An extension for smoothed empirical likelihood confidence intervals for extreme quantiles and small sample sizes

Master Thesis

Author: Oliver Thunich[1;2]
Counseling: Sebastian Schoneberg[2]; Bertram Schäfer[2]
Supervisors: Claus Weihs[1]; David Meintrup[3]

[1]TU Dortmund
[2]Statcon GmbH
[3]TH Ingolstadt

Motivation

technische universität dortmund

STATCON

- Application: Calculating process capability without a distribution assumption.
  - Compute confidence intervals for "extreme" quantiles (e.g. $q = 0.01$).
  - Using a non-parametric method.
- Problem: small sample sizes are desired but lead to
  - Infinite confidence intervals.
  - Bad coverage rates.
- Method: Smoothed empirical likelihood.
  - Capable of returning non symmetrical confidence intervals.
  - Smoothing reduces coverage error.

Existing Methods

technische universität dortmund

STATCON

- Empirical likelihood was first introduced by Owen (1988)
- When quantiles are considered, the log likelihood is dependent on the empirical distribution function F_n (Adimari, 1998):

$$l(\Theta) = 2n\Big[F_n(\Theta)log(\frac{F_n(\Theta)}{q}) + (1 - F_n(\Theta))log(\frac{1 - F_n(\Theta)}{1 - q})\Big] \qquad (1)$$

- Owen (1988) shows that the Wilks theorem is applicable to compute asymptotic confidence intervals: $\lim_{n\to\infty} P(l(\Theta) \leq c) = P(\chi_1^2 \leq c)$
- As this results in a step function, several methods of smoothing have been proposed:
  - Smoothing using a kernel function. (Chen, Hall, 1993)
  - Linear smoothing of F_n (Adimari, 1998)

Linear Smoothing

technische universität dortmund

STATCON

- Let $x_{(1)} \leq ... \leq x_{(n)}$ be the ordered sample.
- The smoothing proposed by Adimari (1998) is achieved by using a linear smoothing F^* of F_n in equation 1.

$$F_n^*(\Theta) = \begin{cases} 0 & \text{if } \Theta < x_{(1)} \\ H(\Theta) & \text{if } \Theta \in [x_{(1)}, x_{(n)}) \\ 1 & \text{if } \Theta \geq x_{(n)} \end{cases}$$

where

$$H(\Theta) = \begin{cases} \frac{2i-1}{2n} & \text{if } \Theta = x_{(i)}; i \in \{1, .., n-1\} \\ (1-\lambda)\frac{2i-1}{2n} + \lambda\frac{2i+1}{2n} & \text{if } \Theta \in (x_{(i)}, x_{(i+1)}); \lambda = \frac{\Theta - x_{(i)}}{x_{(i+1)} - x_{(i)}}; i \in \{1, .., n-1\} \end{cases}$$

Problems:

- Constant likelihood values outside of the observed data can lead to infinite CI's.
- likelihood function still has two jumps (at $x_{(1)}$ and $x_{(n)}$).

Distribution Function

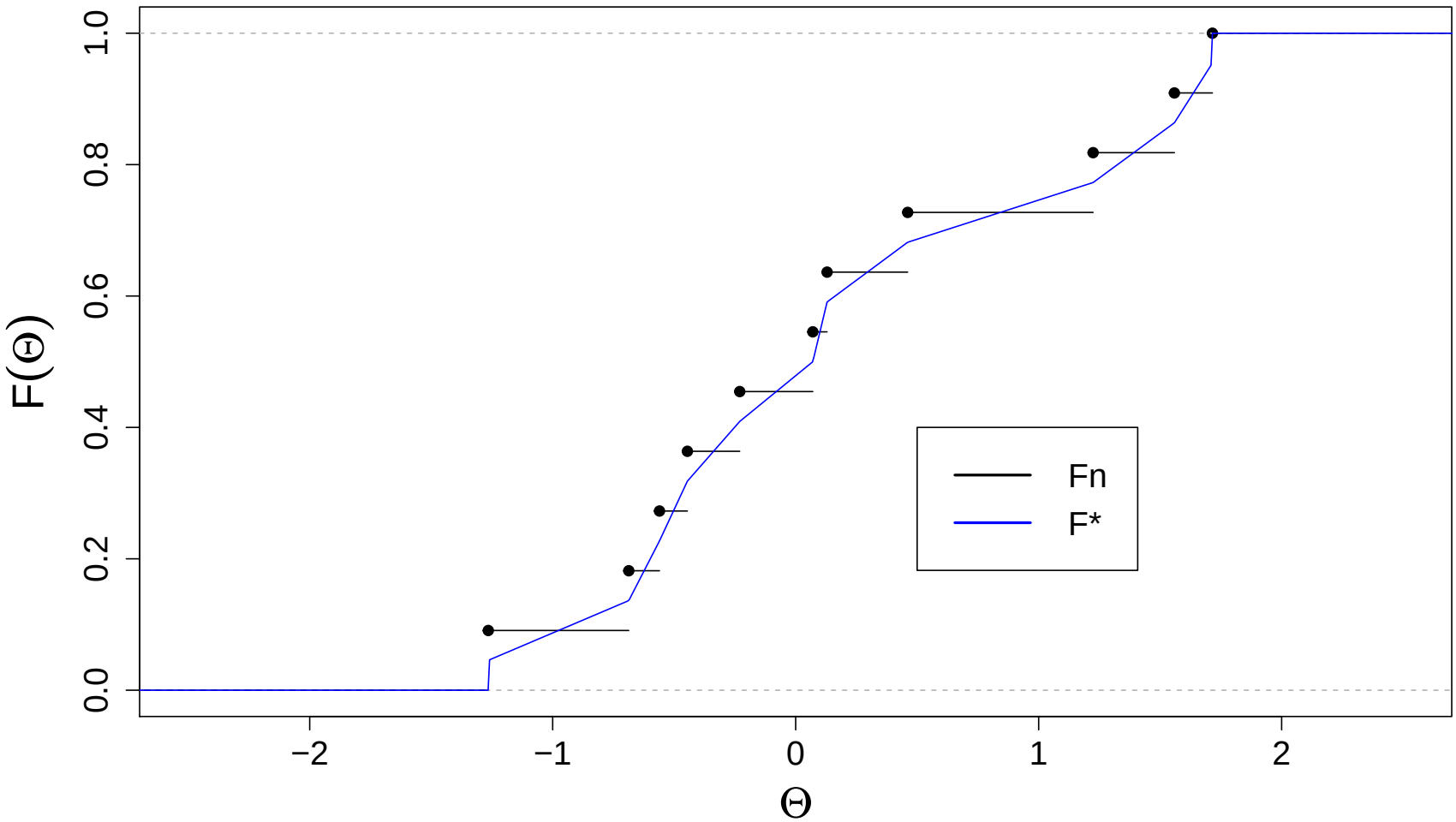

Figure: Smoothing of F_n for 11 observations from a standard normal distribution

Likelihood Function

technische universität dortmund

STATCON

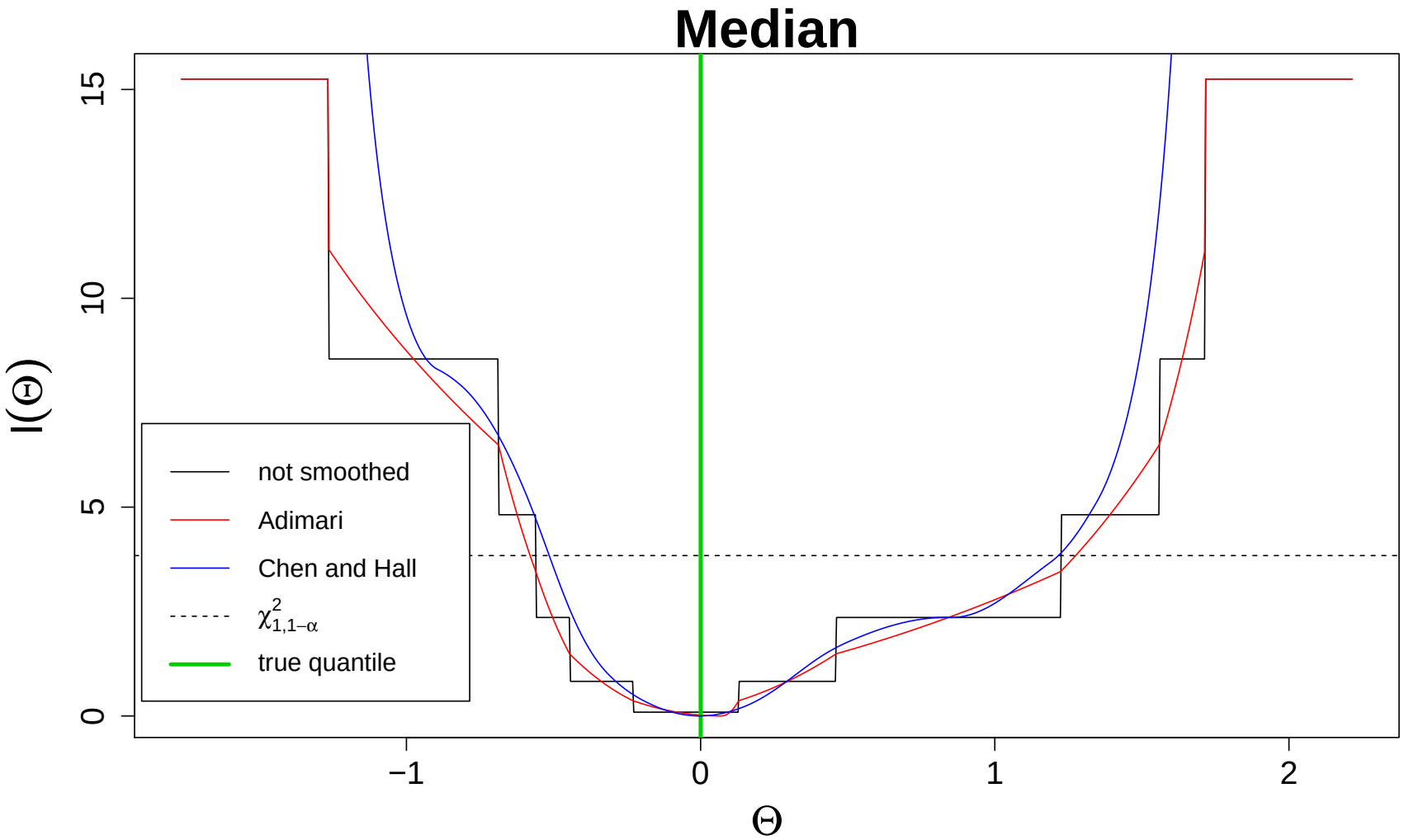

Figure: Smoothed empirical likelihood functions ($n = 11; q = 0.5; \alpha = 0.05$)

Likelihood Function

technische universität dortmund

STATCON

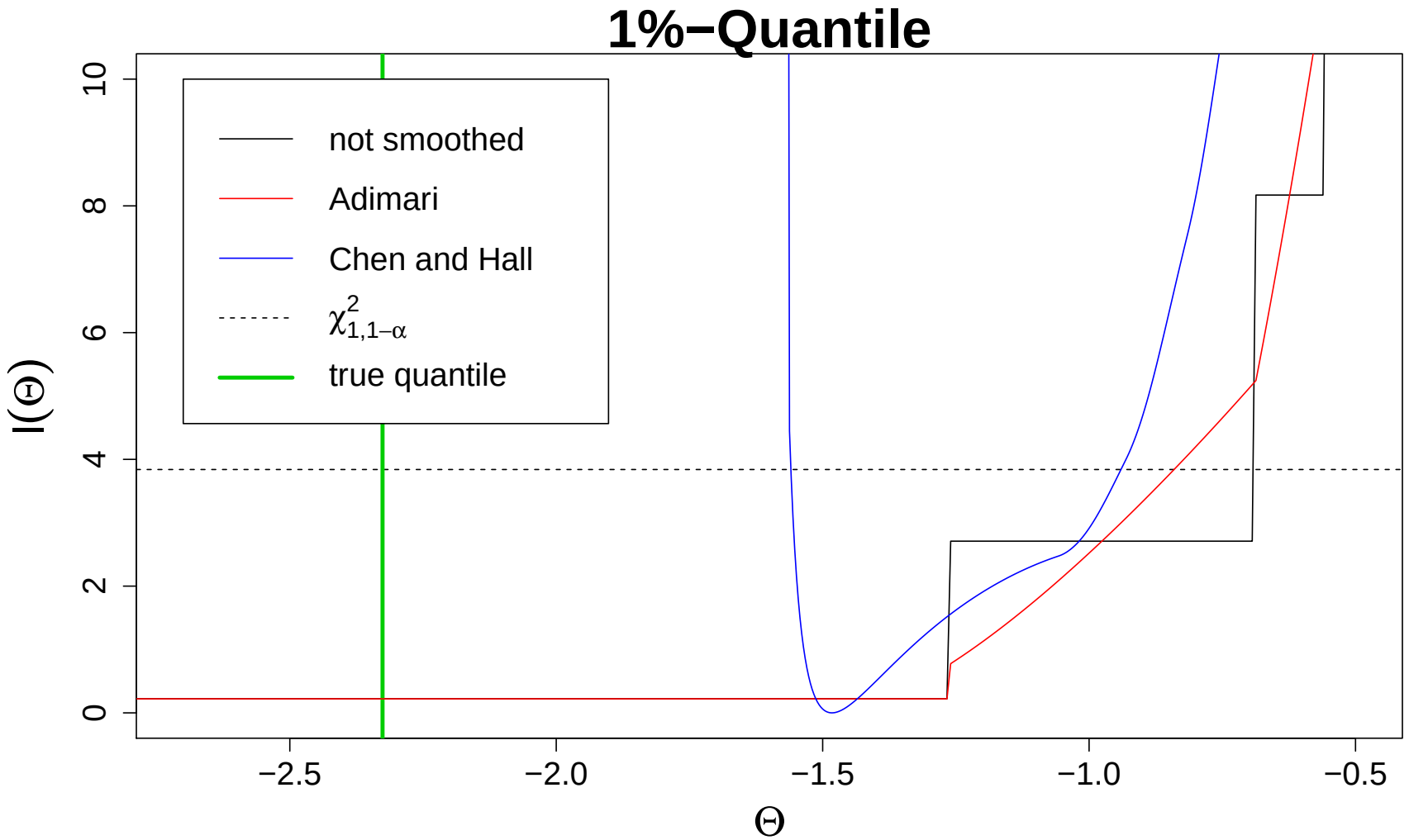

Figure: Smoothed empirical likelihood functions ($n = 11; q = 0.01; \alpha = 0.05$)

Coverage Rates

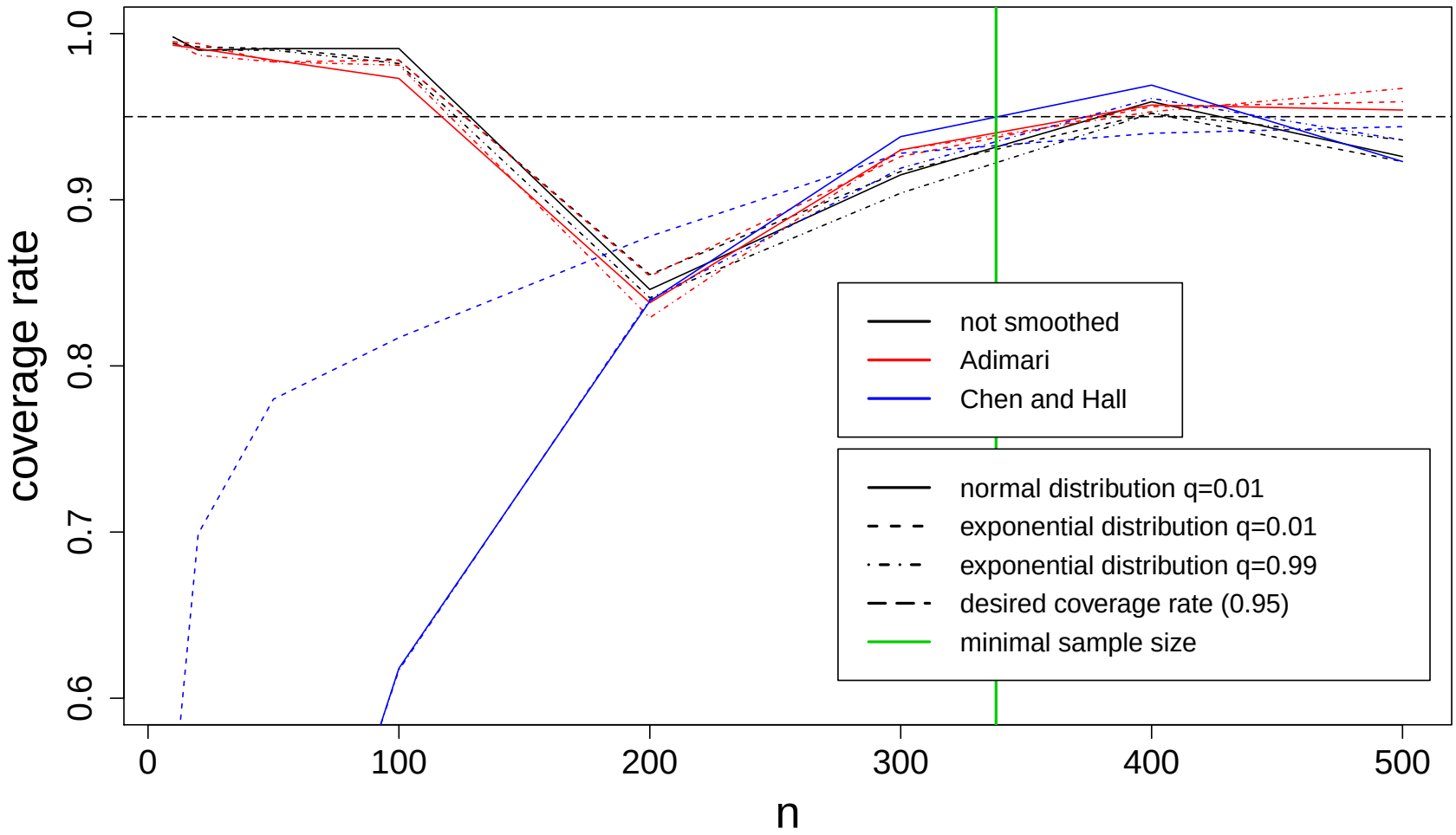

Figure: Estimated coverage rates based on 1000 samples

The Extension

technische universität dortmund

STATCON

Idea:

- Find an extension of the likelihood function for values outside of the observed data so that
  - finite confidence intervals are guaranteed.
  - the desired coverage rate is achieved.

1. extending the smoothed ecdf F^* as follows:

$$F_{ext}(\Theta) = \begin{cases} 0 & \text{if } \Theta \leq x_{(1)} - d_1 c \\ \frac{1}{2n} - \frac{1}{2n * d_1 * c}(x_{(1)} - \Theta) & \text{if } x_{(1)} - d_1 c < \Theta < x_{(1)} \\ H(\Theta) & \text{if } x_{(1)} \leq \Theta \leq x_{(n)} \\ \frac{2n-1}{2n} + \frac{1}{2n * d_2 * c}(\Theta - x_{(n)}) & \text{if } x_{(n)} < \Theta < x_{(n)} + d_2 c \\ 1 & \text{if } \Theta \geq x_{(n)} + d_2 c \end{cases}$$

$$\text{Where } c \geq 1;\ d_1 = \frac{1}{10}\sum_{i=1}^{5}(x_{(i+1)} - x_{(i)}) \text{ and}$$
$$d_2 = \frac{1}{10}\sum_{i=1}^{5}(x_{(n-i+1)} - x_{(n-i)})$$

2. linear extension of the likelihood function.

Visualising F_{ext}

technische universität dortmund

STATCON

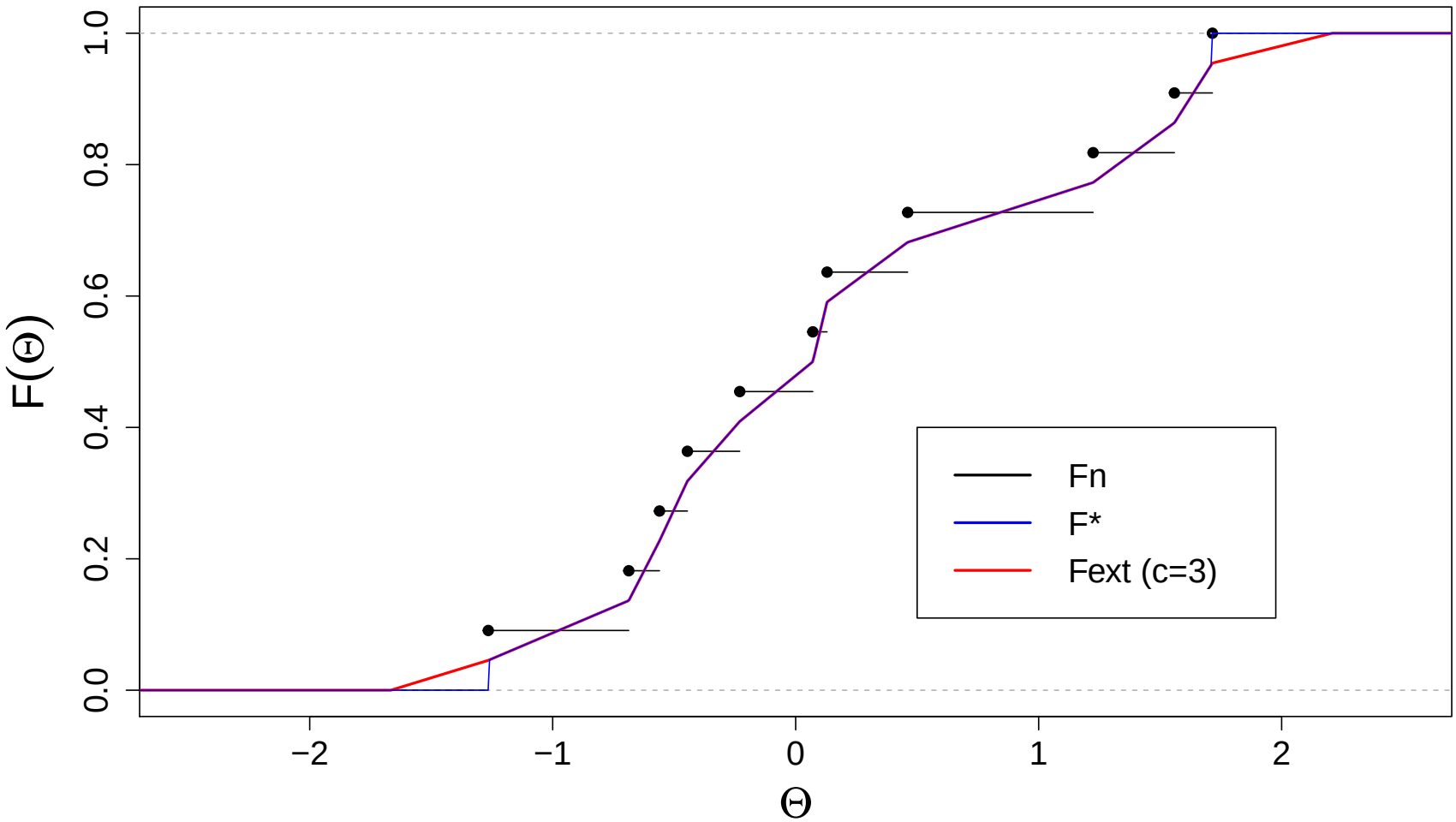

Figure: Smoothing of F_n using F_{ext}

Visualising F_{ext}

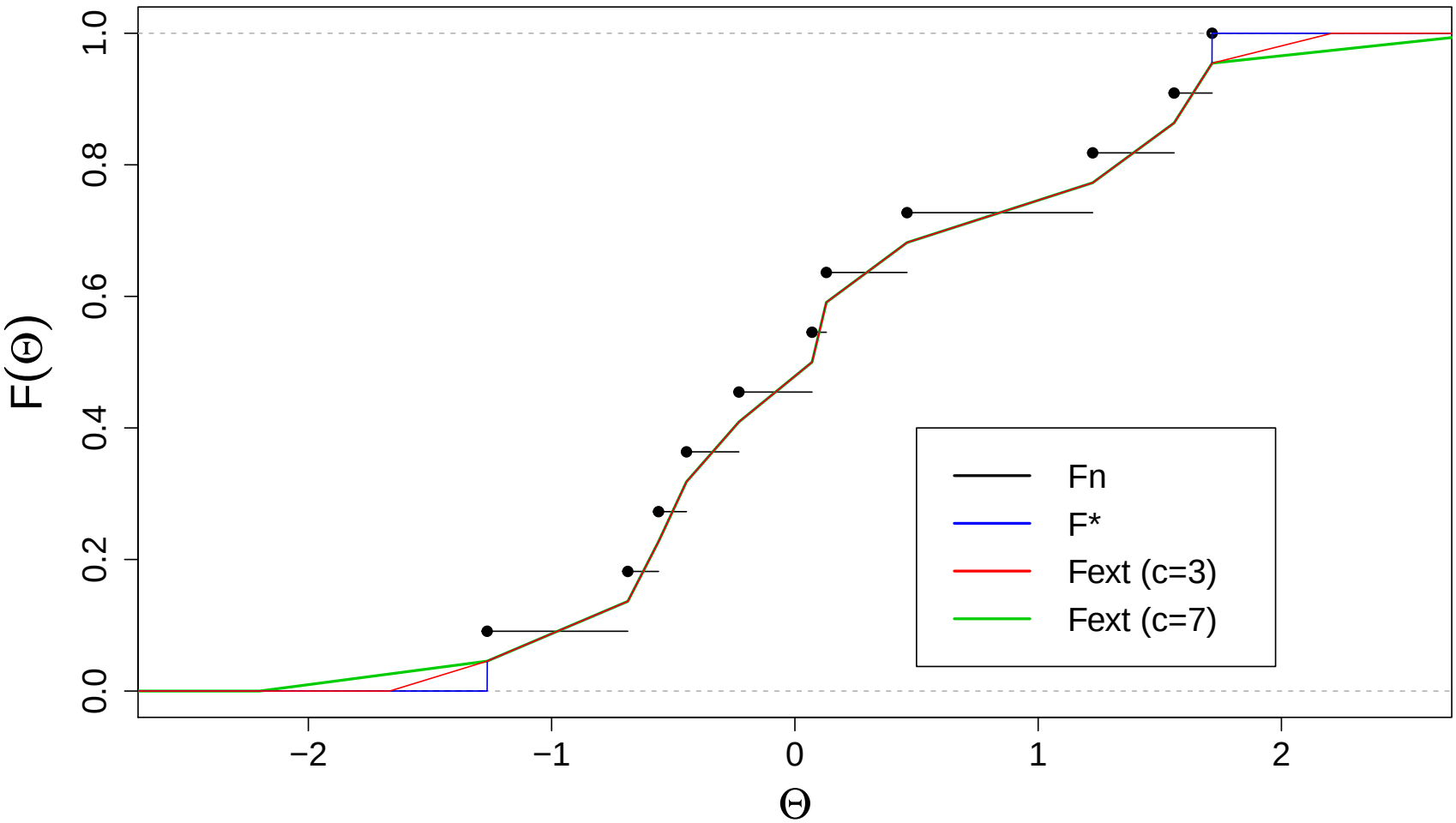

Figure: Visualising the influence of the parameter c on F_{ext}

Visualising F_{ext}

technische universität dortmund

STATCON

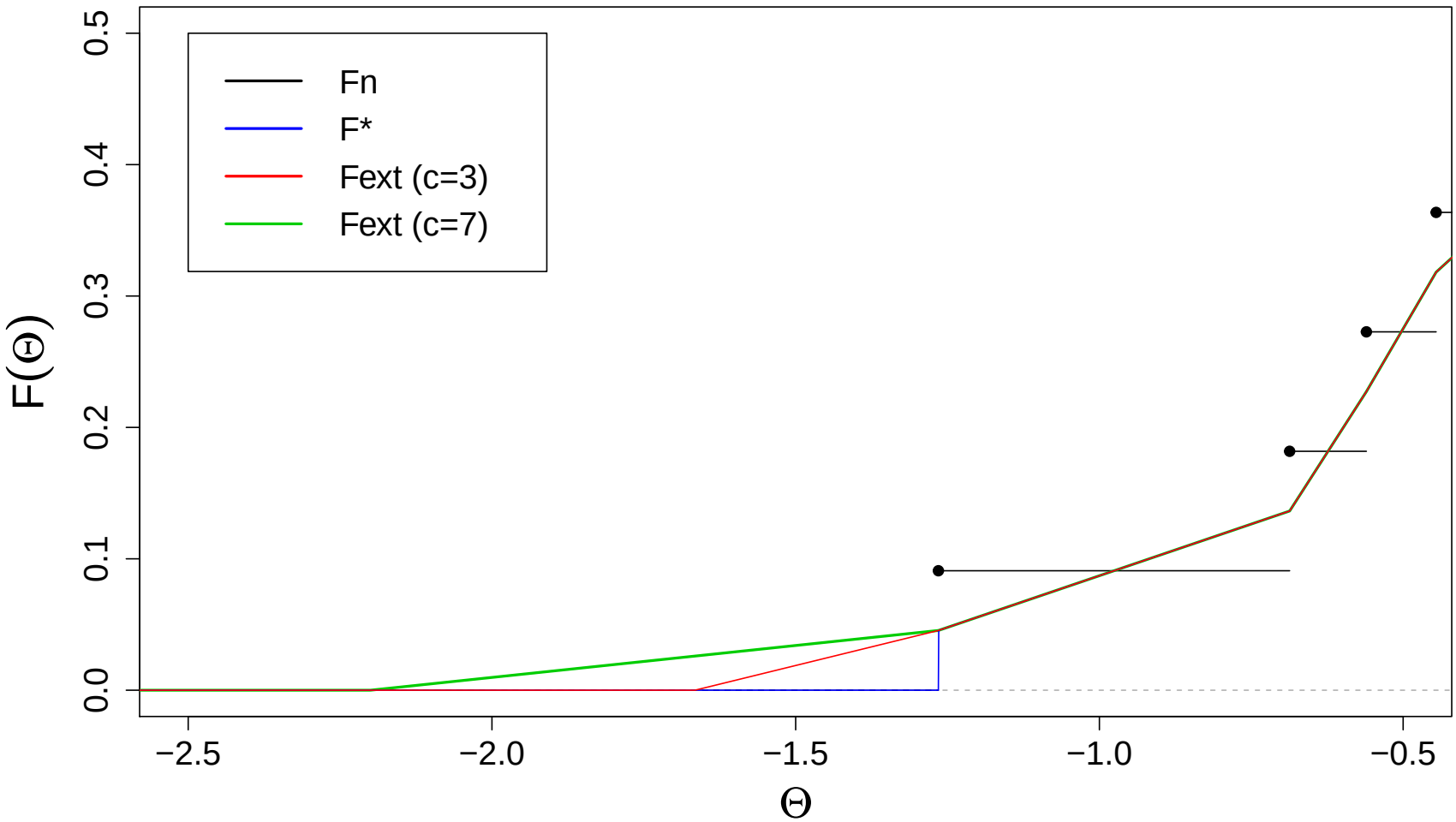

Figure: Visualising the influence of the parameter c on F_{ext}

Visualising $I(\Theta)$

technische universität dortmund

STATCON

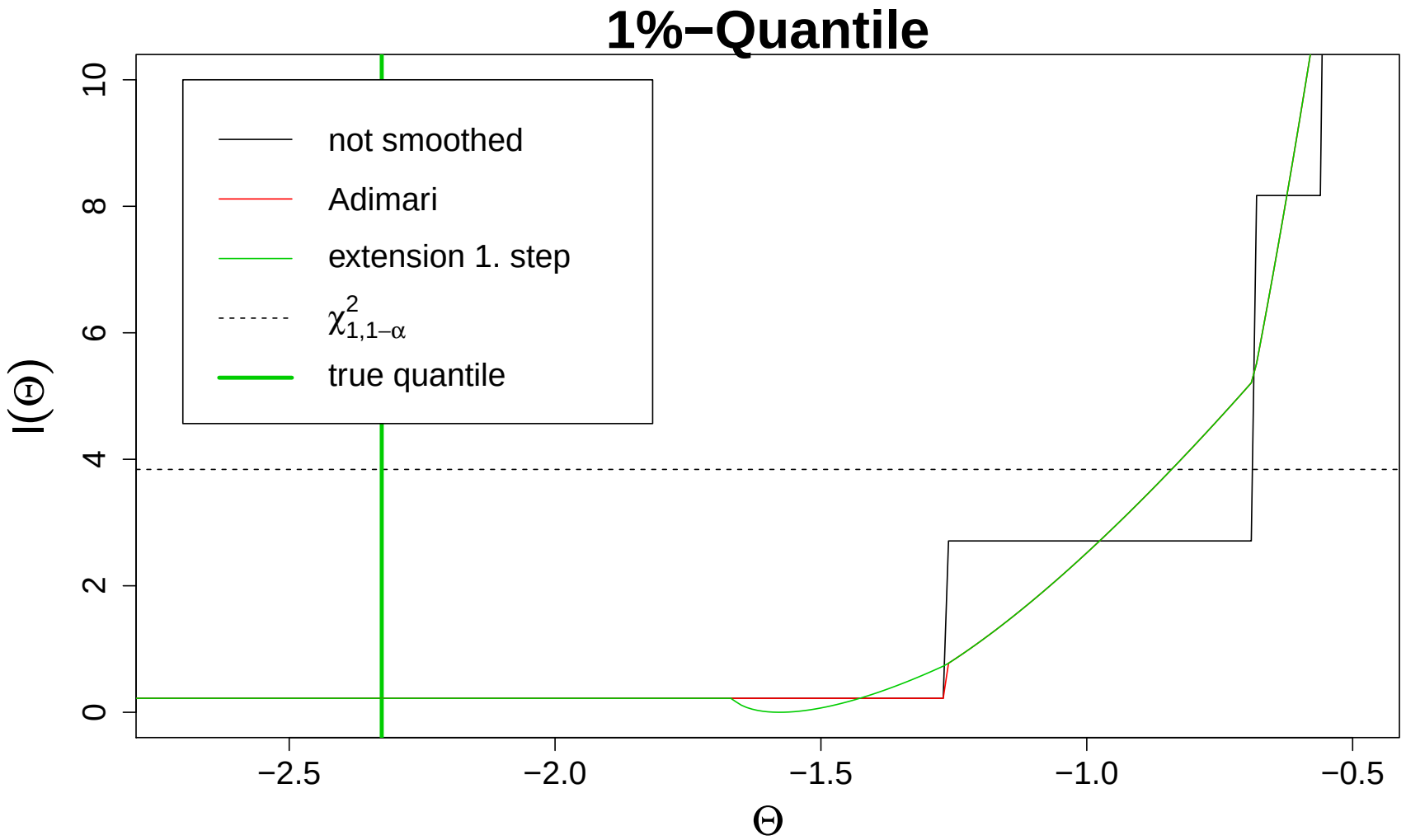

Figure: First step of the extension ($c = 3$)

Visualising $I(\Theta)$

technische universität dortmund

STATCON

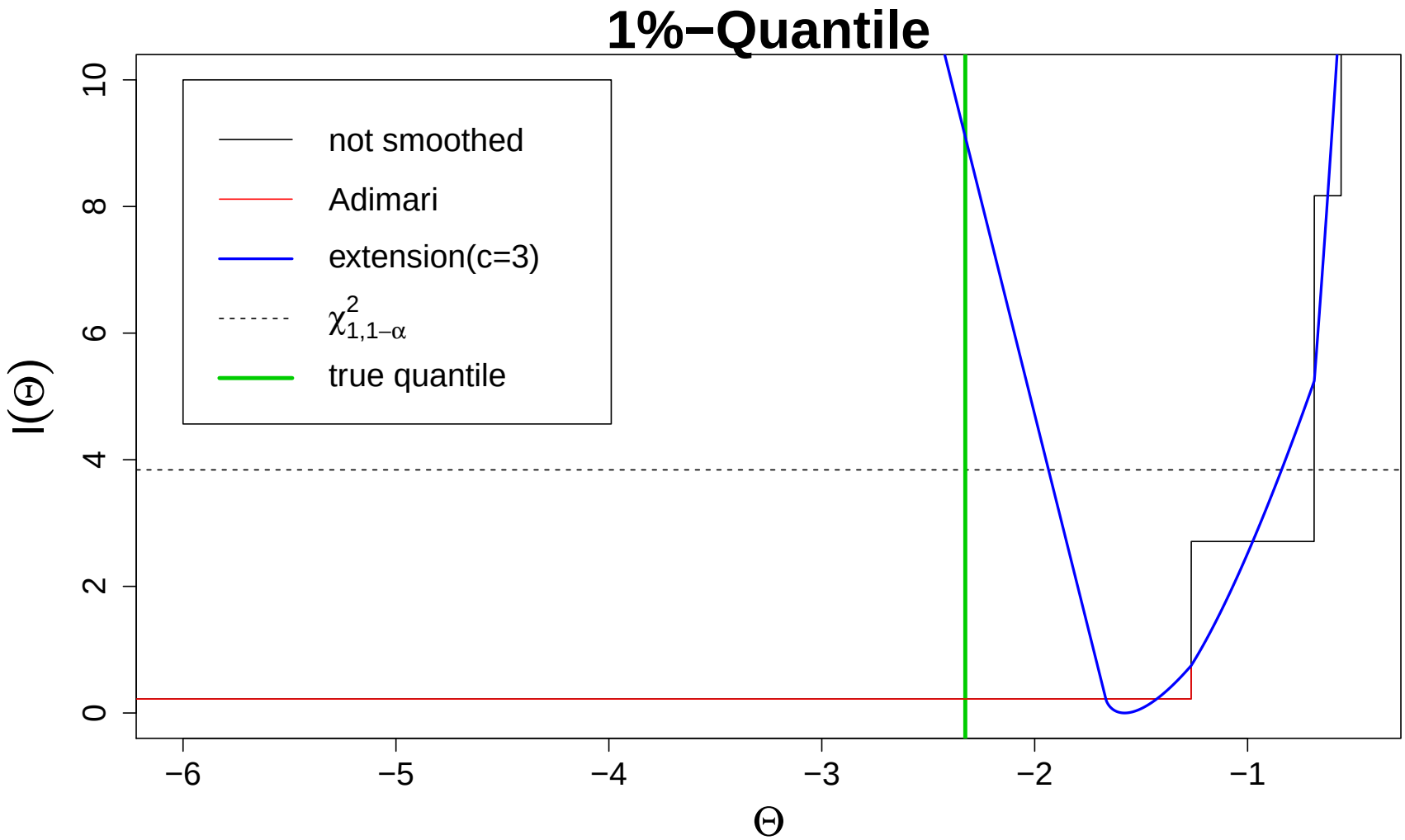

Figure: Second step: further linear extension ($c = 3$)

Visualising $l(\Theta)$

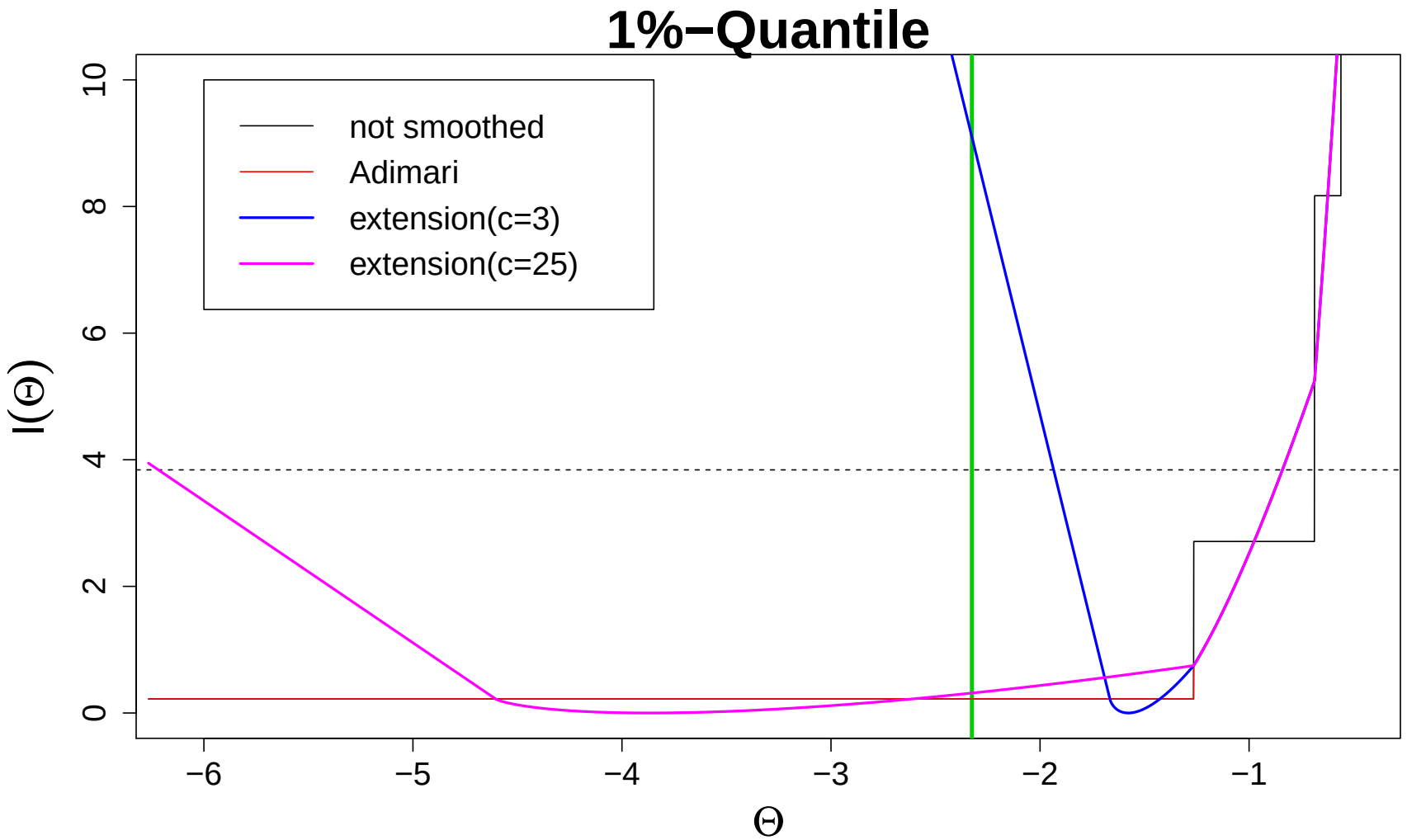

Figure: Fully extended likelihood function for different values of c

Extension parameter

technische universität dortmund

STATCON

- The smallest value of c which results in a coverage rate of at least $(1-\alpha)$ is desired.
- The extension is necessary, if the confidence region extends beyond the observed data.
- The necessity of the extension depends on the significance level α and the product R:

$$R := \begin{cases} q * n & \text{if } q \leq 0.5 \\ (1-q) * n & \text{if } q > 0.5 \end{cases}$$

- Assumption:
  - The required value for c depends on q, n, R and α
  - The required value for c does not depend on the distribution of the data.

Modelling c

technische universität dortmund

STATCON

Simulation study:

- For different values of $n; q$ and α choose the smallest value $c \in \mathbb{N}$ that produces a coverage of at least $(1 - \alpha)$.
- Evaluate coverage using $m = 5000$ samples of size n from a normal distribution.
- Try to find a model for c given q, n, R and α

Modelling c

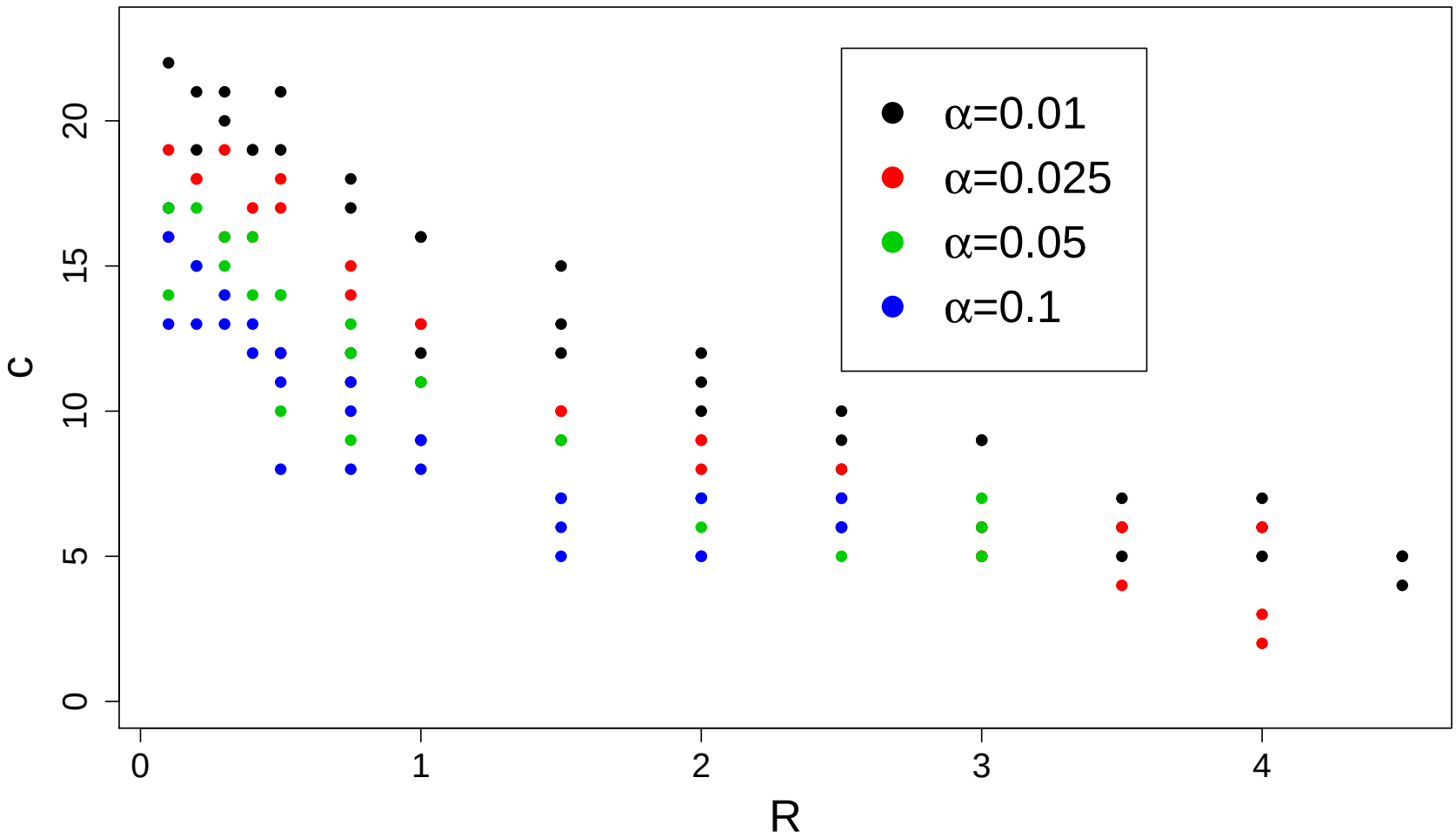

Figure: Chosen value of c for different values of R

Modelling c

technische universität dortmund

STATCON

Simulation study:

- For different values of n, q and α choose the smallest value $c \in \mathbb{N}$ that produces a coverage of at least $(1-\alpha)$.
- Evaluate coverage using $m = 5000$ samples of size n from a normal distribution.

- Model chosen:
  $\hat{c} = 12.344 - 7.082 * \sqrt{R} - 2.454 * log(\alpha) - 75.125 * q - 0.004 * n.$
  - $(1-q)$ is used for $q > 0.5$
- adjusted $R^2 = 0.933$

Modelling c

technische universität dortmund

STATCON

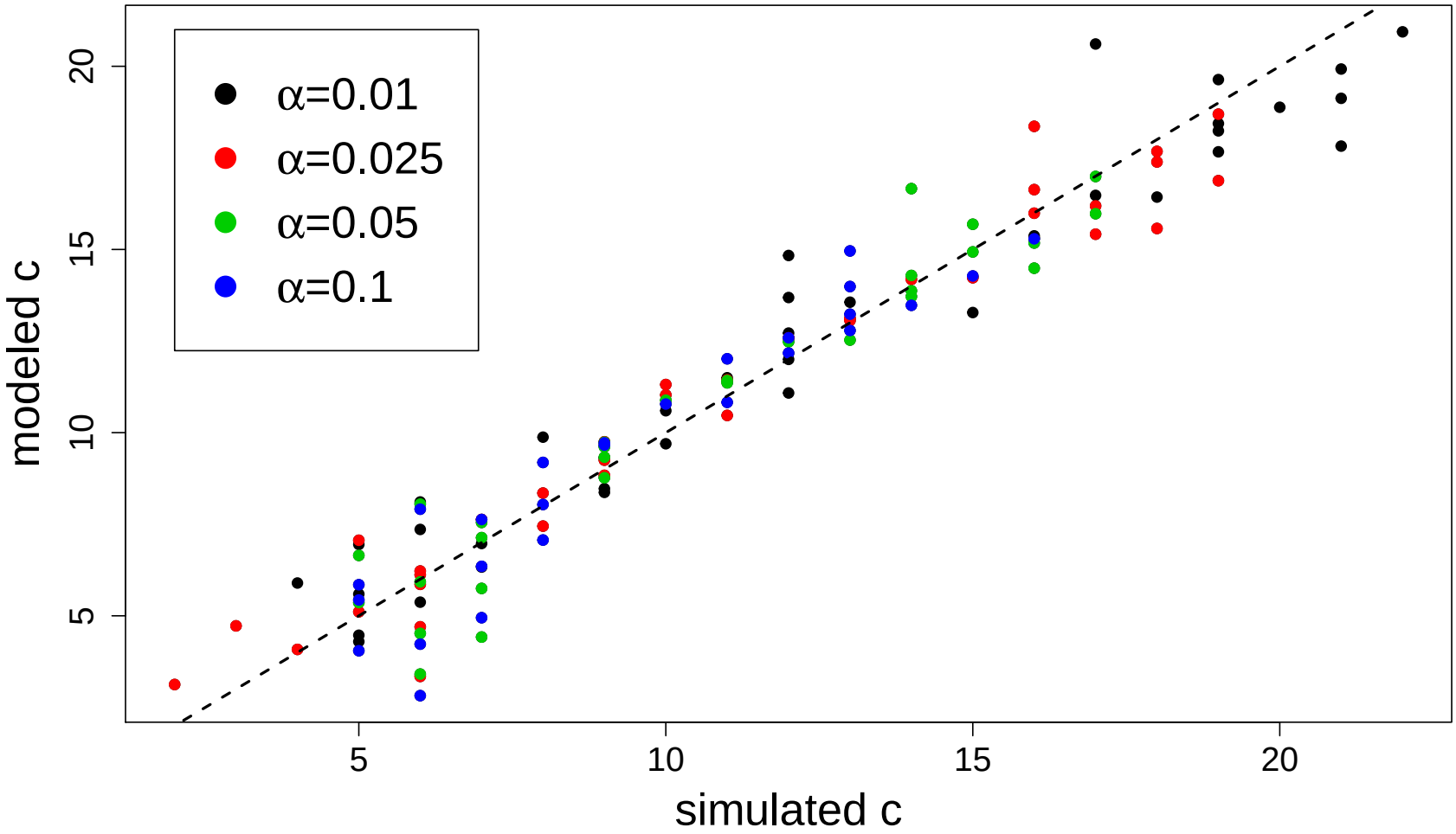

Figure: Visualising the Model

Modelling c

technische universität dortmund

STATCON

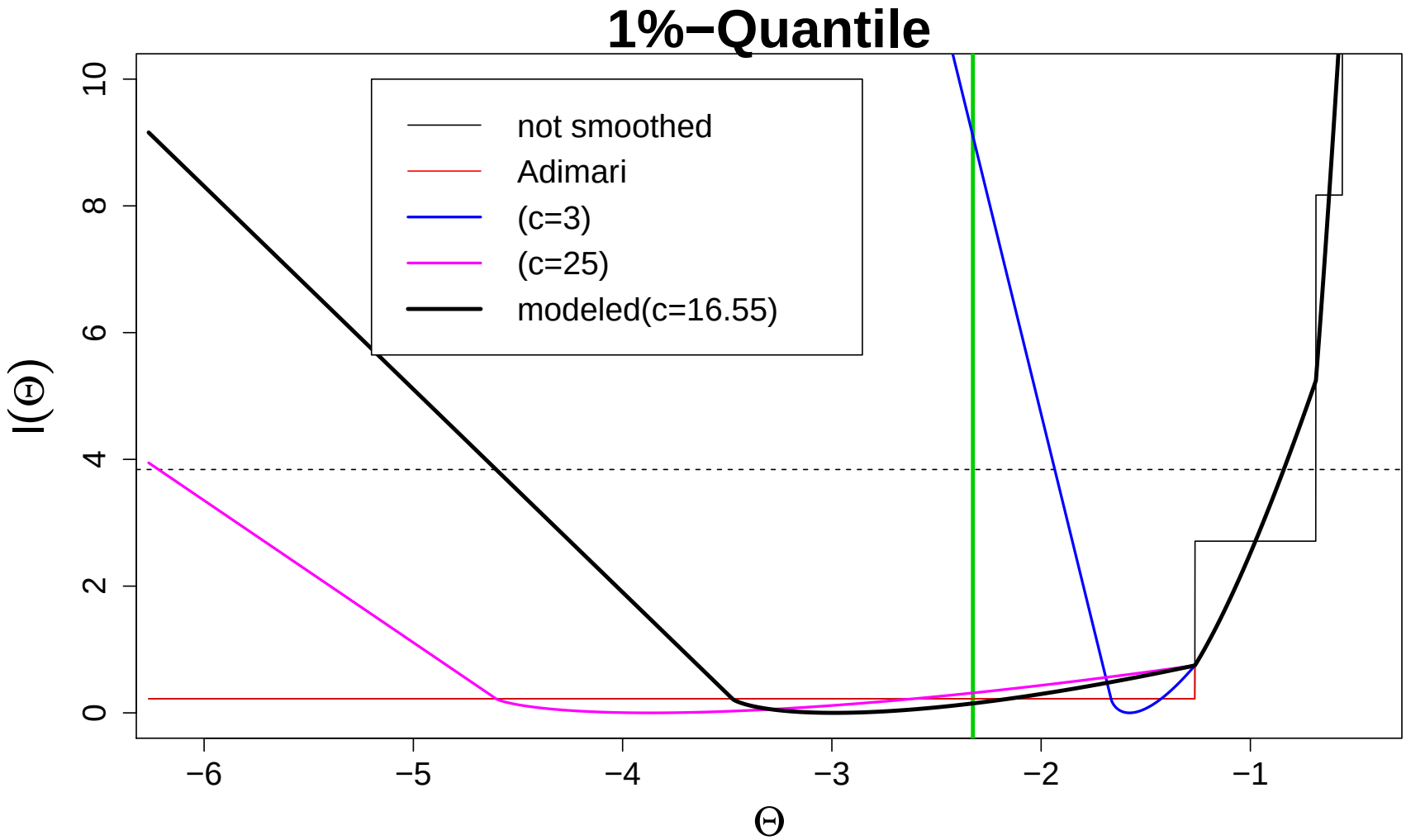

Figure: Example for the modeled value of c ($n = 11$, $q = 0.01$,$\alpha = 0.05$)

Performance: coverage

technische universität dortmund

STATCON

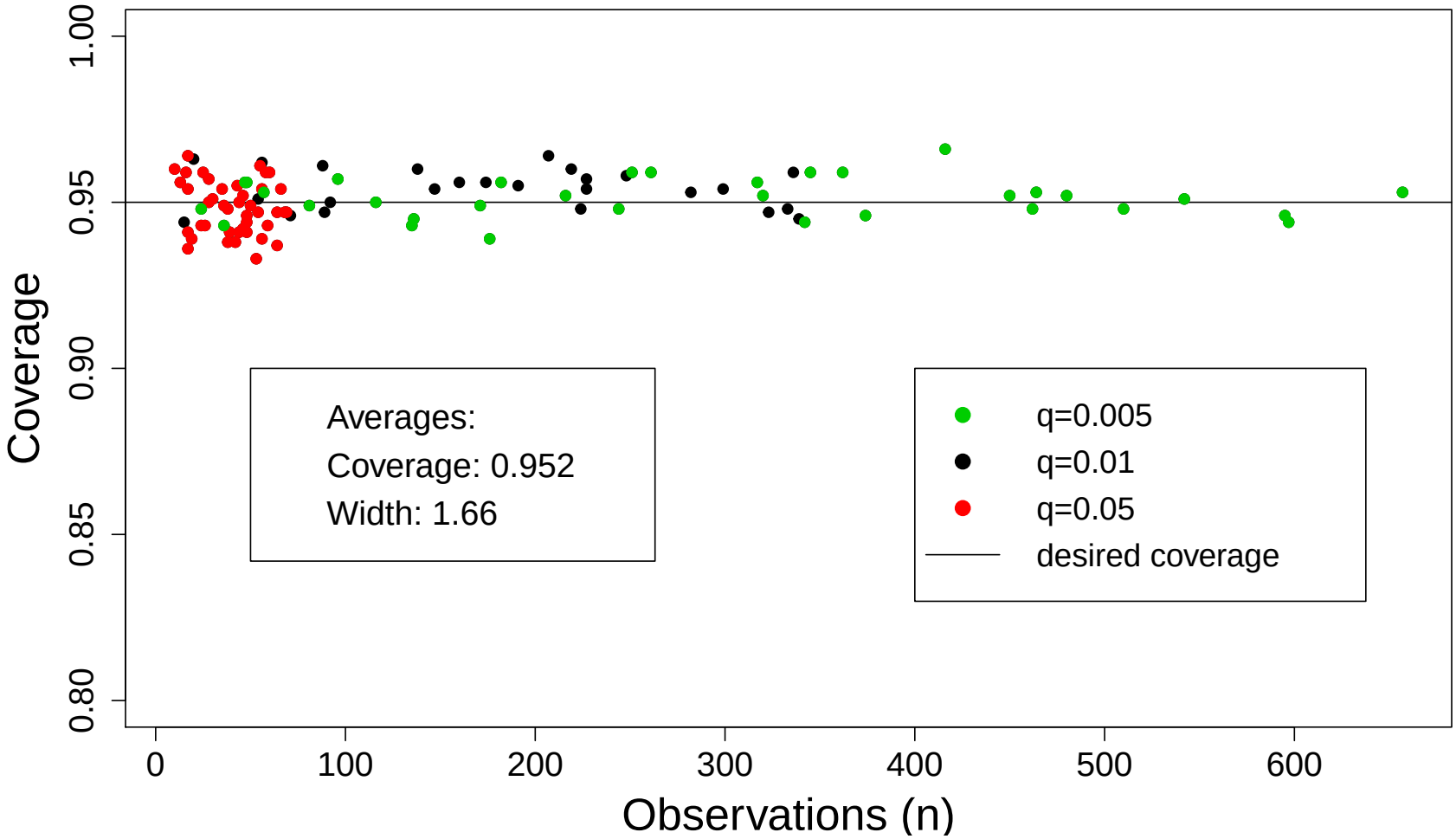

Figure: coverage rates based on 1000 samples for normally distributed data

Performance: coverage

technische universität dortmund

STATCON

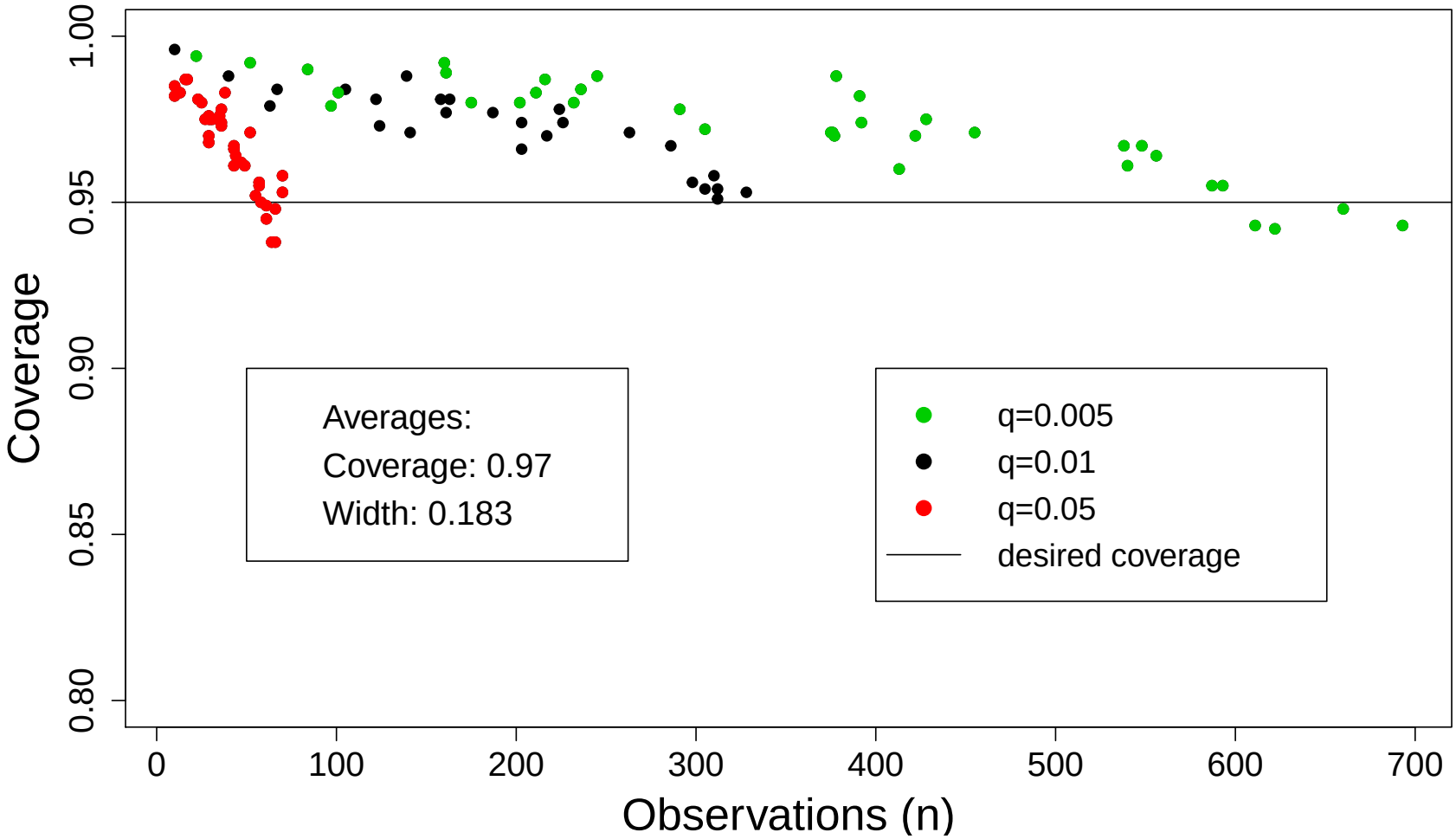

Figure: coverage rates based on 1000 samples for exponentially distributed data

Performance: coverage

technische universität dortmund

STATCON

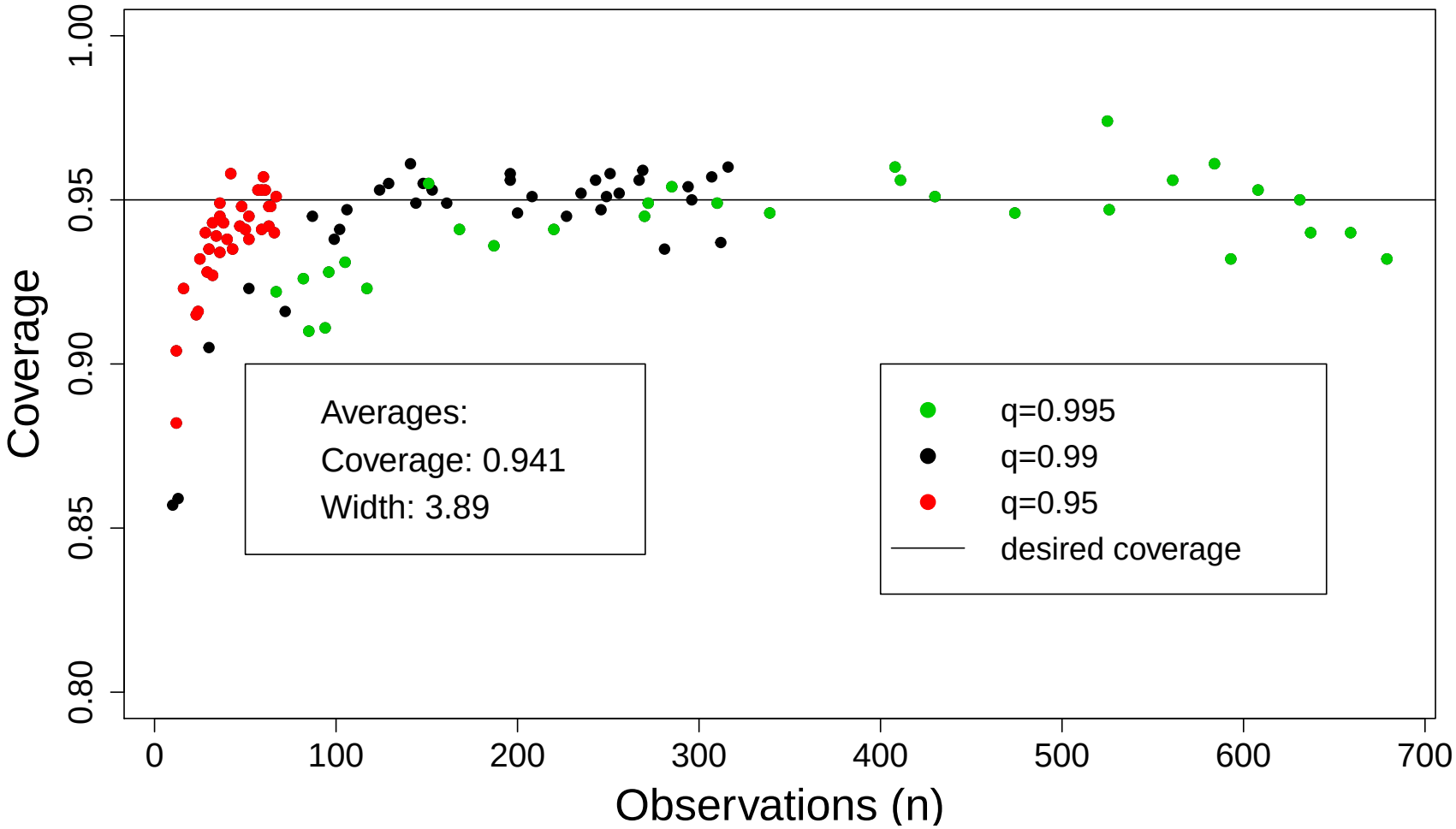

Figure: coverage rates based on 1000 samples for exponentially distributed data

Performance: discussion

technische universität dortmund

STATCON

Table: Average width if coverage is acceptable.
$q = 0.01, 0.99$, $\alpha = 0.05$, based on 1000 samples

	n= 10	n= 20	n= 50	n= 100	n= 200	n= 300	n= 400	n= 500
g-NV	Inf	Inf	Inf	Inf		0.858	0.918	0.664
A-NV	Inf	Inf	Inf	Inf		0.877	0.844	0.729
H-NV						0.856	0.963	0.662
Af-NV	4.554	3.383	2.522	1.953	1.517	1.028	0.849	0.728
g-Exp1	Inf	Inf	Inf	Inf		0.021	0.020	0.016
A-Exp1	Inf	Inf	Inf	Inf		0.022	0.020	0.018
H-Exp1						0.020	0.018	0.016
Af-Exp1	2.043	0.785	0.255	0.109	0.044	0.024	0.020	0.018
g-Exp99	Inf	Inf	Inf	Inf		2.479	2.688	1.886
A-Exp99	Inf	Inf	Inf	Inf		2.573	2.390	2.041
H-Exp99						2.423	2.695	1.806
Af-Exp99			5.463	4.943	4.065	2.882	2.411	2.050

Legend: g-not smoothed; A- Adimari, H-Chen and Hall, Af- Adimari extended, NV-(standard) normal distribution, Exp1-exponential distribution $q = 0.01$, Exp99-exponential distribution $q = 0.99$

Performance: discussion

technische universität dortmund

STATCON

Table: Average width if coverage is acceptable.
$q = 0.01; 0.99;\ \alpha = 0.05$, based on 1000 samples

	n= 10	n= 20	n= 50	n= 100	n= 200	n= 300	n= 400	n= 500
g-NV	Inf	Inf	Inf	Inf		0.858	0.918	0.664
A-NV	Inf	Inf	Inf	Inf		0.877	0.844	0.729
H-NV						0.856	0.963	0.662
Af-NV	4.554	3.383	2.522	1.953	1.517	1.028	0.849	0.728
g-Exp1	Inf	Inf	Inf	Inf		0.021	0.020	0.016
A-Exp1	Inf	Inf	Inf	Inf		0.022	0.020	0.018
H-Exp1						0.020	0.018	0.016
Af-Exp1	2.043	0.785	0.255	0.109	0.044	0.024	0.020	0.018
g-Exp99	Inf	Inf	Inf	Inf		2.479	2.688	1.886
A-Exp99	Inf	Inf	Inf	Inf		2.573	2.390	2.041
H-Exp99						2.423	2.695	1.806
Af-Exp99			5.463	4.943	4.065	2.882	2.411	2.050

Legend: g-not smoothed, A- Adimari, H-Chen and Hall, Af- Adimari extended, NV-(standard) normal distribution, Exp1-exponential distribution $q = 0.01$, Exp99-exponential distribution $q = 0.99$

Performance: discussion

technische universität dortmund

STATCON

Table: Average width if coverage is acceptable.
$q = 0.01, 0.99$, $\alpha = 0.05$, based on 1000 samples

	n= 10	n= 20	n= 50	n= 100	n= 200	n= 300	n= 400	n= 500
g-NV	Inf	Inf	Inf	Inf		0.858	0.918	0.664
A-NV	Inf	Inf	Inf	Inf		0.877	0.844	0.729
H-NV						0.856	0.963	0.662
Af-NV	4.554	3.383	2.522	1.953	1.517	1.028	0.849	0.728
g-Exp1	Inf	Inf	Inf	Inf		0.021	0.020	0.016
A-Exp1	Inf	Inf	Inf	Inf		0.022	0.020	0.018
H-Exp1						0.020	0.018	0.016
Af-Exp1	2.043	0.785	0.255	0.109	0.044	0.024	0.020	0.018
g-Exp99	Inf	Inf	Inf	Inf		2.479	2.688	1.886
A-Exp99	Inf	Inf	Inf	Inf		2.573	2.390	2.041
H-Exp99						2.423	2.695	1.806
Af-Exp99			5.463	4.943	4.065	2.882	2.411	2.050

Legend: g-not smoothed, A- Adimari, H-Chen and Hall, Af- Adimari extended, NV-(standard) normal distribution, Exp1-exponential distribution $q = 0.01$, Exp99-exponential distribution $q = 0.99$

Performance: discussion

technische universität dortmund

STATCON

Table: Average width if coverage is acceptable.
$q = 0.01, 0.99,\ \alpha = 0.05$, based on 1000 samples

	n= 10	n= 20	n= 50	n= 100	n= 200	n= 300	n= 400	n= 500
g-NV	Inf	Inf	Inf	Inf		0.858	0.918	0.664
A-NV	Inf	Inf	Inf	Inf		0.877	0.844	0.729
H-NV						0.856	0.963	0.662
Af-NV	4.554	3.383	2.522	1.953	1.517	1.028	0.849	0.728
g-Exp1	Inf	Inf	Inf	Inf		0.021	0.020	0.016
A-Exp1	Inf	Inf	Inf	Inf		0.022	0.020	0.018
H-Exp1						0.020	0.018	0.016
Af-Exp1	2.043	0.785	0.255	0.109	0.044	0.024	0.020	0.018
g-Exp99	Inf	Inf	Inf	Inf		2.479	2.688	1.886
A-Exp99	Inf	Inf	Inf	Inf		2.573	2.390	2.041
H-Exp99						2.423	2.695	1.806
Af-Exp99			5.463	4.943	4.065	2.882	2.411	2.050

Legend: g-not smoothed, A- Adimari, H-Chen and Hall, Af- Adimari extended, NV-(standard) normal distribution, Exp1-exponential distribution $q = 0.01$, Exp99-exponential distribution $q = 0.99$

Outlook

technische universität dortmund

STATCON

- Better account for the shape of the data by assigning higher weights to extreme observations:
  - Using a weighted mean for computing d_1 and d_2.
- Test of a semi parametric variation:
  - Assume a class of distributions (e.g. exponential) and Modell c using samples from that distribution.

References

Adimari, G. (1998). An empirical likelihood statistic for quantiles. Journal of Statistical Computation and Simulation, 60(1) pages 85-95.

Chen, S. X., Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. The Annals of Statistics, pages 1166-1181.

Owen, A. B. (1988). Empirical Likelihood Ratio Confidence IIntervals for a Single Functional. Biometrika, Vol. 75, No. 2(Jun. 1988), pages 237-249.

Zhu, H. (2007) Smoothed Empirical Likelihood for Quantiles and Some Variations/Extension of Empirical Likelihood for Buckley-James Estimator, Ph.D. dissertation, University of Kentucky.