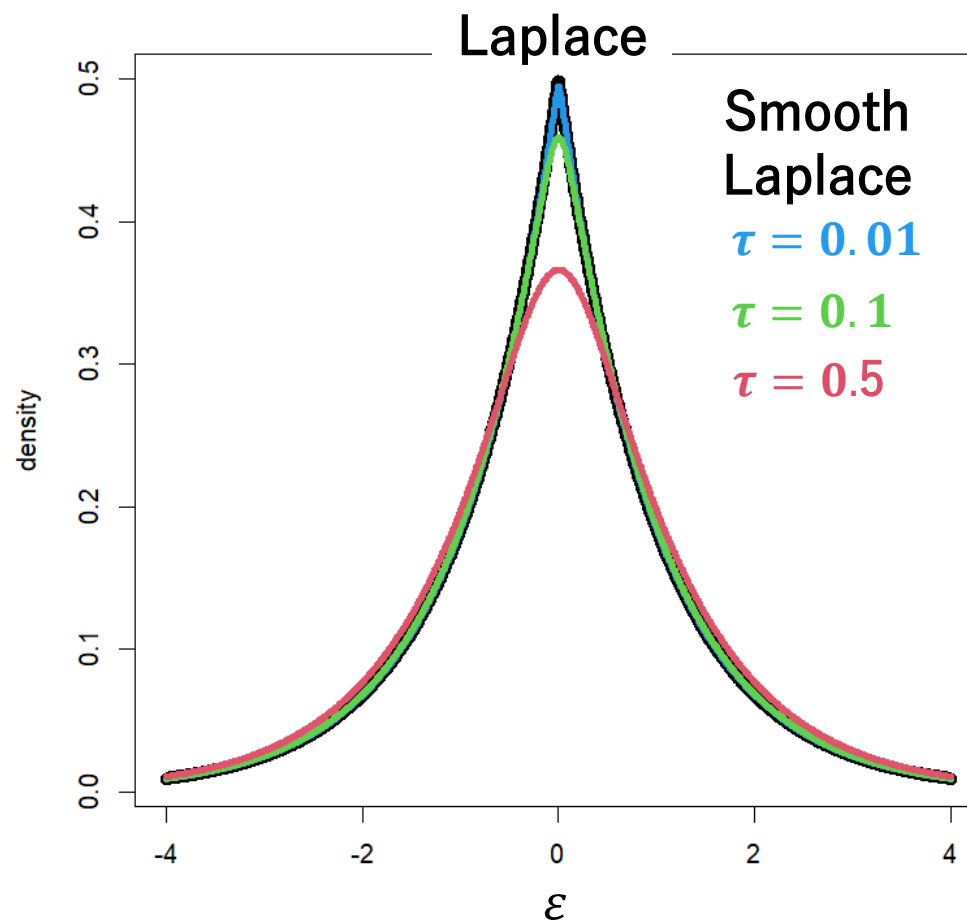


SMOOTH LAPLACE DISTRIBUTION

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

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

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

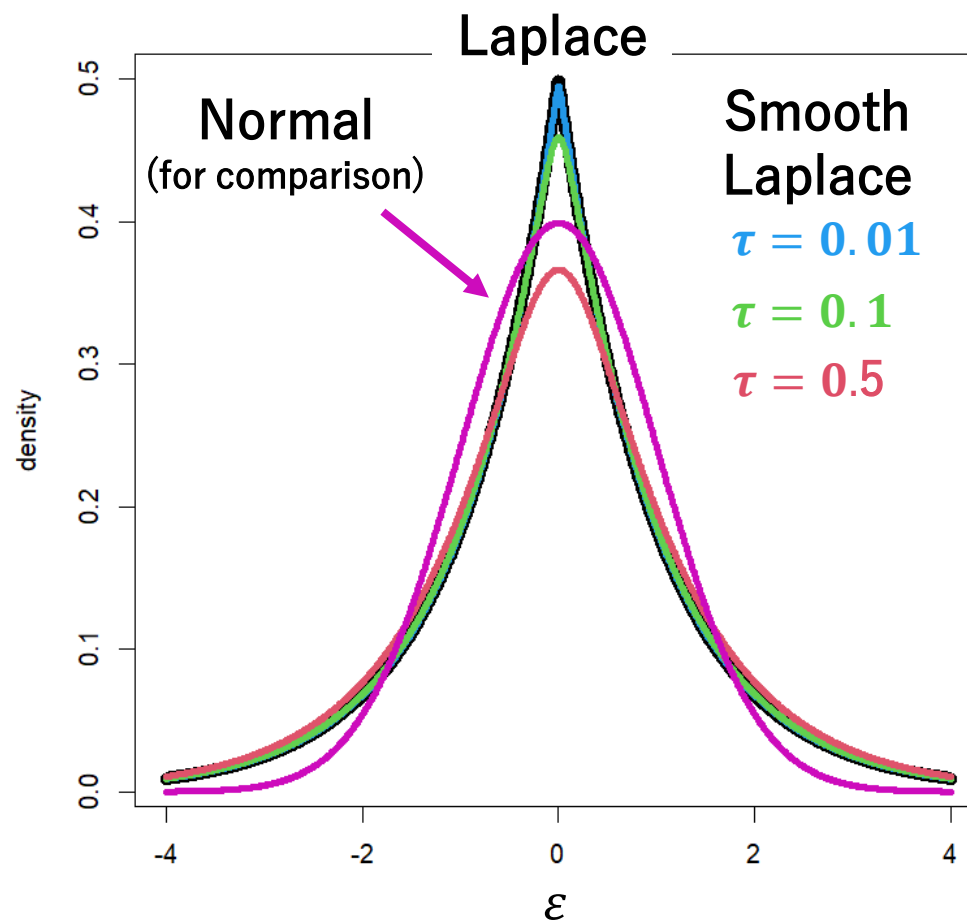


SMOOTH LAPLACE DISTRIBUTION

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

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

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



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

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

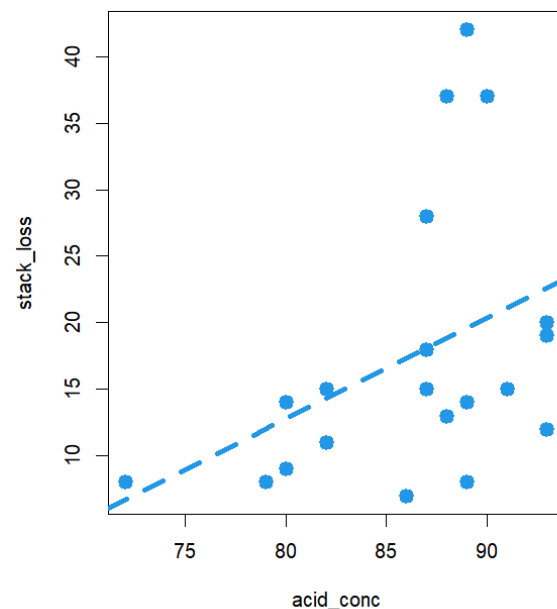
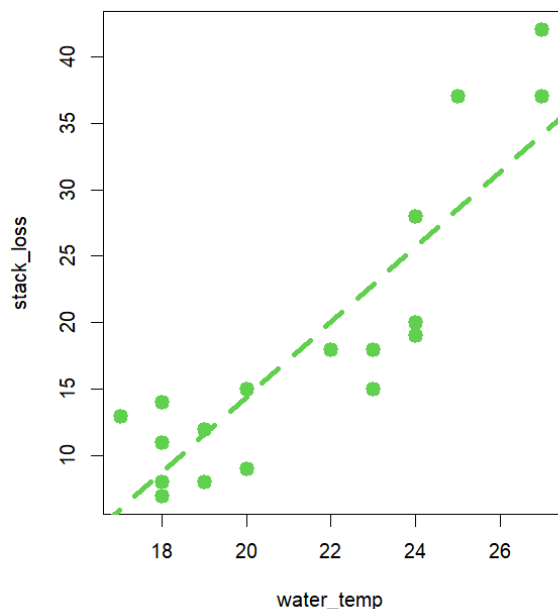
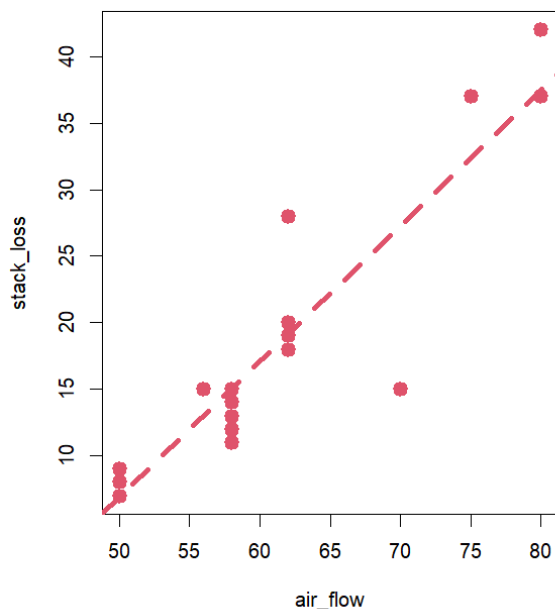
- Differentiable in β and σ
- Standard, gradient-based optimisation can proceed, e.g., `nlm`

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`

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`



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`

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
<code>air_flow</code>	0.83	0.83
<code>water_temp</code>	0.57	0.57
<code>acid_conc</code>	-0.06	-0.06

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`

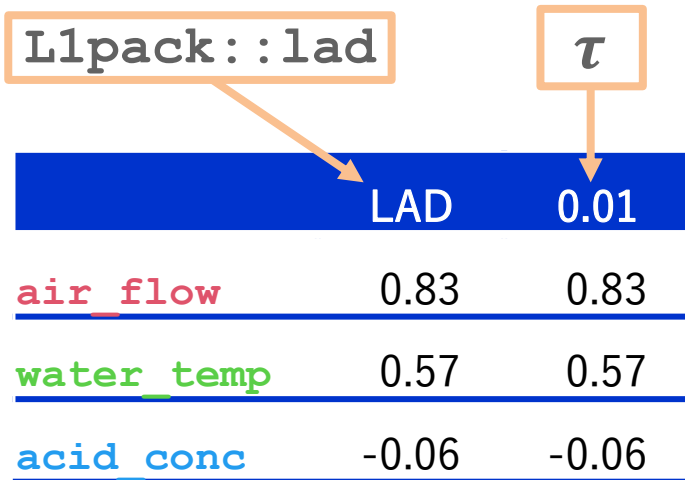
`L1pack::lad`



	LAD	0.01
<code>air_flow</code>	0.83	0.83
<code>water_temp</code>	0.57	0.57
<code>acid_conc</code>	-0.06	-0.06

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`



The diagram illustrates the mapping of variables from R code to a data table. Two boxes at the top, labeled `L1pack::lad` and τ , have arrows pointing to the 'LAD' and '0.01' headers of the table below. The table has three data rows: `air_flow` (0.83), `water_temp` (0.57), and `acid_conc` (-0.06). The input variable names are color-coded to match the list above: red for `air_flow`, green for `water_temp`, and blue for `acid_conc`.

	LAD	0.01
<code>air_flow</code>	0.83	0.83
<code>water_temp</code>	0.57	0.57
<code>acid_conc</code>	-0.06	-0.06

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`

The diagram illustrates the relationship between the L1pack::lad function and the tau parameter. It shows a table of results for different tau values (0.01, 0.1, 0.5) and for larger tau values. The table has four columns: LAD, 0.01, 0.1, and 0.5. The rows represent the input variables: air_flow, water_temp, and acid_conc. The values for air_flow are constant across all tau values, while the values for water_temp and acid_conc increase as tau increases. A bracket labeled 'larger tau' groups the 0.1 and 0.5 columns.

	LAD	τ 0.01	larger τ 0.1	0.5
<code>air_flow</code>	0.83	0.83	0.83	0.83
<code>water_temp</code>	0.57	0.57	0.59	0.69
<code>acid_conc</code>	-0.06	-0.06	-0.07	-0.10

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`

	<code>L1pack::lad</code> LAD	τ 0.01	larger τ 0.1 0.5		even larger τ 2 5 10		
<code>air_flow</code>	0.83	0.83	0.83	0.83	0.81	0.77	0.75
<code>water_temp</code>	0.57	0.57	0.59	0.69	0.92	1.09	1.18
<code>acid_conc</code>	-0.06	-0.06	-0.07	-0.10	-0.12	-0.14	-0.14

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`

	<code>L1pack::lad</code> LAD	τ 0.01	larger τ 0.1 0.5		even larger τ 2 5 10			lm
<code>air_flow</code>	0.83	0.83	0.83	0.83	0.81	0.77	0.75	0.72
<code>water_temp</code>	0.57	0.57	0.59	0.69	0.92	1.09	1.18	1.30
<code>acid_conc</code>	-0.06	-0.06	-0.07	-0.10	-0.12	-0.14	-0.14	-0.15

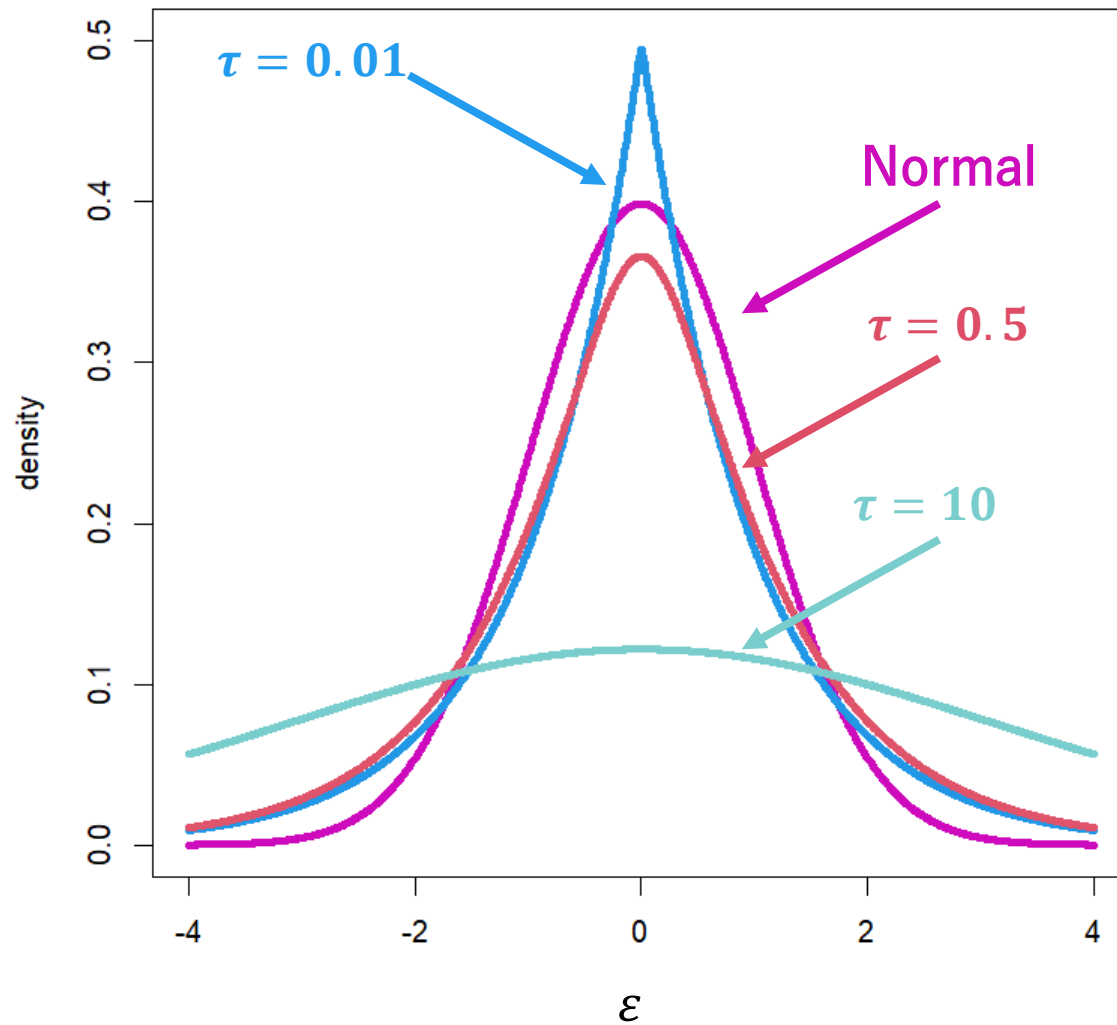
STACK LOSS DATA FIT

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

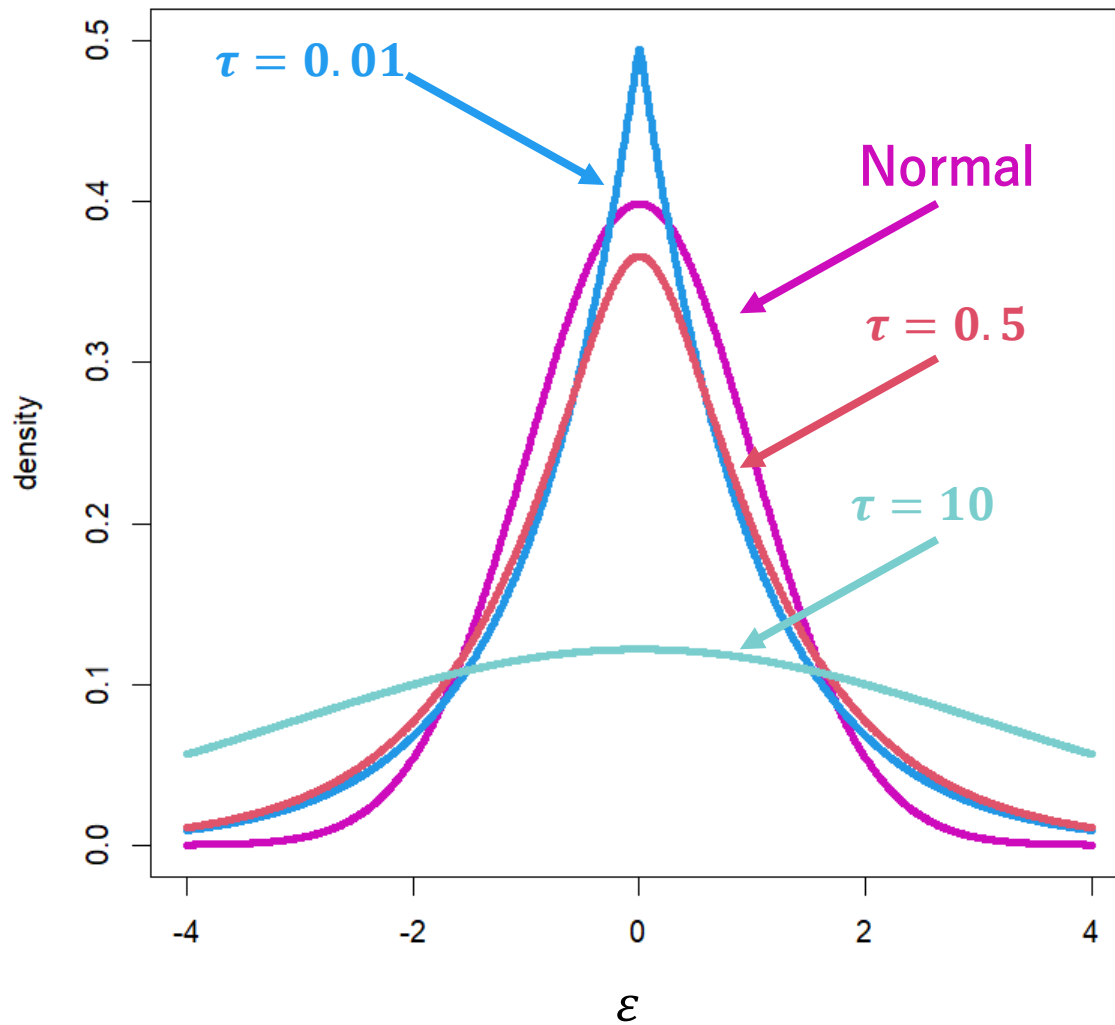
	<code>L1pack::lad</code> LAD	τ 0.01	larger τ 0.1 0.5		even larger τ 2 5 10			lm
<code>air_flow</code>	0.83	0.83	0.83	0.83	0.81	0.77	0.75	0.72
<code>water_temp</code>	0.57	0.57	0.59	0.69	0.92	1.09	1.18	1.30
<code>acid_conc</code>	-0.06	-0.06	-0.07	-0.10	-0.12	-0.14	-0.14	-0.15

→
..tending to least squares?

INCREASING τ

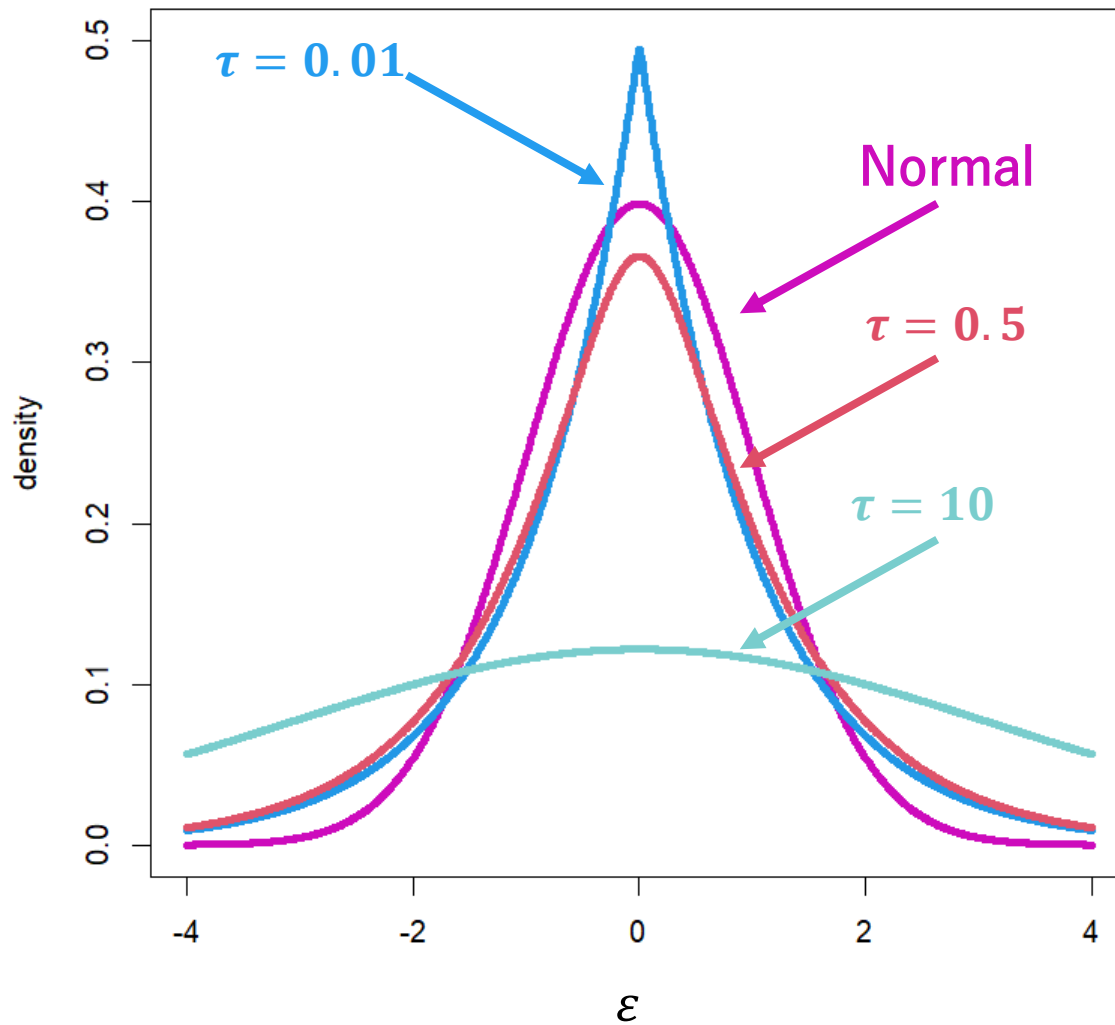


INCREASING τ



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

INCREASING τ

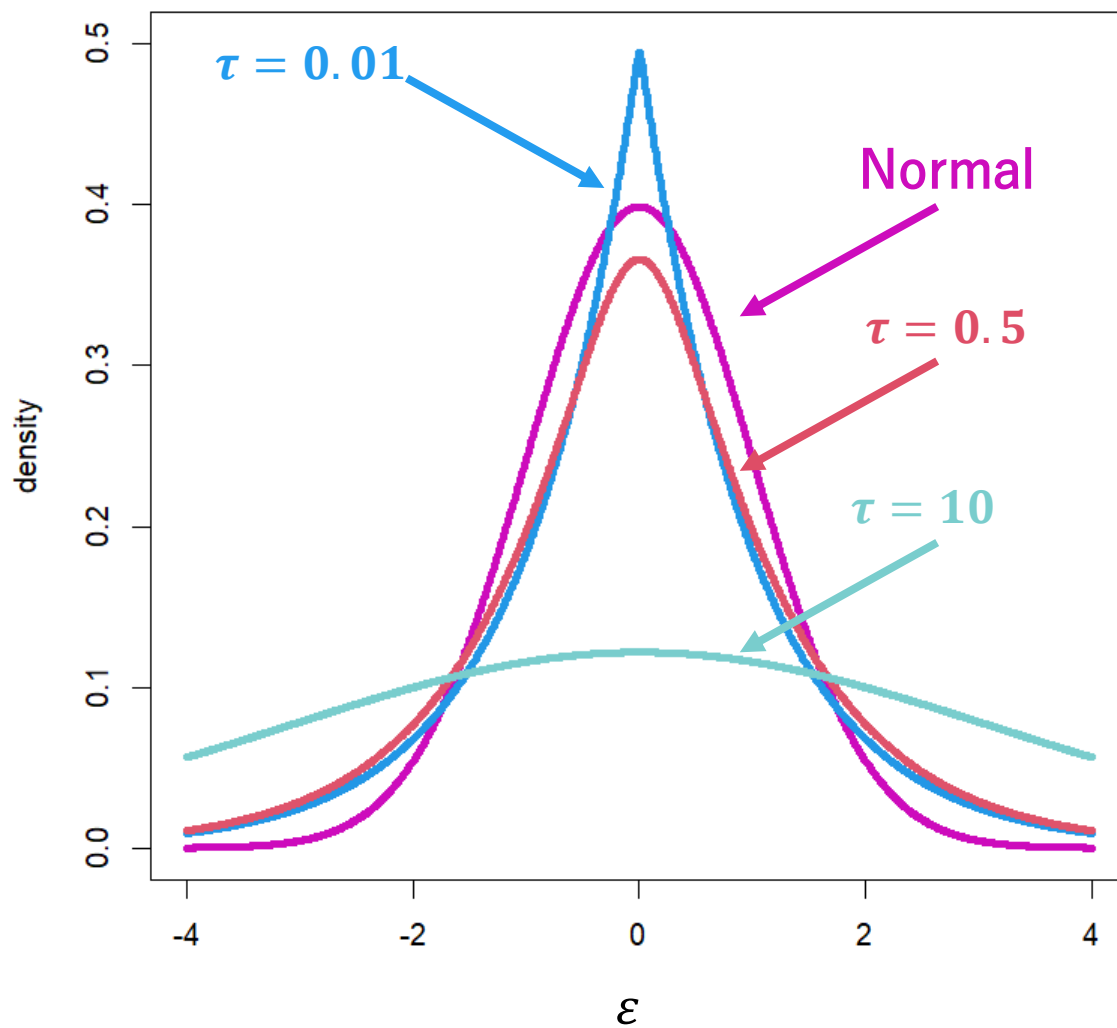


$$f_{\tau}(\epsilon) = c_{\tau}e^{-a_{\tau}(\epsilon)}$$

Small τ

$$a_{\tau}(\epsilon) \sim |\epsilon|$$

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: $\tilde{a}_\tau(\varepsilon) = (\tau + 1) a_\tau(\varepsilon)$
 - $\lim_{\tau \rightarrow 0} \tilde{a}_\tau(\varepsilon) = |\varepsilon|$

BEHAVIOUR FOR LARGE τ

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

BEHAVIOUR FOR LARGE τ

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

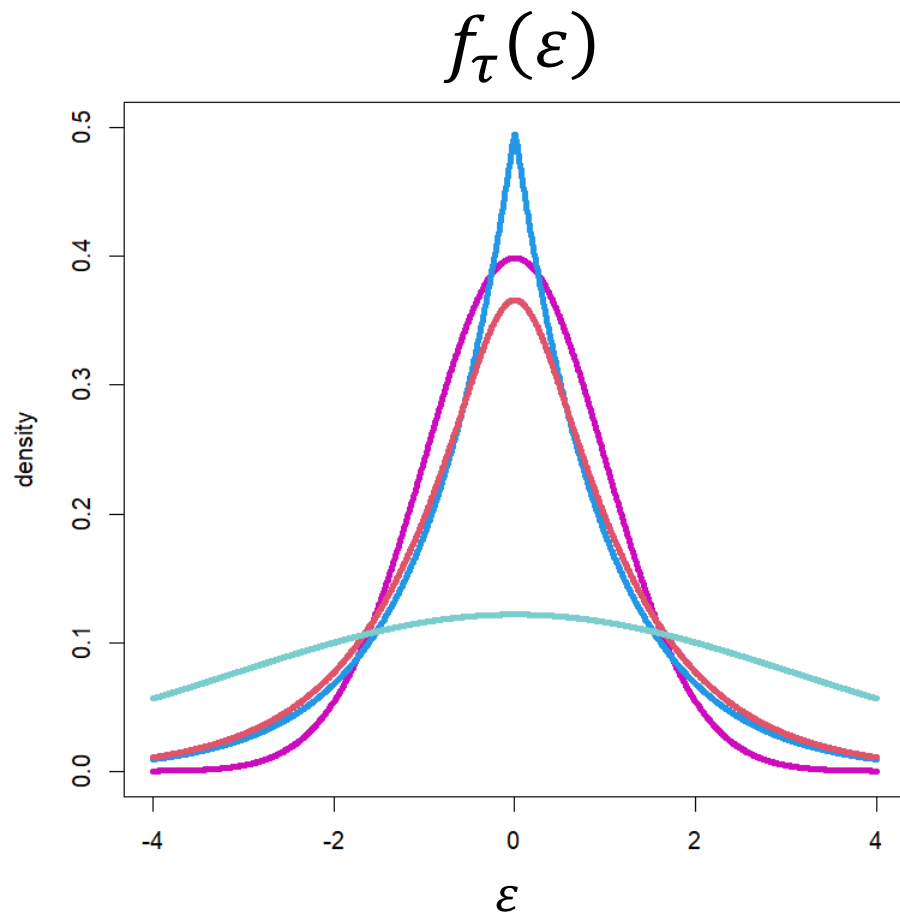
BEHAVIOUR FOR LARGE τ

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

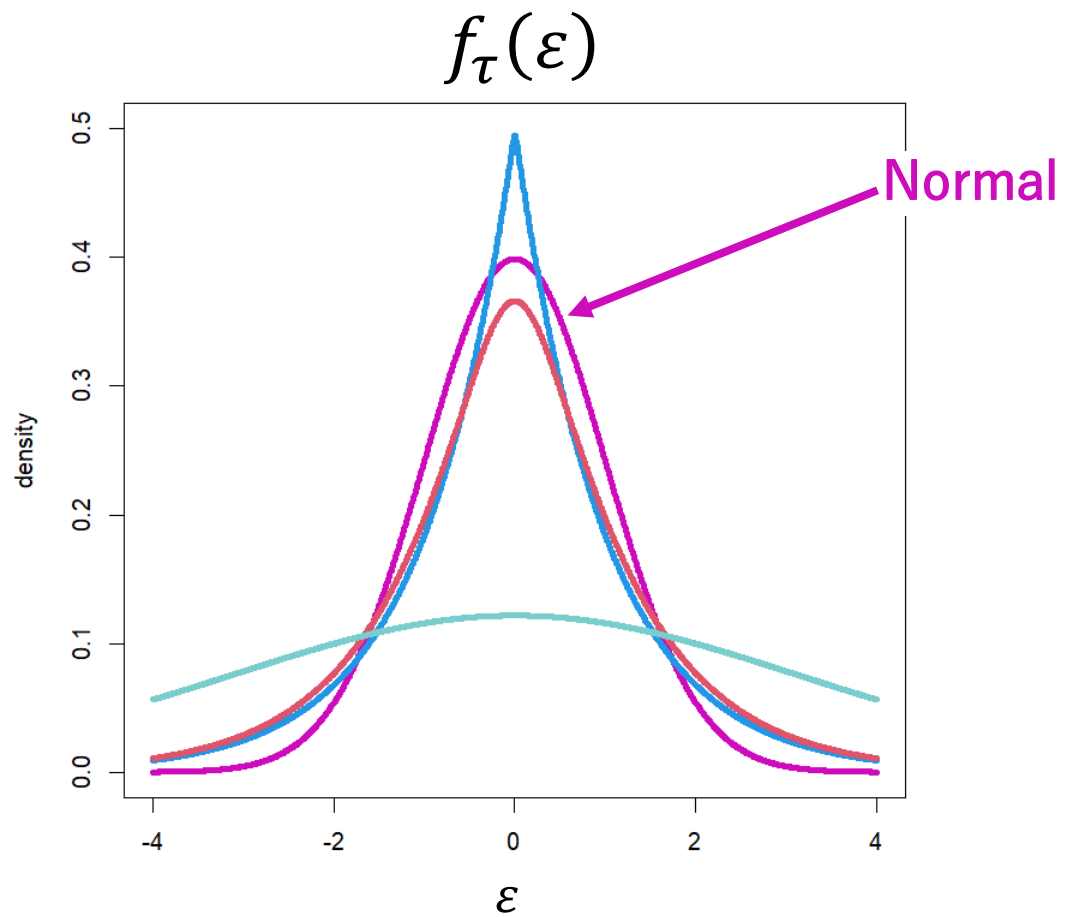
BEHAVIOUR FOR LARGE τ

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

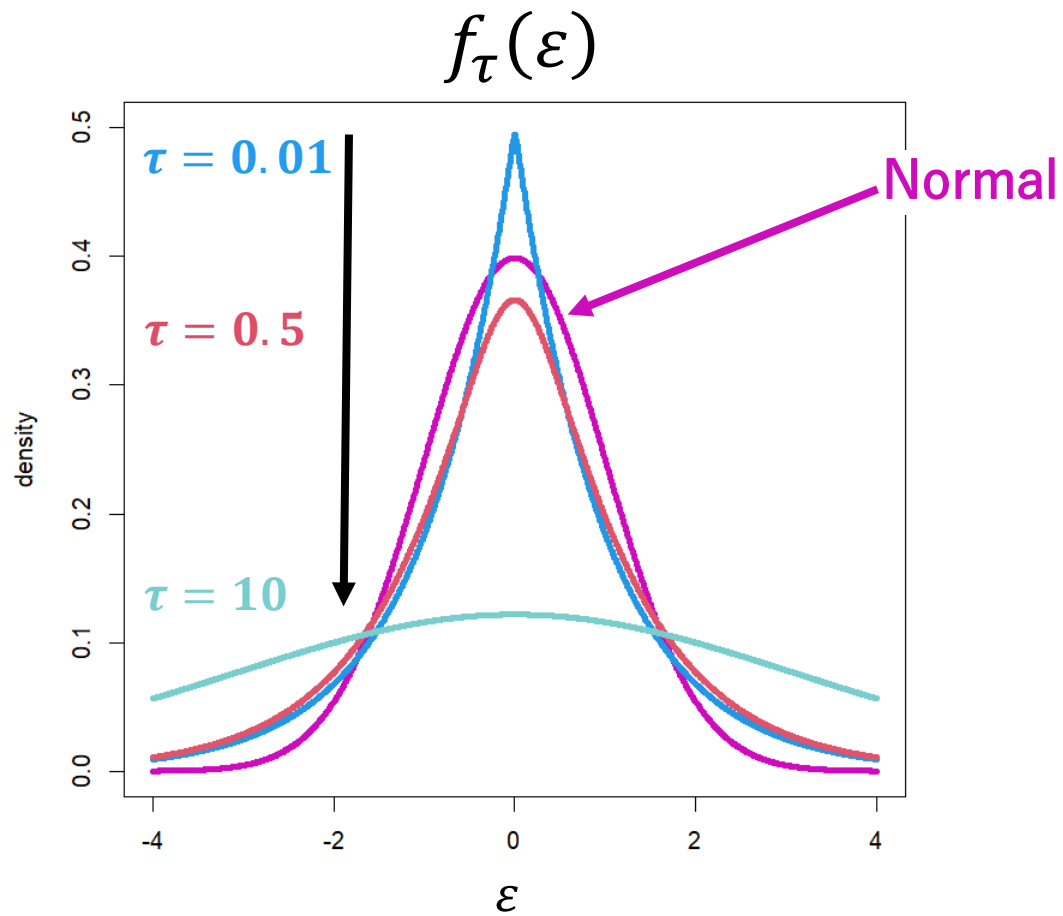
NEW PARAMETERISATION



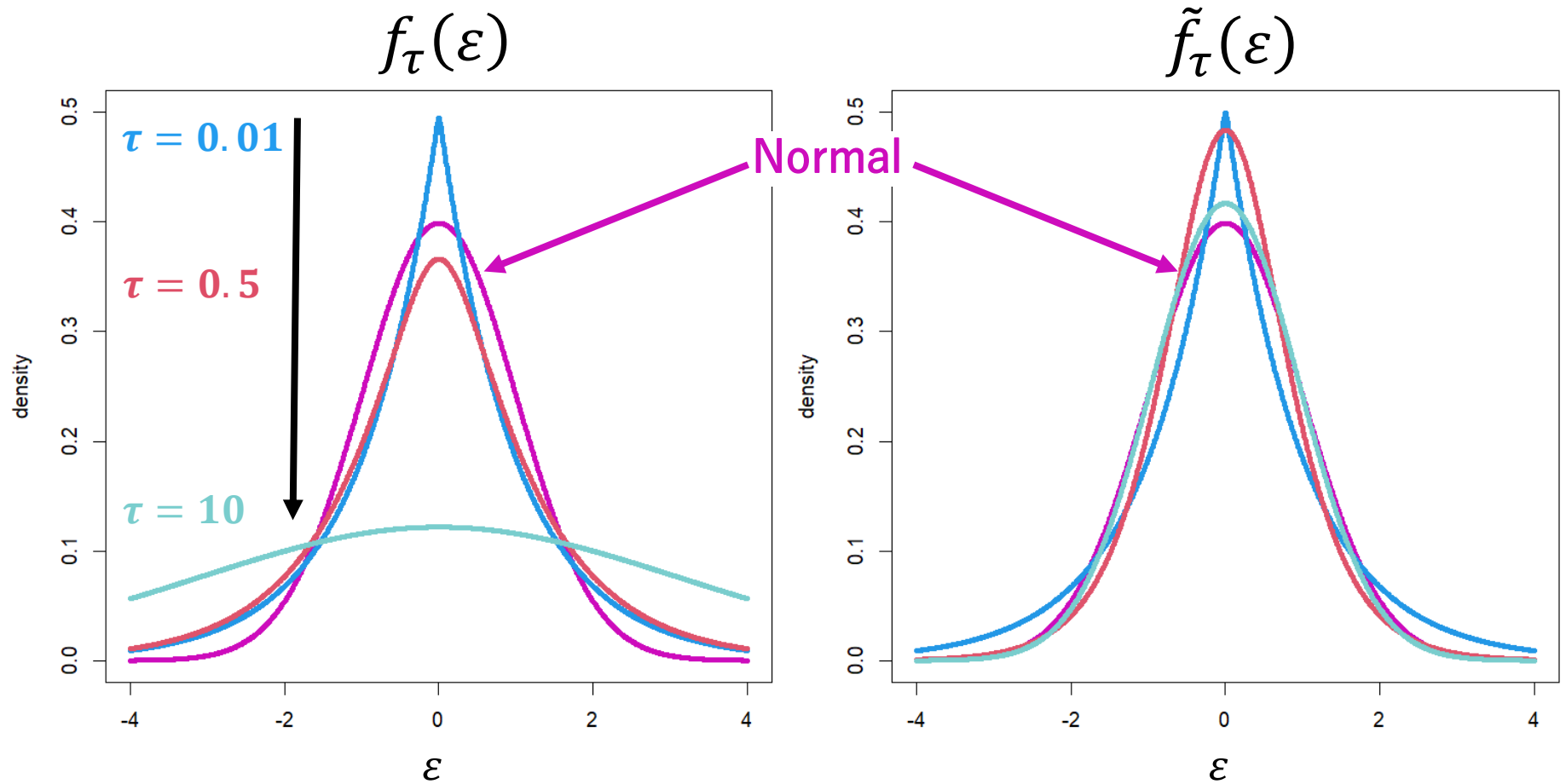
NEW PARAMETERISATION



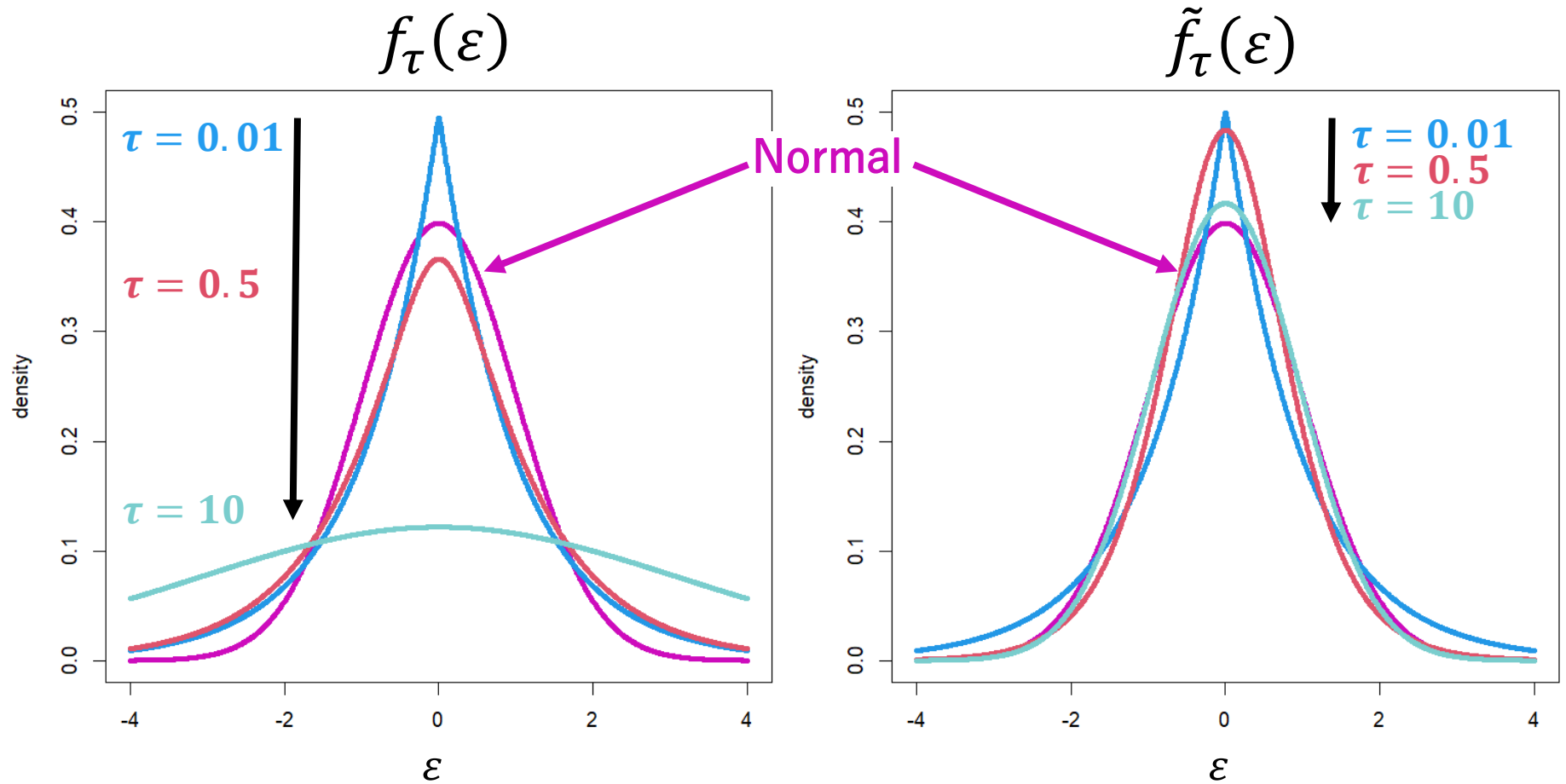
NEW PARAMETERISATION



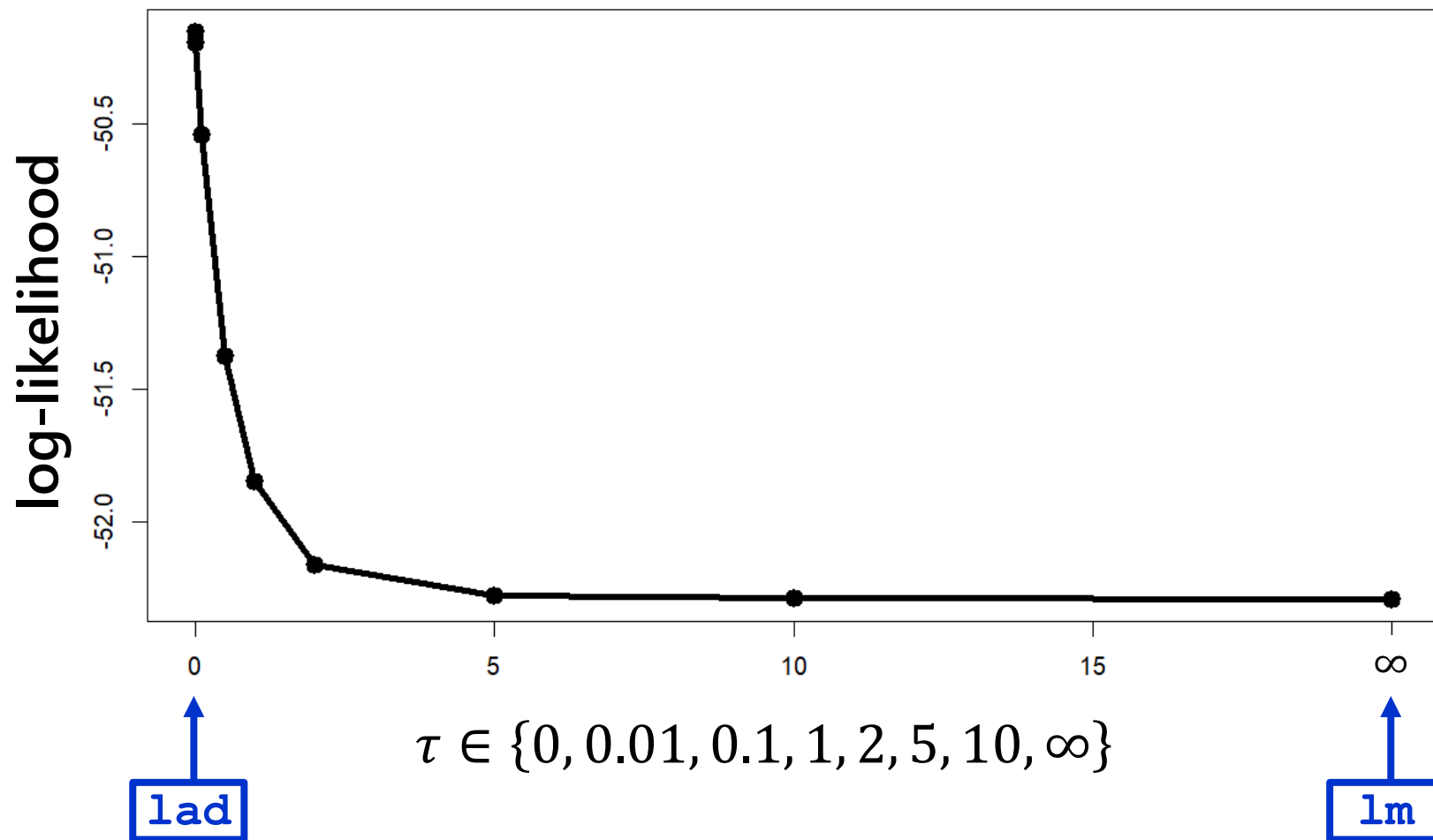
NEW PARAMETERISATION



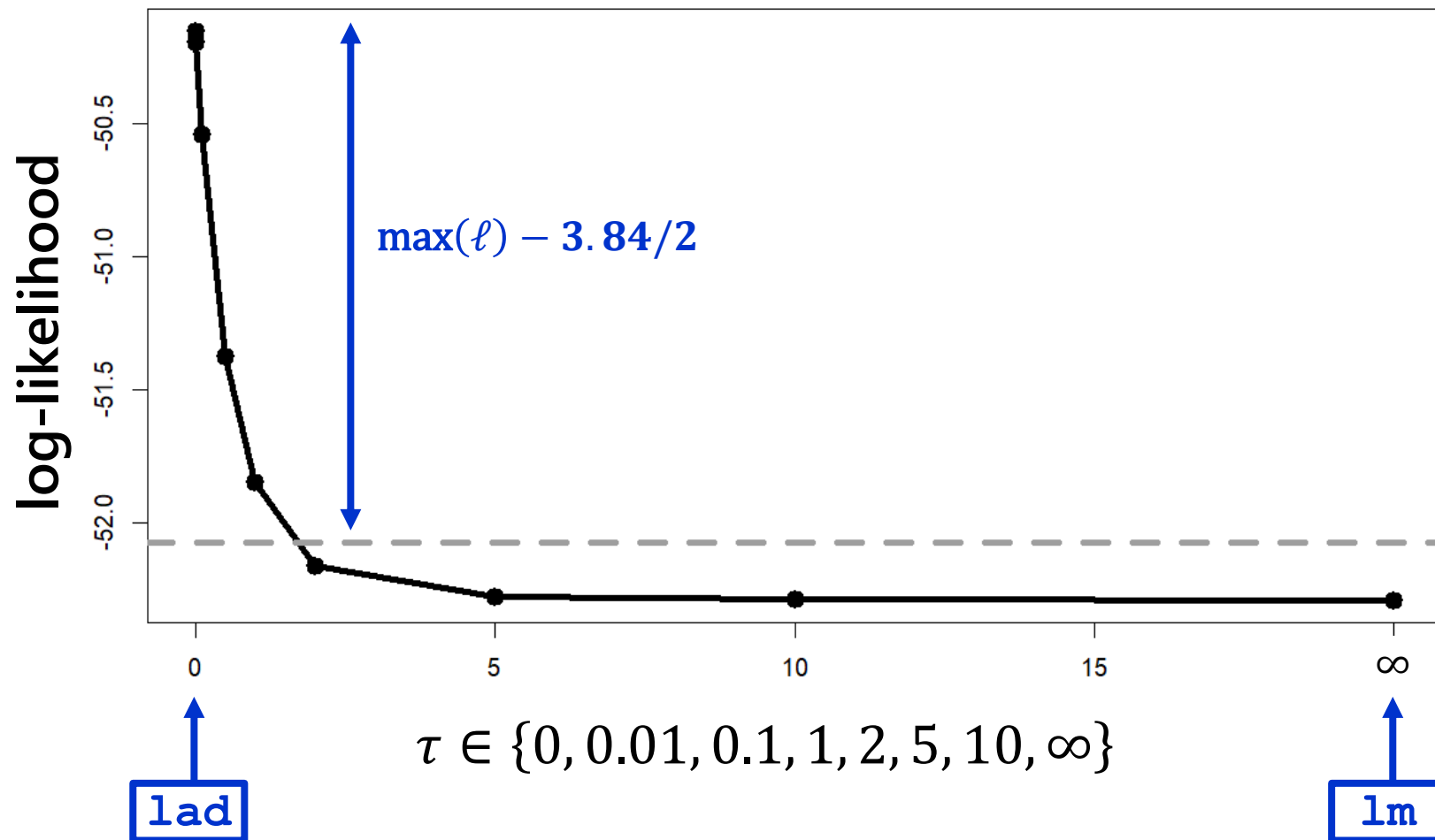
NEW PARAMETERISATION



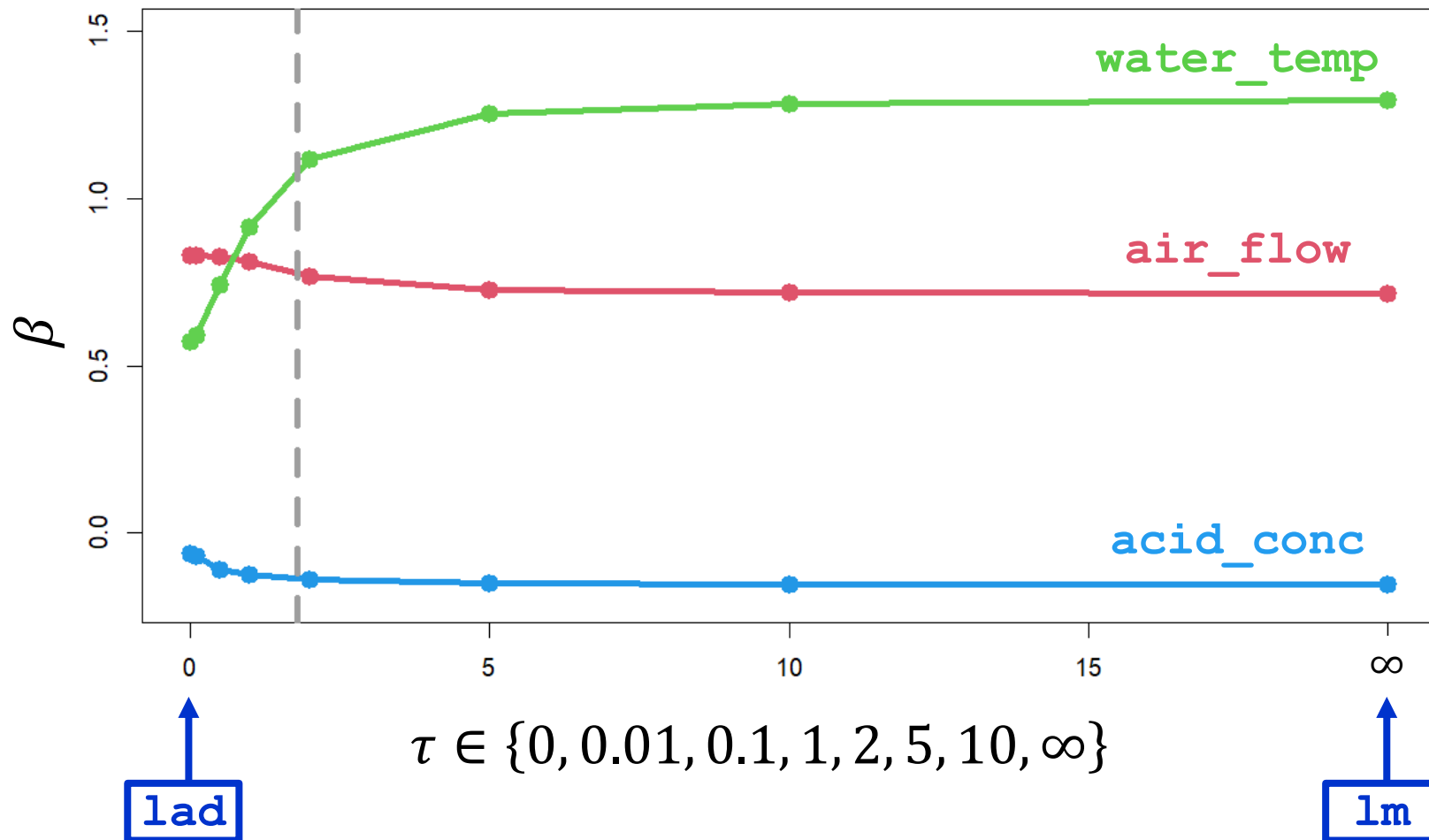
STACKLOSS: LOG-LIKELIHOOD



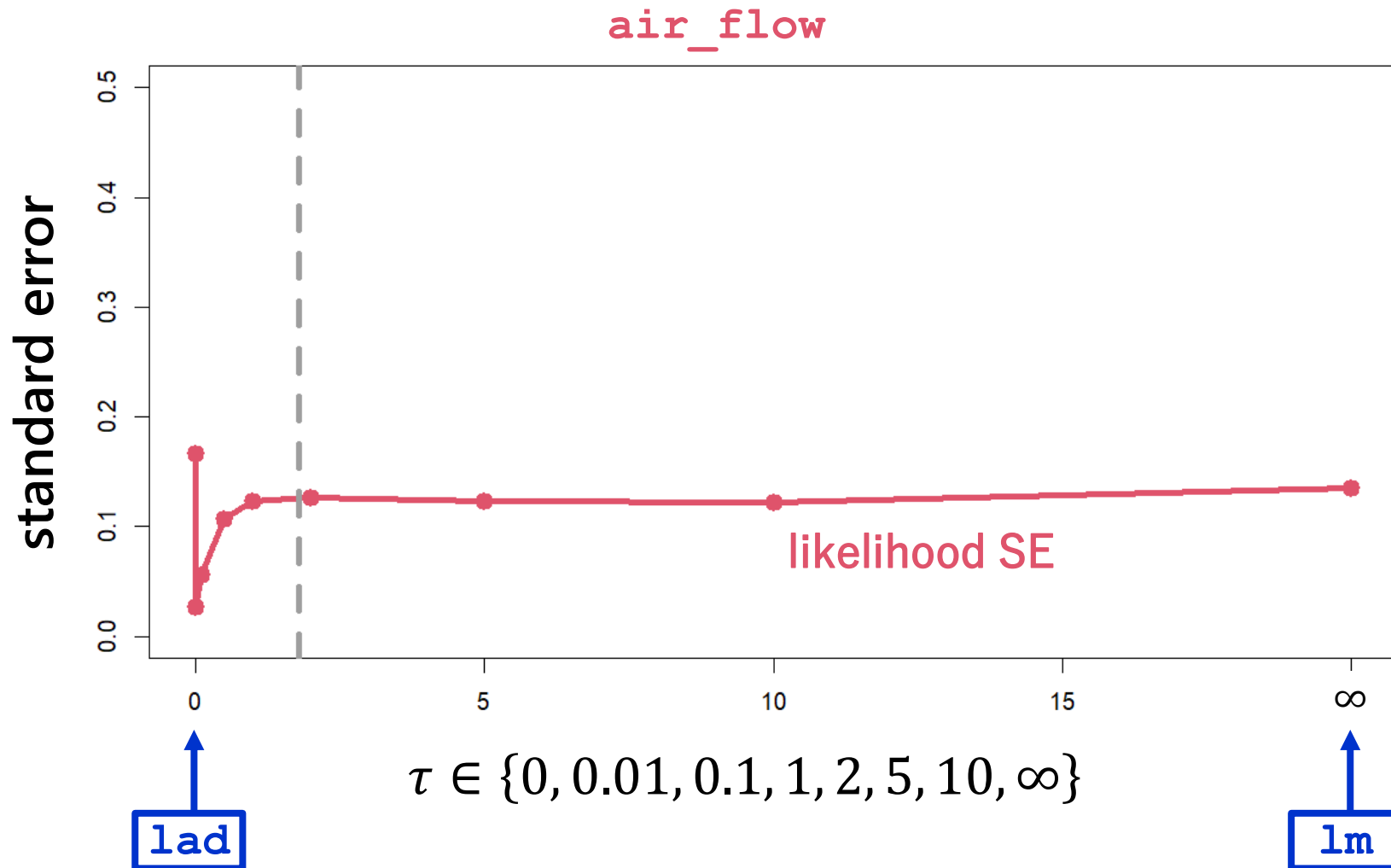
STACKLOSS: LOG-LIKELIHOOD



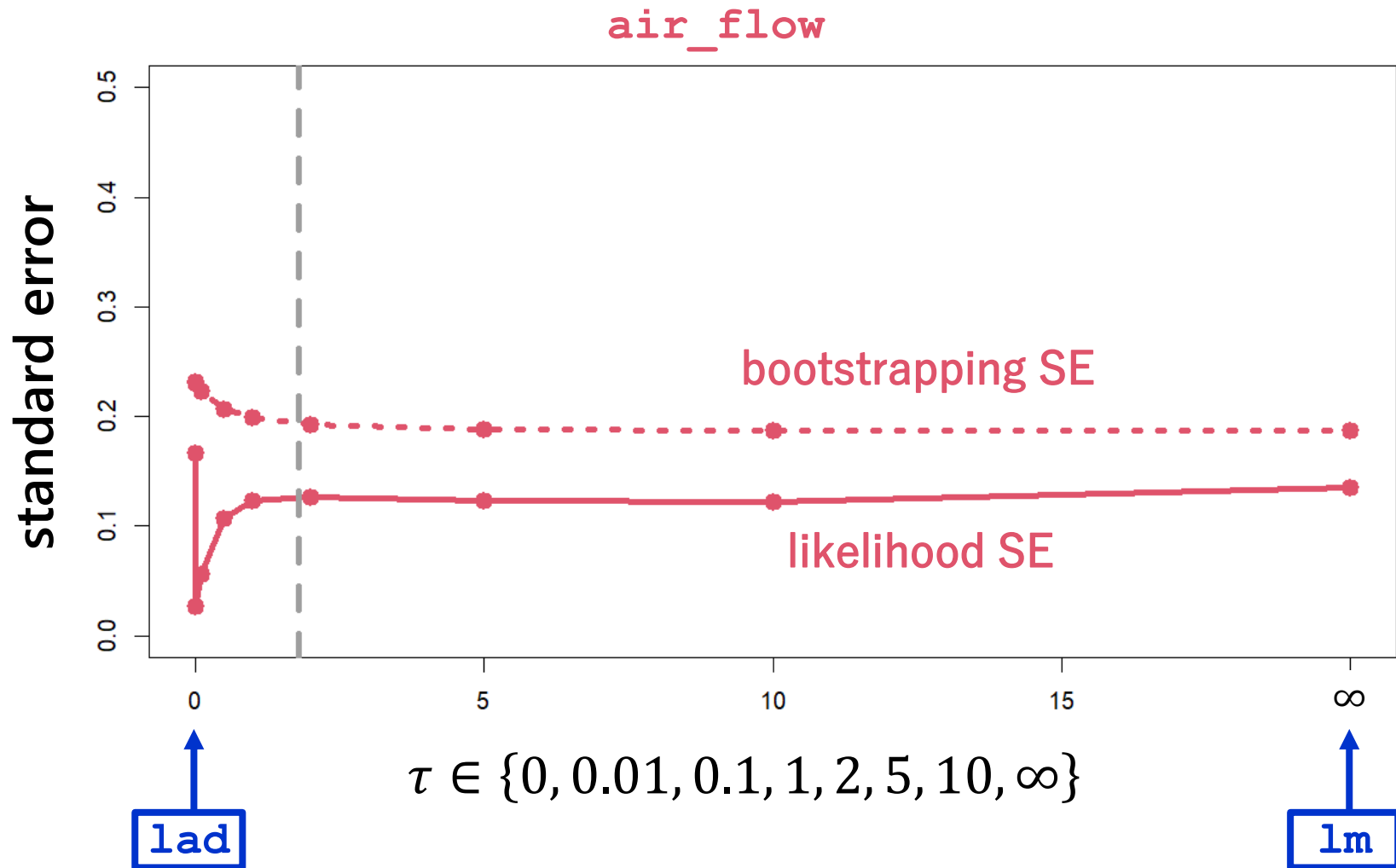
STACKLOSS: BETA COEFFICIENTS



STACKLOSS: STANDARD ERRORS



STACKLOSS: STANDARD ERRORS

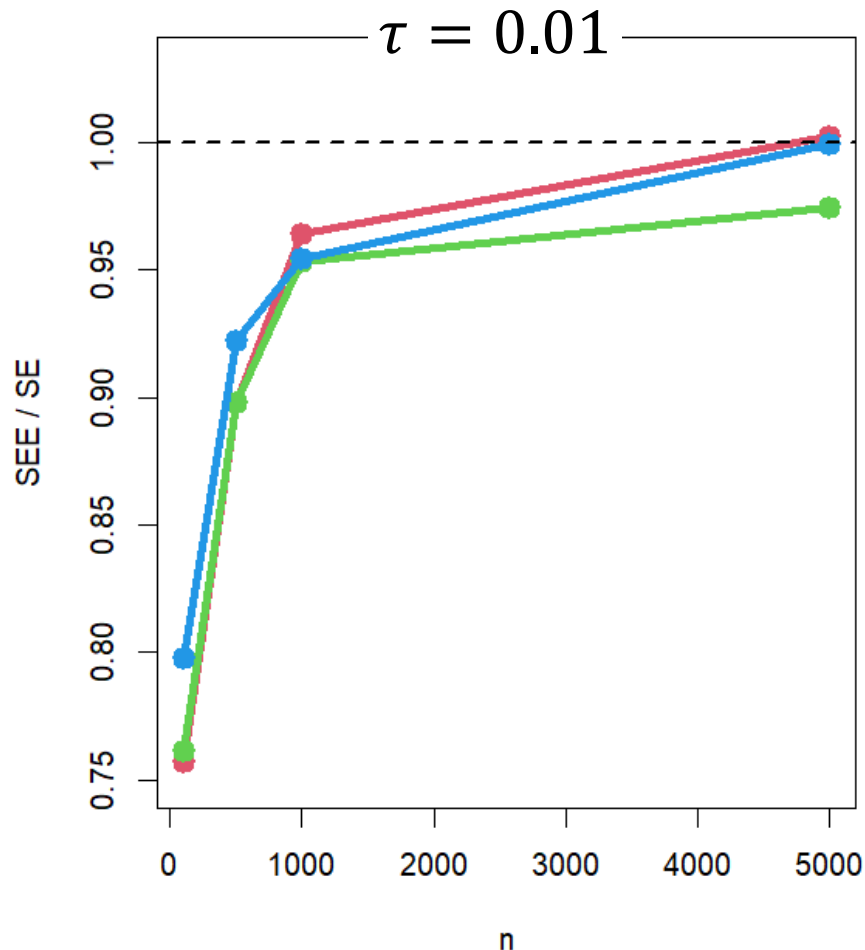


SIMULATION: SE ESTIMATION

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

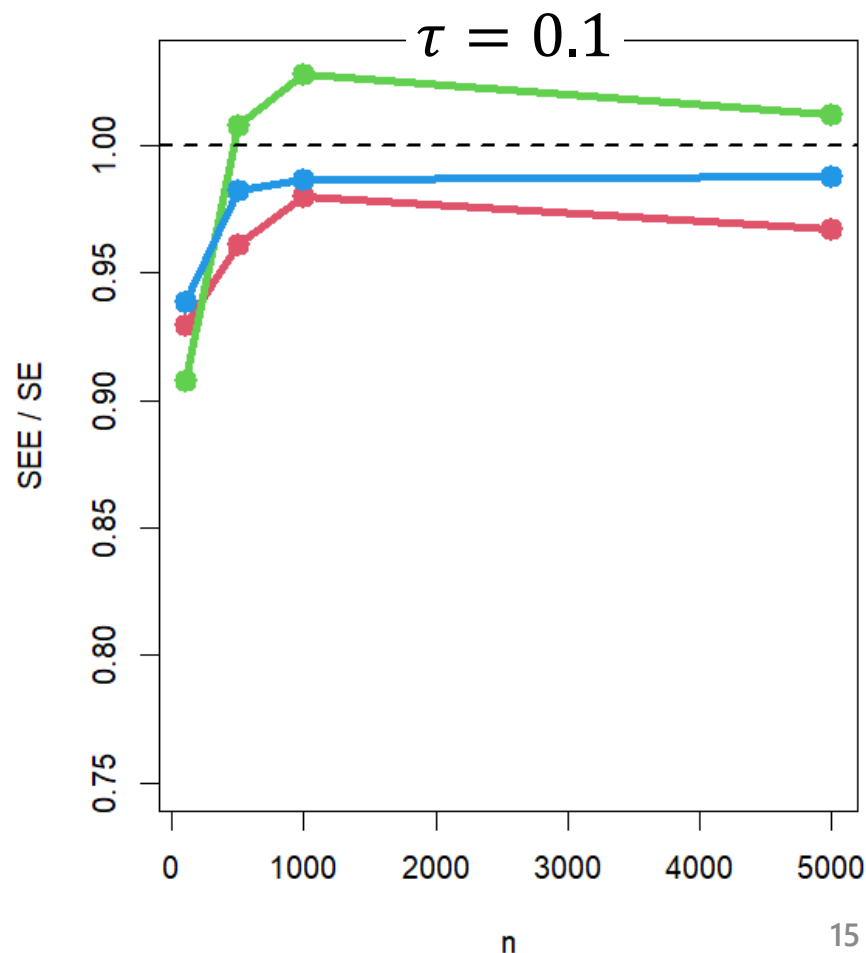
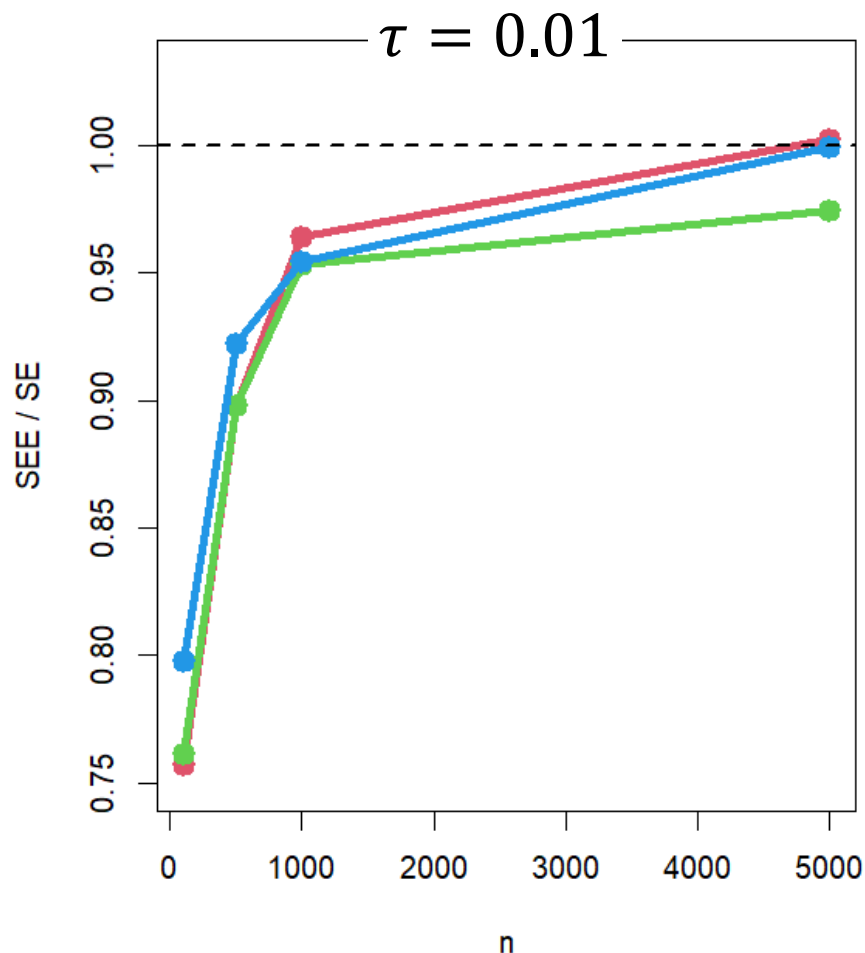
SIMULATION: SE ESTIMATION

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



SIMULATION: SE ESTIMATION

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

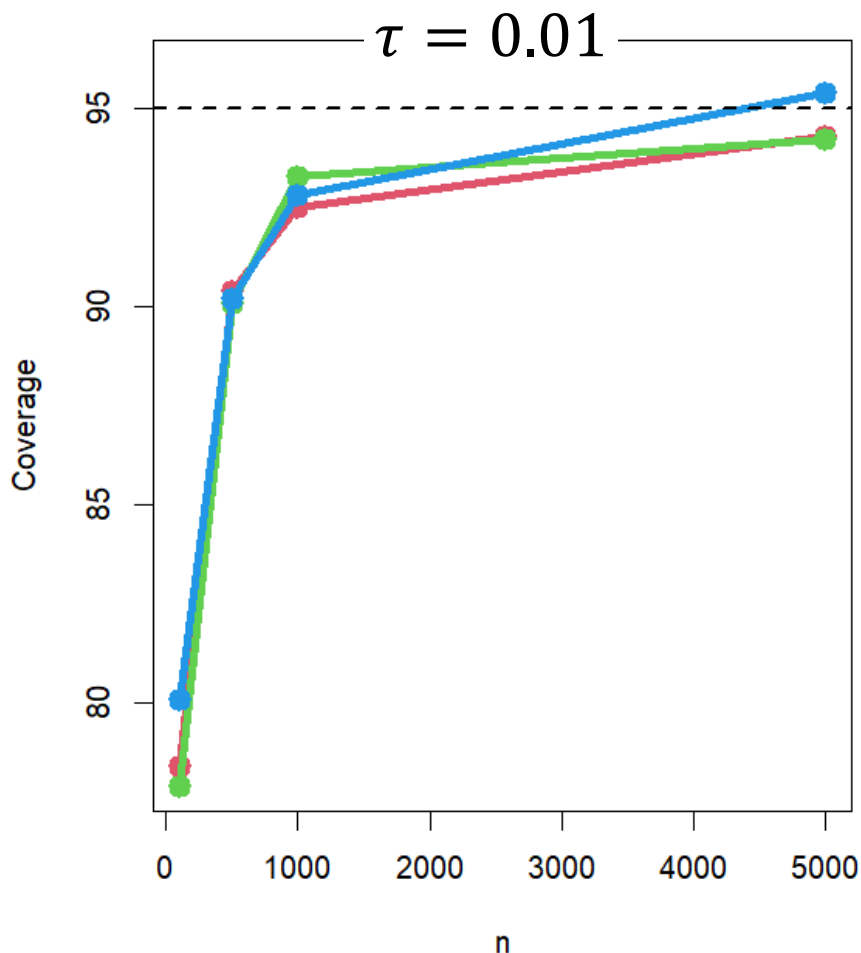


SIMULATION: 95% CI COVERAGE

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

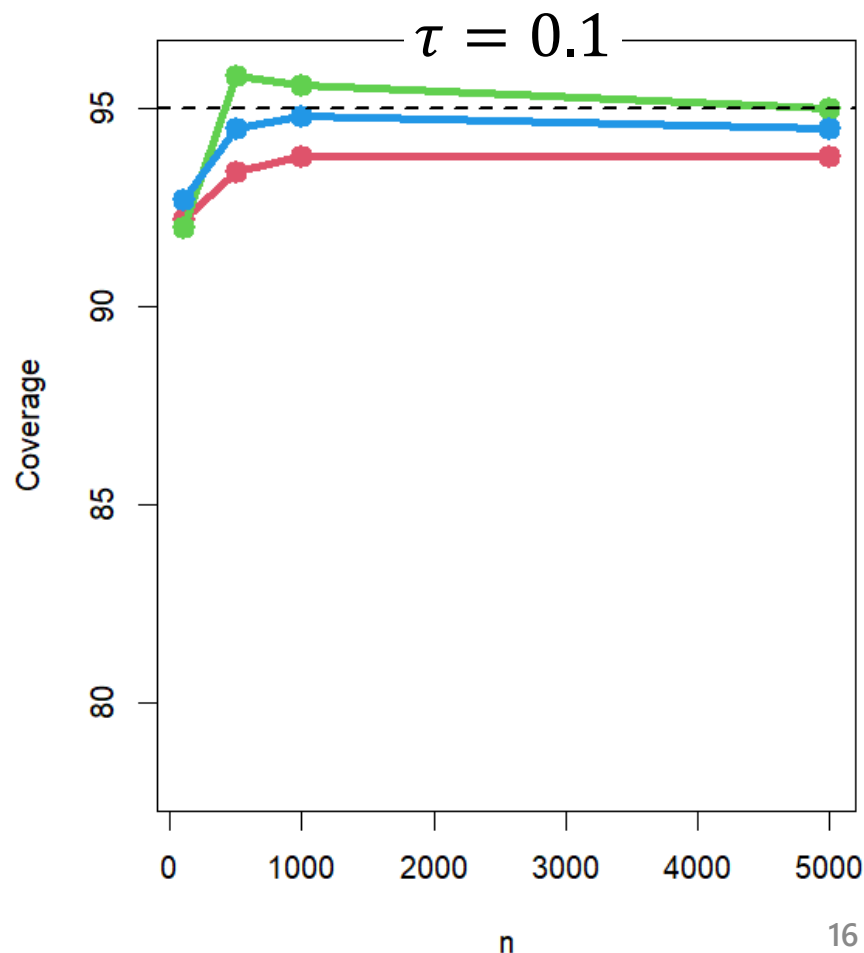
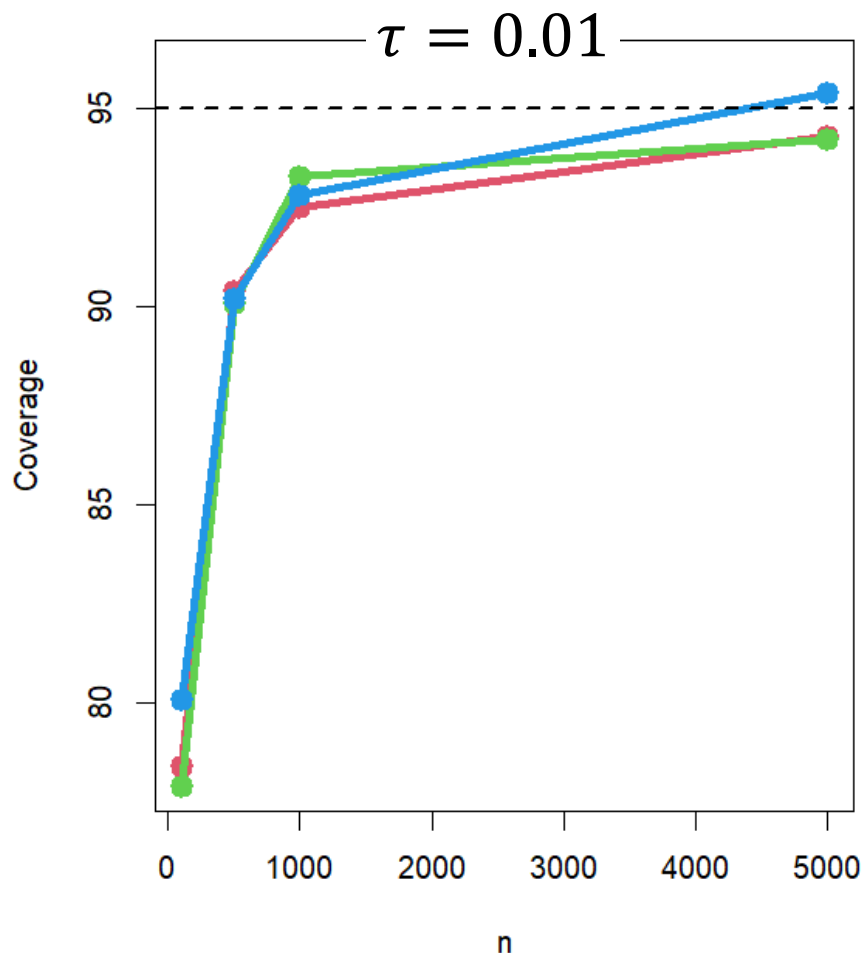
SIMULATION: 95% CI COVERAGE

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



SIMULATION: 95% CI COVERAGE

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



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. O'Neill, K. Burke

[Distributional regression models with automatic variable selection](#)