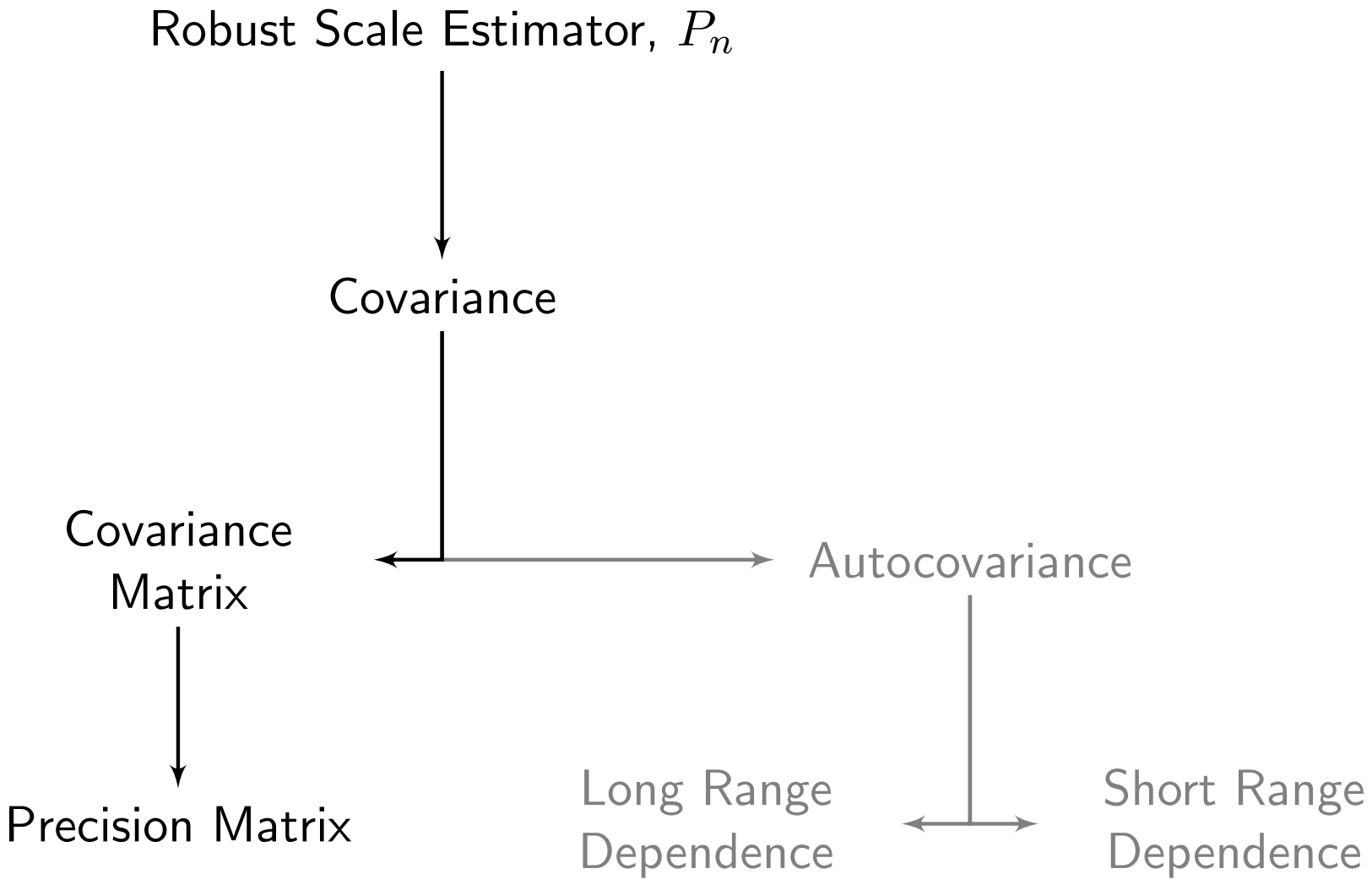
Robust estimation of scale and covariance with P_n and its application to precision matrix estimation

Garth Tarr, Samuel Müller and Neville Weber

USYD 2013

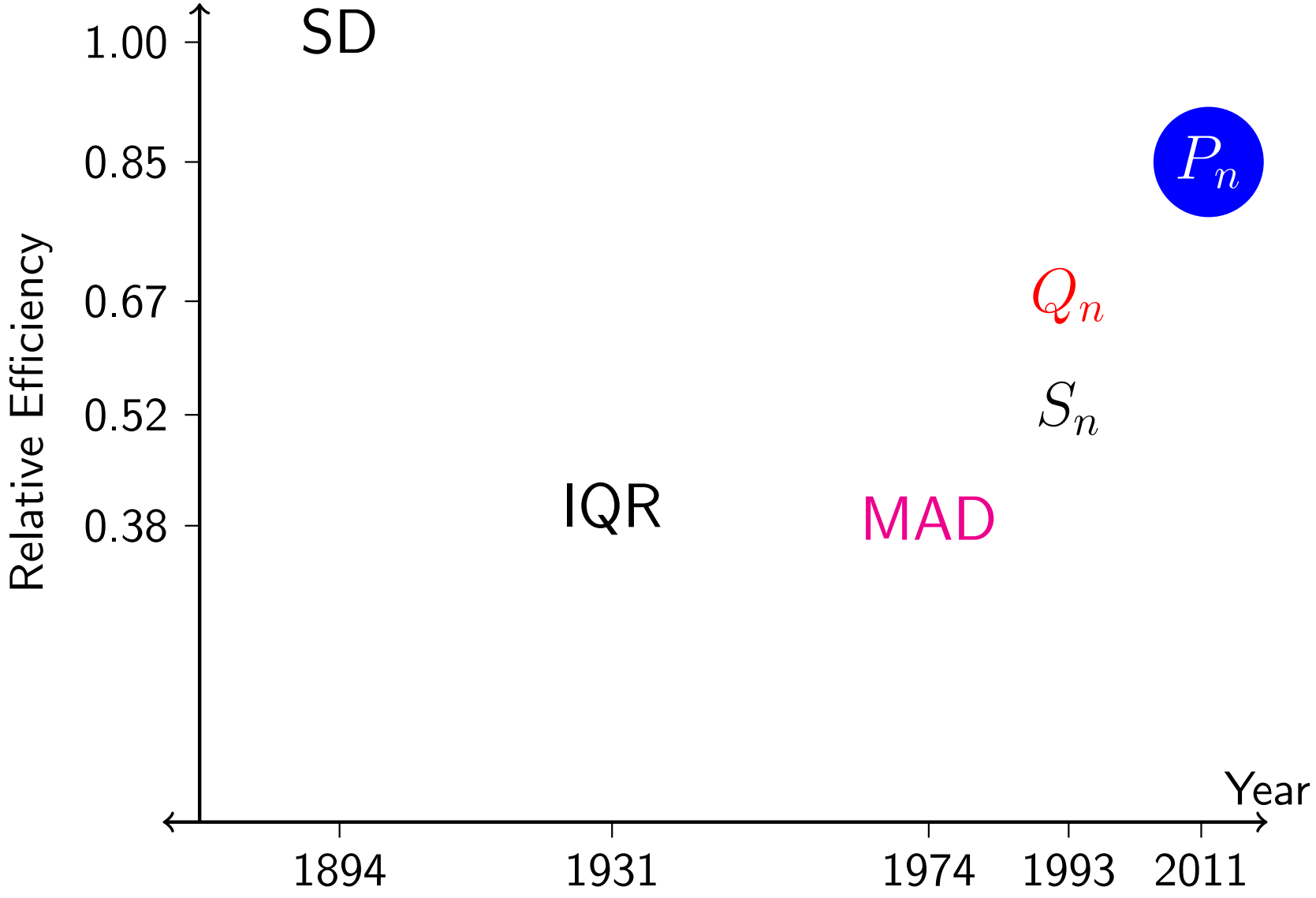
School of Mathematics and Statistics
THE UNIVERSITY OF SYDNEY

Outline


Robust Scale Estimator, P_n
Covariance
Covariance Matrix
Precision Matrix
Autocovariance
Long Range Dependence
Short Range Dependence

Outline

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:

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

- 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:

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

Generalised L-statistics

A generalised 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**: $h(x) = x$,

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

- **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: $h(x, y) = |x - y|$,

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

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:

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

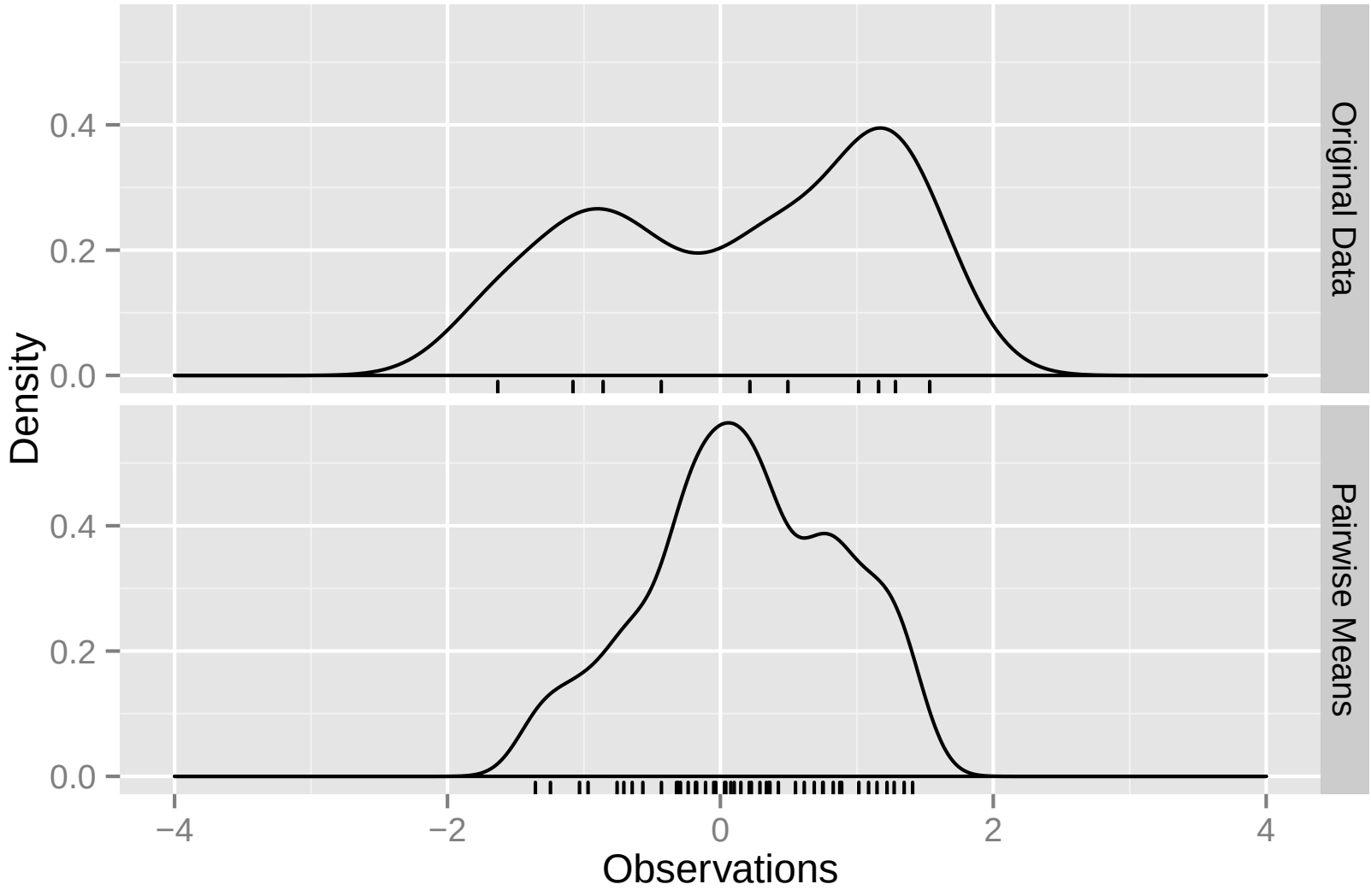
Definition

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

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

Why pairwise means?

Consider 10 observations from $\mathcal{N}(0, 1)$.



Influence curve

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

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

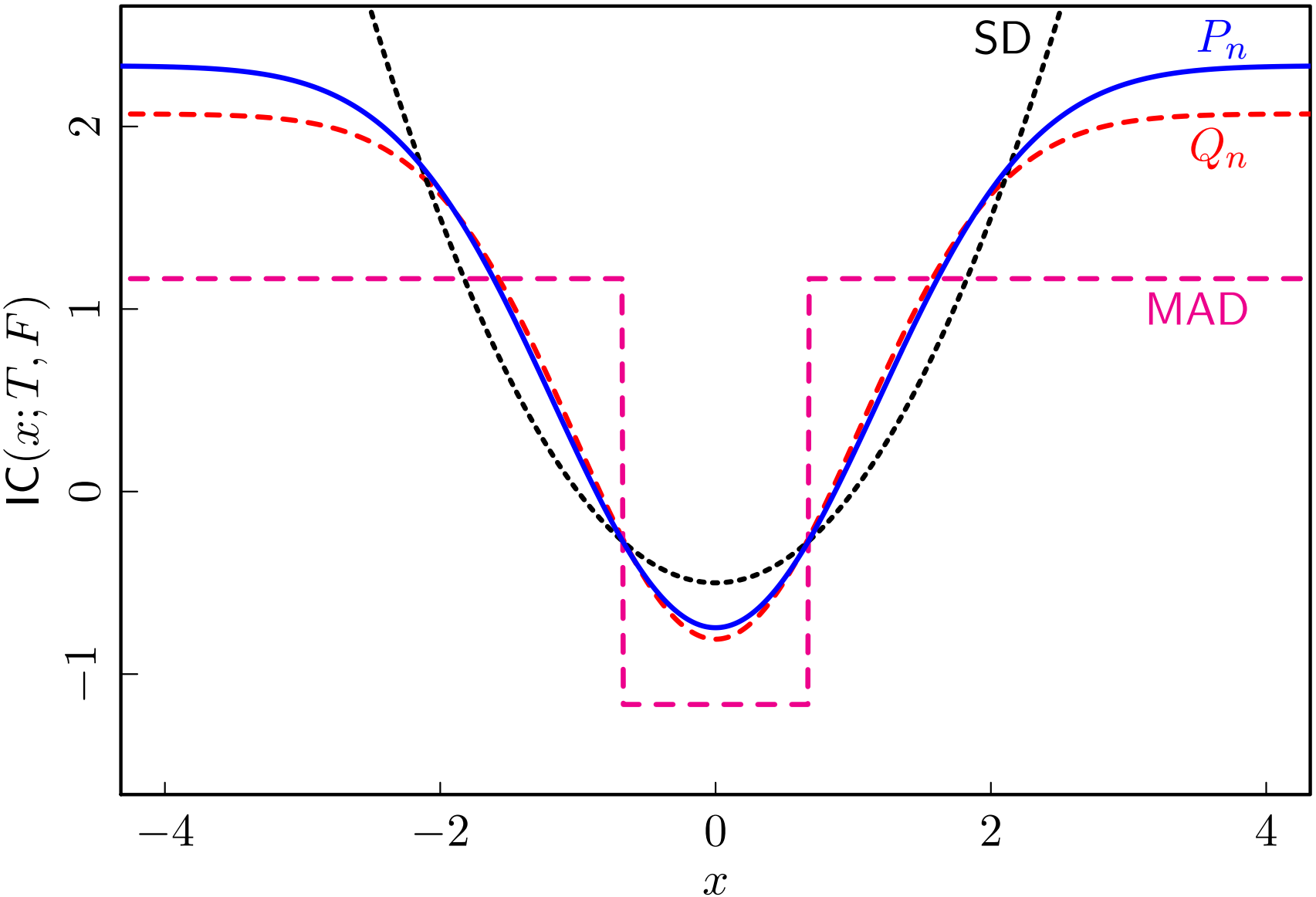
  where δ_x has all its mass at x.
- Serfling (1984) outlines the IC for GL-statistics.

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 $F = \Phi$



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:

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

- 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,
$$P_n = c\left[H_n^{-1}(3/4) - H_n^{-1}(1/4)\right].$$
- 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.
- When the underlying data are Gaussian,
$$V = \int \mathrm{IC}(x, P_n, \Phi)^2 \,\mathrm{d}\,\Phi(x) = 0.579.$$
- 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**.

Breakdown value

Definition

The breakdown value is the smallest fraction of contamination that can cause arbitrarily corruption to an estimator.

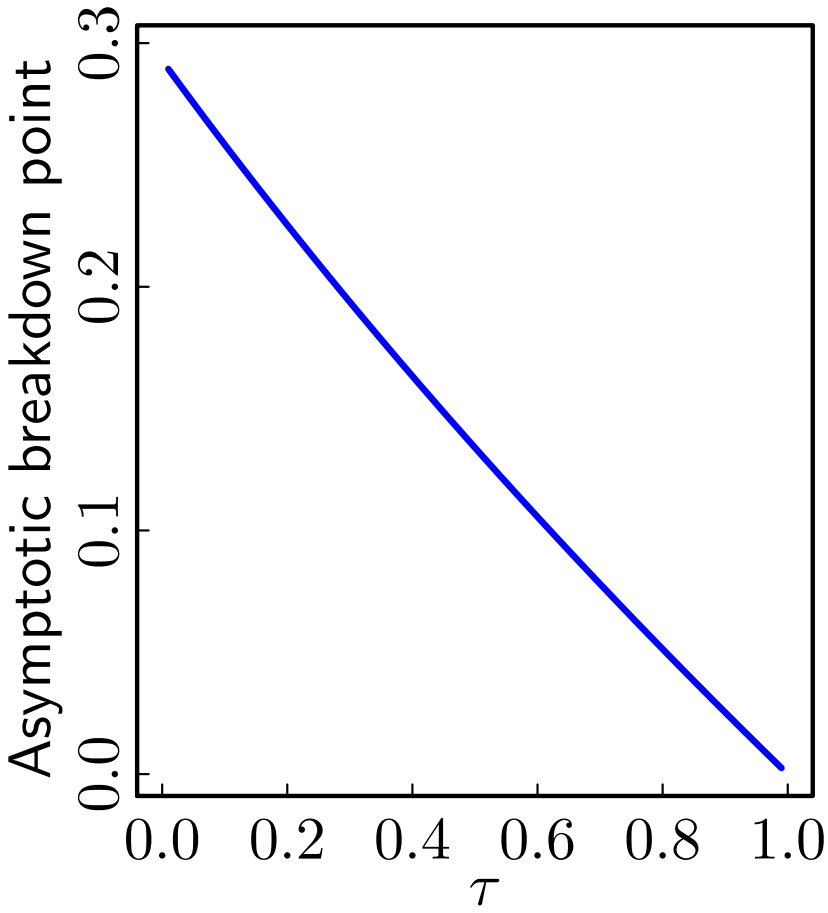
- P_n will breakdown if 25% of the pairwise means are contaminated.
- Arbitrarily changing m of the original observations leaves $n - m$ uncontaminated and $\binom{n-m}{2}$ pairwise means remain uncontaminated.
- P_n will remain bounded so long as more than 75% of the pairwise means are uncontaminated, i.e.

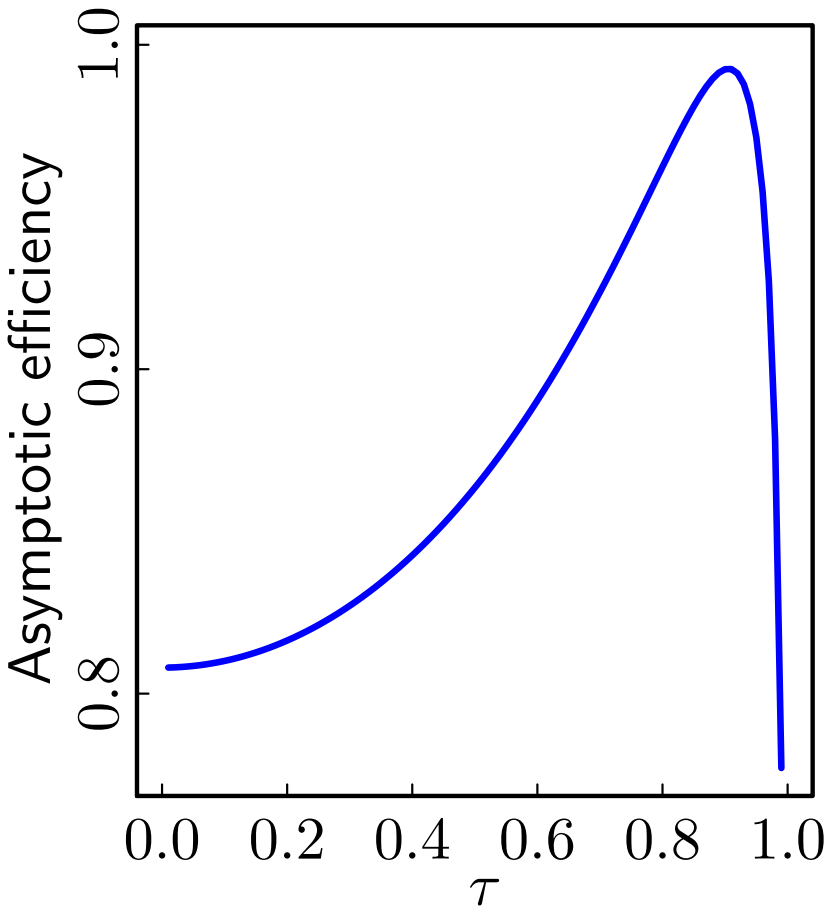
$$\binom{n-m}{2} > 0.75\binom{n}{2}.$$

- The asymptotic breakdown value of P_n is 13.4%.

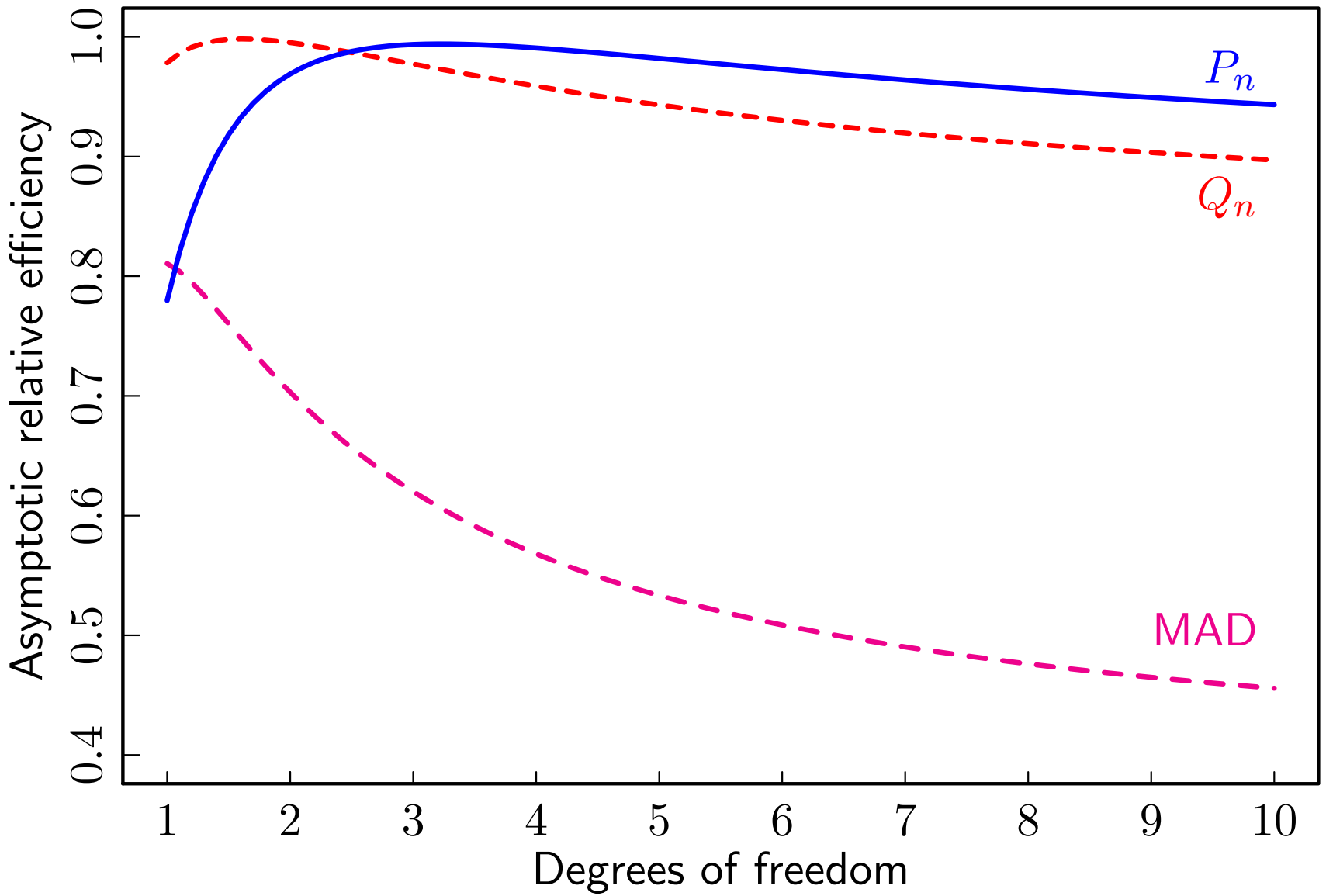
Why the interquartile range?

- Consider, $P_n(\tau) = H_n^{-1}\left(\frac{1+\tau}{2}\right) - H_n^{-1}\left(\frac{1-\tau}{2}\right)$.
- The asymptotic breakdown point is: $\varepsilon^* \approx 1 - \sqrt{\frac{1+\tau}{2}}$.





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



Adaptive trimming: $\tilde{P}_n$

- 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 $d = 5$.

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

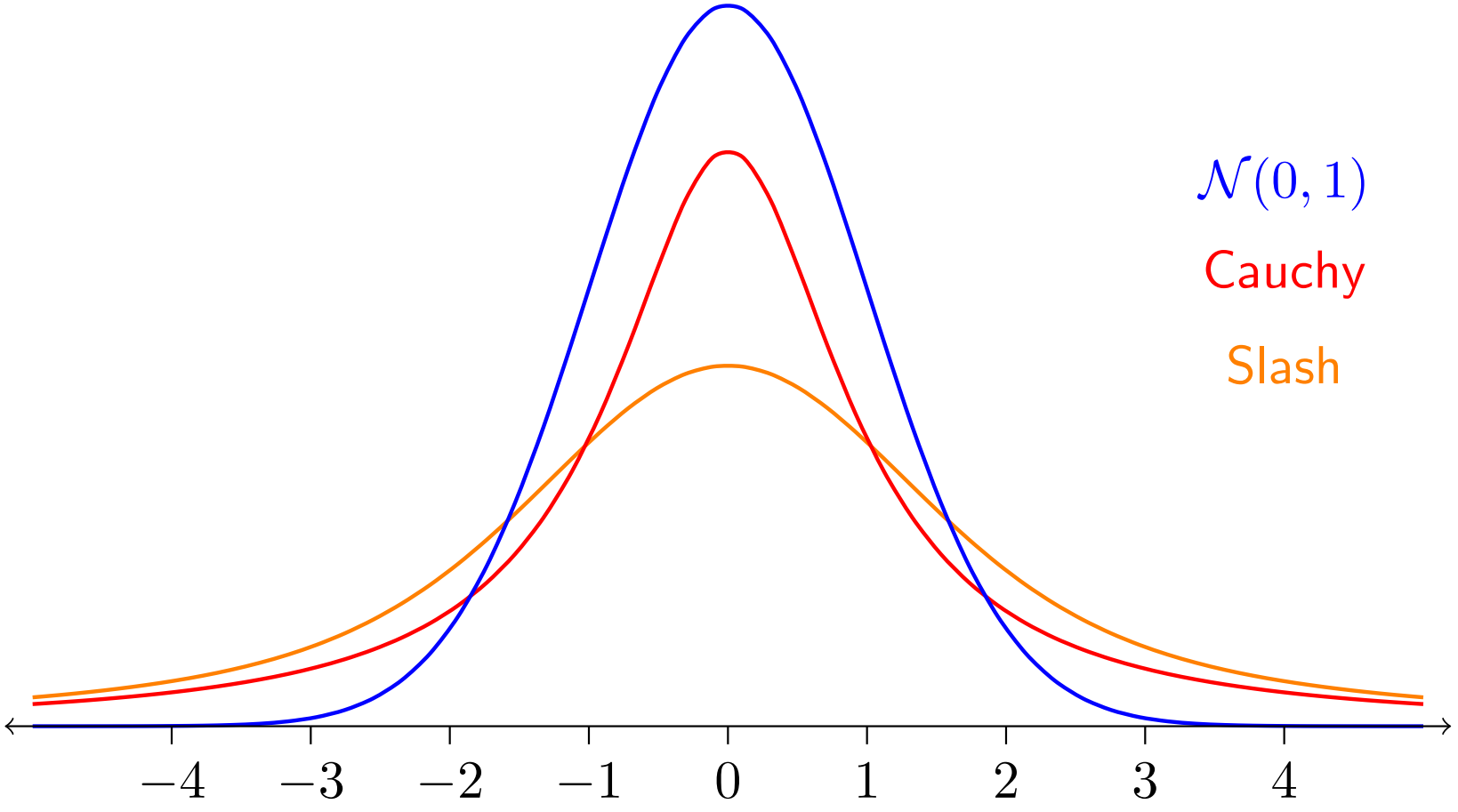
- $\mathbf{X}_i$ are independent samples of size n,
- $\hat{\sigma}_i$ is the ML scale estimate, and
- $T(\mathbf{X}_i)$ is the proposed scale estimate.

Distributions considered

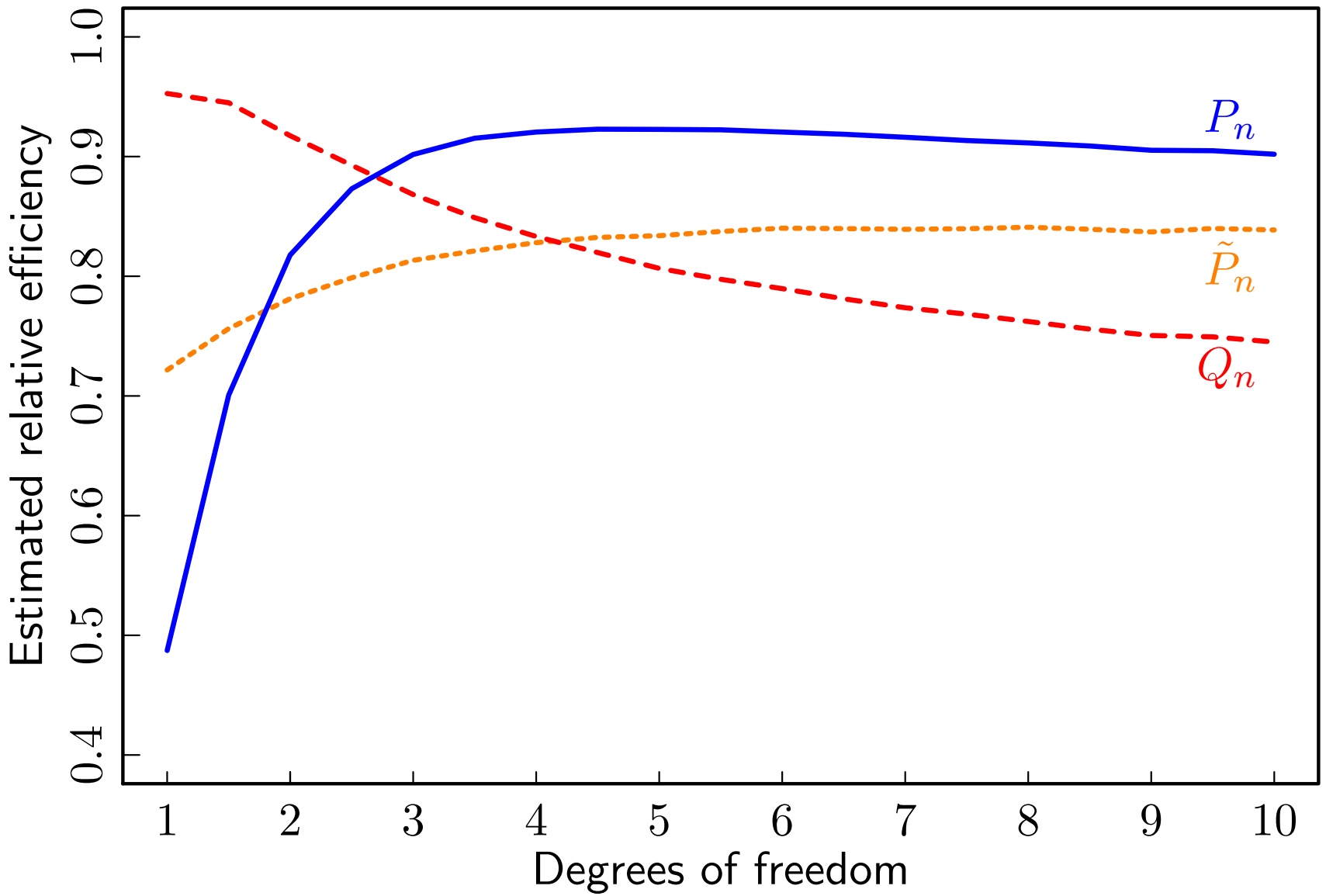
- t distributions with degrees of freedom between 1 and 10.
- Configural polysampling using Tukey's 3 corners: **Gaussian**, **One-wild** and **Slash**.

Aside: the slash distribution

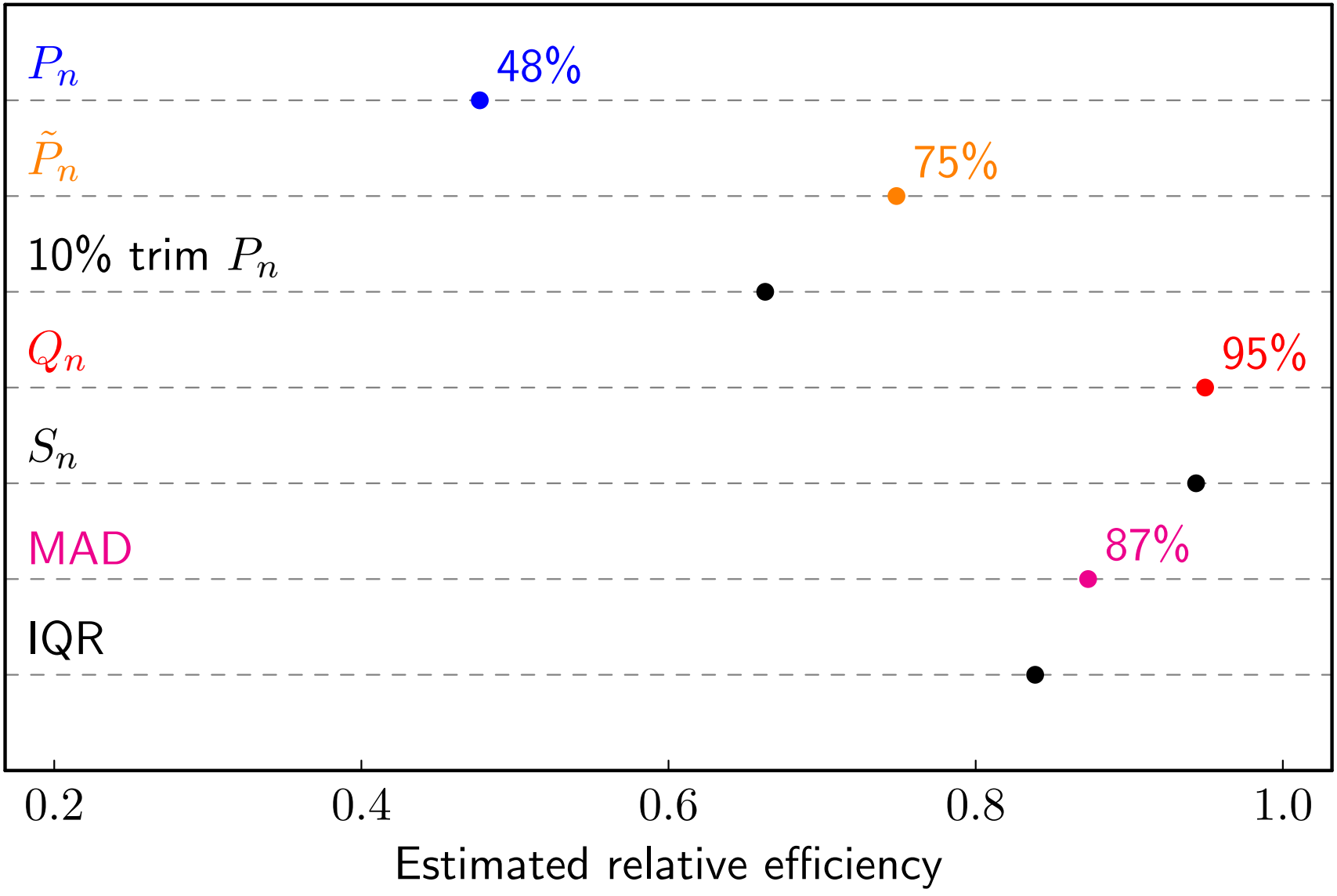
If $Z \sim \mathcal{N}(0,1)$ and $U \sim \mathcal{U}(0,1)$ then $X = \mu + \sigma Z/U$ follows the slash distribution. When $\mu = 0$ and $\sigma = 1$ we have the standard slash distribution.



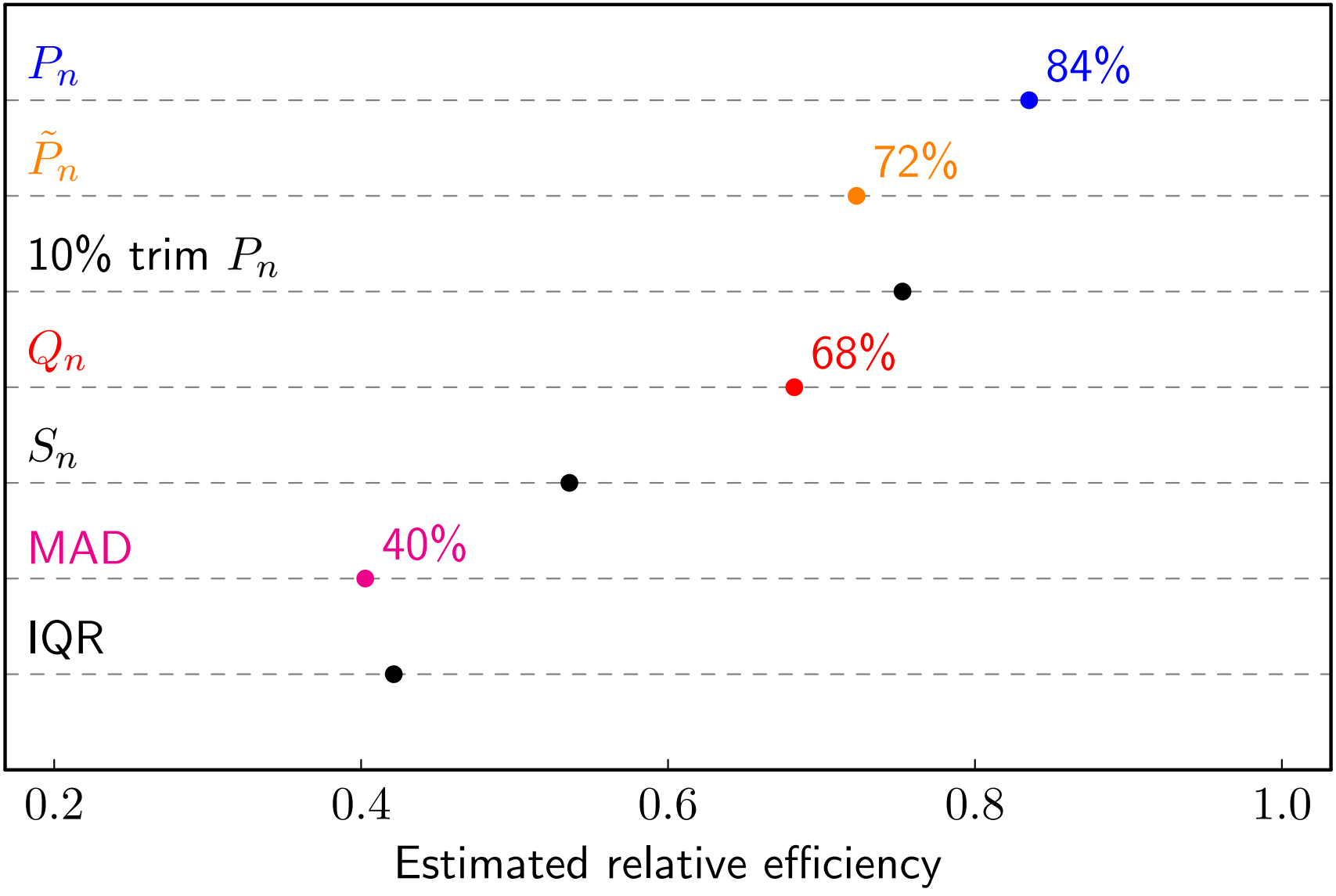
Relative efficiencies: $f = t_\nu$ for $\nu \in [1, 10]$ and $n = 20$



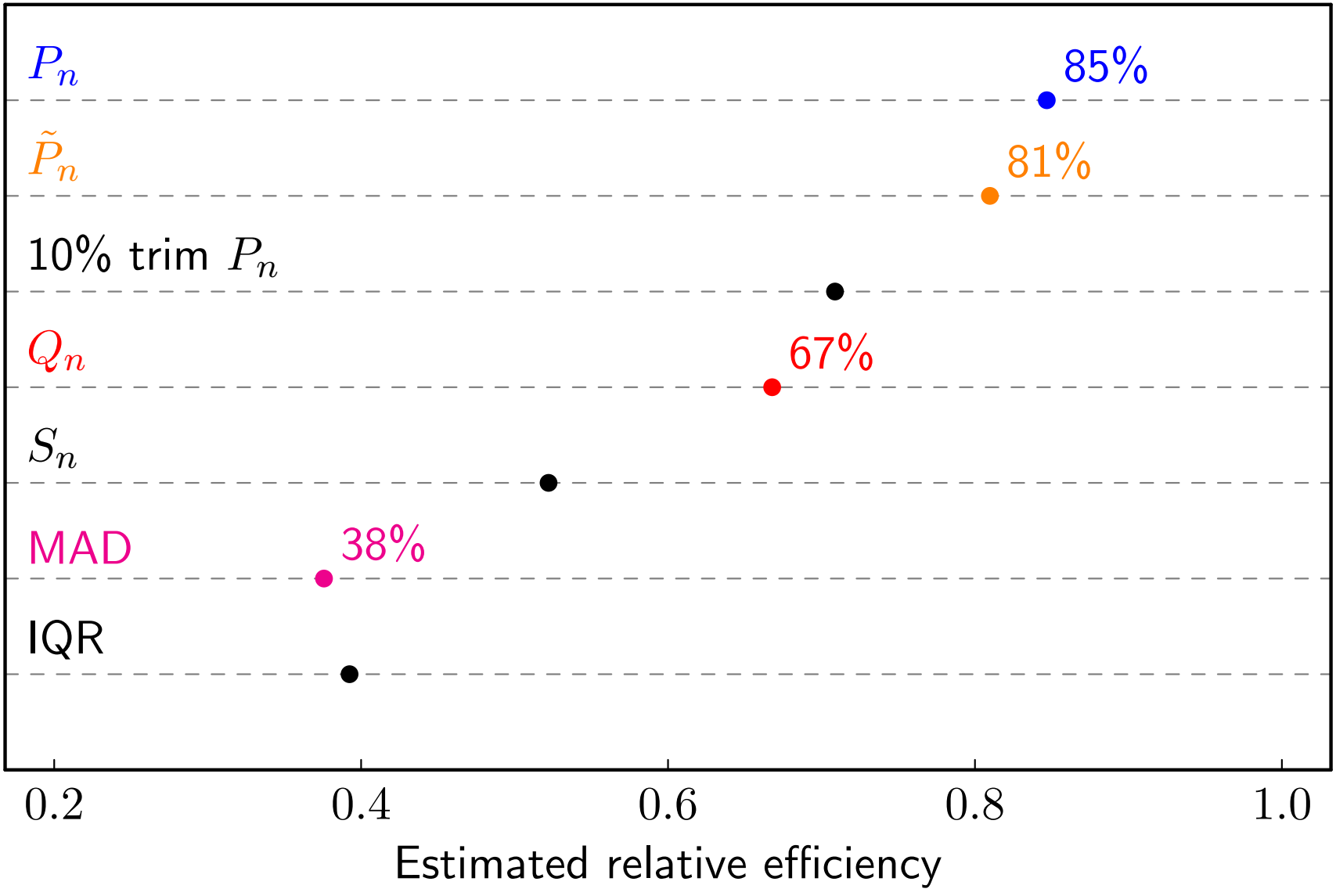
Relative efficiencies at the Slash corner ($n = 20$)



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



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



Outline

From scale to covariance: the GK device

- Gnanadesikan and Kettenring (1972) relate scale and covariance using the following identity,

$$\mathrm{cov}(X, Y) = \frac{1}{4\alpha\beta}\left[\mathrm{var}(\alpha X + \beta Y) - \mathrm{var}(\alpha X - \beta Y)\right],$$

where X and Y are random variables.

- In general, X and Y can have different units, so we set $\alpha = 1/\sqrt{\mathrm{var}(X)}$ and $\beta = 1/\sqrt{\mathrm{var}(Y)}$.

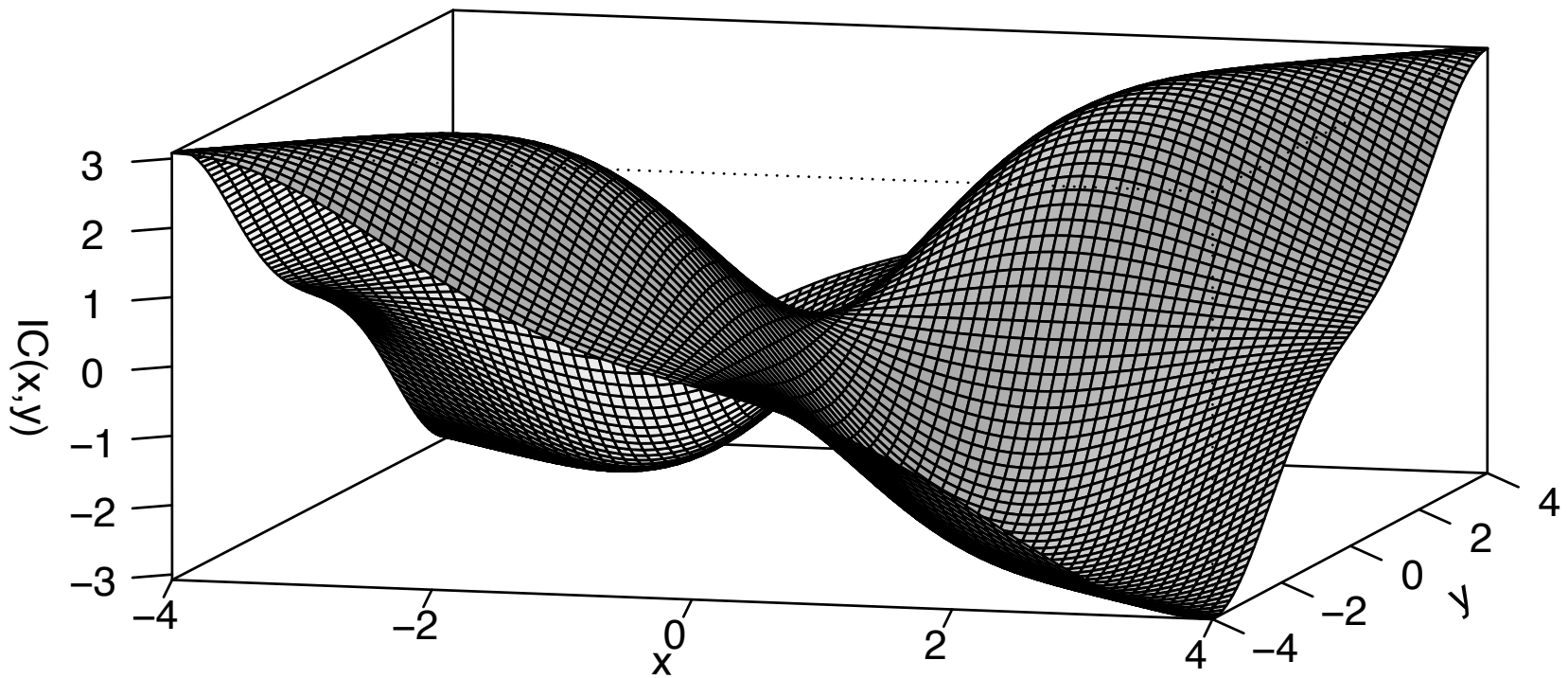
- Replacing **variance** with $\boldsymbol{P_n^2}$ we can similarly construct,

$$\gamma_P(X, Y) = \frac{1}{4\alpha\beta}\left[P_n^2(\alpha X + \beta Y) - P_n^2(\alpha X - \beta Y)\right],$$

where $\alpha = 1/P_n(X)$ and $\beta = 1/P_n(Y)$.

Robustness properties

- γ_P inherits the **13.4%** breakdown value from P_n.
- Following Genton and Ma (1999), we have shown that **influence curve** and therefore the **gross error sensitivity** of γ_P can be derived from the IC of P_n.

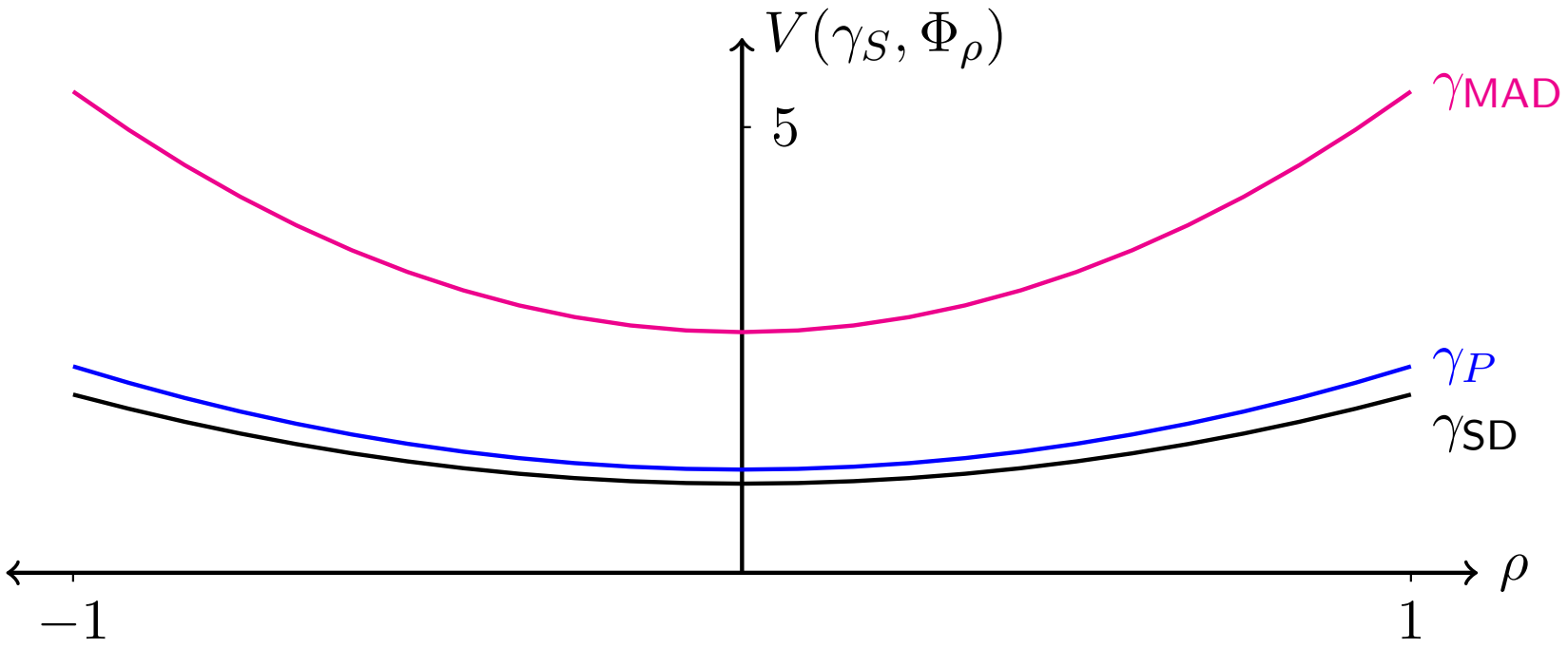


! IC is **bounded** and hence the gross error sensitivity is **finite**.

Asymptotic variance

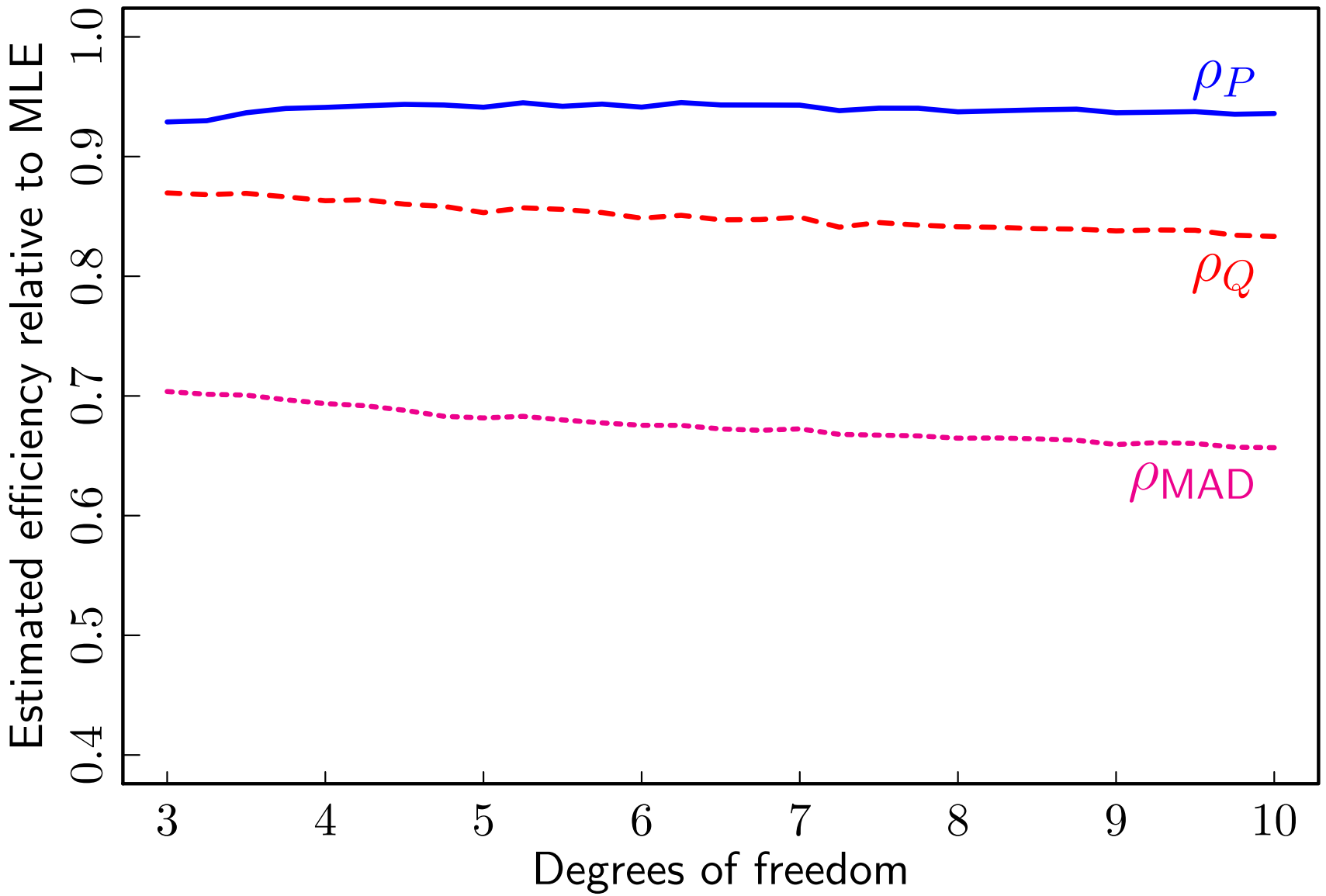
- Genton and Ma (1999) show that the asymptotic variance is directly proportional to that of the underlying scale estimator:

$$V(\gamma_S, \Phi_\rho) = 2(1 + \rho^2)V(S, \Phi)$$

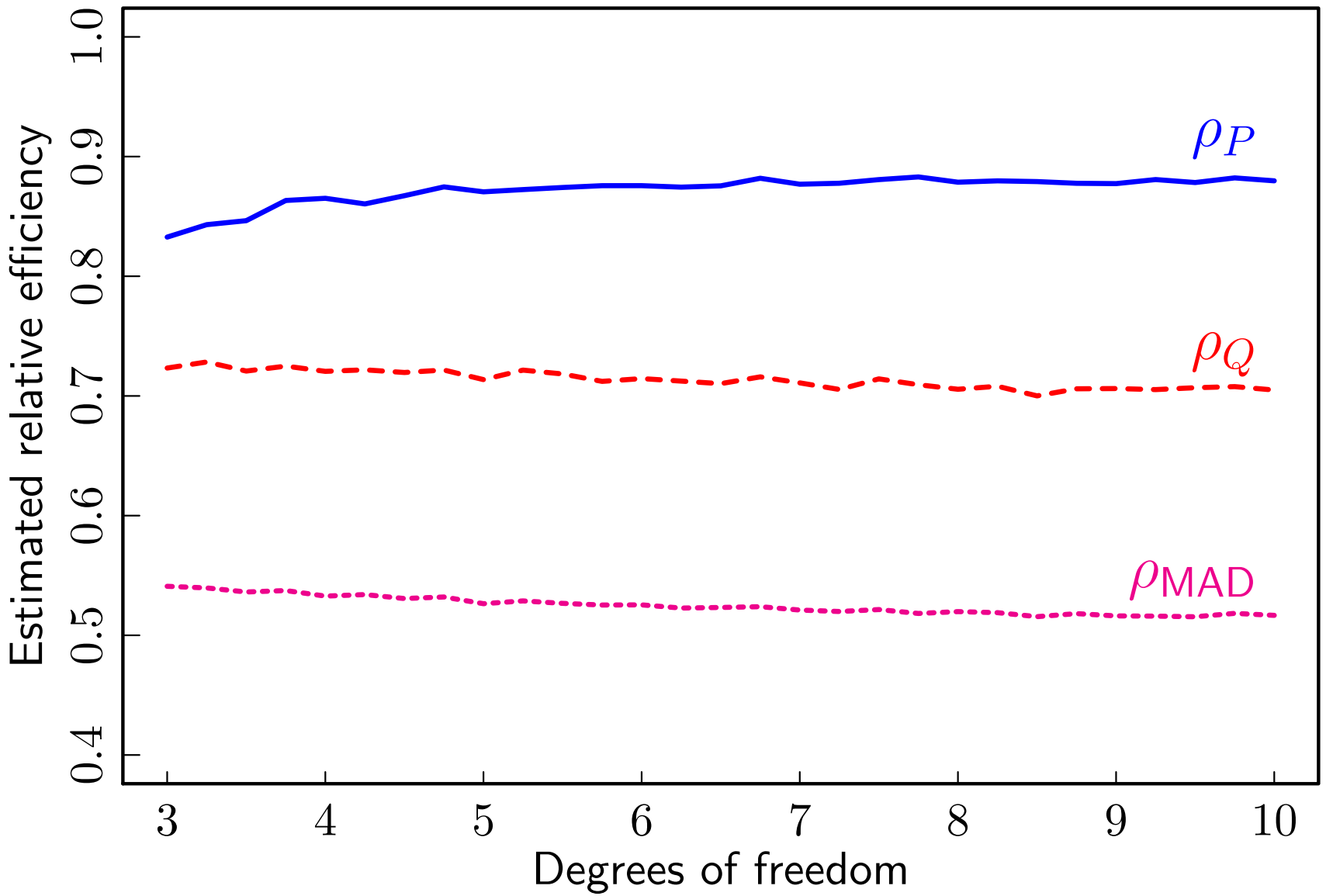


! Asymptotic efficiency of γ_P relative to the standard deviation at the Gaussian remains 86%.

Bivariate t_ν for $\nu \in [3, 10]$, $n = 20$ and $\rho = 0.5$.



Bivariate t_ν for $\nu \in [3, 10]$, $n = 20$ and $\rho = 0.9$.



Outline

Existing robust covariance matrix estimation methods

- M-estimates have BDP $\leq 1/(1+p)$
- Minimum volume ellipsoid (MVE) converges at $n^{-1/3}$
- S-estimates behave like MLE for large p
- Minimum covariance determinant (MCD) requires subsampling, so slow for large p

! Try constructing a covariance matrix using pairwise robust covariance estimates.

Covariance matrices

"Good" covariance matrices should be:

1. Positive-semidefinite.
2. Affine equivariant. That is, if $\hat{\mathbf{C}}$ is a covariance matrix estimator, $\mathbf{A}$ and $\mathbf{a}$ are constants and $\mathbf{X}$ is random then,

$$\hat{\mathbf{C}}(\mathbf{AX} + \mathbf{a}) = \mathbf{A}\hat{\mathbf{C}}(\mathbf{X})\mathbf{A}'.$$

! How can we ensure that a matrix full of pairwise covariances is a "good" covariance matrix?

- Maronna and Zamar (2002) outline the OGK routine which ensures that the resulting covariance matrix is positive-definite and approximately affine-equivariant matrix.

OGK procedure (Maronna and Zamar, 2002)

Let $s(\cdot)$ be a scale function and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{n \times p}$ be a data matrix.

1. Let $\mathbf{D} = \mathrm{diag}(s(\mathbf{x}_1), \ldots, s(\mathbf{x}_p))$ and define $\mathbf{Y} = \mathbf{X}\mathbf{D}^{-1}$.
2. Compute the "correlation matrix" applying $s(\cdot)$ to the columns of $\mathbf{Y}$:

$$\mathbf{U} = [u_{jk}] = \begin{cases} \frac{1}{4}(s(\mathbf{y}_j + \mathbf{y}_k)^2 - s(\mathbf{y}_j - \mathbf{y}_k)^2) & j \neq k \\ 1 & j = k \end{cases}$$

3. Compute the eigendecomposition $\mathbf{U} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}'$.
4. Project the data onto the basis eigenvectors: $\mathbf{Z} = \mathbf{Y}\mathbf{E}$.
5. Estimate the variances in the coordinate directions: $\mathbf{\Gamma} = \mathrm{diag}(s(\mathbf{z}_1)^2, \ldots, s(\mathbf{z}_p)^2)$.
6. The estimated covariance matrix is then,

$$\hat{\mathbf{\Sigma}} = \mathbf{D}^2\mathbf{E}\mathbf{\Gamma}\mathbf{E}'.$$

Precision matrices

- In many practical applications, the covariance matrix is not what is really required.
- PCA, Mahalanobis distance, LDA, etc. use the inverse covariance matrix: the precision matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$.
- Precision matrices are also of interest in modelling Gaussian Markov random fields, where zeros in the correspond to conditional independence between variables.

Sparsity!

In many applications, it is often useful to impose a level of sparsity on the estimated precision matrix.

Regularisation techniques

Graphical lasso (glasso) (Friedman, Hastie, Tibshirani, 2007) minimises the penalised negative Gaussian log-likelihood:

$$f(\mathbf{\Theta}) = \mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log|\mathbf{\Theta}| + \lambda||\mathbf{\Theta}||_1,$$

where $||\mathbf{\Theta}||_1$ is the L_1 norm and λ is a tuning parameter for the amount of shrinkage.

Constrained ℓ_1-minimisation for inverse matrix estimation (CLIME) (Cai, Liu and Luo, 2011) solves the following objective function:

$$\min||\mathbf{\Theta}||_1 \quad \text{subject to: } ||\hat{\mathbf{\Sigma}}\mathbf{\Theta} - \mathbf{I}||_\infty \leq \lambda.$$

Quadratic Inverse Covariance (QUIC) (Hsieh, et. al. 2011) solves the same minimisation problem as the glasso but uses a second order approach.

QUadratic Inverse Covariance (QUIC)

Given a regularisation parameter $\lambda \geq 0$, the ℓ_1 regularised Gaussian MLE for $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ can be estimated by solving a regularised log determinant program:

$$\underset{\mathbf{\Theta} \succ \mathbf{0}}{\operatorname{argmin}} f(\mathbf{\Theta}) = \underset{\mathbf{\Theta} \succ \mathbf{0}}{\operatorname{argmin}} \{ \underbrace{\operatorname{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log|\mathbf{\Theta}|}_{g(\mathbf{\Theta})} + \underbrace{\lambda||\mathbf{\Theta}||_1}_{h(\mathbf{\Theta})} \}$$

where $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix and $||\mathbf{\Theta}||_1 = \sum_{i,j} |\theta_{ij}|$.

Key features

- The ℓ_1 regularisation promotes sparsity in $\mathbf{\Theta}$
- $g(\mathbf{\Theta})$ is twice differentiable and strictly convex $\Rightarrow$ quadratic approximation
- $h(\mathbf{\Theta})$ is convex but not differentiable

Algorithm 1 QUadratic Inverse Covariance

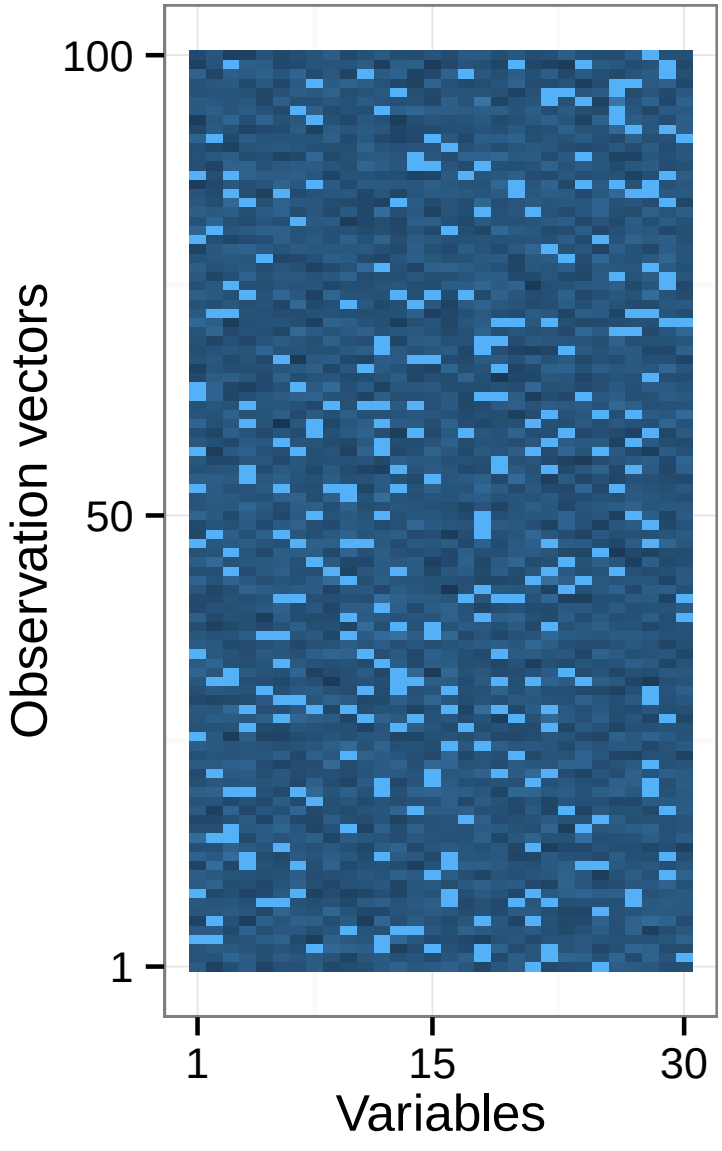
Input: symmetric matrix $\hat{\mathbf{\Sigma}}$, scalar λ, stopping tolerance ϵ

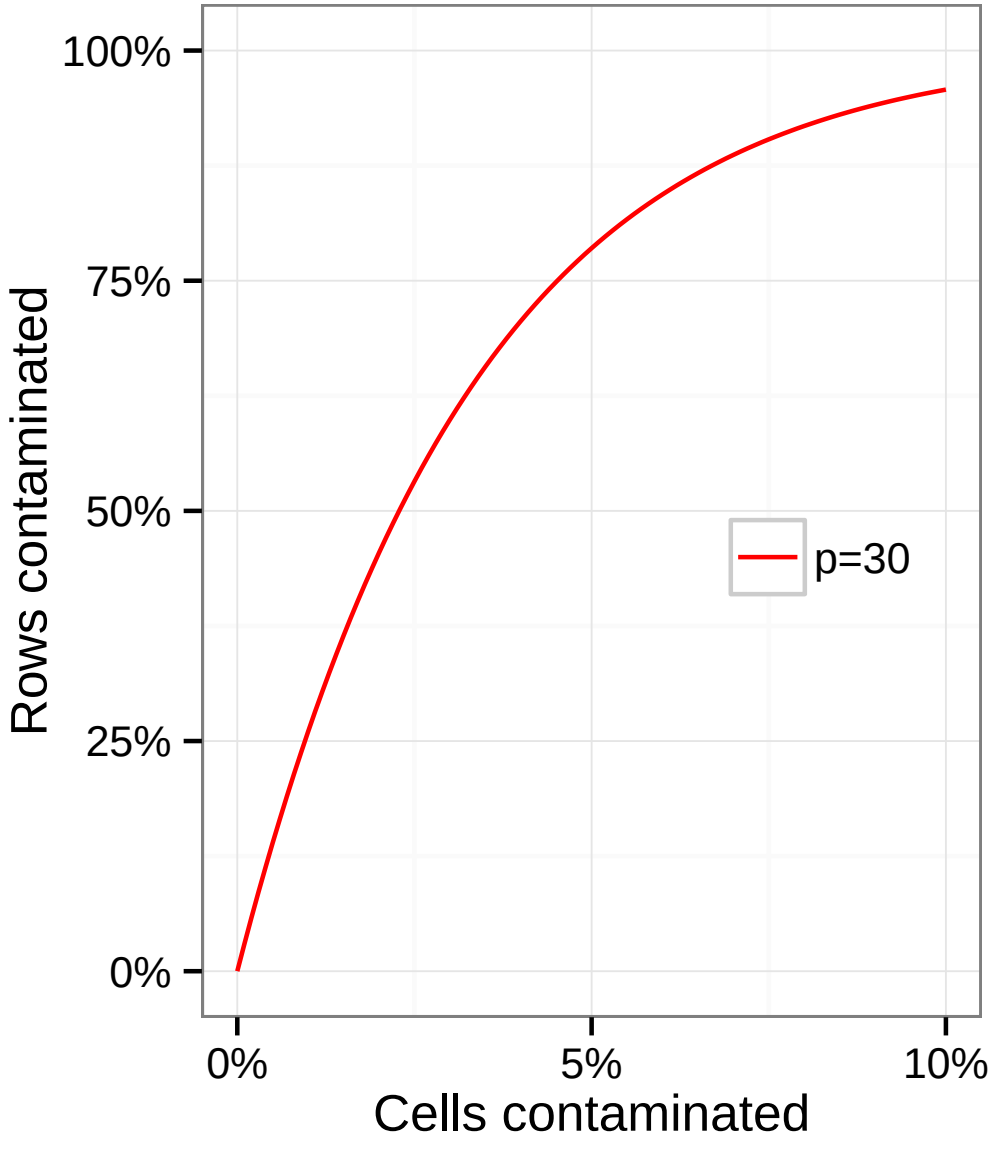
1: **for** $i = 0, 1, \ldots$ **do**
2: Compute $\mathbf{\Sigma}_t = \mathbf{\Theta}_t^{-1}$.
3: Find second order approximation $\bar{f}(\mathbf{\Delta}; \mathbf{\Theta}_t)$ to $f(\mathbf{\Theta}_t + \mathbf{\Delta})$.
4: Partition variables into free and fixed sets.
5: Find Newton direction $\mathbf{D}_t$ using coordinate descent over the free variable set (Lasso problem).
6: Determine step-size α such that $\mathbf{\Theta}_{t+1} = \mathbf{\Theta}_t + \alpha \mathbf{D}_t$ is positive definite and the objective function sufficiently decreases.
7: **end for**

Output: sequence of $\mathbf{\Theta}_t$

R package `QUIC`

Randomly scattered contamination





Simulation design

Two scenarios: **sparse** and **dense** precision matrices, $\mathbf{\Theta}$.

sample size	$n = 120$
variables	$p = 30$
data	$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$
sparsity parameter	$\lambda = 0.1$
contamination	0% to 3% randomly scattered
replications	$N = 1000$

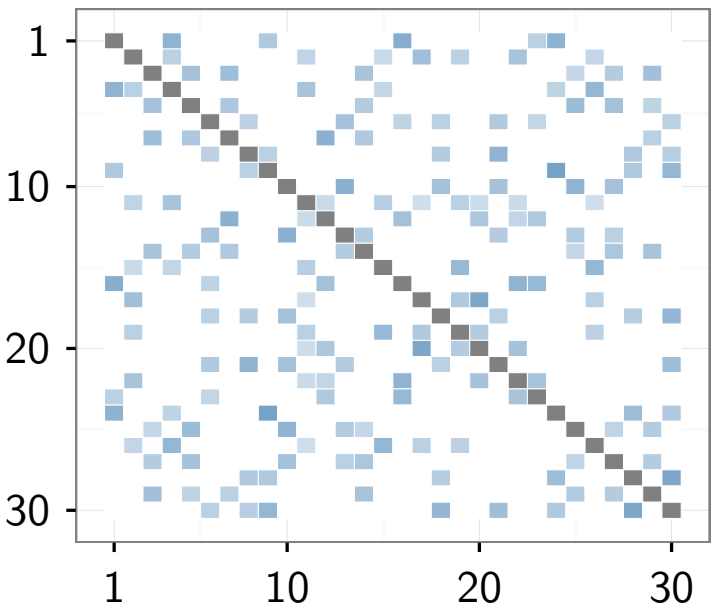
Performance metric

The entropy loss

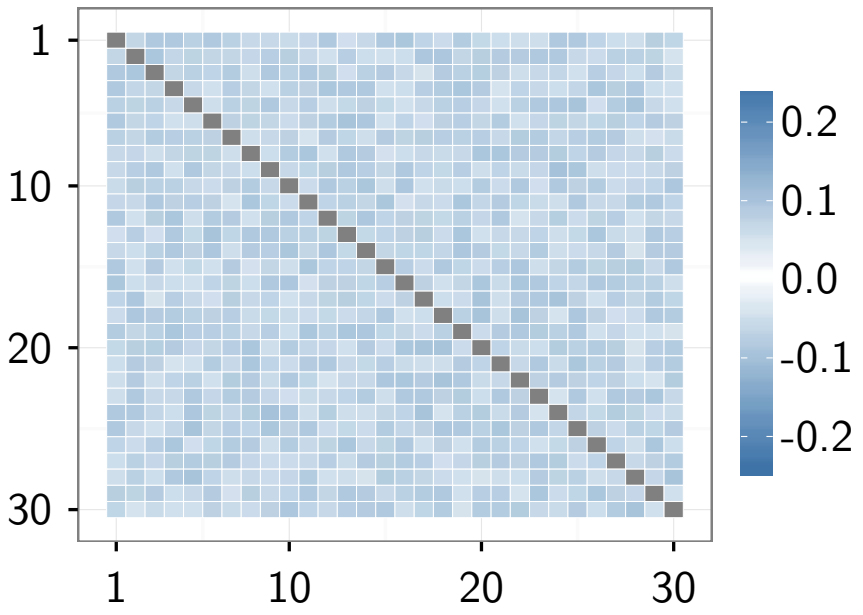
$$L(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) = \text{tr}(\mathbf{\Theta}^{-1}\hat{\mathbf{\Theta}}) - \log|\mathbf{\Theta}^{-1}\hat{\mathbf{\Theta}}| - p.$$

Simulation design

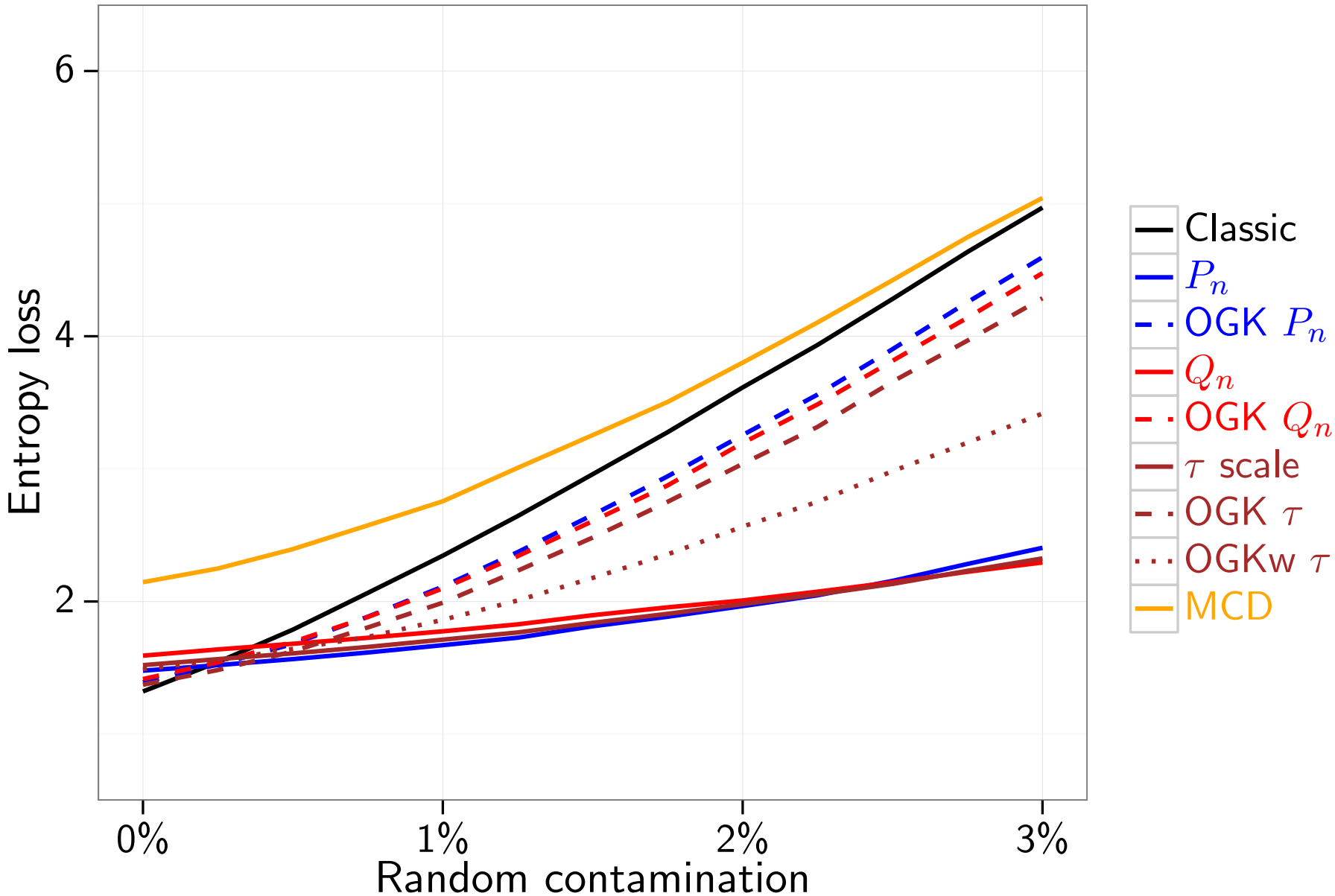
Sparse Θ



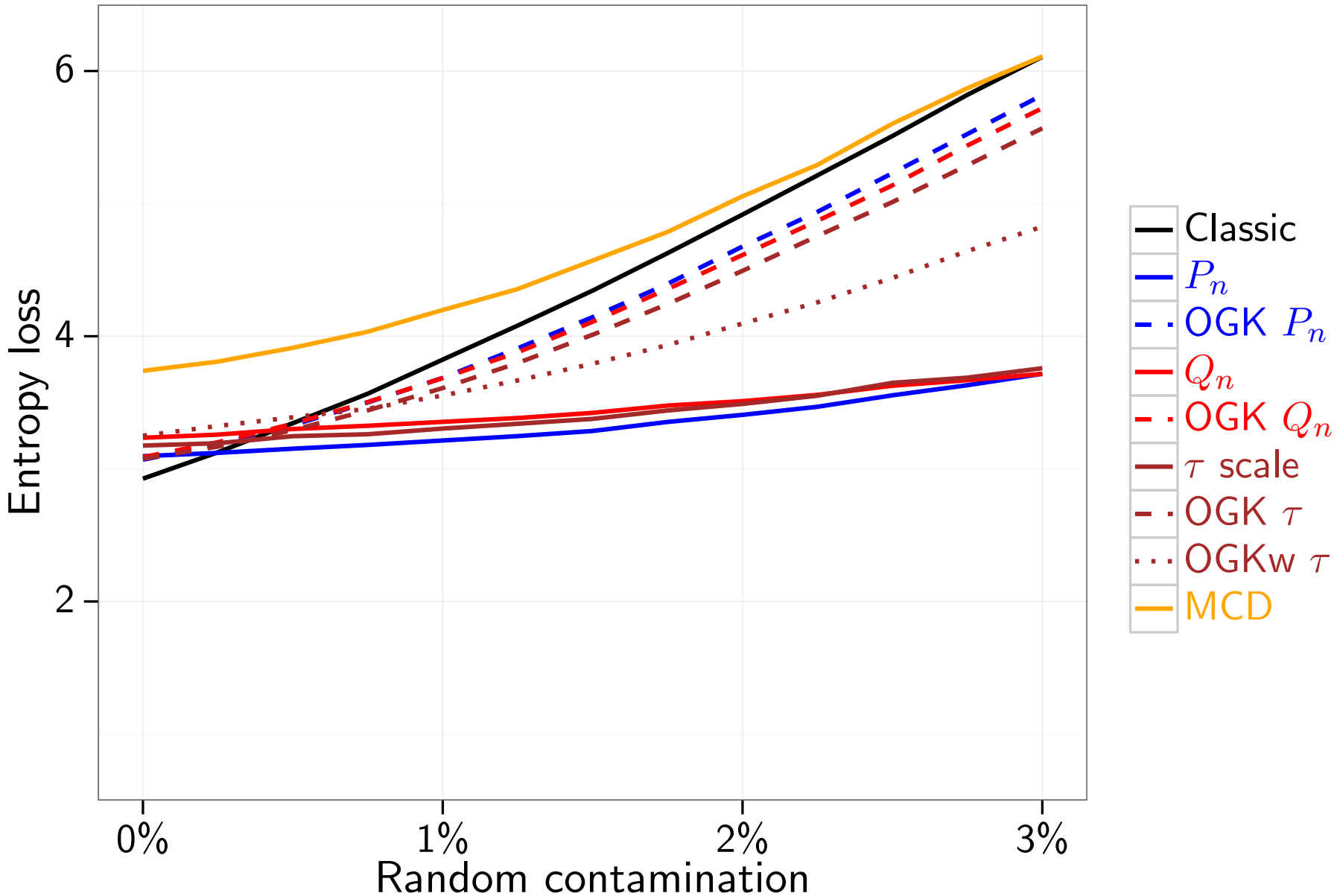
Dense Θ



Sparse Θ, outliers at 5



Dense Θ, outliers at 5



Outline

Summary

1. Aim

- Efficient, robust and widely applicable scale and covariance estimators.

2. Method

- P_n scale estimator with the GK identity and an orthogonalisation procedure for pairwise covariance matrices or QUIC procedure for (sparse) precision matrix estimation.

3. Results

- 86% asymptotic efficiency at the Gaussian distribution.
- Robustness properties flow through to covariance estimates and beyond.

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.

Genton, M. and Ma, Y. (1999).
Robustness properties of dispersion estimators.
Statistics & Probability Letters, 44(4):343–350.

References

Hsieh, C-J., Sustik, M.A., Dhillon I.S. and Ravikumar, P.K. (2011).
Sparse inverse covariance matrix estimation using quadratic approximation.
Advances in Neural Information Processing Systems 24, 2330–2338.

Gnanadesikan, R. and Kettenring J. R. (1972).
Robust estimates, residuals and outlier detection with multiresponse data.
Biometrics, 28(1):81–124.

Maronna, R. and Zamar, R., (2002).
Robust estimates of location and dispersion for high-dimensional datasets.
Technometrics, 44(4):307–317.

Tarr, G., Müller, S. and Weber, N.C., (2012).
A robust scale estimator based on pairwise means.
Journal of Nonparametric Statistics, 24(1):187–199.