Confidence distributions based on M-estimators

Sam-Erik Walker

Dept. of Math., Univ. of Oslo, 13 May 2015

focusstat

FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

Outline

1. M-estimators
2. Proper scoring rules
3. Asymptotics
4. Robustness
5. Numerical example
6. Concluding remarks

M-estimators

- Forms a broad class of estimators obtained as the minimum (or maximum) of sums of functions of the data, and was proposed in 1964 by Peter J. Huber for generalizing MLE

$$\rho\text{-type}: \hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(x_i, \theta\right)$$

$$\psi\text{-type}: \sum_{i=1}^{n} \psi\left(x_i, \hat{\theta}\right) = \sum_{i=1}^{n} \nabla_{\theta} \rho\left(x_i, \hat{\theta}\right) = 0$$

- Examples of M-estimators:
  - Least squares estimators
  - MLE
  - Huber's M-estimator
  - Estimators derived from proper scoring rules

Huber's M-estimator

$$\sum_{i=1}^{n} \psi_k\left(\frac{x_i - \hat{\theta}_n}{\hat{\sigma}_n}\right) = 0 \quad \text{where} \quad \psi_k(x) = \begin{cases} +k & x \geq +k \\ x & -k \leq x \leq +k \\ -k & x \leq -k \end{cases}$$

- Here k is a tuning constant determining the degree of robustness
- If $k \to \infty$ then $\hat{\theta}_n$ is the mean; If $k \to 0$ then $\hat{\theta}_n$ is the median
- Huber's motivation was that unrestricted ψ -functions have undesirable properties, being unstable to outliers
- $\hat{\sigma}_n$ is usually a robust estimate of scale, e.g., $\text{MAD} = \text{median}\left(\left|x_i - \text{median}(x_i)\right|\right)$
- Robust start value for location: Median

Proper scoring rules

- A scoring rule is a *real* function $S(x,F)$ where x is a sample from some outcome space, and F is a model distribution over this outcome space
- Define $S(G,F)=E_G\{S(X,F)\}$ where $X \sim G$ the data distribution
- The scoring rule is said to be *proper* iff $S(G,F) \geq S(G,G)$ for all F
- It is said to be *strictly proper* iff $S(G,F) > S(G,G)$ for all $F \neq G$
- If S is proper then $cS+h(x)$ is also proper if $c>0$, $h(x)$ arbitrary function

$$D(G,F)=S(G,F)-S(G,G)$$

DIVERGENCE SCORING RULE ENTROPY

Proper scoring rules: Inference

- If the model distribution is *parametric*, i.e., $F = F_\theta$, and if the unknown data distribution G is approximated by $\hat{G}_n(x) = \sum_{i=1}^{n} I(x_i \le x)/n$, we obtain the following *empirical proper scoring rule estimator* of θ

$$\hat{\theta}_n = \arg\min_\theta D\left(\hat{G}_n, F_\theta\right) = \arg\min_\theta \left\{S\left(\hat{G}_n, F_\theta\right) - S\left(\hat{G}_n, \hat{G}_n\right)\right\} = \arg\min_\theta S\left(\hat{G}_n, F_\theta\right)$$

$$= \arg\min_\theta E_{\hat{G}_n}\left\{S\left(X, F_\theta\right)\right\} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} S\left(x_i, F_\theta\right) = \arg\min_\theta \sum_{i=1}^{n} S\left(x_i, F_\theta\right)$$

Proper scoring rules: Inference

- Thus, S plays the role of the ρ-function in M-estimation
- Further, if S is differentiable w.r.t. θ, we have

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} S\left(x_i, F_\theta\right) \quad \Leftrightarrow \quad \sum_{i=1}^{n} s\left(x_i, F_{\hat{\theta}_n}\right) = 0$$

i.e., the optimal parameters represent the solution of the so-called *scoring rule estimating equation(s*), p equations if there are p parameters, thus s plays the role of the ψ-function in M-estimation

Proper scoring rules: Examples

- The most prominent proper scoring rule is the *logarithmic score* (Good, 1952)

$$S\left(x, F_\theta\right) = -\log f\left(x, \theta\right)$$

  leading empirically to MLE

- The associated divergence is the *Kullback-Leibler* divergence, and the associated entropy is the *Shannon* entropy

$$D\left(G, F_\theta\right) = S\left(G, F_\theta\right) - S\left(G, G\right) = \int \left\{\log g\left(x\right) - \log f\left(x, \theta\right)\right\} g\left(x\right) dx$$

Proper scoring rules: Examples

- The logarithmic score is a special case of a general so called *separable Bregman score* construction (Dawid, 2007)

$$S\left(x, F_\theta\right) = -\varphi'\left\{f\left(x,\theta\right)\right\} - \int\left[\varphi\left\{f\left(u,\theta\right)\right\} - f\left(u,\theta\right)\varphi'\left\{f\left(u,\theta\right)\right\}\right] du$$

where the defining function $\varphi : \mathbb{R}^+ \to \mathbb{R}$ is *convex* and *differentiable*

- By setting $\varphi(t) = t \ln t$, which is convex and differentiable, we obtain the logarithmic score

Proper scoring rules: Examples

- Another important special case of this construction is the *Tsallis score* obtained by setting $\varphi(t) = t^q$ for $q > 1$. This yields

$$S_q\left(x, F_\theta\right) = -q \cdot f\left(x,\theta\right)^{q-1} + \left(q-1\right)\int f\left(u,\theta\right)^q du$$
$$= \left\{-q \cdot f\left(x,\theta\right)^{q-1} + \left(q-1\right)\int f\left(u,\theta\right)^q du + q - 1\right\} \Big/ (q-1)$$

- Incidentally this is the same as the *Basu-Harris-Hjort-Jones power divergence* estimator (Basu et al., 1998) with parameter $\alpha = q - 1 > 0$ obtained by setting $\varphi(t) = t \cdot t^{(\alpha)}$ where $t^{(\alpha)}$ is the Box-Cox power transf.
- When $q \to 1+$ ($\alpha \to 0+$) we obtain the logarithmic score

Proper scoring rules: Examples

- Finally, an interesting so-called *local* scoring rule (which depends on F_θ only through its value at x , and the value of a finite number of its derivatives at x) is the *Hyvärinen score* (Hyvärinen, 2005)

$$S_H\left(x, F_\theta\right) = 2\frac{\partial^2}{\partial x^2}\ln f\left(x,\theta\right) + \left|\frac{\partial}{\partial x}\ln f\left(x,\theta\right)\right|^2$$

- This has the convenient property that it can be computed without knowledge of the normalising constant of the density
- It can also be defined for multivariate distributions

Asymptotics

Theorem. Under suitable regularity conditions, e.g., Barndorff-Nielsen and Cox (1994), an M-estimator (proper scoring rule) estimator $\hat{\theta}_n$ is asymptotically normal with mean θ_G (least false parameter), and variance V, i.e.,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_G\right) \xrightarrow{D} N\left(0, V\right)$$

where

$$V = J^{-1} K J^{-T}; \quad J = E_G\left\{\frac{\partial s(\theta)}{\partial \theta^T}\right\}\Bigg|_{\theta=\theta_G}; \quad K = E_G\left\{s\left(\theta_G\right) s\left(\theta_G\right)^T\right\}$$

When $G = F_{\theta_0}$, the above theorem holds with $\theta_G = \theta_0$ (true parameter)

Asymptotics

- The matrix V^{-1} is known as the Godambe information matrix (Godambe, 1960)
- In general for M-estimators (scoring rule estimators) $J \neq K$
- In the special case of the logarithmic score (MLE), and when the model is exact, i.e., $G = F_\theta$, we have that

$$V^{-1} = J = K$$

is the Fisher information matrix. In this case

$$V_M = J_M^{-1} K_M J_M^{-T} \geq V_{MLE} = J_{MLE}^{-1}$$

Robustness

- The *influence function* (IF) of an estimator measures the effect on it of a small contamination at a point x
- The supremum of the IF over the data space measures the worst influence of such contamination, so supplying a measure of gross-error sensitivity
- A desirable property for a statistical procedure is that this gross-error sensitivity is *finite*, i.e. that IF is *bounded*
- This is termed *B-robustness*

Robustness

- From the general theory of M-estimators (Huber and Ronchetti, 2009), the IF of an M- or scoring rule estimator $\hat{\theta}$ is given by

$$\mathrm{IF}\left(x; s, G\right) = J^{-1} s\left(x, \theta_G\right)$$

- Thus, if $s(x,\theta)$ is bounded in x for each θ, then the M- or scoring rule estimator $\hat{\theta}$ is B-robust
- The IF can also be used to evaluate the asymptotic variance of $\hat{\theta}$ since

$$V = E_G\left\{\mathrm{IF}\left(X; s, G\right)\mathrm{IF}\left(X; s, G\right)^T\right\}$$

Robustness: Bregman divergence estimator

- A necessary and sufficient condition for B-robustness of the Bregman estimator, where s is given by

$$-s(x,\theta) = \lambda(x,\theta) - E_\theta \lambda(X,\theta) \quad \text{with} \quad \lambda(x,\theta) = \varphi''\{f(x,\theta)\}\nabla_\theta f(x,\theta)$$

  is the following (Dawid et al. (2014a), Condition 5.1):

  For all θ, $\lambda(x,\theta)$ *is a bounded function of* x

- This condition *combines* properties of the Bregman divergence generating function φ, and the form of the model $f(x,\theta)$

Robustness: Bregman divergence estimator

- A sufficient condition for B-robustness of the Bregman estimator is

  Dawid et al. (2014a), Condition 5.2:
  (i) The Bregman gauge φ'' *is locally bounded (on all intervals (0,M))*
  (ii) Both $f(x,\theta)$ *and* $\nabla_\theta f(x,\theta)$ *are bounded in* x*, for each* θ

- Tsallis score: $\varphi(t) = t^q;\ \varphi''(t) \propto t^{q-2}$ locally bounded for $q \geq 2$
- Log score: $\varphi(t) = t \ln t;\ \varphi''(t) = 1/t$ *not* locally bounded (on (*0,M*))

Numerical example

- We have tested the performance of the Tsallis score/BHHJ power divergence estimator for estimating the mean and scale (sdev) for a Normal distribution using uncontaminated and contaminated data
- We draw 100 samples from a N(0,1) (uncontaminated case) and from a contaminated distribution 0.95*N(0,1) + 0.05*N(3,1)
- Repeating this 1000 times and look at the statistics of estimates

- We compare Tsallis/BHHJ estimates for q = 1.2, 1.5 and 2.0 with MLE

Numerical example (cont.)

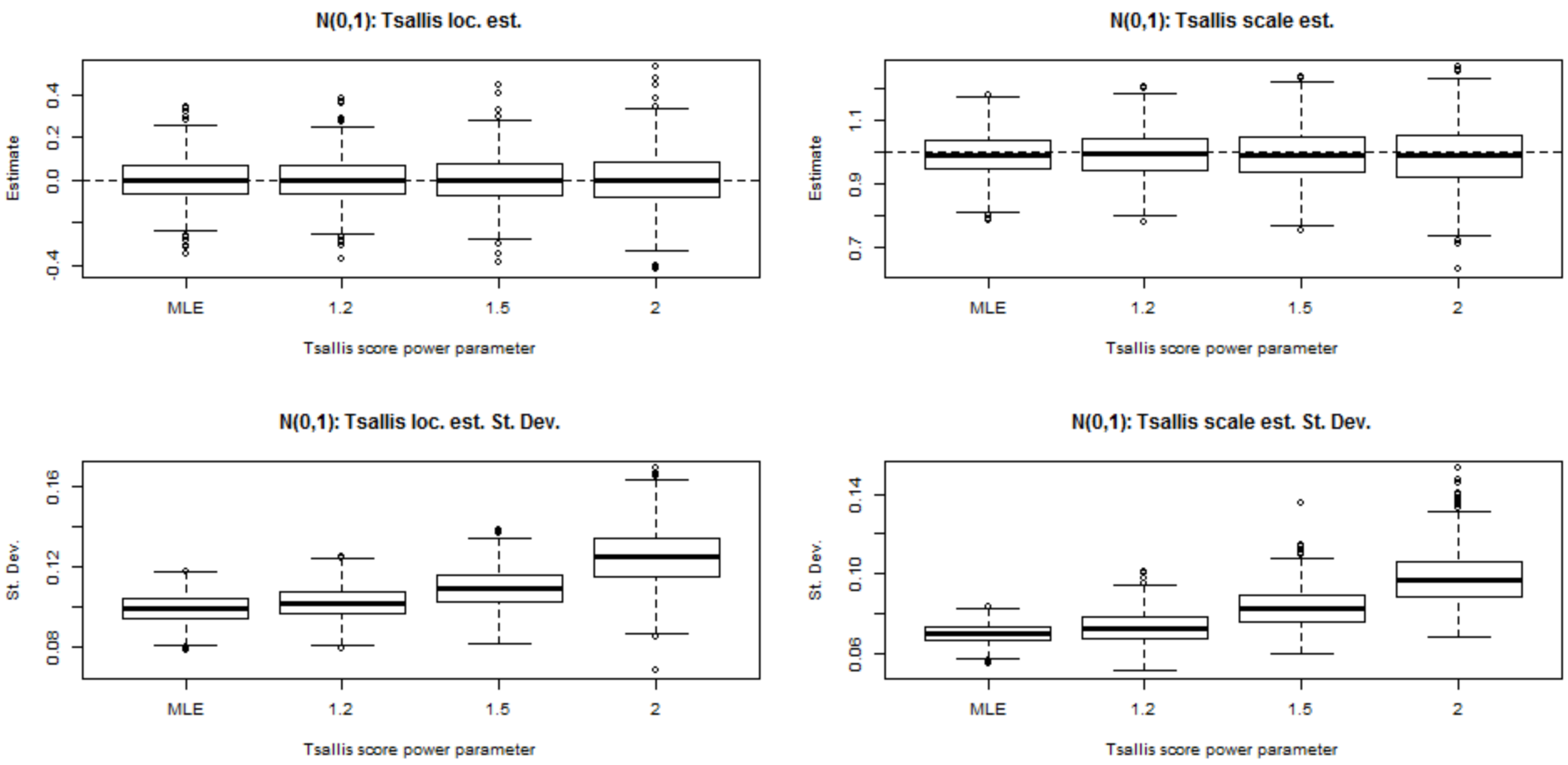
- We use the asymptotic normality of this particular M-estimator to construct coverage distributions and curves using

$$\mathrm{cc}(\theta) = \left| 1 - 2\Phi\left(\frac{\theta - \hat{\theta}_n}{\hat{\sigma}_n} \right) \right|$$

where $\hat{\sigma}_n = \mathrm{sqrt}\left(\mathrm{diag}\left(\hat{V}_n\right)_{ii}\right)$ for $i = 1, 2$, with $\hat{V}_n = \frac{1}{n} \hat{J}_n^{-1} \hat{K}_n \hat{J}_n^{-T}$.

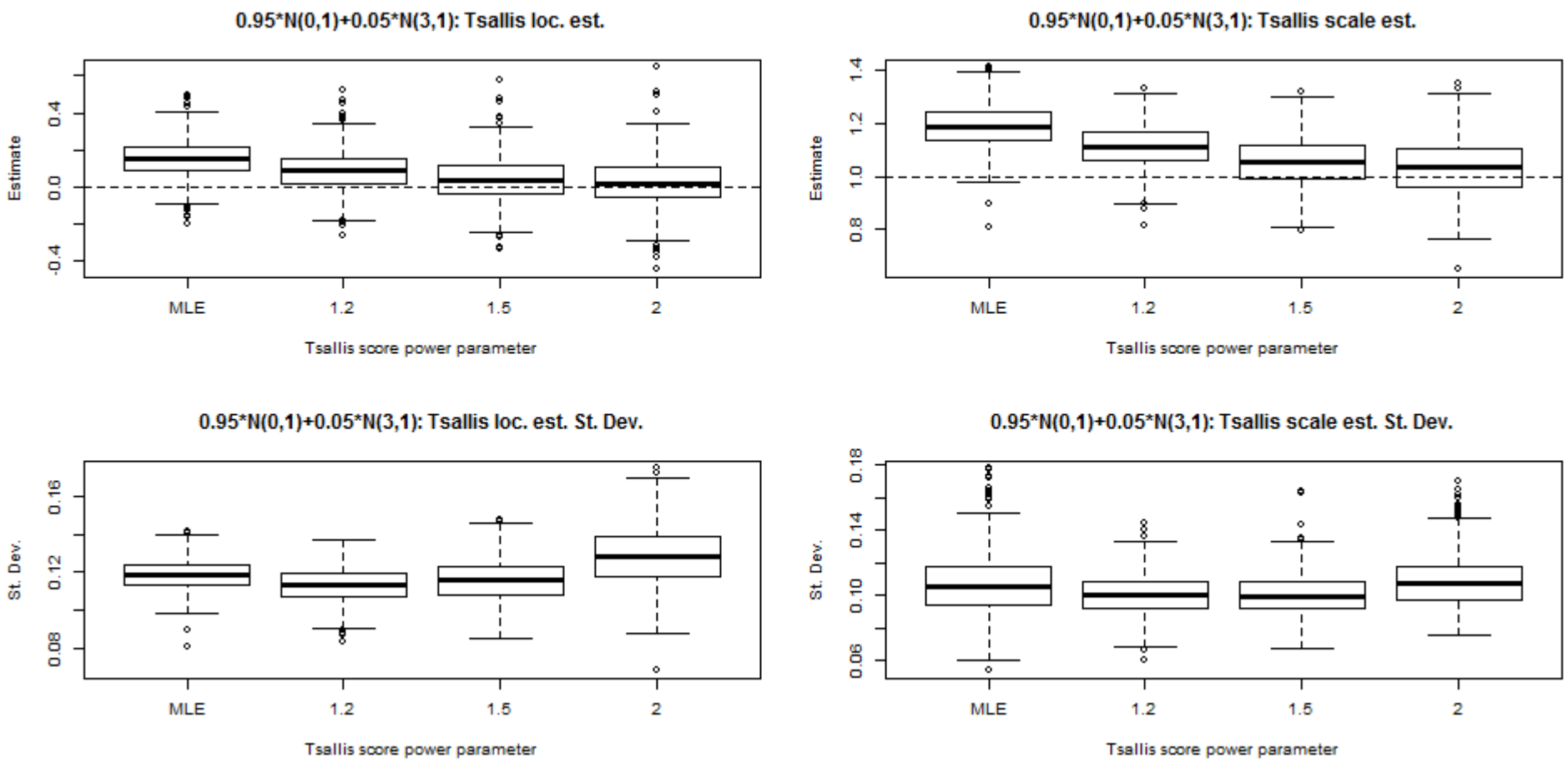
- Note that we use the “sandwich” matrix also for MLE for contaminated data since the model is not exact!

Uncontaminated case: N(0,1)



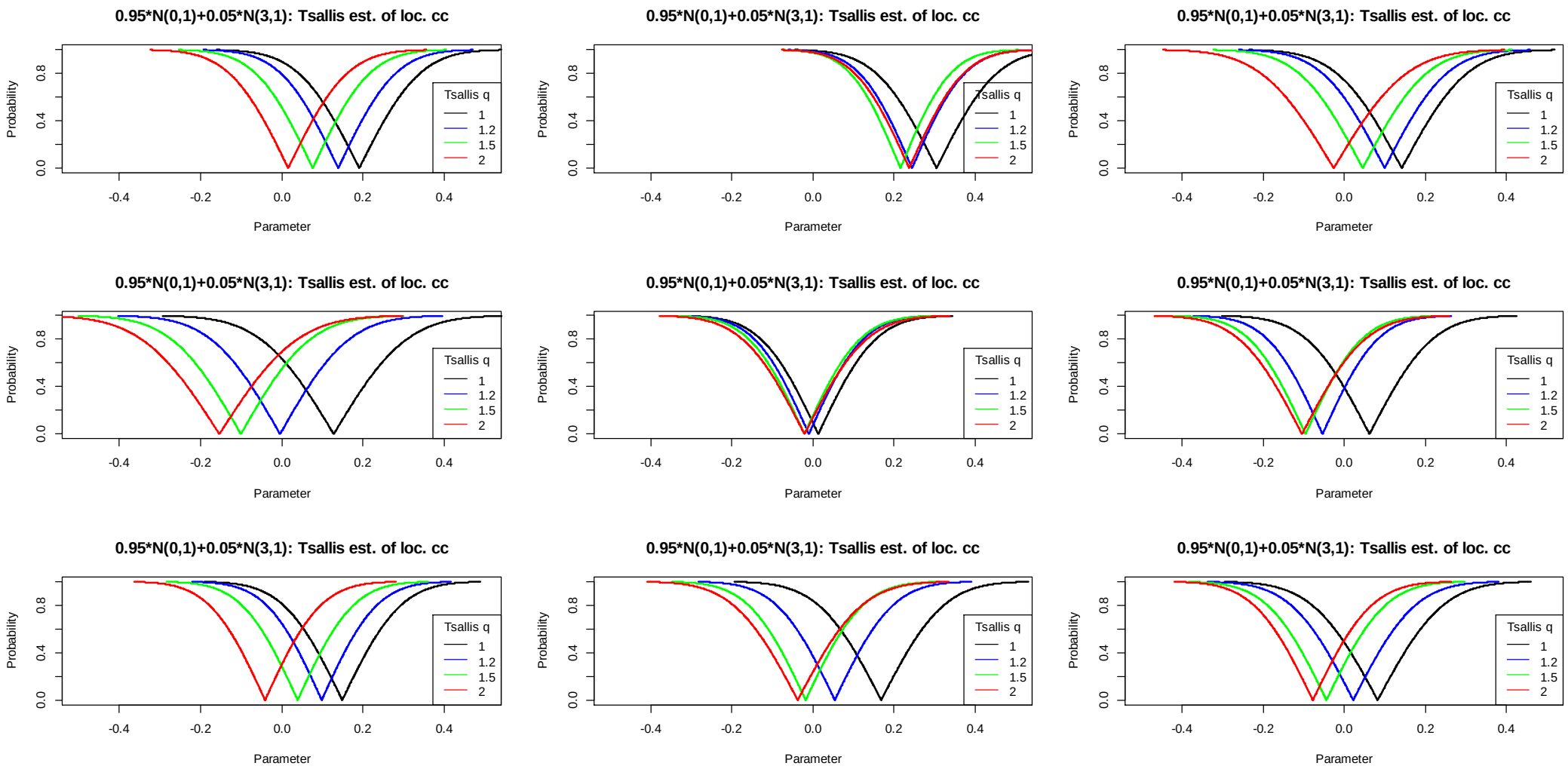
Location n = 100 #sim = 1000 Scale

Contaminated case: 0.95*N(0,1) + 0.05*N(3,1)



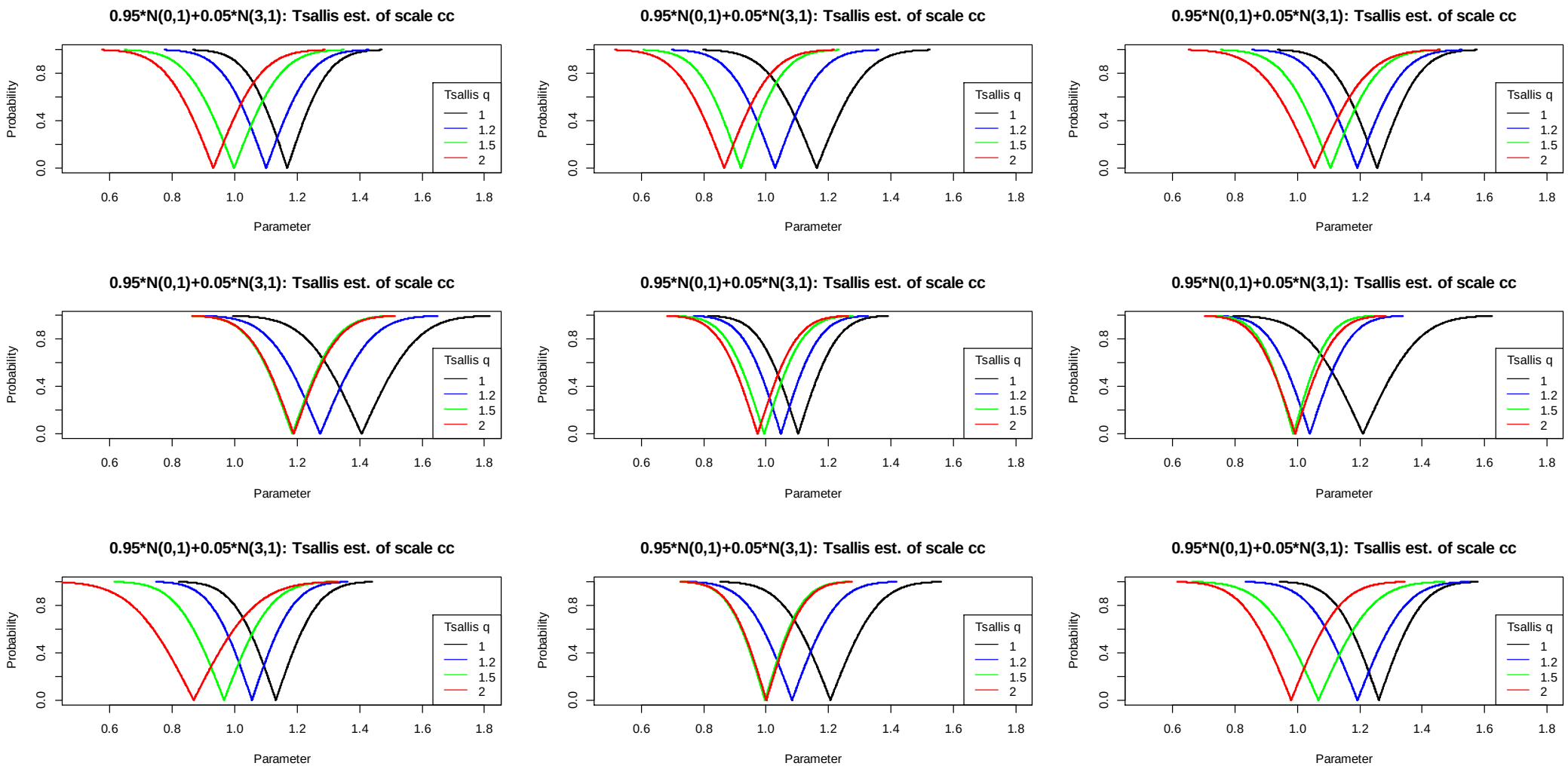
Location n = 100 #sim = 1000 Scale

0.95*N(0,1) + 0.05*N(3,1): Confidence curves for location parameter



n = 100 sim = 1-9

0.95*N(0,1) + 0.05*N(3,1): Confidence curves for scale parameter



n = 100 sim = 1-9

0.95*N(0,1) + 0.05*N(3,1)
Coverage probabilities in % for location = 0

Cov. Prob. %	MLE	Tsallis q = 1.2	Tsallis q = 1.5	Tsallis q = 2.0
50	21.4	40.7	49.7	50.5
75	42.9	63.5	73.3	76.2
90	67.4	82.6	88.3	90.1
95	78.5	90.1	93.3	93.3
99	93.1	97.1	98.3	98.7

n = 100 #sim = 1000

0.95*N(0,1) + 0.05*N(3,1)
Coverage probabilities in % for scale = 1

Prob. %	MLE	Tsallis q = 1.2	Tsallis q = 1.5	Tsallis q = 2.0
50	5.4	27.2	42.6	49.4
75	17.2	48.9	70.5	73.2
90	42.1	75.2	87.6	89.5
95	61.7	87.1	94.1	95.3
99	90.7	97.4	98.8	99.2

n = 100 #sim = 1000

Concluding remarks

- M-estimators represent a very general and versatile approach to the problem of parametric estimation, with an important sub-class formed by proper and strictly proper scoring rules (Dawid et al., 2014a, b; Gneiting et al., 2007)

- We may lose some efficiency as compared with MLE, but obtain *improved robustness,* and in some cases they may offer *computational advances*

- Most of the theory of MLE also applies to M- or scoring rule estimation with *little or no modification*, and thus can be used to construct hypothesis tests and confidence intervals or distributions (Dawid et al., 2014a)

Concluding remarks

- Recent simulation studies (Dawid et al., 2014a) indicate that (adjusted) scoring and likelihood-ratio type statistics yield confidence regions whose coverage properties are satisfactory

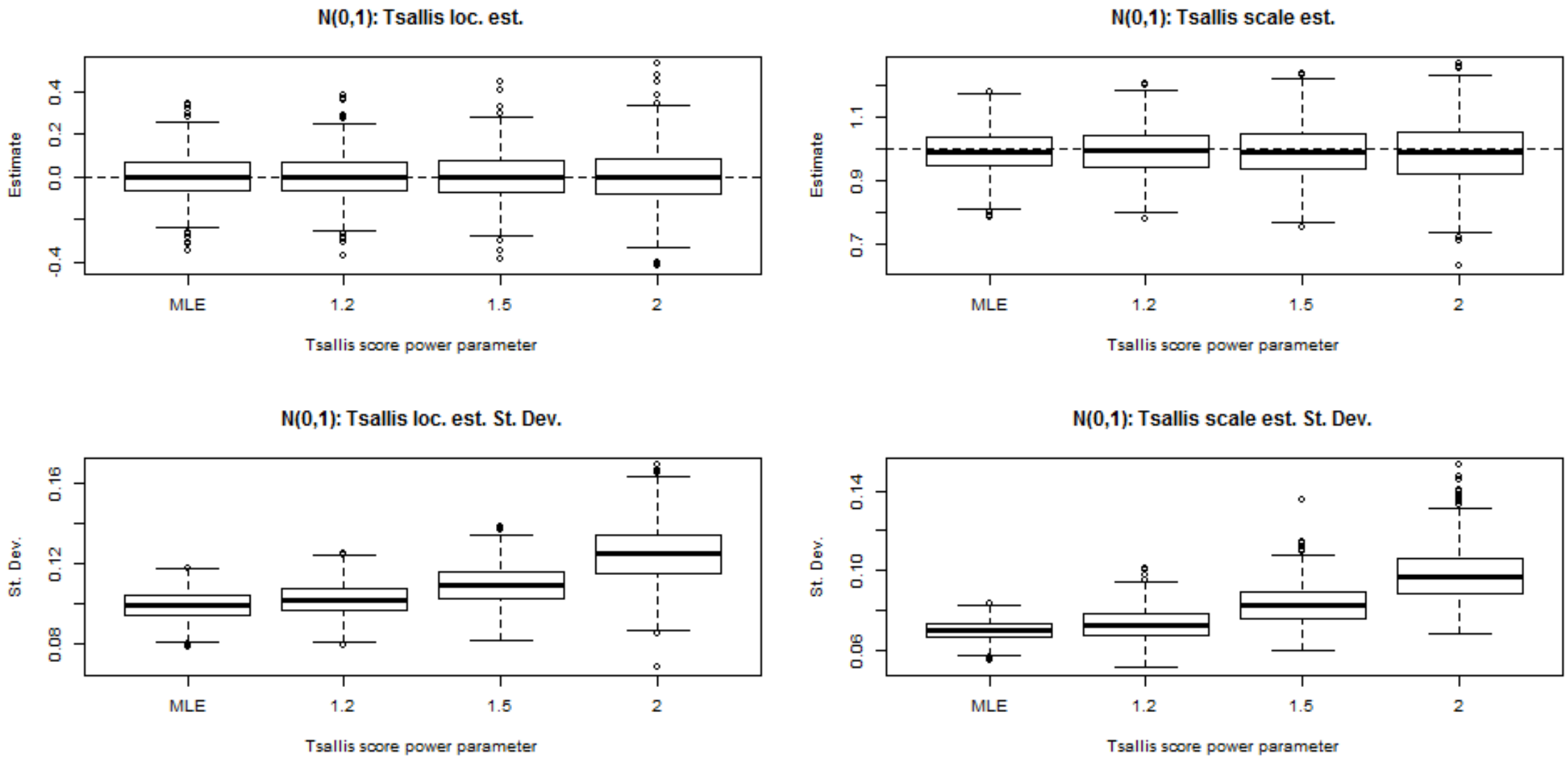
- The Tsallis score / BHHJ power divergence estimators offers *robustness with negligible loss of efficiency* for a large class of both univariate and multivariate parametric distributions

- The Hyvärinen score offers to estimate parameters in situations where the normalising constant is unavailable, or difficult to extract, but is *not* robust

References

- Mameli, V., Ventura, V. (2015) Higher-order asymptotics for scoring rules. *J. Stat. Plan. Inference*.
- Dawid, P., Musio, M., Ventura, L. (2014a) Minimum Scoring Rule Inference. arXiv:1403.3920v1 [math.ST] 16 Mar 2014.
- Dawid, P., Musio, M (2014b) Theory and applications of proper scoring rules, arXiv: 1401.0398v1 [math.ST] 02 Jan 2014
- Gneiting, T., Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc*., Vol. 102, No. 477.
- Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C. (1998) Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85, 549-559.

Thanks for your attention!

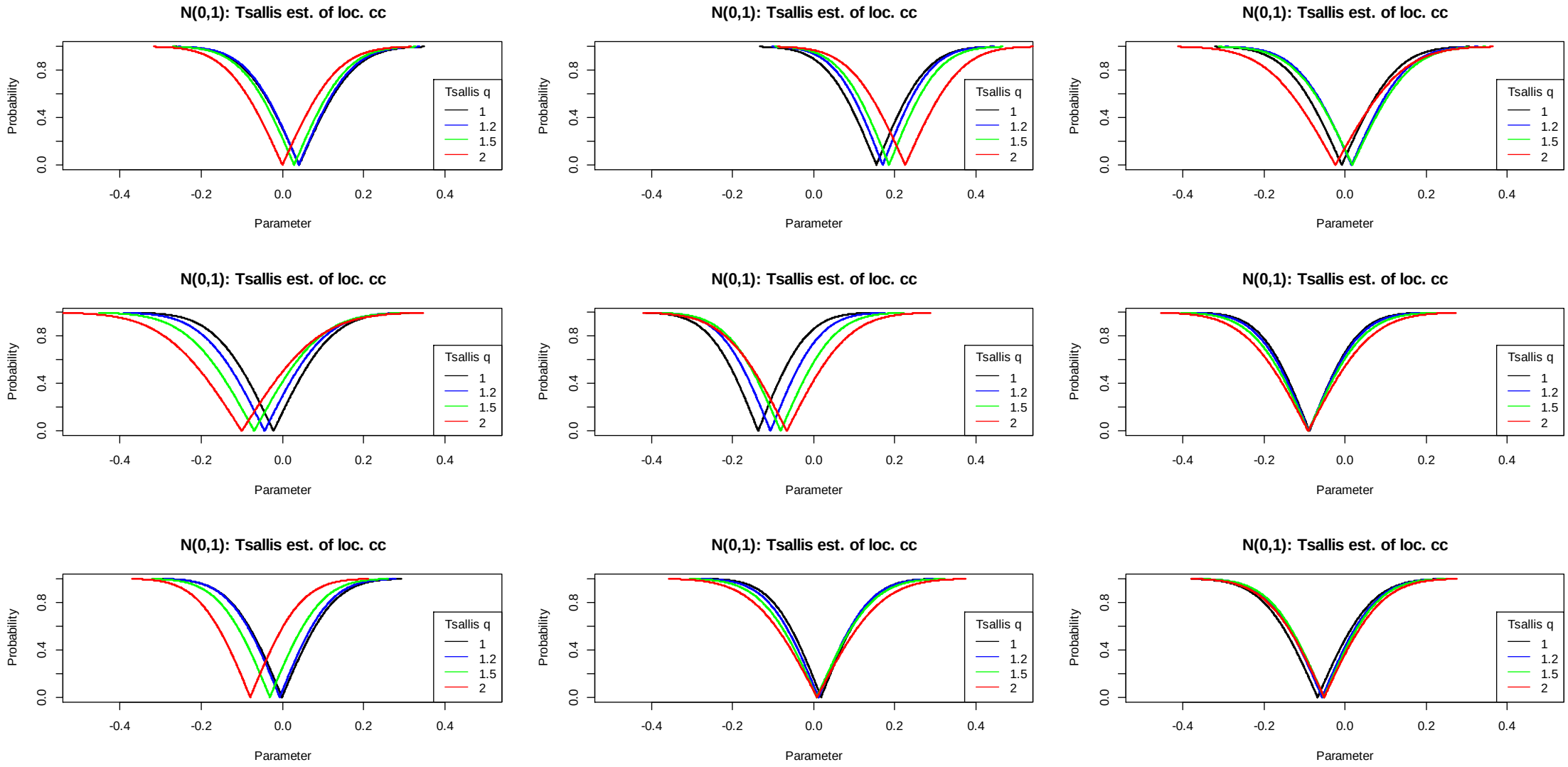

N(0,1): Tsallis loc. est.
Estimate
-0.4
0.0
0.2
0.4
MLE
1.2
1.5
2
Tsallis score power parameter
N(0,1): Tsallis scale est.
Estimate
0.7
0.9
1.1
MLE
1.2
1.5
2
Tsallis score power parameter
N(0,1): Tsallis loc. est. St. Dev.
St. Dev.
0.08
0.12
0.16
MLE
1.2
1.5
2
Tsallis score power parameter
N(0,1): Tsallis scale est. St. Dev.
St. Dev.
0.06
0.10
0.14
MLE
1.2
1.5
2
Tsallis score power parameter

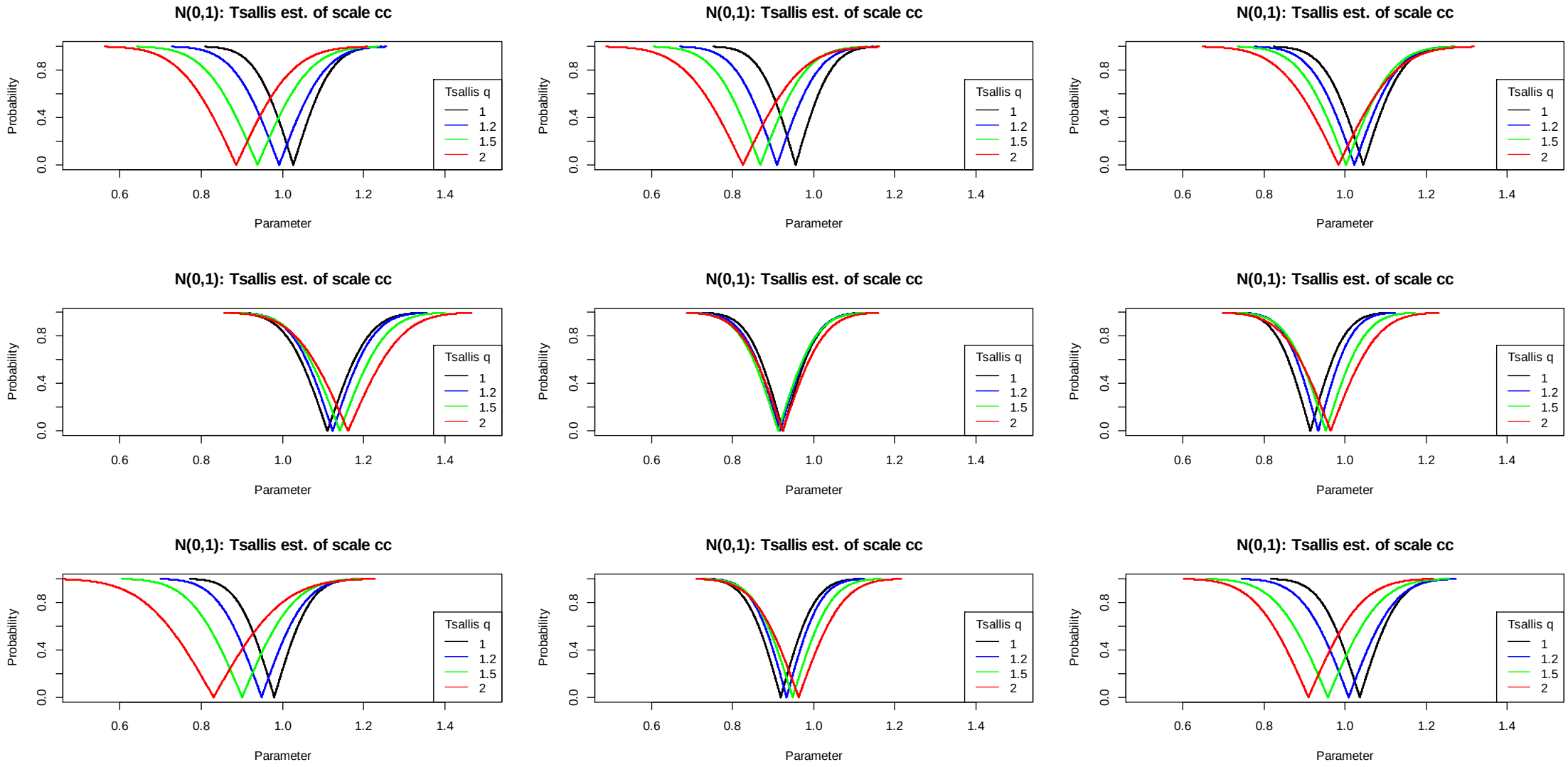
Uncontaminated data
Coverage probabilities in % for location

Cov. Prob. %	MLE	Tsallis q = 1.2	Tsallis q = 1.5	Tsallis q = 2.0
50	51.0	51.8	51.1	51.2
75	74.9	75.1	77.0	77.3
90	88.9	90.1	89.8	90.2
95	94.2	94.3	94.6	94.4
99	98.8	98.3	98.5	98.2

Uncontaminated data
Coverage probabilities in % for scale

Prob. %	MLE	Tsallis q = 1.2	Tsallis q = 1.5	Tsallis q = 2.0
50	50.1	48.7	47.7	50.2
75	75.7	75.3	73.0	73.8
90	89.9	90.3	89.9	89.1
95	94.8	94.9	94.8	94.5
99	98.7	98.7	99.0	99.1


N(0,1): Tsallis est. of loc. cc
Probability
Parameter
Tsallis q
1
1.2
1.5
2


N(0,1): Tsallis est. of scale cc
Probability
Parameter
Tsallis q
1
1.2
1.5
2