Efficient and Robust Scale Estimation

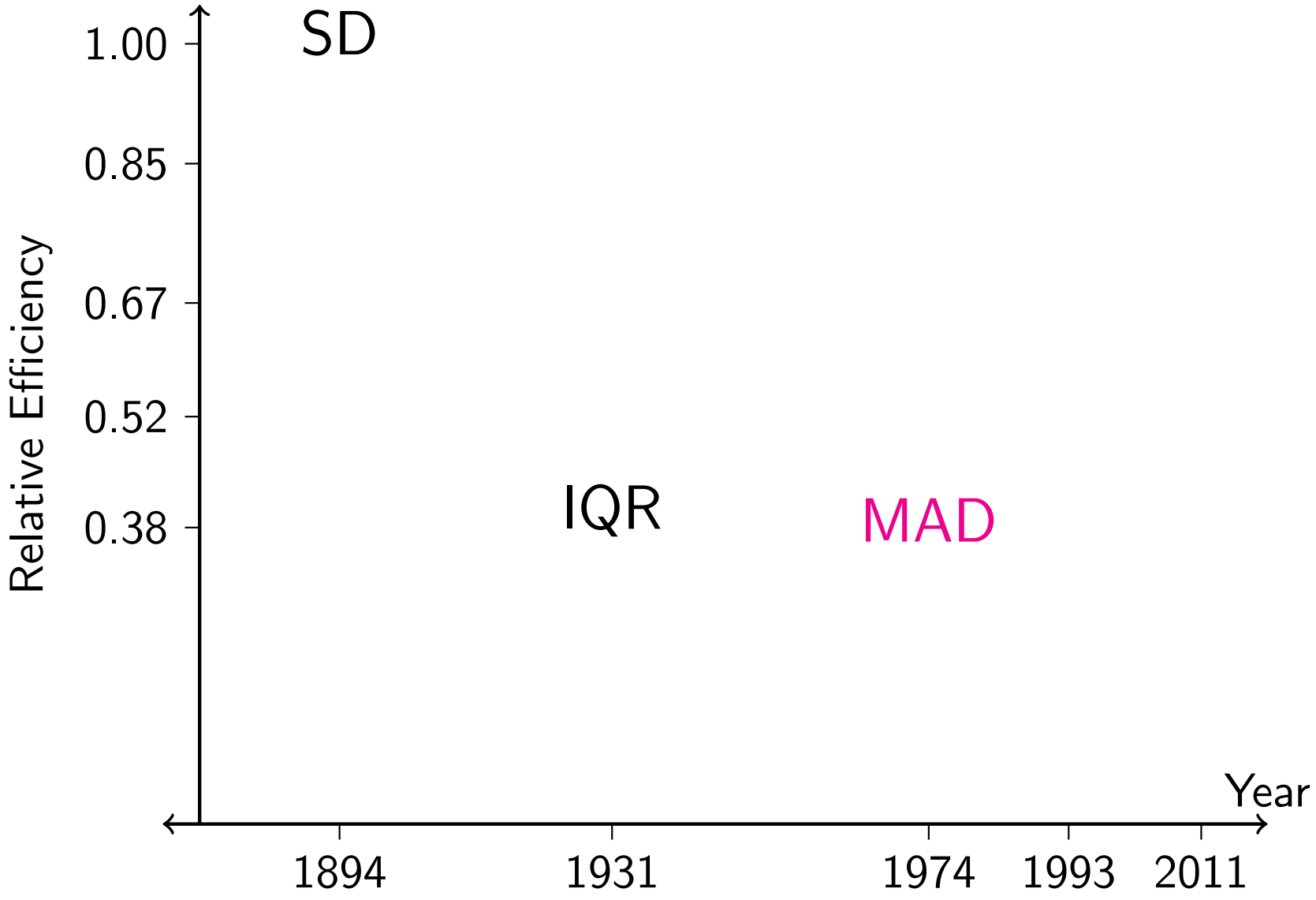
Garth Tarr, Samuel Müller and Neville Weber

School of Mathematics and Statistics
THE UNIVERSITY OF SYDNEY

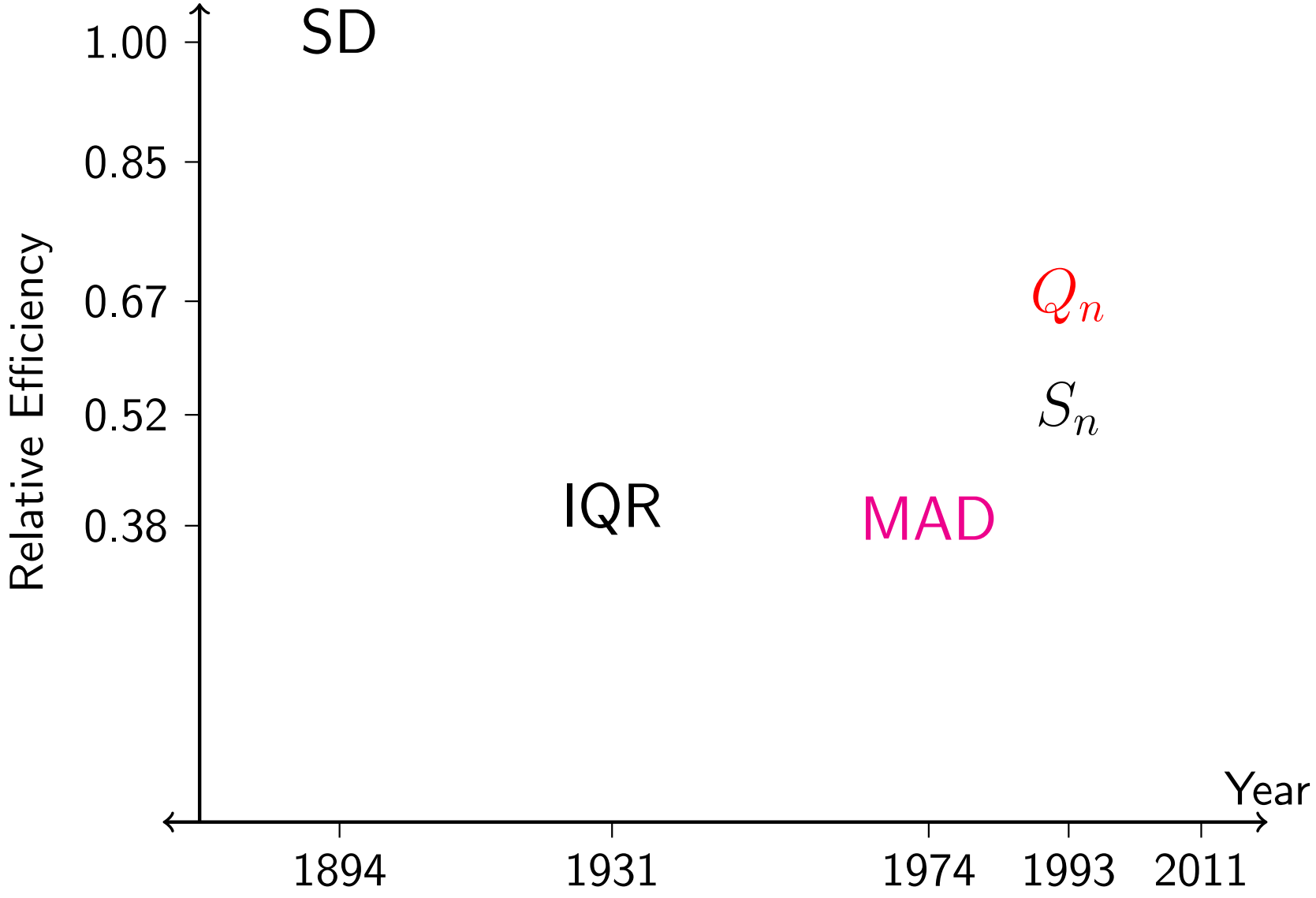
Outline

History of relative efficiencies at the Gaussian ($n = 20$)

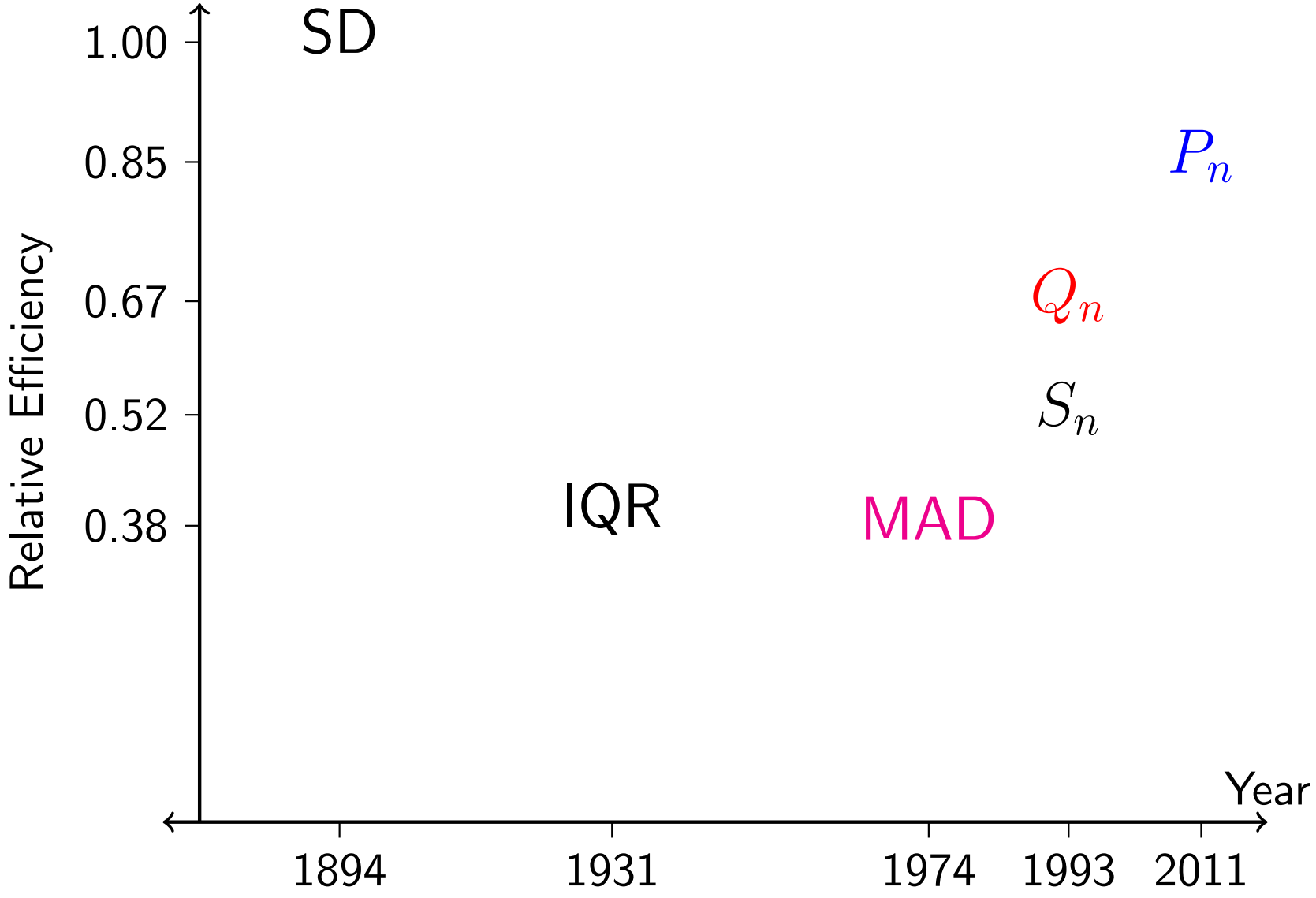
History of relative efficiencies at the Gaussian ($n = 20$)



History of relative efficiencies at the Gaussian ($n = 20$)



History of relative efficiencies at the Gaussian ($n = 20$)



U-quantile statistics

- Given data $\mathbf{X} = (X_1, \ldots, X_n)$ and a symmetric kernel $h : \mathbb{R}^2 \mapsto \mathbb{R}$ a U-statistic of order 2 is defined as:

$$U_n(\mathbf{X}) := \binom{n}{2}^{-1} \sum_{i<j} h(X_i, X_j). \tag{1}$$

U-quantile statistics

- Given data $\mathbf{X} = (X_1, \ldots, X_n)$ and a symmetric kernel $h : \mathbb{R}^2 \mapsto \mathbb{R}$ a U-statistic of order 2 is defined as:

$$U_n(\mathbf{X}) := \binom{n}{2}^{-1} \sum_{i<j} h(X_i, X_j). \tag{1}$$

- Let $H(t) = P(h(X_i, X_j) \leq t)$ be the cdf of the kernels with corresponding empirical distribution function,

$$H_n(t) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{I}\{h(X_i, X_j) \leq t\}, \quad \text{for } t \in \mathbb{R}. \tag{2}$$

U-quantile statistics

- Given data $\mathbf{X} = (X_1, \ldots, X_n)$ and a symmetric kernel $h : \mathbb{R}^2 \mapsto \mathbb{R}$ a U-statistic of order 2 is defined as:

$$U_n(\mathbf{X}) := \binom{n}{2}^{-1} \sum_{i<j} h(X_i, X_j). \tag{1}$$

- Let $H(t) = P(h(X_i, X_j) \leq t)$ be the cdf of the kernels with corresponding empirical distribution function,

$$H_n(t) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{I}\{h(X_i, X_j) \leq t\}, \quad \text{for } t \in \mathbb{R}. \tag{2}$$

- For $0 < p < 1$, the corresponding sample U-quantile is:

$$H_n^{-1}(p) := \inf\{t : H_n(t) \geq p\}. \tag{3}$$

Generalized L-statistics

A generalized linear (GL) statistic can be defined as

$$T_n(H_n) = \int_I J(p) H_n^{-1}(p) dp + \sum_{j=1}^{d} a_j H_n^{-1}(p_j).$$

where

- J is function for smooth weighting of $H_n^{-1}(p)$
- $I \in [0, 1]$ is some interval
- a_j are discrete coefficients for $H_n^{-1}(p_j)$

(Serfling, 1984)

Examples of GL-statistics

- **Interquartile range**: $h(x) = x$,

$$\mathrm{IQR} = H_n^{-1}(0.75) - H_n^{-1}(0.25)$$

Examples of GL-statistics

- **Interquartile range**: $h(x) = x$,

$$\mathrm{IQR} = H_n^{-1}(0.75) - H_n^{-1}(0.25)$$

- **Variance**: $h(x, y) = \frac{1}{2}(x - y)^2$,

$$\int_0^1 H_n^{-1}(p)dp$$

Examples of GL-statistics

- **Interquartile range**: $h(x) = x$,

$$\mathrm{IQR} = H_n^{-1}(0.75) - H_n^{-1}(0.25)$$

- **Variance**: $h(x, y) = \frac{1}{2}(x - y)^2$,

$$\int_0^1 H_n^{-1}(p)dp$$

- **Winsorized variance**: $h(x, y) = \frac{1}{2}(x - y)^2$,

$$\int_0^{0.75} H_n^{-1}(p)dp + 0.25H_n^{-1}(0.75)$$

Examples of GL-statistics

- **Interquartile range**: $h(x) = x$,

$$\text{IQR} = H_n^{-1}(0.75) - H_n^{-1}(0.25)$$

- **Variance**: $h(x, y) = \frac{1}{2}(x - y)^2$,

$$\int_0^1 H_n^{-1}(p)dp$$

- **Winsorized variance**: $h(x, y) = \frac{1}{2}(x - y)^2$,

$$\int_0^{0.75} H_n^{-1}(p)dp + 0.25H_n^{-1}(0.75)$$

- **Rousseeuw and Croux's** Q_n: $h(x, y) = |x - y|$,

$$H_n^{-1}(0.25)$$

Outline

Pairwise mean scale estimator: P_n

- Consider the set of $\binom{n}{2}$ pairwise means:

$$\{h(X_i, X_j), 1 \leq i < j \leq n\}$$

where $h(X_1, X_2) = (X_1 + X_2)/2$.

Pairwise mean scale estimator: P_n

- Consider the set of $\binom{n}{2}$ pairwise means:

$$\{h(X_i, X_j), 1 \leq i < j \leq n\}$$

  where $h(X_1, X_2) = (X_1 + X_2)/2$.
- Let H_n be the corresponding empirical distribution function:

$$H_n(t) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{I}\{h(X_i, X_j) \leq t\}, \quad \text{for } t \in \mathbb{R}.$$

Pairwise mean scale estimator: P_n

- Consider the set of $\binom{n}{2}$ pairwise means:

$$\{h(X_i, X_j), 1 \leq i < j \leq n\}$$

  where $h(X_1, X_2) = (X_1 + X_2)/2$.
- Let H_n be the corresponding empirical distribution function:

$$H_n(t) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{I}\{h(X_i, X_j) \leq t\}, \quad \text{for } t \in \mathbb{R}.$$

Definition

P_n is defined as

$$P_n = c\left[H_n^{-1}(0.75) - H_n^{-1}(0.25)\right],$$

where $c \approx 1.048$ is a correction factor to make P_n consistent for the standard deviation when the underlying observations are Gaussian.

Influence curve

- The influence curve for a functional T at distribution F is

$$\mathrm{IC}(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}$$

  where δ_x has all its mass at x.
- Serfling (1984) outlines the IC for GL-statistics.

Influence curve

- The influence curve for a functional T at distribution F is

$$\mathrm{IC}(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}$$

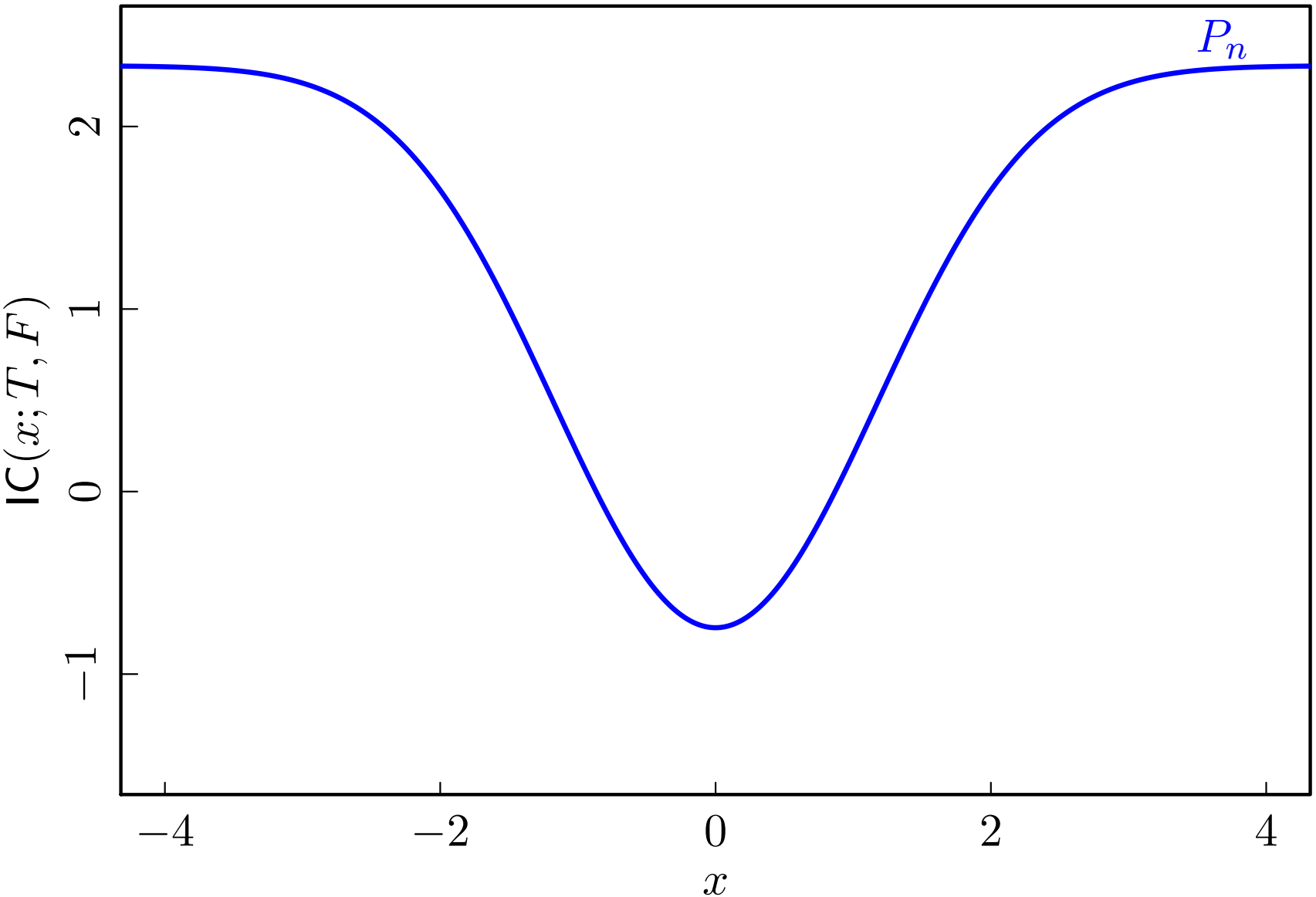
  where δ_x has all its mass at x.
- Serfling (1984) outlines the IC for GL-statistics.

Influence curve for P_n

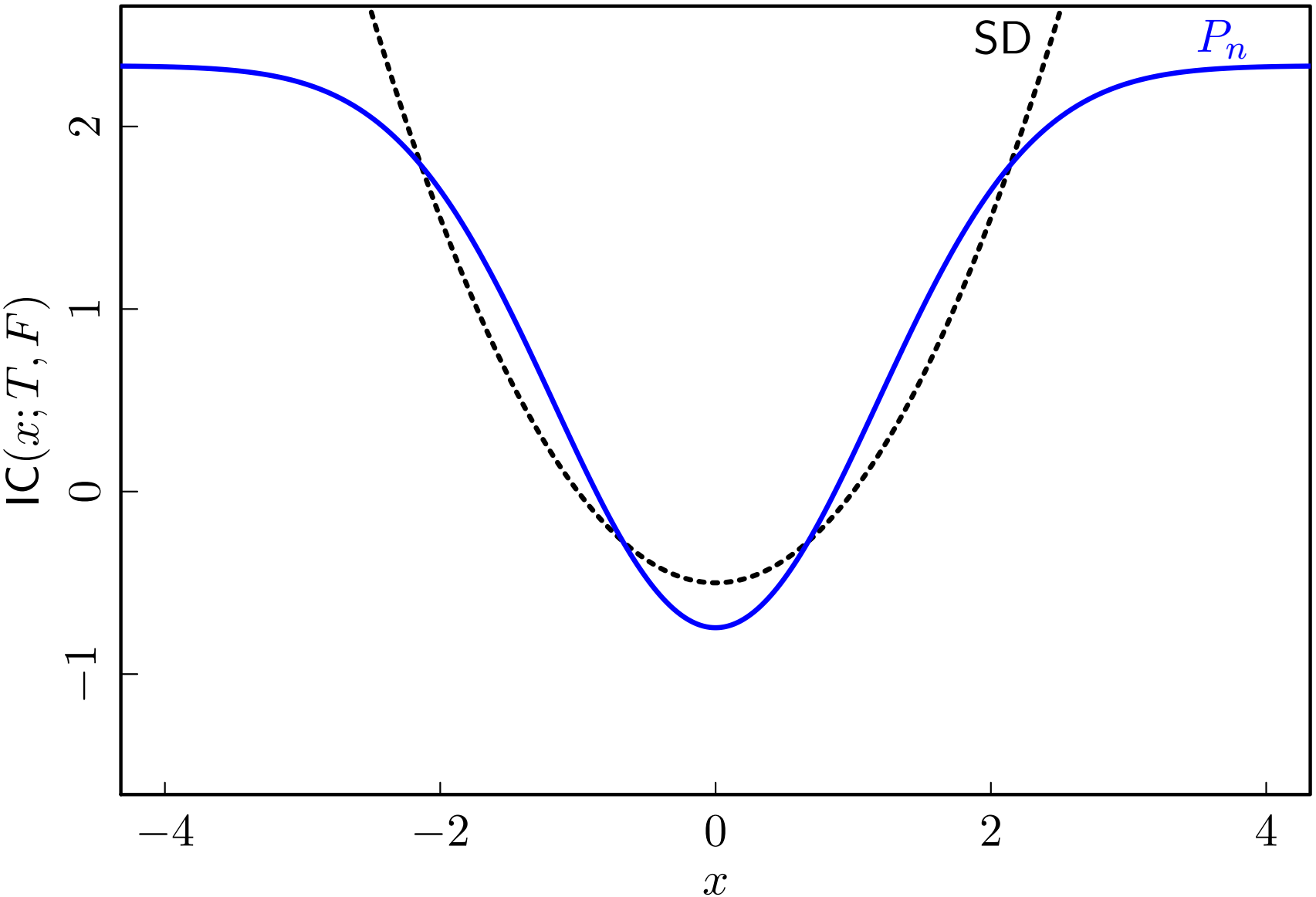
Assuming that F has derivative $f > 0$ on $[F^{-1}(\epsilon), F^{-1}(1-\epsilon)]$ for all $\epsilon > 0$,

$$\mathrm{IC}(x; P_n, F) = c \left[\frac{0.75 - F(2H_F^{-1}(0.75) - x)}{\int f(2H_F^{-1}(0.75) - x) f(x) dx} - \frac{0.25 - F(2H_F^{-1}(0.25) - x)}{\int f(2H_F^{-1}(0.25) - x) f(x) dx} \right].$$

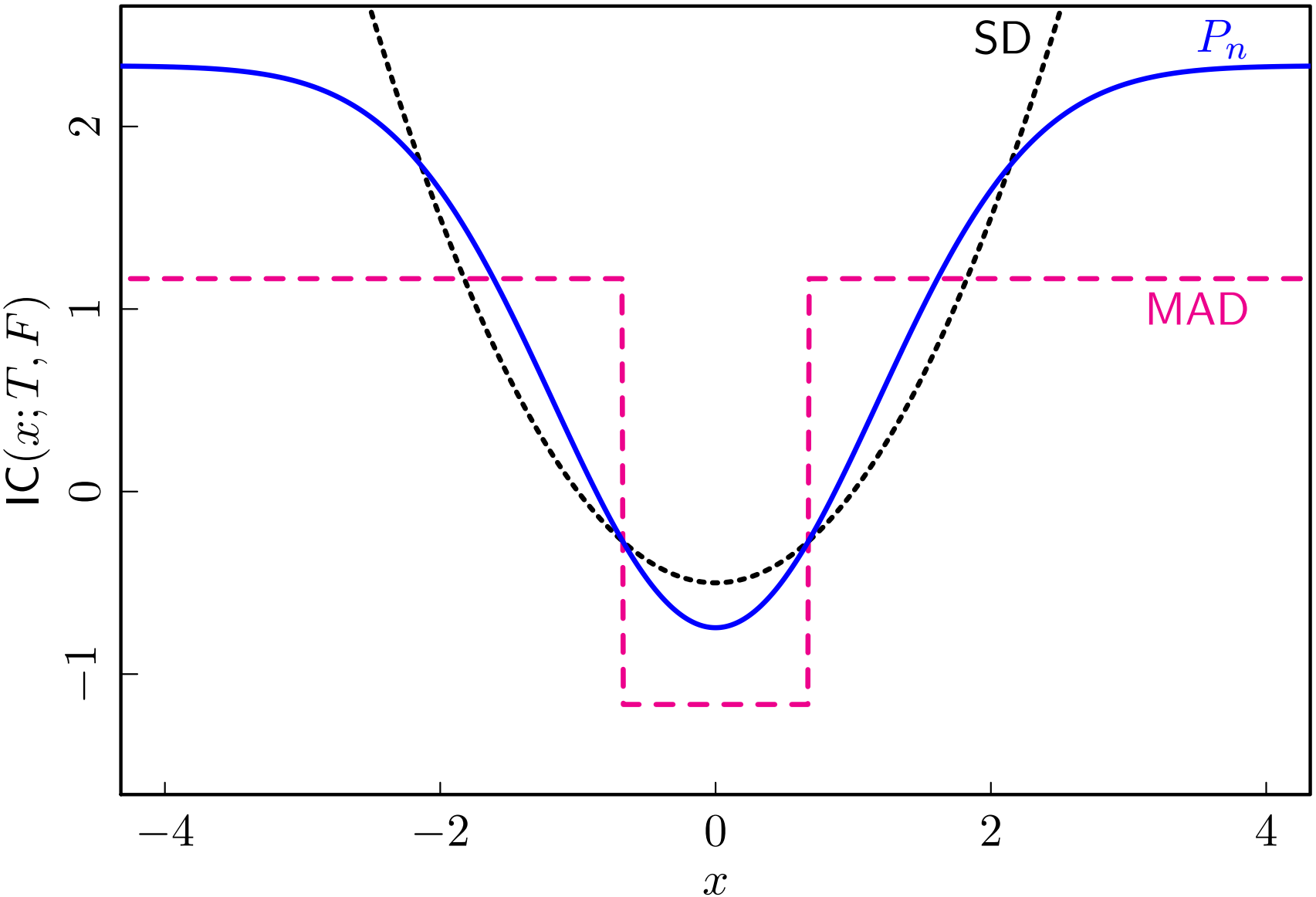
Influence curves when $F = \Phi$



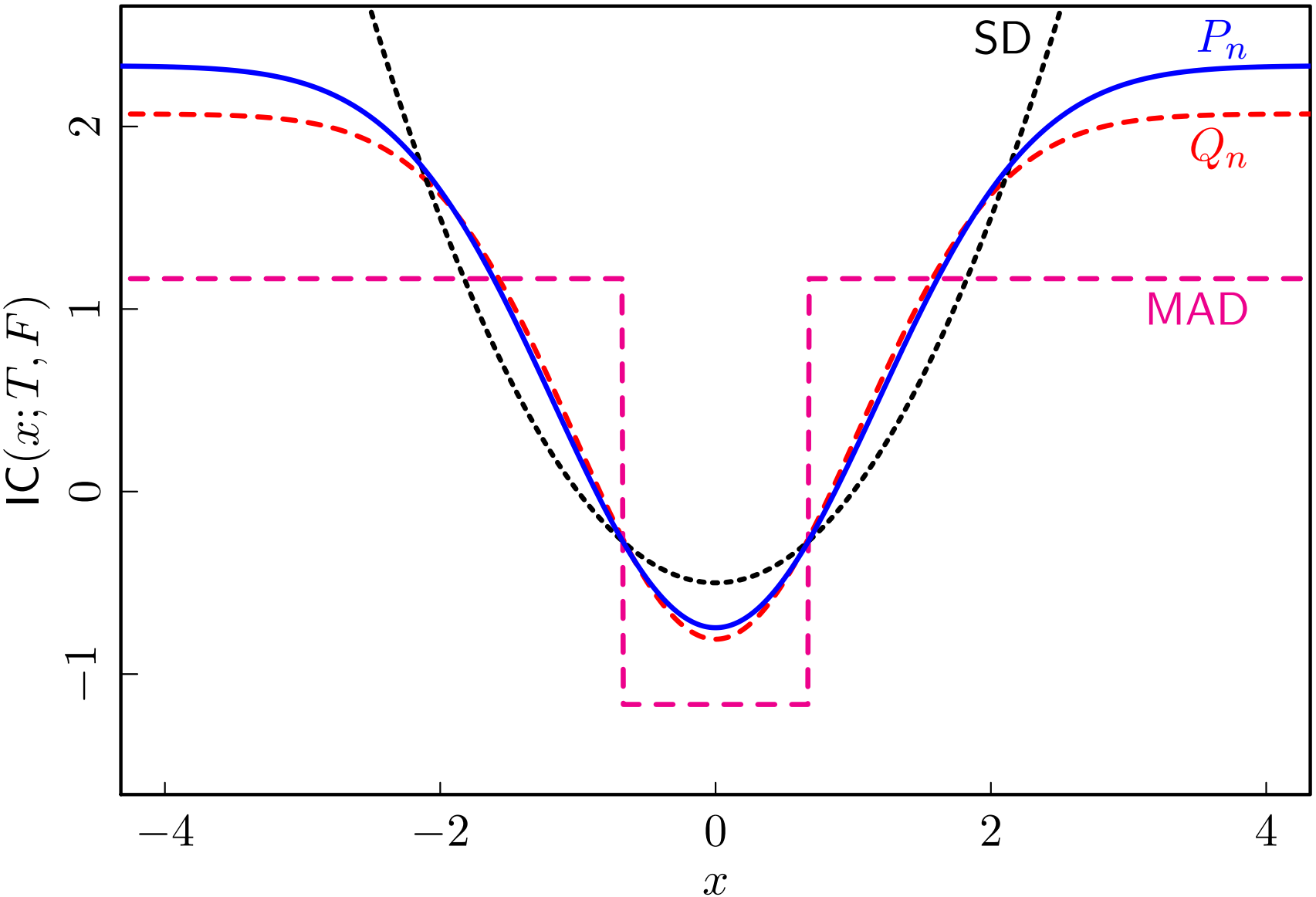
Influence curves when $F = \Phi$



Influence curves when $F = \Phi$



Influence curves when $F = \Phi$



Asymptotic normality

- The empirical U-process is

$$\left(\sqrt{n}\left(H_n(t) - H(t)\right)\right)_{t\in\mathbb{R}}.$$

Asymptotic normality

- The empirical U-process is

$$\left(\sqrt{n}\left(H_n(t) - H(t)\right)\right)_{t\in\mathbb{R}}.$$

- Silverman (1976) proved that in this context, $\sqrt{n}(H_n(\cdot) - H(\cdot))$ converges weakly to an almost sure continuous zero-mean Gaussian process W with covariance function:

$$\mathbb{E}W(s)W(t) = 4P(h(X_1, X_2) \leq s, h(X_1, X_3) \leq t) - 4H(s)H(t).$$

Asymptotic normality

- The empirical U-process is

$$\left(\sqrt{n}\left(H_n(t) - H(t)\right)\right)_{t\in\mathbb{R}}.$$

- Silverman (1976) proved that in this context, $\sqrt{n}(H_n(\cdot) - H(\cdot))$ converges weakly to an almost sure continuous zero-mean Gaussian process W with covariance function:

$$\mathbb{E}W(s)W(t) = 4P(h(X_1, X_2) \leq s, h(X_1, X_3) \leq t) - 4H(s)H(t).$$

- For $0 < p < q < 1$, if H', the derivative of H, is strictly positive on the interval $[H^{-1}(p) - \varepsilon, H^{-1}(q) + \varepsilon]$ for some $\varepsilon > 0$, then we can use the inverse map to show

$$\sqrt{n}\left(H_n^{-1}(\cdot) - H^{-1}(\cdot)\right) \xrightarrow{\mathcal{D}} \frac{W(H^{-1}(\cdot))}{H'(H^{-1}(\cdot))}.$$

Asymptotic normality

- Recall,

$$P_n = c\left[H_n^{-1}(3/4) - H_n^{-1}(1/4)\right].$$

Asymptotic normality

- Recall,

$$P_n = c\left[H_n^{-1}(3/4) - H_n^{-1}(1/4)\right].$$

- Hence,

$$\sqrt{n}(P_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

where

$$\theta = c\left(H^{-1}(3/4) - H^{-1}(1/4)\right)$$

and V both depend on the underlying distribution.

Asymptotic variance and relative efficiency

- The asymptotic variance follows from the previous limit theorem or from the expected square of the influence function.

Asymptotic variance and relative efficiency

- The asymptotic variance follows from the previous limit theorem or from the expected square of the influence function.
- When the underlying data are Gaussian, numerical integration yields,

$$V = \int \mathrm{IC}(x, P_n, \Phi)^2 d\Phi(x) = 0.579.$$

Asymptotic variance and relative efficiency

- The asymptotic variance follows from the previous limit theorem or from the expected square of the influence function.
- When the underlying data are Gaussian, numerical integration yields,

$$V = \int \mathrm{IC}(x, P_n, \Phi)^2 d\Phi(x) = 0.579.$$

- This equates to an asymptotic efficiency of **0.86** as compared with **0.82** for Q_n and **0.37** for the MAD at the **normal**.

Asymptotic variance and relative efficiency

- The asymptotic variance follows from the previous limit theorem or from the expected square of the influence function.
- When the underlying data are Gaussian, numerical integration yields,

$$V = \int \mathrm{IC}(x, P_n, \Phi)^2 d\Phi(x) = 0.579.$$

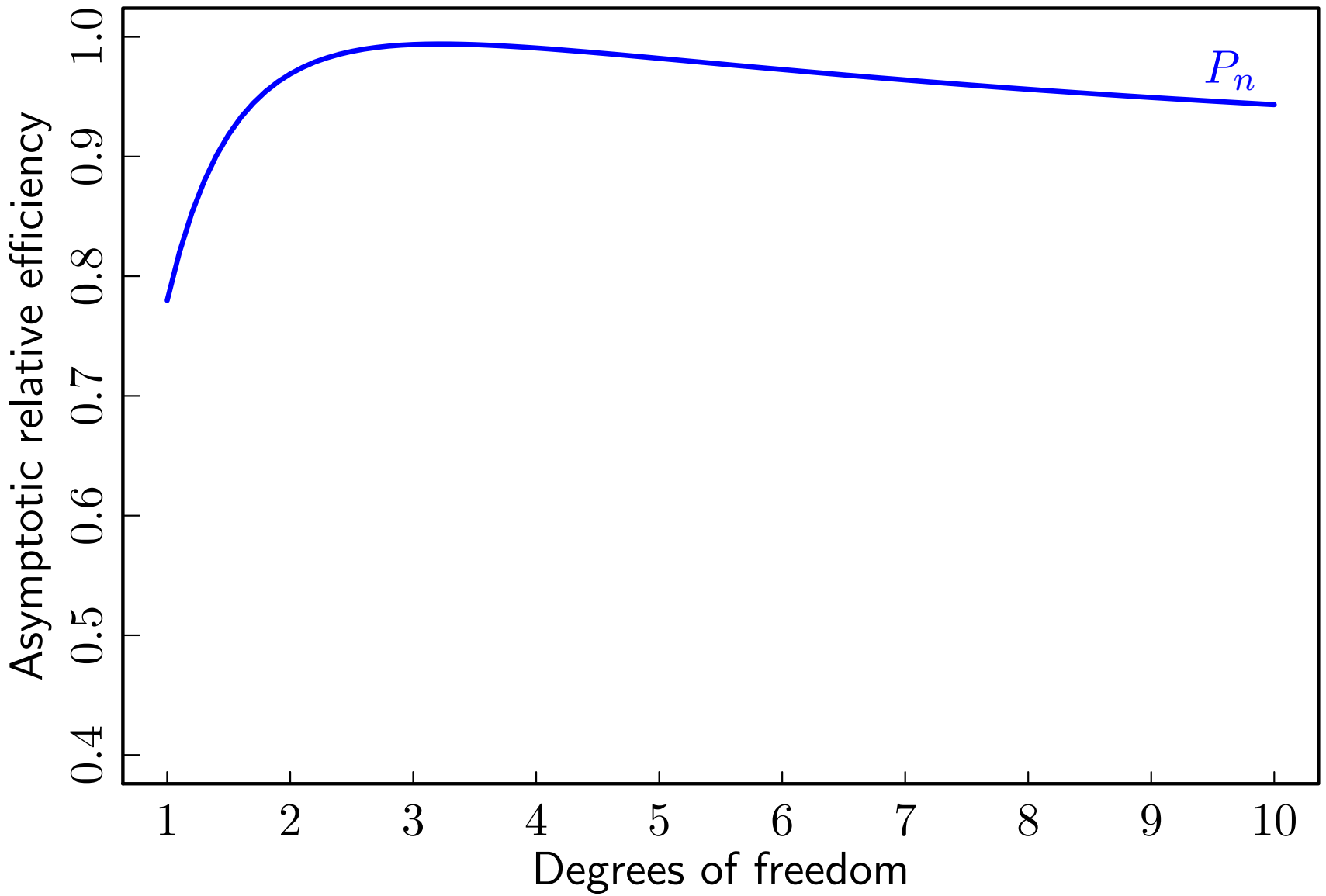
- This equates to an asymptotic efficiency of **0.86** as compared with **0.82** for Q_n and **0.37** for the MAD at the **normal**.

Result

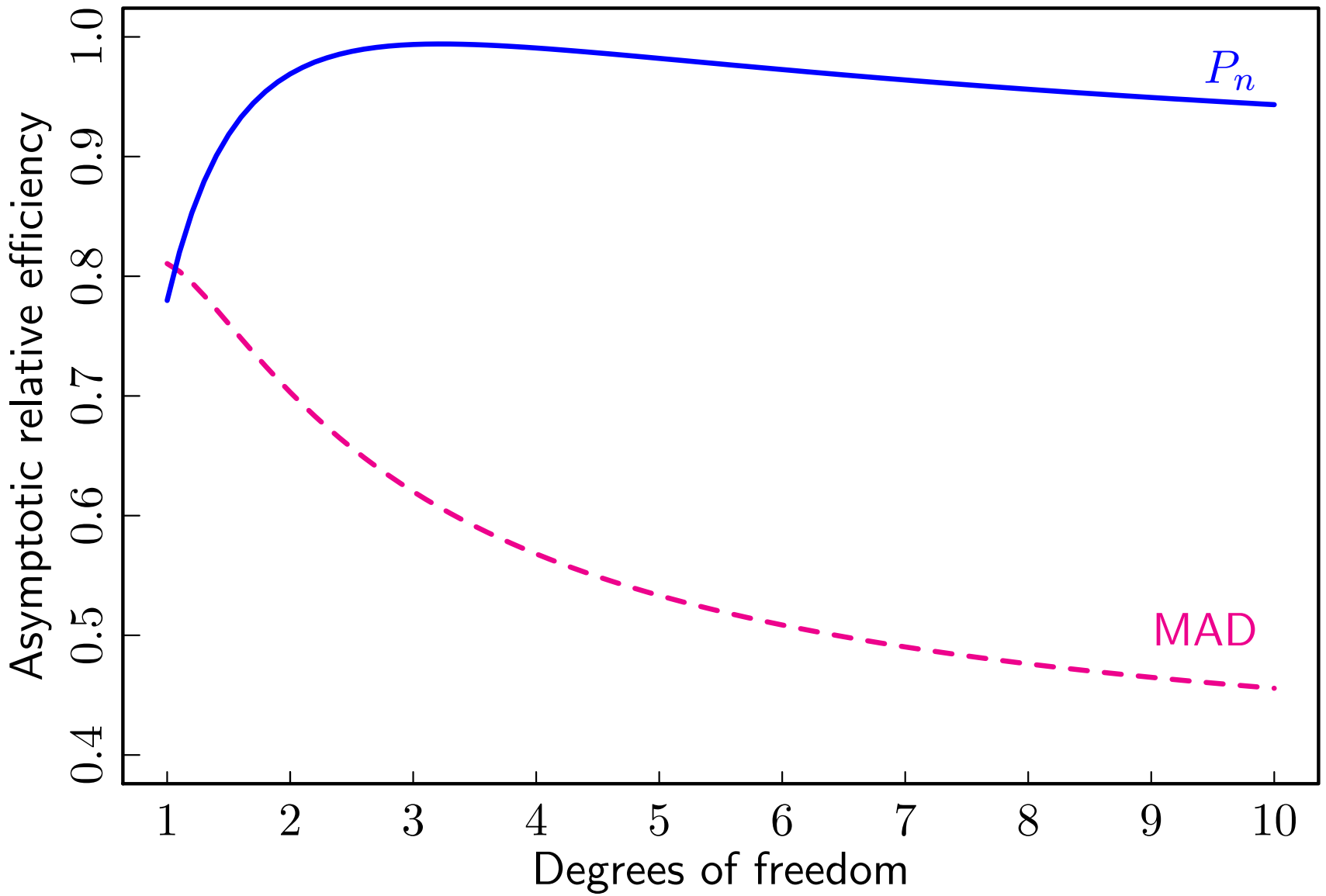
We have a robust estimator that is asymptotically more efficient than Q_n at the normal.

But how does it compare at heavier tailed distributions?

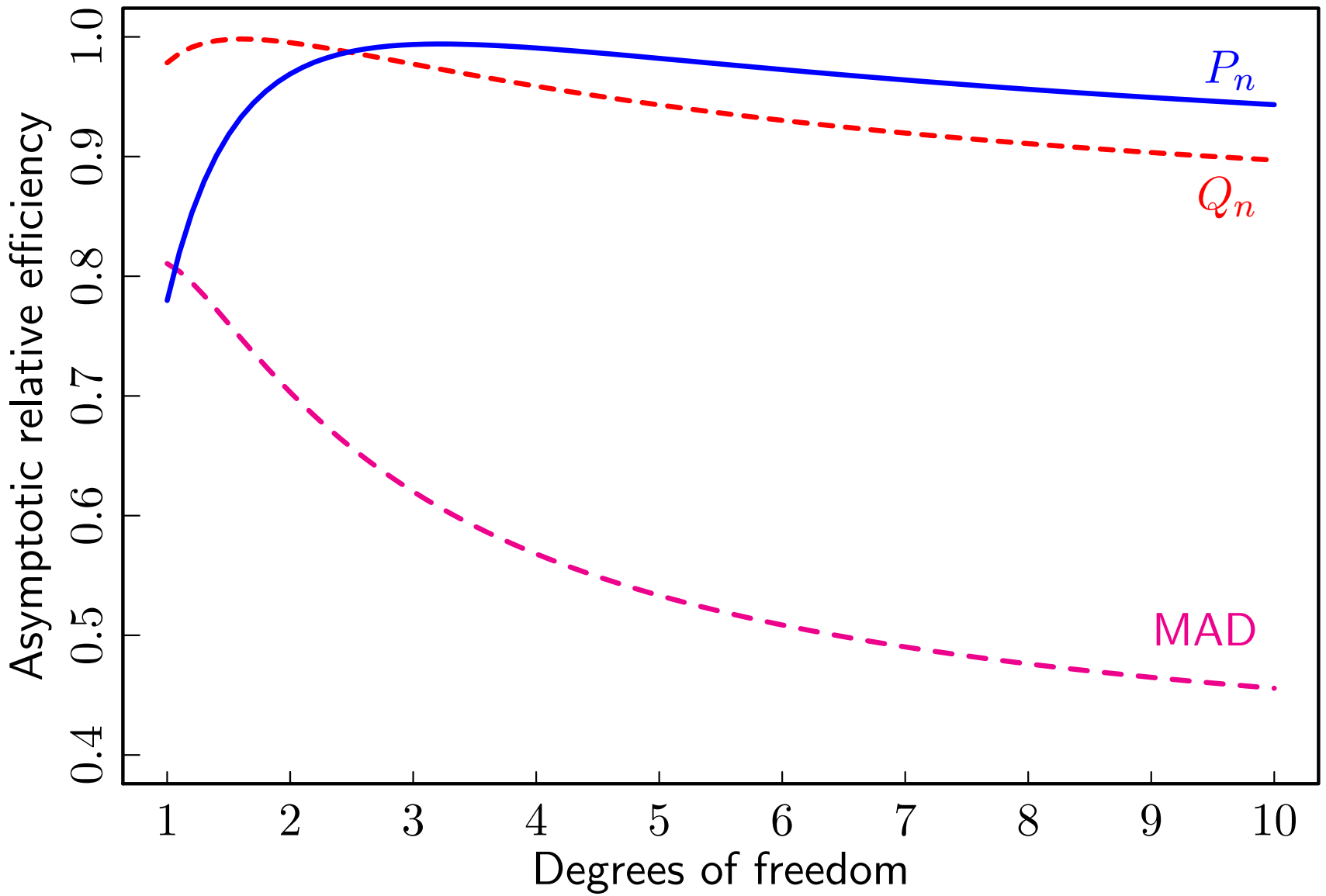
Asymptotic relative efficiency when $f = t_\nu$ for $\nu \in [1, 10]$



Asymptotic relative efficiency when $f = t_\nu$ for $\nu \in [1, 10]$



Asymptotic relative efficiency when $f = t_\nu$ for $\nu \in [1, 10]$



Discrete distributions

- For $P_n = 0$, the interquartile range of the pairwise means must be equal to zero.

Discrete distributions

- For $P_n = 0$, the interquartile range of the pairwise means must be equal to zero.
- I.e. more than 50% of the pairwise means must be equal.

Discrete distributions

- For $P_n = 0$, the interquartile range of the pairwise means must be equal to zero.
- I.e. more than 50% of the pairwise means must be equal.
- For a discrete random variable with possible outcomes, $x_1, x_2, \ldots$, and $\mathrm{P}(X = x_j) = p_j$, for $j = 1, 2, \ldots$, the expected proportion of **pairwise differences** equal to zero is $\sum_j p_j^2$.
- In particular,

$$\sum_j p_j^2 > 0.25 \iff \lim_{n\to\infty} \mathrm{P}(Q_n = 0) = 1.$$

Discrete distributions

- For $P_n = 0$, the interquartile range of the pairwise means must be equal to zero.
- I.e. more than 50% of the pairwise means must be equal.
- For a discrete random variable with possible outcomes, $x_1, x_2, \ldots$, and $\mathrm{P}(X = x_j) = p_j$, for $j = 1, 2, \ldots$, the expected proportion of **pairwise differences** equal to zero is $\sum_j p_j^2$.
- In particular,

$$\sum_j p_j^2 > 0.25 \iff \lim_{n\to\infty} \mathrm{P}(Q_n = 0) = 1.$$

Example ($X \sim \mathrm{Poisson}(1)$)

For the Poisson(1), $\sum_j p_j^2 \approx 0.31$ and so in the limit, $Q_n = 0$ with high probability, whereas $c^{-1}P_n = 1$.

Discrete distributions

- In finite samples, apart from some trivial cases,

$$P_n = 0 \implies Q_n = 0.$$

Discrete distributions

- In finite samples, apart from some trivial cases,

$$P_n = 0 \implies Q_n = 0.$$

Example ($X \sim \text{Binomial}(6, 0.4)$)

- In the limit neither Q_n nor P_n will converge to zero.
- In samples of size $n = 20$, Q_n will return a scale estimate of zero, on average **12%** of the time.
- In contrast, P_n returns zero less than **0.1%** of the time.

Adaptive trimming: $\tilde{P}_n$

(Potential) Cons with P_n

- P_n has a breakdown value of 13%.
- P_n is not very efficient at the Cauchy.

Adaptive trimming: $\tilde{P}_n$

(Potential) Cons with P_n

- P_n has a breakdown value of 13%.
- P_n is not very efficient at the Cauchy.

- For preliminary high breakdown location and scale estimates, $m(\mathbf{X})$ and $s(\mathbf{X})$ respectively, an observation, X_i, is trimmed if

$$\frac{|X_i - m(\mathbf{X})|}{s(\mathbf{X})} > d, \tag{4}$$

where d is the tuning parameter.

Adaptive trimming: $\tilde{P}_n$

(Potential) Cons with P_n

- P_n has a breakdown value of 13%.
- P_n is not very efficient at the Cauchy.

- For preliminary high breakdown location and scale estimates, $m(\mathbf{X})$ and $s(\mathbf{X})$ respectively, an observation, X_i, is trimmed if

$$\frac{|X_i - m(\mathbf{X})|}{s(\mathbf{X})} > d, \tag{4}$$

  where d is the tuning parameter.
- Achieves the best possible breakdown value for a sensible choice of tuning parameter.

Adaptive trimming: $\tilde{P}_n$

(Potential) Cons with P_n

- P_n has a breakdown value of 13%.
- P_n is not very efficient at the Cauchy.

- For preliminary high breakdown location and scale estimates, $m(\mathbf{X})$ and $s(\mathbf{X})$ respectively, an observation, X_i, is trimmed if

$$\frac{|X_i - m(\mathbf{X})|}{s(\mathbf{X})} > d, \tag{4}$$

  where d is the tuning parameter.
- Achieves the best possible breakdown value for a sensible choice of tuning parameter.

Definition

Denote $\tilde{P}_n$ as the adaptively trimmed P_n with $d = 5$.

Outline

Efficiency of P_n in finite samples

Following Randal (2008) efficiencies are estimated over m independent samples as

$$\widehat{\text{eff}}(T) = \frac{\widehat{\text{Var}}\left(\ln \hat{\sigma}_1, \ldots, \ln \hat{\sigma}_m\right)}{\widehat{\text{Var}}\left(\ln T(\mathbf{X}_1), \ldots, \ln T(\mathbf{X}_m)\right)}. \tag{5}$$

Efficiency of P_n in finite samples

Following Randal (2008) efficiencies are estimated over m independent samples as

$$\widehat{\text{eff}}(T) = \frac{\widehat{\text{Var}}\left(\ln \hat{\sigma}_1, \ldots, \ln \hat{\sigma}_m\right)}{\widehat{\text{Var}}\left(\ln T(\mathbf{X}_1), \ldots, \ln T(\mathbf{X}_m)\right)}. \tag{5}$$

For each $i = 1, 2, \ldots, m$,

- $\mathbf{X}_i$ are independent samples of size n,
- $\hat{\sigma}_i$ is the ML scale estimate, and
- $T(\mathbf{X}_i)$ is the proposed scale estimate.

Efficiency of P_n in finite samples

Following Randal (2008) efficiencies are estimated over m independent samples as

$$\widehat{\text{eff}}(T) = \frac{\widehat{\text{Var}}\left(\ln \hat{\sigma}_1, \ldots, \ln \hat{\sigma}_m\right)}{\widehat{\text{Var}}\left(\ln T(\mathbf{X}_1), \ldots, \ln T(\mathbf{X}_m)\right)}. \tag{5}$$

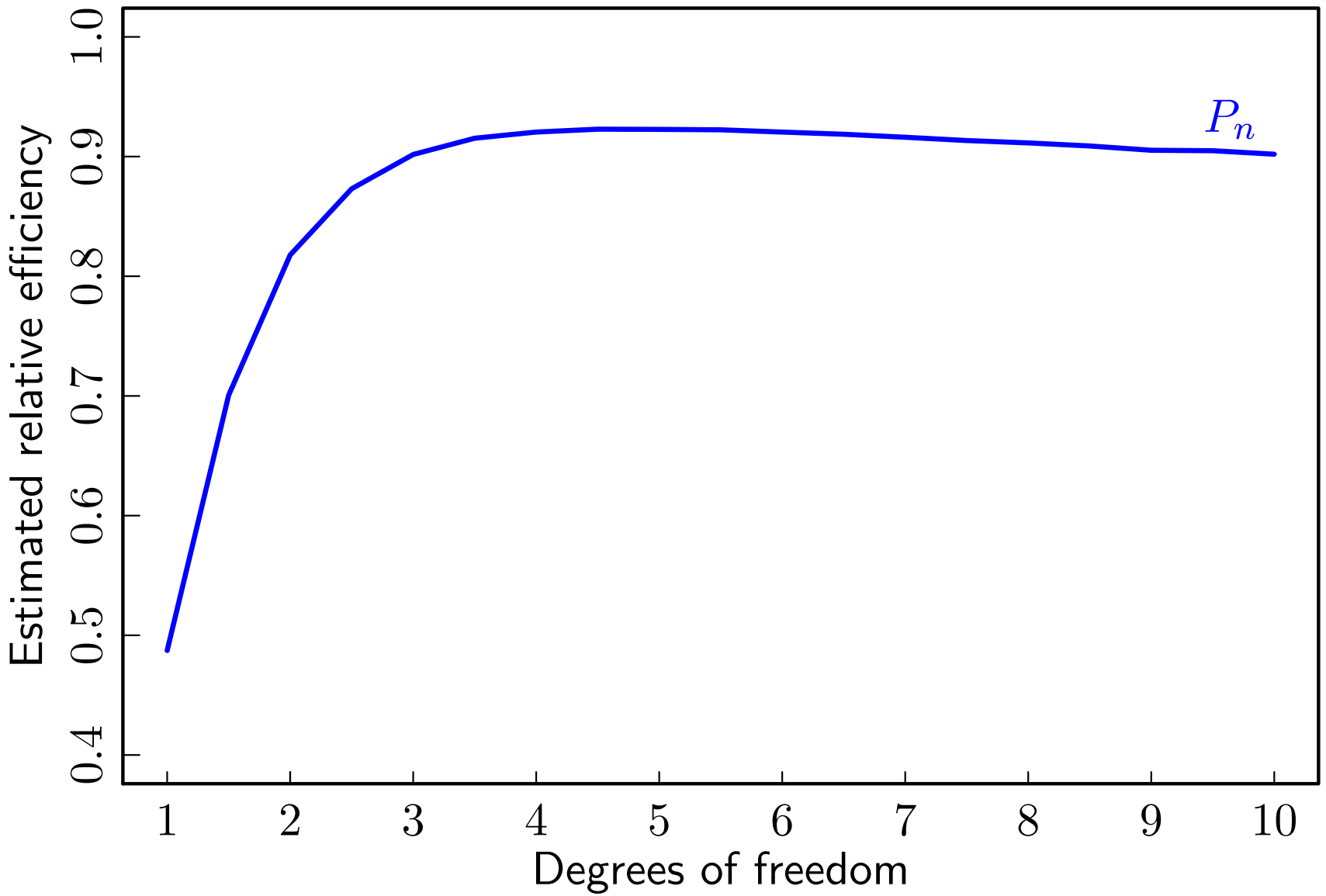
For each $i = 1, 2, \ldots, m$,

- $\mathbf{X}_i$ are independent samples of size n,
- $\hat{\sigma}_i$ is the ML scale estimate, and
- $T(\mathbf{X}_i)$ is the proposed scale estimate.

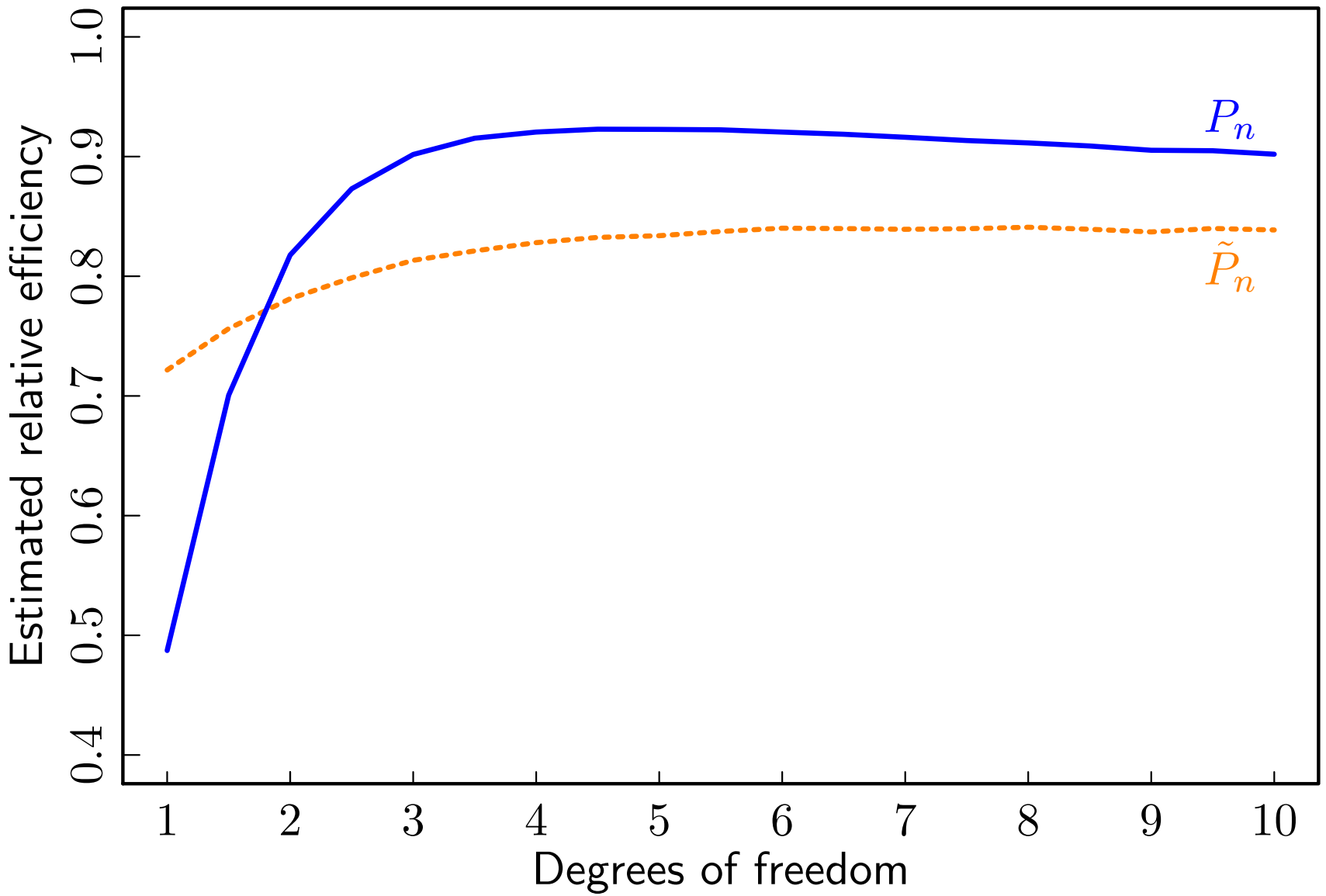
Distributions considered

- t distributions with degrees of freedom between 1 and 10.
- Configural polysampling using Tukey's 3 corners: **Gaussian**, **One-wild** and **Slash**.

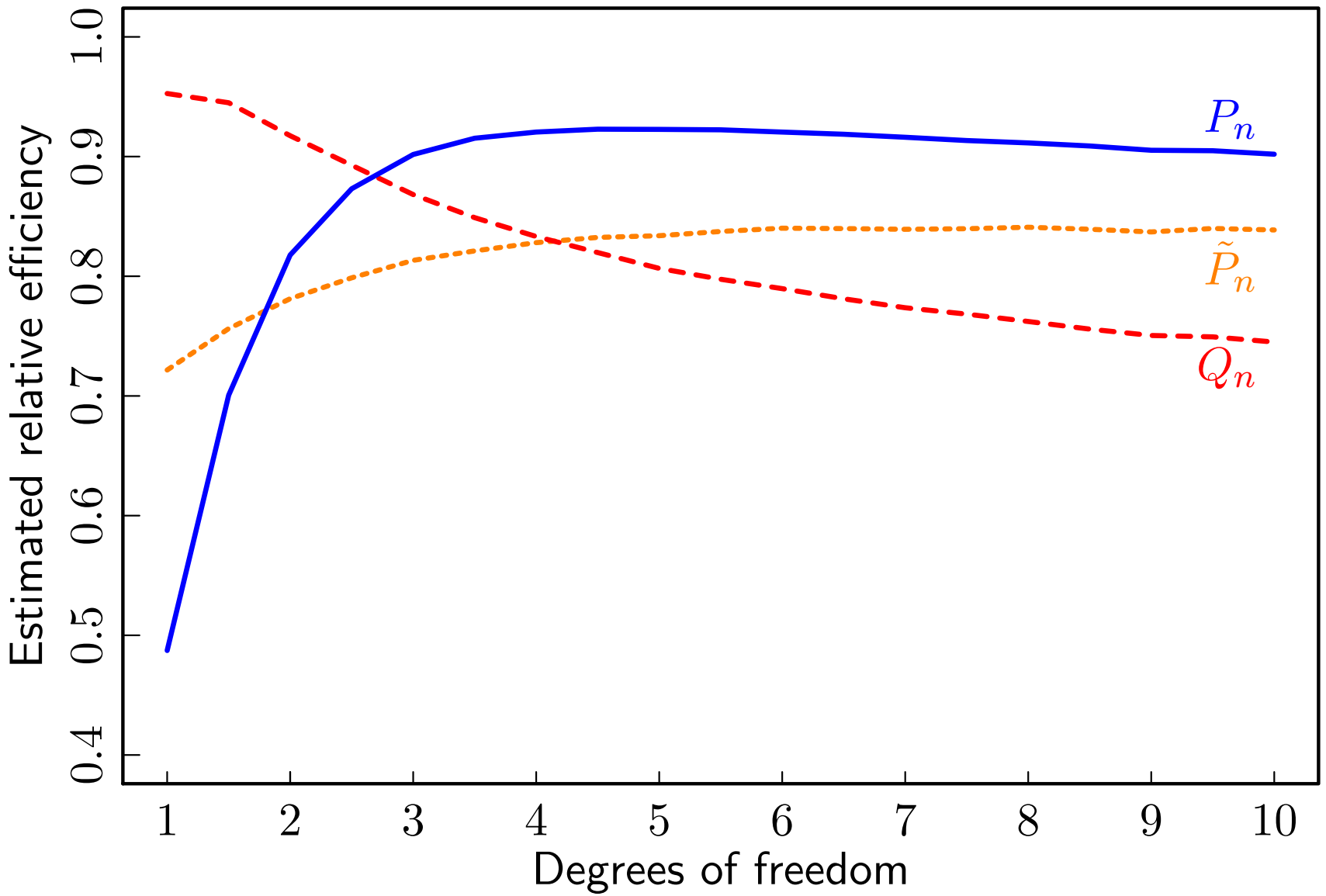
Relative efficiencies: $f = t_\nu$ for $\nu \in [1, 10]$ and $n = 20$



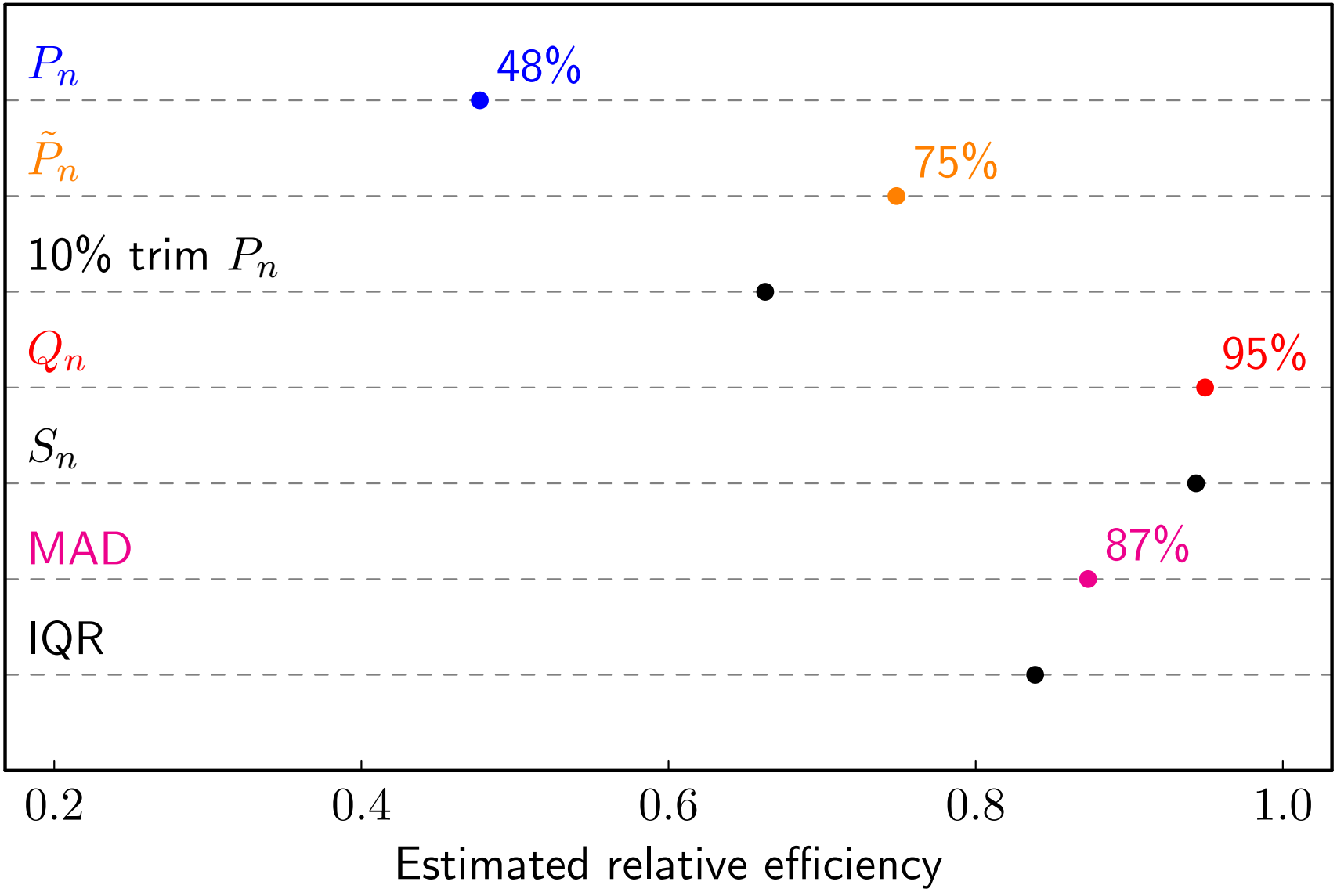
Relative efficiencies: $f = t_\nu$ for $\nu \in [1, 10]$ and $n = 20$



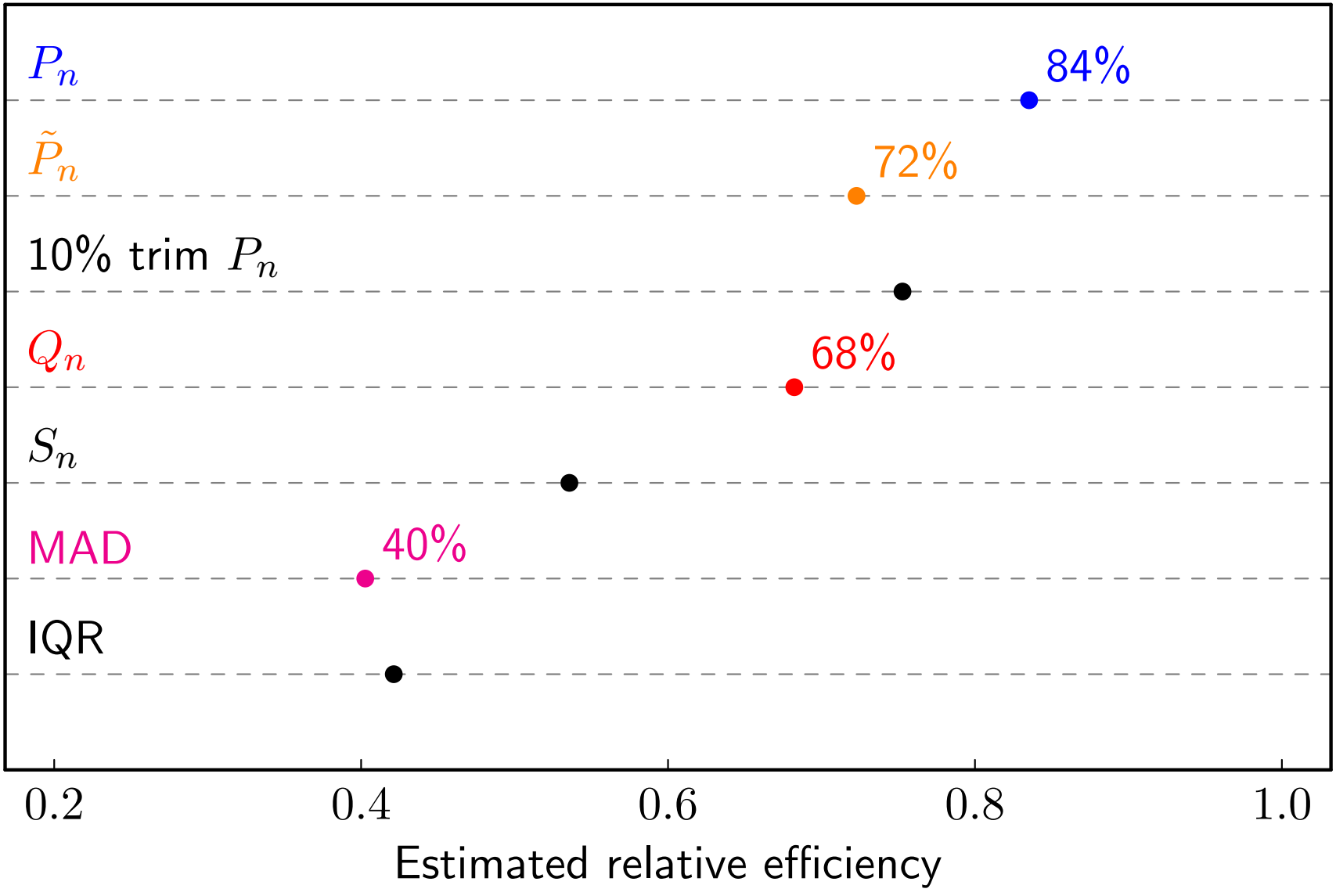
Relative efficiencies: $f = t_\nu$ for $\nu \in [1, 10]$ and $n = 20$



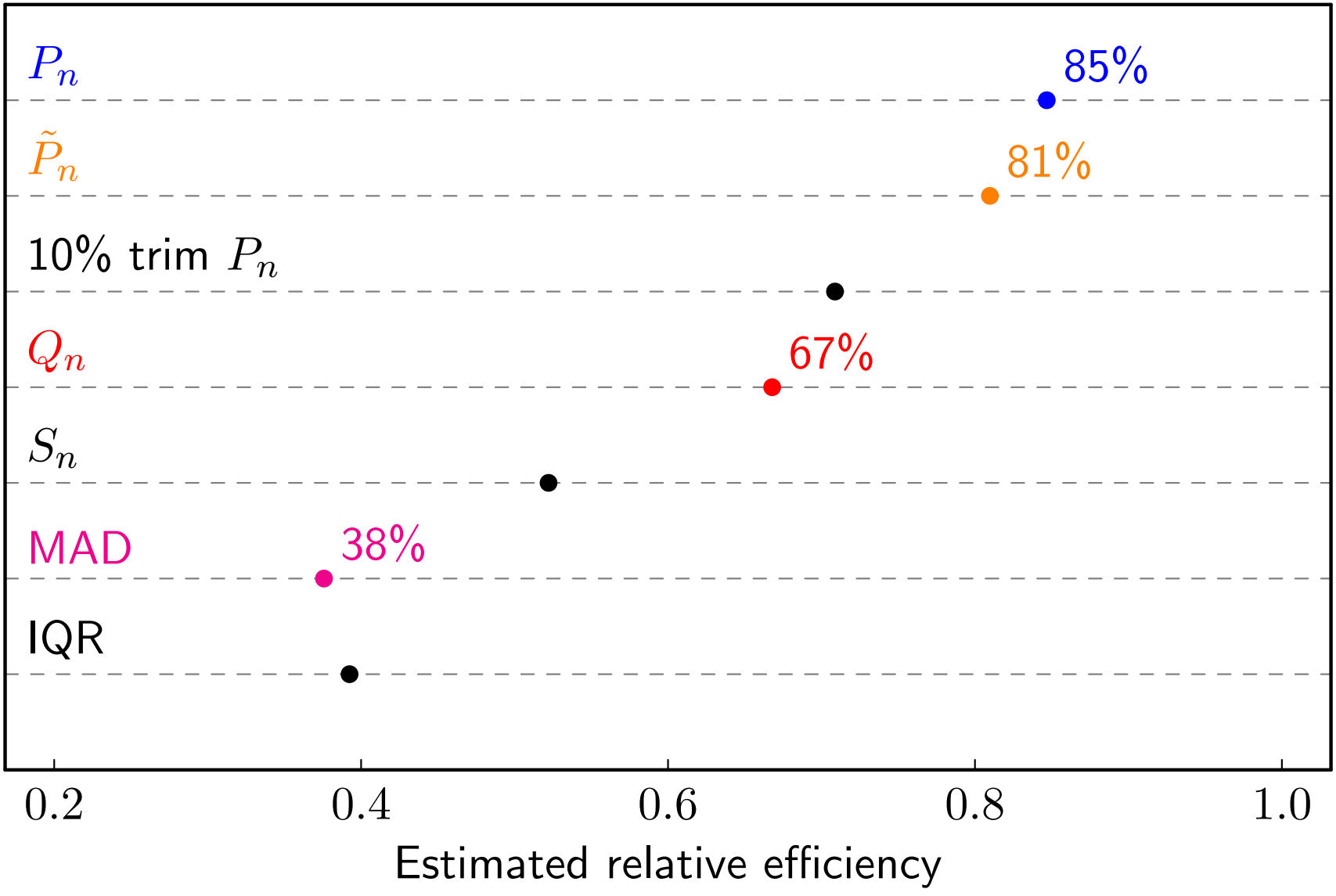
Relative efficiencies at the Slash corner ($n = 20$)



Relative efficiencies at the One-wild corner ($n = 20$)



Relative efficiencies at the Gaussian corner ($n = 20$)



Outline

Summary

1. Aim

- An efficient, robust and widely applicable scale estimator.

Summary

1. Aim

- An efficient, robust and widely applicable scale estimator.

2. Method

- P_n scale estimator – a GL-statistic.

Summary

1. Aim

- An efficient, robust and widely applicable scale estimator.

2. Method

- P_n scale estimator – a GL-statistic.

3. Results

- 86% asymptotic efficiency at the normal.
- Highly efficient even when the underlying distribution has quite heavy tails.
- Less likely to fail for discrete distributions than Q_n.

References

Hampel, F. (1974).
The influence curve and its role in robust estimation.
Journal of the American Statistical Association, 69(346):383–393.

Randal, J. (2008).
A reinvestigation of robust scale estimation in finite samples.
Computational Statistics & Data Analysis, 52(11):5014–5021.

Rousseeuw, P. and Croux, C. (1993).
Alternatives to the median absolute deviation.
Journal of the American Statistical Association, 88(424):1273–1283.

Serfling, R. J. (1984).
Generalized L-, M-, and R-statistics.
The Annals of Statistics, 12(1):76–86.

Silverman, B. (1976).
Limit theorems for dissociated random variables.
Advances in Applied Probability, 8(4):806–819.