



A smooth ROC curve estimator based on log-concave density estimates

Kaspar Rufibach

Division of Biostatistics

Institute for Social and Preventive Medicine

University of Zurich

Setup

Diagnostic test:

- **Continuous** outcome, e.g. measurement of biomarker.
- Measured in non-diseased patients, X_1, \dots, X_m (controls).
- Measured in diseased patients Y_1, \dots, Y_n (cases).

Assess diagnostic accuracy of test via **estimation of ROC curve**.

Desired properties of a ROC curve (estimate):

- **Invariance** against monotone transformations of X_i, Y_j .
- **Unbiasedness**: $R(t) \geq t$ for all $t \in [0, 1] \Rightarrow$ ROC curve should not fall below diagonal.
- **Propriety**: $R(t)$ is concave.

Methods to estimate a ROC curve

Empirical ROC curve: Plot **true positives** $\{Y_j > s\}/n$ vs. **false positives** $\{X_i > s\}/m$.

Can be written as $R(t; \mathbb{F}_m, \mathbb{G}_n)$ where

$$R(t; F, G) = 1 - G(F^{-1}(1 - t))$$

where

- $t \in [0, 1]$: false-positive proportion corresponding to a positivity cutoff s of the test.
- $\mathbb{F}_m, \mathbb{G}_n$: empirical CDFs, induce empirical ROC curve $\mathbb{R}_{m,n}$.

Alternative ROC curve estimators: **Plug in** estimates for F and G .

Properties of estimates: empirical

Empirical ROC curve:

- **Nonparametric** estimate: Displays data pattern well.
- Not smooth.
- Suffers from **potential large variability** at a point t .
- **Horizontal** stretches \Rightarrow inconvenient if interested in finding a particular FPF at a specified TPF.
- Invariant.
- Strongly consistent under mild assumptions. [Hsieh and Turnbull \(1996\)](#)

Properties of estimates: binormal

Binormal ROC curve:

- Assume **normal distribution** for X_i and $Y_j \Rightarrow R(t) = \Phi(a + b\Phi^{-1}(t))$.
- Many ways to estimate a, b . Often complicated, **no implementation** available.
- (Semi-)parametric \Rightarrow not robust.
- Not invariant, biased, only proper in special case.

Properties of estimates: kernel

Kernel estimate:

- Hall and Hyndman (2003): select bandwidths for F and G to minimize MSE of estimation of R directly.
- **Smooth**, typically not invariant, may be biased, not necessarily proper.
- **Choice** of kernel (not critical), **bandwidth** (critical) to estimate F and G .

Only binormal and empirical ROC curve estimate routinely implemented \Rightarrow most commonly used.

ROC curve estimation assuming log-concave densities

Estimate densities of F and G via **nonparametric maximum likelihood** assuming **log-concavity**:

$$f = \exp \phi \quad g = \exp \gamma,$$

ϕ, γ concave functions $\mathbb{R} \rightarrow [-\infty, \infty)$.

Rufibach (2007), Balabdaoui et al. (2009), Dümbgen and Rufibach (2009),
Dümbgen and Rufibach (2011)...

ROC curve estimation assuming log-concave densities

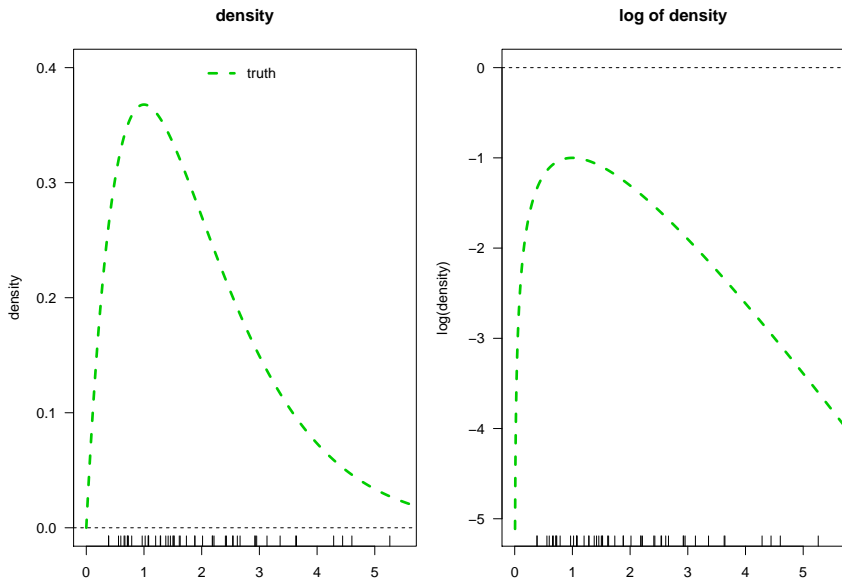
Merits of log-concave estimator \hat{f}_n :

- Normal, Gamma, Beta, Uniform, ...
- Flexible nonparametric assumption: Allows for **skewness**, surrogate for **unimodality**.
- Favorable theoretical properties.
- **Efficient computation**: R package **logcondens**
Dümbgen and Rufibach (2011).

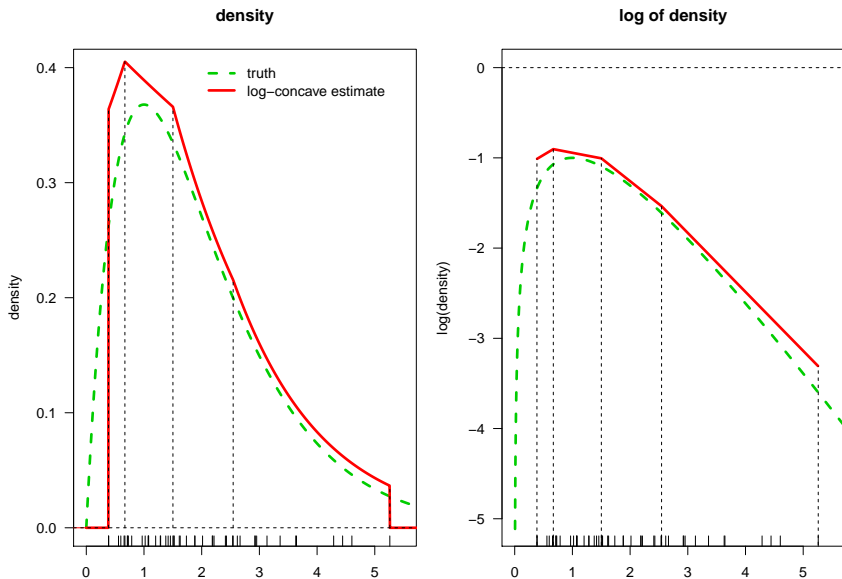
Shape of \hat{f}_n :

- Log-density **piecewise linear**.
- Knots only at (few) observations.
- Supported on $[X_1, X_n]$.

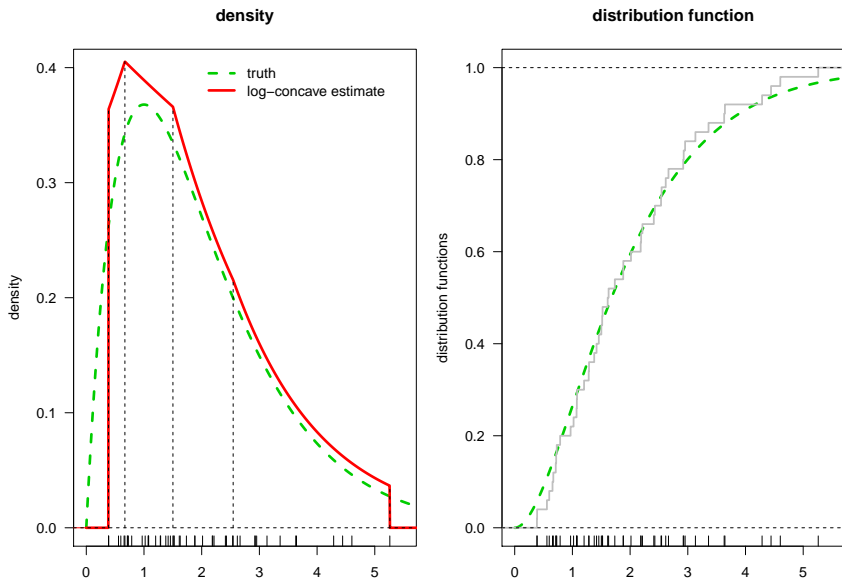
Example: Gamma(2, 1) density, $n = 50$



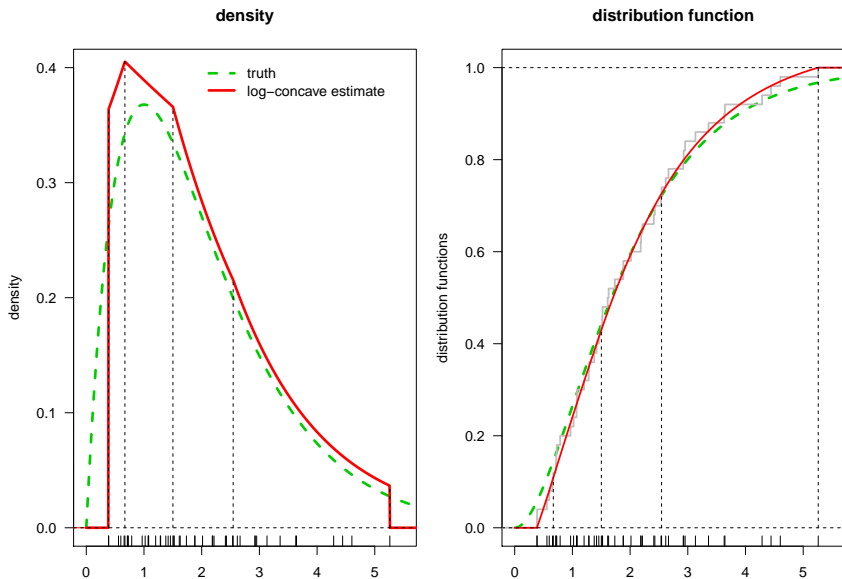
Example: Gamma(2, 1) density, $n = 50$



Example: Gamma(2, 1) density, $n = 50$



Example: Gamma(2, 1) density, $n = 50$



ROC curve estimation assuming log-concave densities

Under **general assumptions** on f :

$$\sqrt{n} \max_{t \in T_n} |\hat{F}_n(t) - \mathbb{F}_n(t)| \xrightarrow{P} 0 \quad (n \rightarrow \infty).$$

for some T_n .

\hat{F}_n is **asymptotically equivalent** to \mathbb{F}_n .

\hat{F}_n can be considered a **smoother** of \mathbb{F}_n .

ROC curve estimation assuming log-concave densities

Define new ROC curve estimate:

$$\begin{aligned}\widehat{R}_{m,n}(t) &:= R(t; \widehat{F}_m, \widehat{G}_n) \\ &= 1 - \widehat{G}_n(\widehat{F}_m^{-1}(1-t)), \quad t \in [0, 1].\end{aligned}$$

Main features:

- **Flexible** nonparametric smooth estimate of R .
- **Fully automatic**: No choice of kernel, bandwidth, or any other regularization parameter.
- **Simple computation**: function `logConROC` in package **logcondens**.
- Not necessarily invariant or unbiased or proper \Rightarrow same as for **binormal**, but typically less pronounced.

$\widehat{R}_{m,n}$ is asymptotically equivalent to $\mathbb{R}_{m,n}$

Under **reasonable** assumptions on f and g we have

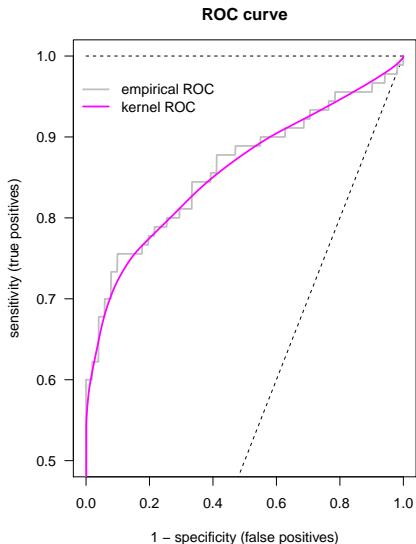
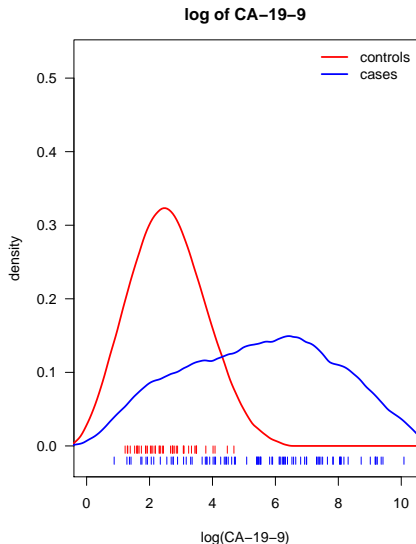
$$\sqrt{n} \sup_{t \in J} \left(\widehat{R}_{m,n}(t) - \mathbb{R}_{m,n}(t) \right) \rightarrow_{\mathbb{P}} 0 \quad (n \rightarrow \infty)$$

on some J and suitable $m = m(n)$.

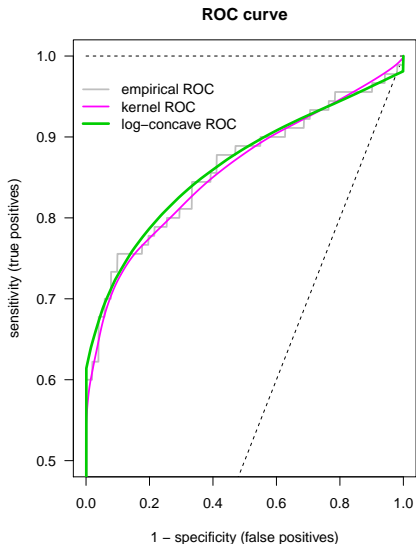
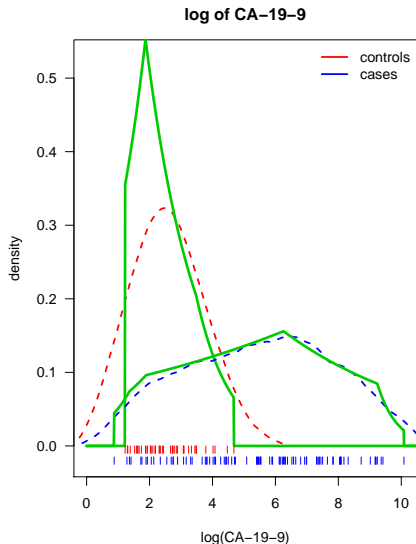
$\widehat{R}_{m,n}$ can be considered a **smoother** of $\mathbb{R}_{m,n}$!

Strong consistency and limiting distribution for $\mathbb{R}_{m,n}$: [Hsieh and Turnbull \(1996\)](#).

Example: pancreatic cancer serum biomarker



Example: pancreatic cancer serum biomarker



Assessment of estimation accuracy

Average square error (ASE) for a ROC curve estimate \widehat{R} :

$$ASE(\widehat{R}) = (1/100) \sum_{k=1}^{100} \left(\widehat{R}(u_k) - R(u_k) \right)^2$$

for 100 equidistant grid points u_j .

Provide boxplot of

$$\sqrt{ASE(\widehat{R})_j} / \sqrt{ASE(\mathbb{R}_{m,n})_j}$$

for ROC curve estimates in $j = 1, \dots, 500$ simulation runs.

Competitors in simulation study

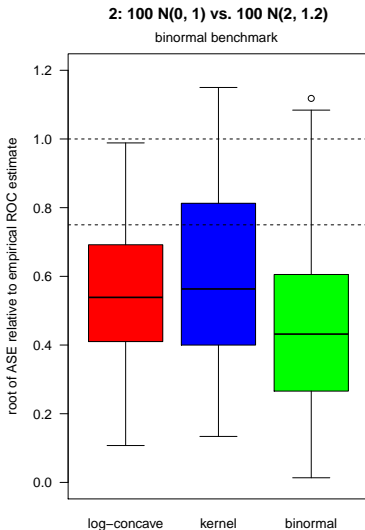
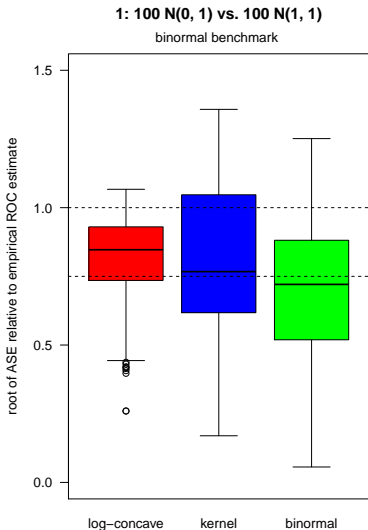
- **Empirical** (= benchmark),
- fully parametric **binormal**,
- **kernel** by Hall and Hyndman (2003),
- the new estimate $\hat{R}_{m,n}$ based on **log-concave** densities.

Simulation scenarios: correctly specified case

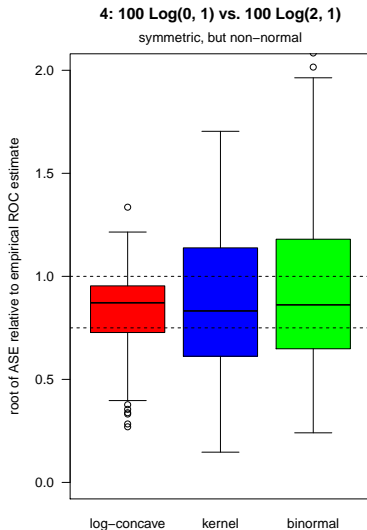
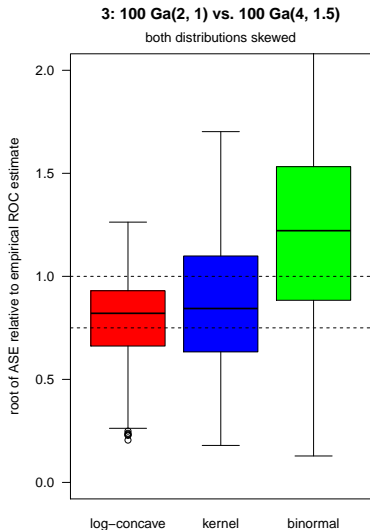
Scenario	F	m	G	n	rationale
1	$N(0, 1)$	100	$N(1, 1)$	100	binormal benchmark
2	$N(0, 1)$	100	$N(2, 1.2)$	100	binormal benchmark
3	$Ga(2, 1)$	100	$Ga(4, 1.5)$	100	both distributions skewed
4	$Log(0, 1)$	100	$Log(2, 1)$	100	symmetric, but non-normal

Correctly specified: all these distributions have **log-concave densities**.

Simulation results: Scenarios 1 & 2



Simulation results: Scenarios 3 & 4



Messages from simulations: correctly specified case

Main points: The log-concave estimate $\widehat{R}_{m,n}$ is

- generally **more efficient than empirical**.
- performs often better than kernel ROC estimate.
- Loss to **fully parametric** binormal remarkably **small** in Normal scenario.
- Non-normal scenario: binormal ROC performs poorly.

Take home messages

- Log-concavity **often sensible** to assume for diagnostic test data.
- **No choice** of regularization parameter such as e.g. bandwidth.
- $\widehat{R}_{m,n}$ **asymptotically equivalent** to empirical ROC curve estimate.
- **Superior to empirical** for finite n if log-concavity holds.
- Loss compared to binormal model **modest** if normality holds.
- **Robust** to moderate deviations from log-concavity (simulations in paper).
- Might be biased and not invariant.
- Simple and efficient computation:
logConROC(cases, controls, ...) using **logcondens**.

Final suggestion for *your* diagnostic test analysis

I suggest to use the new estimate $\hat{R}_{m,n}$:

- whenever the **log-concavity** assumption is sensible and
- you would either use the empirical or the binormal.

Thank you for your attention.

References

- ▶ Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.* **37** 1299–1331.
- ▶ Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607.
- ▶ Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function. *Bernoulli* **15** 40–68.
- ▶ Dümbgen, L. and Rufibach, K. (2011). **logcondens**: Computations related to univariate log-concave density estimation. *Journal of Statistical Software* **39** 1–28.
- ▶ Hall, P. G. and Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statist. Probab. Lett.* **64** 181–189.
- ▶ Hazelton, M. L. (2011). Assessing log-concavity of multivariate densities. *Statist. Probab. Lett.* **81** 121–125.
- ▶ Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.* **24** 25–40.
- ▶ Rufibach, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Statist. Comp. Sim.* **77** 561–574.
- ▶ Rufibach, K. (2011). A smooth roc curve estimator based on log-concave density estimates. Tech. rep., Division of Biostatistics, University of Zurich.
- ▶ Rufibach, K. and Dümbgen, L. (2011). *logcondens: Estimate a Log-Concave Probability Density from iid Observations*. R package version 2.0.3.
- ▶ Wieand, S., Gail, M. H., James, B. R. and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76** 585–592.

$\widehat{R}_{m,n}$ is asymptotically equivalent to $\mathbb{R}_{m,n}$

Asymptotic equivalence implies, as $n \rightarrow \infty$:

$$\sqrt{n}(\widehat{R}_{m,n}(t) - R(t)) \rightarrow_d \mathbb{B}_1(1 - R(t; F, G)) + \sqrt{\lambda} \frac{g(F^{-1}(1-t))}{f(F^{-1}(1-t))} \mathbb{B}_2(1-t)$$

uniformly on some J , for $\mathbb{B}_1(t)$ and $\mathbb{B}_2(t)$ two independent Brownian Bridges.

Result for $\mathbb{R}_{m,n}$: [Hsieh and Turnbull \(1996\)](#).

Robustness: misspecified case

Scenario	F	m	G	n
5	Lomax(3, 7)	100	Lomax(5, 3)	100
6	t(5, 0)	20	t(5, 2)	20
7	t(5, 0)	100	t(5, 2)	100
8	N(0, 1)	100	$0.75 \cdot N(2.5, 1) + 0.25 \cdot N(2.5, 3)$	100

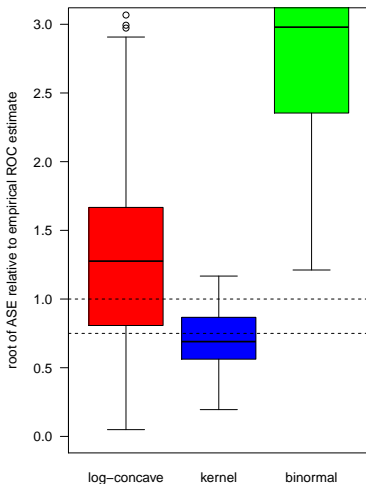
One or both densities **not log-concave**, but all unimodal.

Lomax: even **log-convex**.

Simulation results: Scenarios 5 & 6

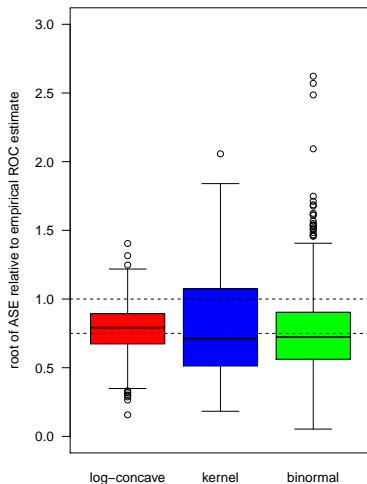
5: 100 Lomax(3, 7) vs. 100 Lomax(5, 3)

often used parametric unbiased model

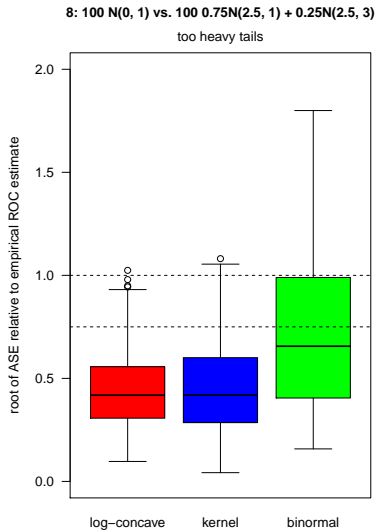
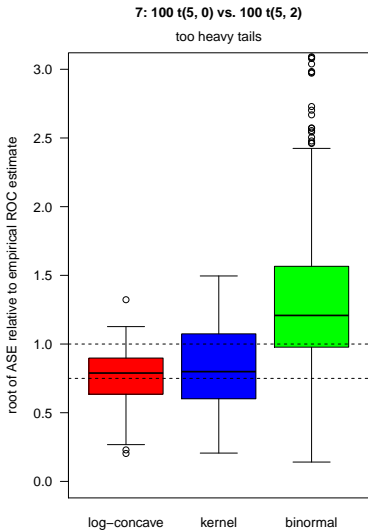


6: 20 t(5, 0) vs. 20 t(5, 2)

too heavy tails, small samples



Simulation results: Scenarios 7 & 8



Messages from simulations: misspecified case

Main points for robustness of $\hat{R}_{m,n}$:

- Log-concave estimate often **still more efficient** than empirical.
- Performance comparable to kernel ROC estimate, except for Lomax.

Further work

Further points discussed in [Rufibach \(2011\)](#):

- **Testing for log-concavity**: [Hazelton \(2011\)](#).
- Exploratory alternative: compare log-concave to kernel density estimate.
- Results \approx transform to estimation of AUC. All methods generally inferior to empirical, kernel rather poor.