ROBUST ESTIMATION OF CRONBACH'S ALPHA

Andreas Christmann
University of Dortmund, HRZ

Stefan Van Aelst
University of Ghent

ICORS, Vancouver, May 2002

APPLICATION

- Evaluation: teaching abilities of german teachers
  - A. Nolle, University of Dortmund, IFS
  - knowledge of different learning strategies
  - social skills
  - approx. 800 questionnaires
  - hundreds of variables: ordinal, some metric
- Aims: scoring, measuring reliability, comparison of subgroups

CRONBACH'S ALPHA

- Cronbach (1951)
- summarizing the information of several items
- questionnaires
- scoring
- popular method to measure reliability for a set of items
- implemented in: SAS, SPSS, ...
- easy to compute in R, S-PLUS

Items $Y_j = T_j + \varepsilon_j$ for $j = 1, \ldots, p$, where

- T_j true unobservable test scores
- ε_j error terms, $\mathrm{E}(\varepsilon_j) = 0$
- T_j and ε_j indep.

$$\text{Score}: \quad Z = \sum_{j=1}^{p} Y_j$$

Definition: Cronbach's alpha

$$\alpha_n^C = \frac{p}{p-1} \frac{\mathrm{Var}\left(\sum_{j=1}^{p} Y_j\right) - \sum_{j=1}^{p} \mathrm{Var}\left(Y_j\right)}{\mathrm{Var}\left(\sum_{j=1}^{p} Y_j\right)}$$

$$= \frac{p}{p-1} \frac{\sum \sum_{i \neq j} \mathrm{Cov}(Y_i, Y_j)}{\mathrm{Var}\left(\sum_{j=1}^{p} Y_j\right)}$$

$$= \frac{p}{p-1} \left[1 - \frac{\sum_{j=1}^{p} \mathrm{Var}(Y_j)}{\sum_{j=1}^{p} \sum_{k=1}^{p} \mathrm{Cov}(Y_j, Y_k)} \right]$$

Gutman (1953): Cronbach's alpha is a lower bound of reliability

CLASSICAL ESTIMATION OF CRONBACH'S ALPHA

- α_C^S based on empirical covariance matrix

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^t$$

- S non-robust (e.g. Hampel et al. '86)

- α_C^S non-robust:
  - does not measure the reliability of the bulk of the data
  - can be completely misleading if outliers are present

ROBUST ESTIMATION

- α_C^R based on robust estimator of the scatter matrix
  - measuring the reliability of the bulk of the data
  - robust against outliers
  - investigate: consistency, IF function, empirical IF, SC_n
- M-estimator based on $t-$distribution (Kent & Tyler '91)
- RMCD (Rousseeuw & Leroy '87, Rousseeuw & Van Driessen '99)
- S-estimator with Tukey's biweight function

  (Rousseeuw & Yohai '84)

ROBUST ESTIMATORS OF LOCATION AND SCATTER

data $\{y_i; 1 \leq i \leq n\}$ with $y_i = (y_{i1}, \ldots, y_{ip})^t \in \mathbb{R}^p$.

M-estimator T3

- Kent & Tyler '91: M-estimator based on t_3-distribution
- good efficiency
- existence of a unique solution of the objective criterion
- always converging iterative algorithm; fast
- can easily be explained to non-statisticians
- reasonable robustness, low breakdown point for large p

MCD (Rousseeuw & Leroy '87)

- Fix $\lceil n/2 \rceil \leq h \leq n$. Here $h = \lceil 0.75n \rceil$.

  Determine optimal subset: $\det \left(\mathrm{Cov}(\{y_{i_1}, \ldots, y_{i_h}\}) \right) = \inf$!

- $t_n^0 = \frac{1}{h} \sum_{j=1}^h y_{i_j}$, $\quad C_n^0 = c_n \frac{1}{h} \sum_{j=1}^h (y_{i_j} - t_n^0)(y_{i_j} - t_n^0)^t$

  Pison et al. '02 : c_n

- high breakdown point, bounded IF, but low asympt. efficiency

RMCD (Rousseeuw & van Zomeren '90)

- $t_n^1 = \frac{\sum_{i=1}^n w(d_i) y_i}{\sum_{i=1}^n w(d_i)}$, $C_n^1 = d_n \frac{\sum_{i=1}^n w(d_i)(y_i - t_n^1)(y_i - t_n^1)^t}{\sum_{i=1}^n w(d_i)}$.
- weights $w(d_i) = I(d_i \leq q_\delta)$ where $\delta = 0.025$, $q_\delta = \chi^2_{p+q,1-\delta}$, robust MCD distance $d_i = \sqrt{(y_i - t_n^0)^t (C_n^0)^{-1} (y_i - t_n^0)}$.
- high breakdown point, bounded IF, improved efficiency

  (Lopuhaä & Rousseeuw '91, Lopuhaä '99, Croux & Haesbroeck '99)
- can easily be explained to non-statisticians
- S-PLUS, SAS

  Rousseeuw & Van Driessen '99: Fast-MCD, SAS/IML

S-estimator S_{bw}

- (t_n^S, C_n^S) that minimizes $\det(C_n)$ under the constraint
  $\frac{1}{n}\sum_{i=1}^n \rho(\sqrt{(y_i - t_n)^t C_n^{-1}(y_i - t_n)}\,) \leq b$,
  $t_n \in \mathbb{R}^p$ and $C_n \in \mathrm{PDS}(p)$
- Tukey : $\rho_c(t) = \min\left(\frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4}, \frac{c^2}{6}\right)$, $t \in \mathbb{R}$.
- asympt. normal, bounded IF, improved asympt. efficiency
- algorithms: Ruppert '92, Rocke & Woodruff '93

ROBUST CRONBACH'S ALPHA

- unimodal elliptically symmetric distributions $F_{\mu,V}$

  $f_{\mu,V}(y) = \frac{g(y-\mu)^t V^{-1}(y-\mu)}{\sqrt{\det(V)}}$, $\mu \in \mathbb{R}^p$, $V \in \mathrm{PDS}(p)$, $g' < 0$

- $\alpha = \frac{p}{p-1} \frac{\sum\sum_{j \neq k} V_{jk}}{\sum\sum_{j,k} V_{jk}}$

- consistency of $\hat{V}$:

  MCD: Butler, Davies & Jhun '93

  RMCD: Lopuhaä '99

  S: Davies '87, Lopuhaä '89

- Consistency:

  If $C_n \to V$ (a.s. or P) then $\alpha_n^R(C_n) \to \alpha$ (a.s. or P)

- Functional C: $C(G) \in \mathrm{PDS}(p)$ such that $C(F_n) = C_n$

- Fisher consistency:

  $\alpha_C^R(F_{\mu,V}) = \alpha \quad$ if $\quad C(F_{\mu,V}) = V$

- Influence function

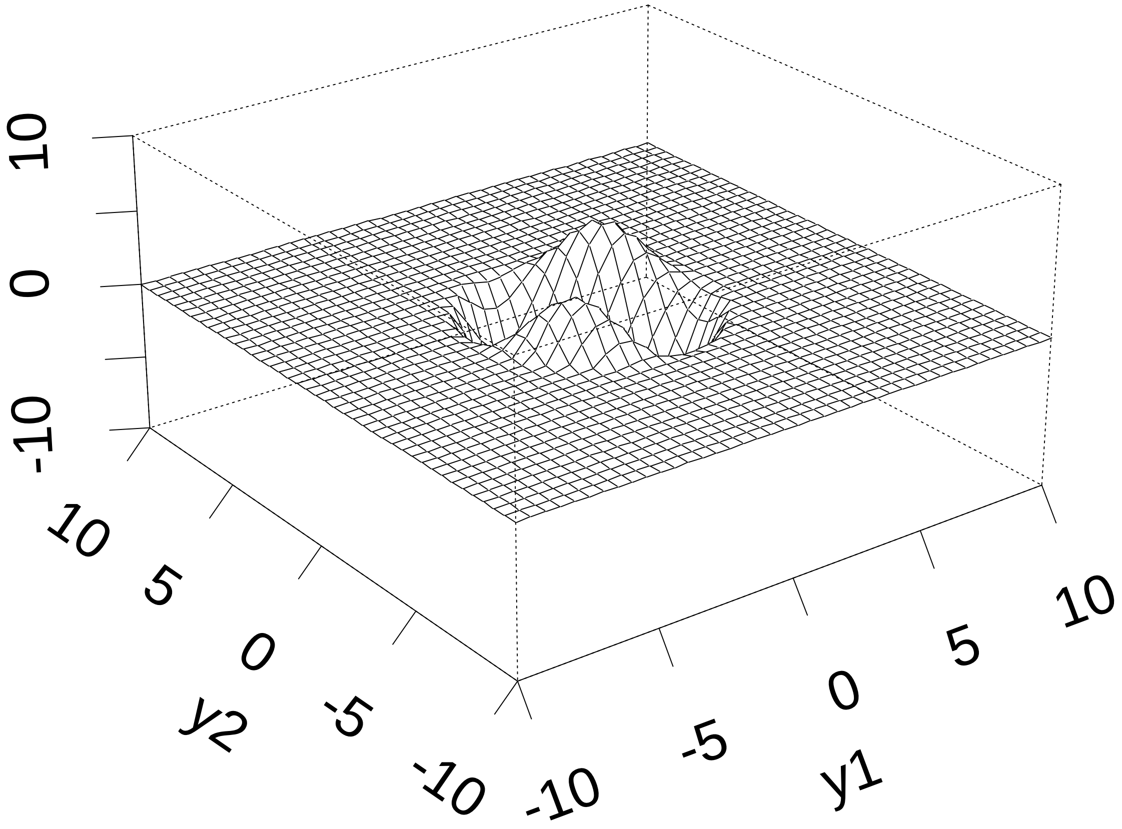
  $IF(y; \alpha_C^R, F_{\mu,V}) = \lim_{\varepsilon \downarrow 0} \frac{\alpha_C^R((1-\varepsilon)F_{\mu,V}+\varepsilon\Delta_y)-\alpha_C^R(F_{\mu,V})}{\varepsilon}$

- If the functional C for the scatter matrix has an IF then the IF of α_C^R at elliptically symmetric distributions $F := F_{\mu,V}$ is given by

$$IF(y; \alpha_C^R, F) = \frac{\frac{p}{p-1} \sum \sum_{j \neq k} \mathrm{IF}(y; c_{j,k}, F) - \alpha_C^R(F) \sum \sum_{j,k} \mathrm{IF}(y; c_{jk}, F)}{\sum \sum_{j,k} V_{jk}}.$$

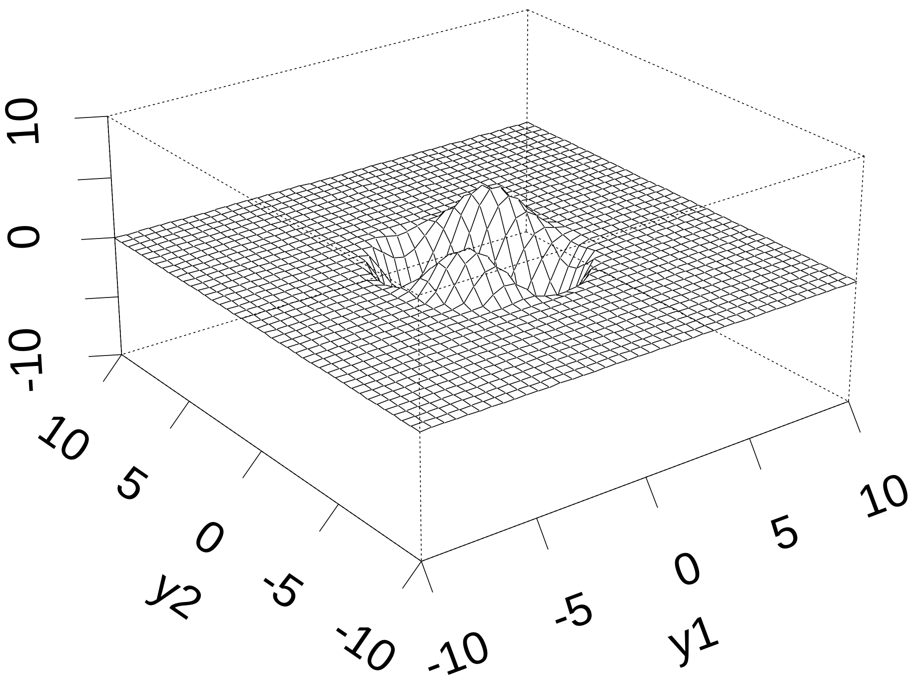
- IF of α_C^R is bounded $\quad$ if $\quad$ IF of C is bounded.

IF of Cronbach's alpha based on S_{bw}: $p = 2$, $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$

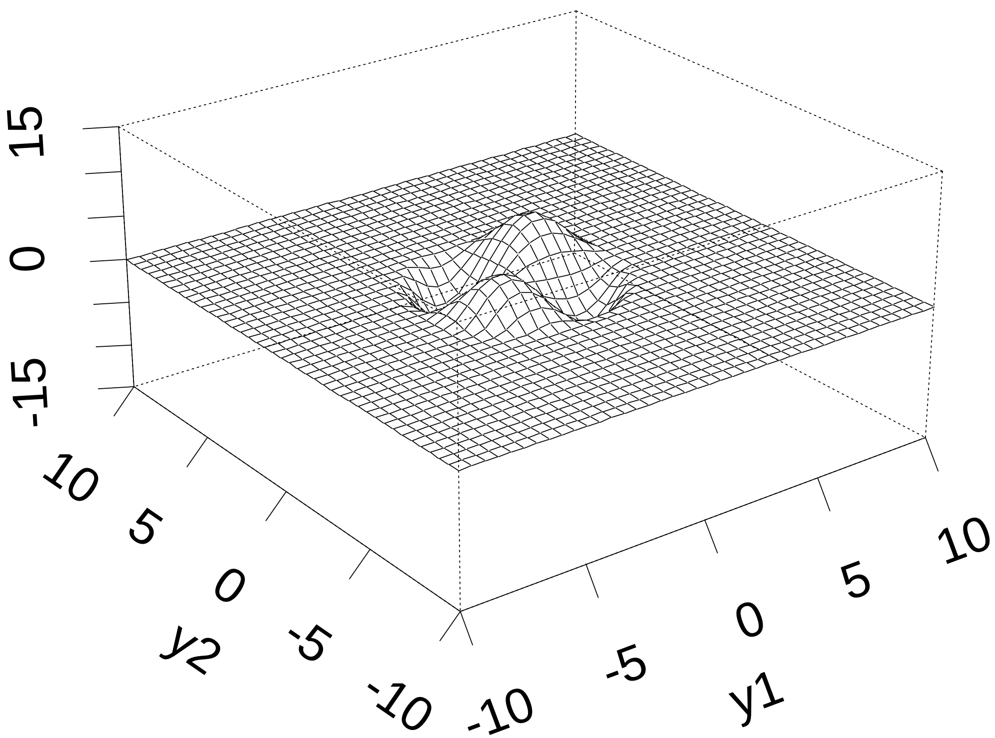


Similar to IF of largest EV in robust PCA based on S_{bw} (Croux & Haesbroeck, 2000)

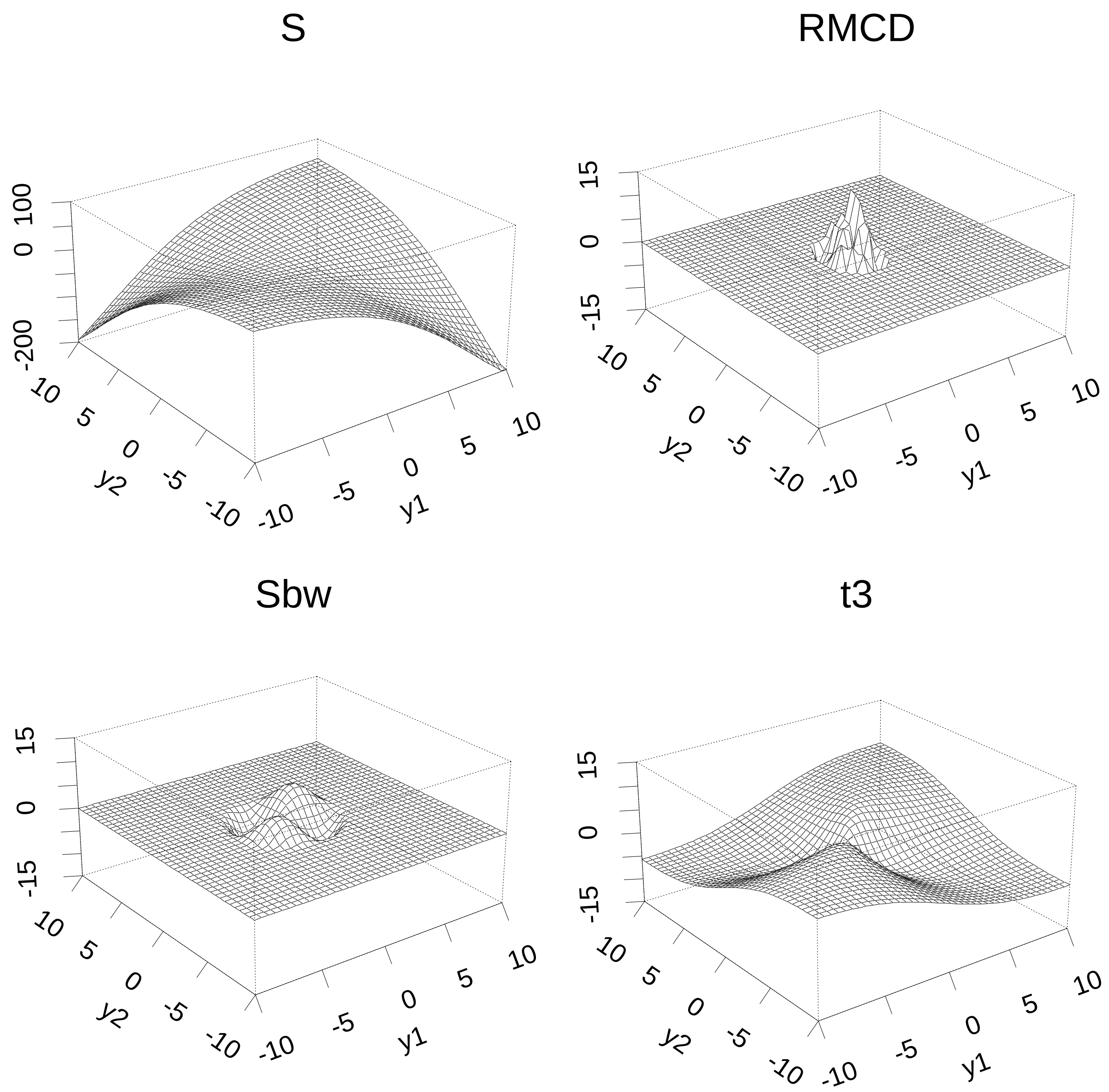
IF: S_{bw}, $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$



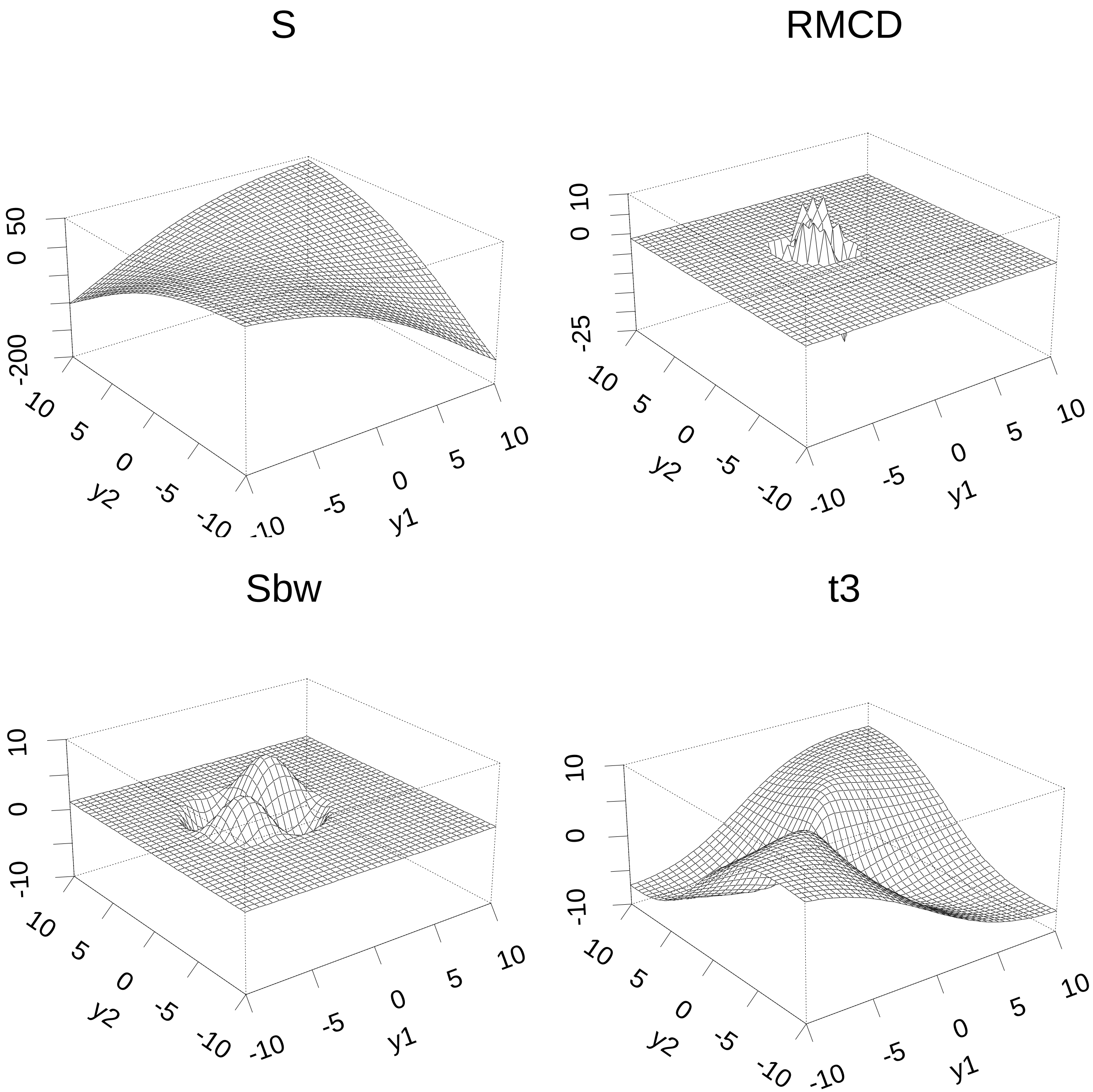
Sensitivity curve: S_{bw}, $n = 100$



Sensitivity curves: $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$, $p = 2$, $n = 100$



Sensitivity curves: $F = 10\% M3$, $\rho = 0$, $p = 2$, $n = 100$



SIMULATIONS

- Models:

  $\mathrm{N}(\mu, \Sigma)$

  $t_3(\mu, \Sigma)$

  $\delta\%$ **M1**: $(1-\delta)\mathrm{N}(\mu, \Sigma) + \delta\mathrm{N}(\mu, \Sigma_1)$

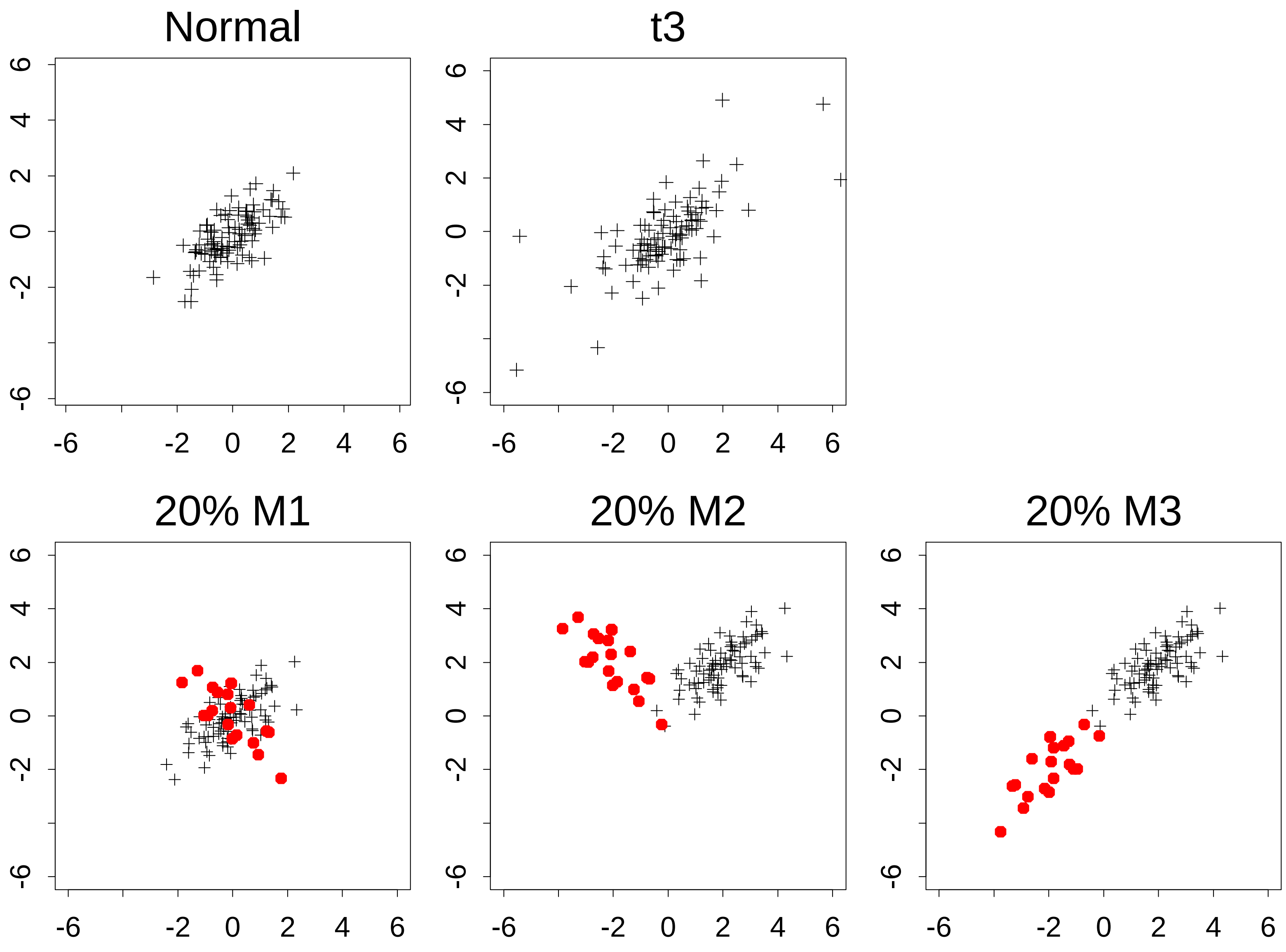
  $\delta\%$ **M2**: $(1-\delta)\mathrm{N}(\mu, \Sigma) + \delta\mathrm{N}(\mu_1, \Sigma_1)$

  $\delta\%$ **M3**: $(1-\delta)\mathrm{N}(\mu, \Sigma) + \delta\mathrm{N}(\mu_1, \Sigma)$

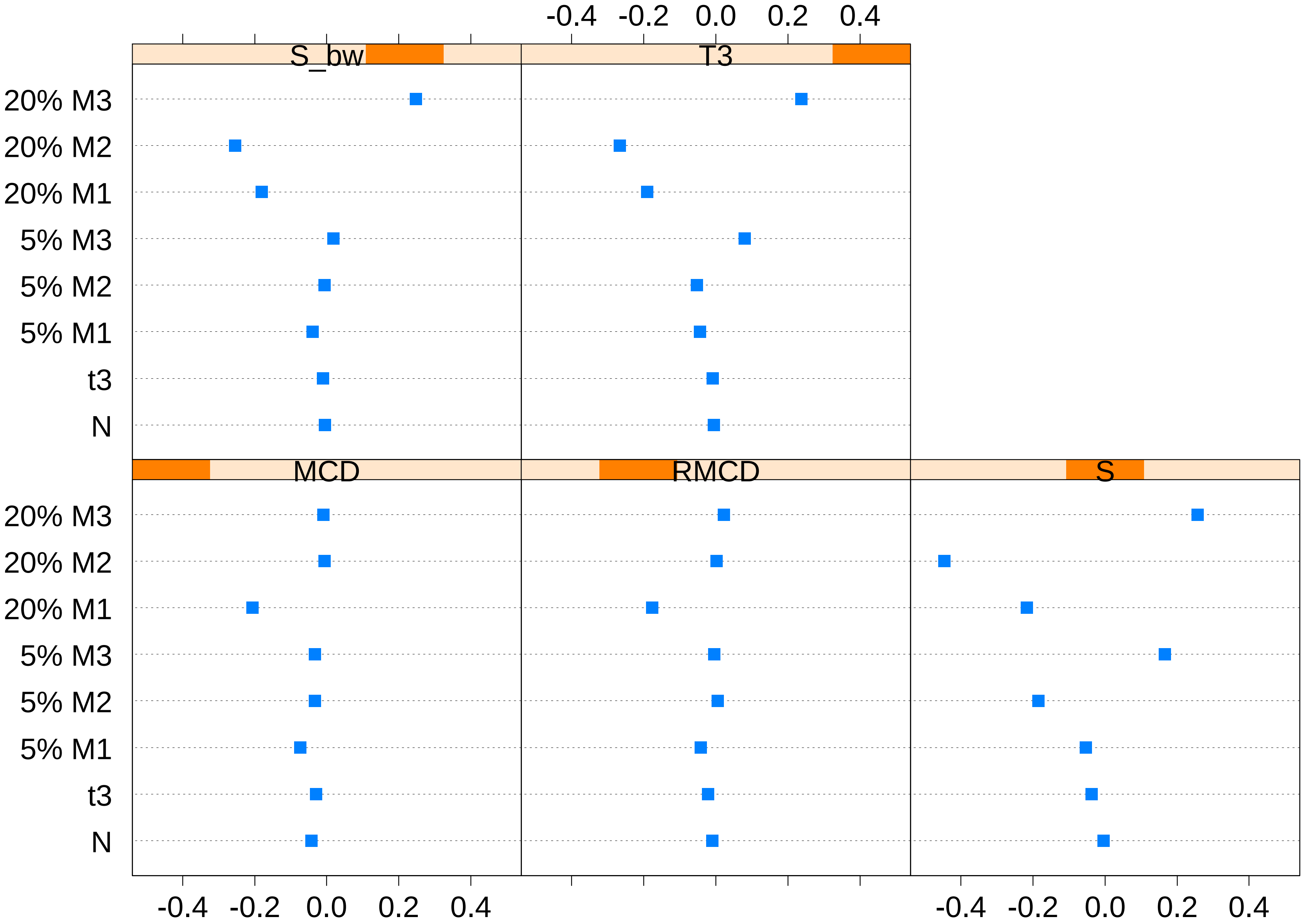
- $p = 2, 10,$ $\quad n = 40, 100, 500,$ $\quad 500$ replications

- $\delta = 5\%, 10\%, 20\%,$ $\quad$ correlation $\rho = 0, 0.2, 0.5, 0.8$

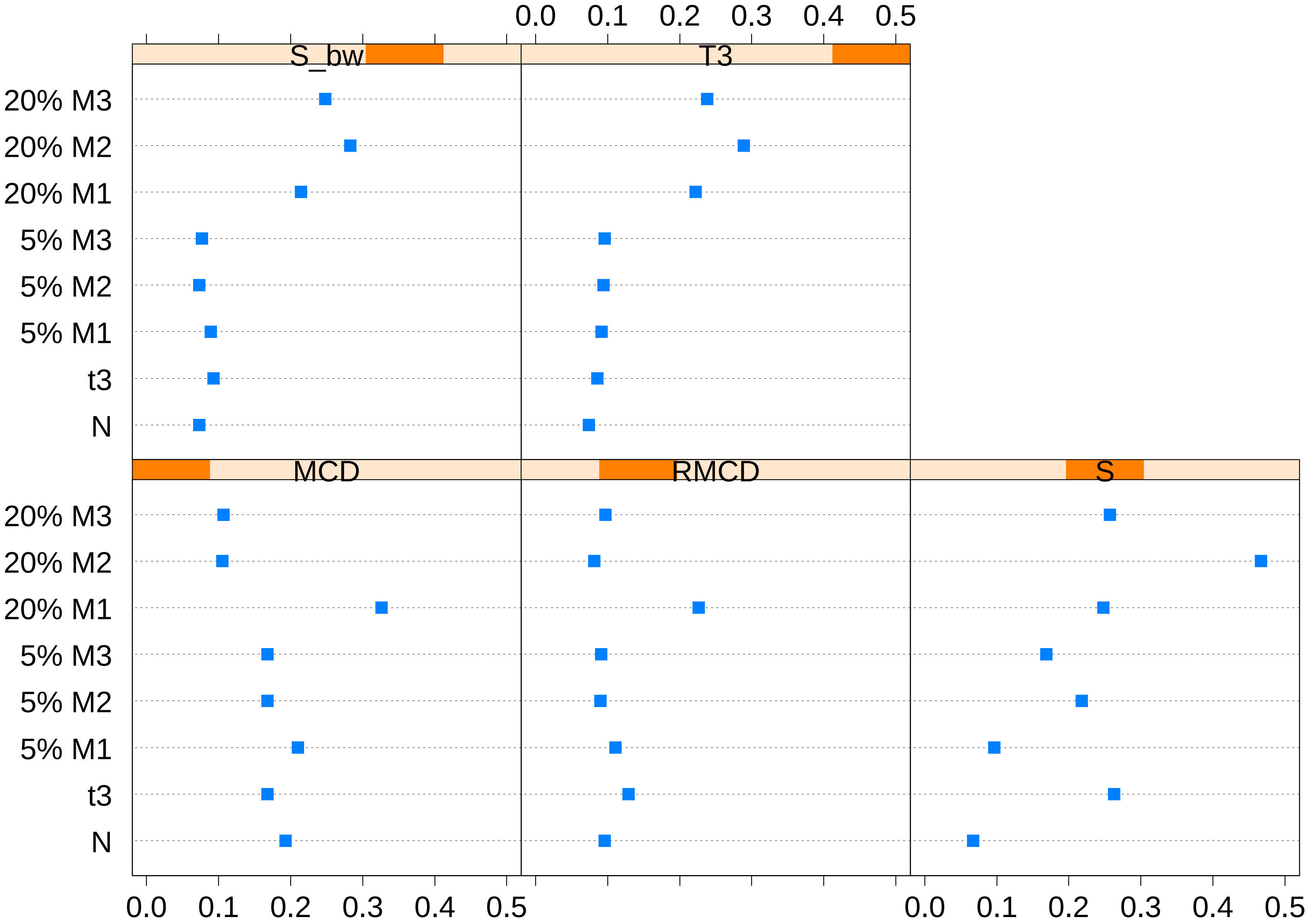
Scatterplots of simulated data: $p = 2$, $n = 100$, $\rho = 0.8$

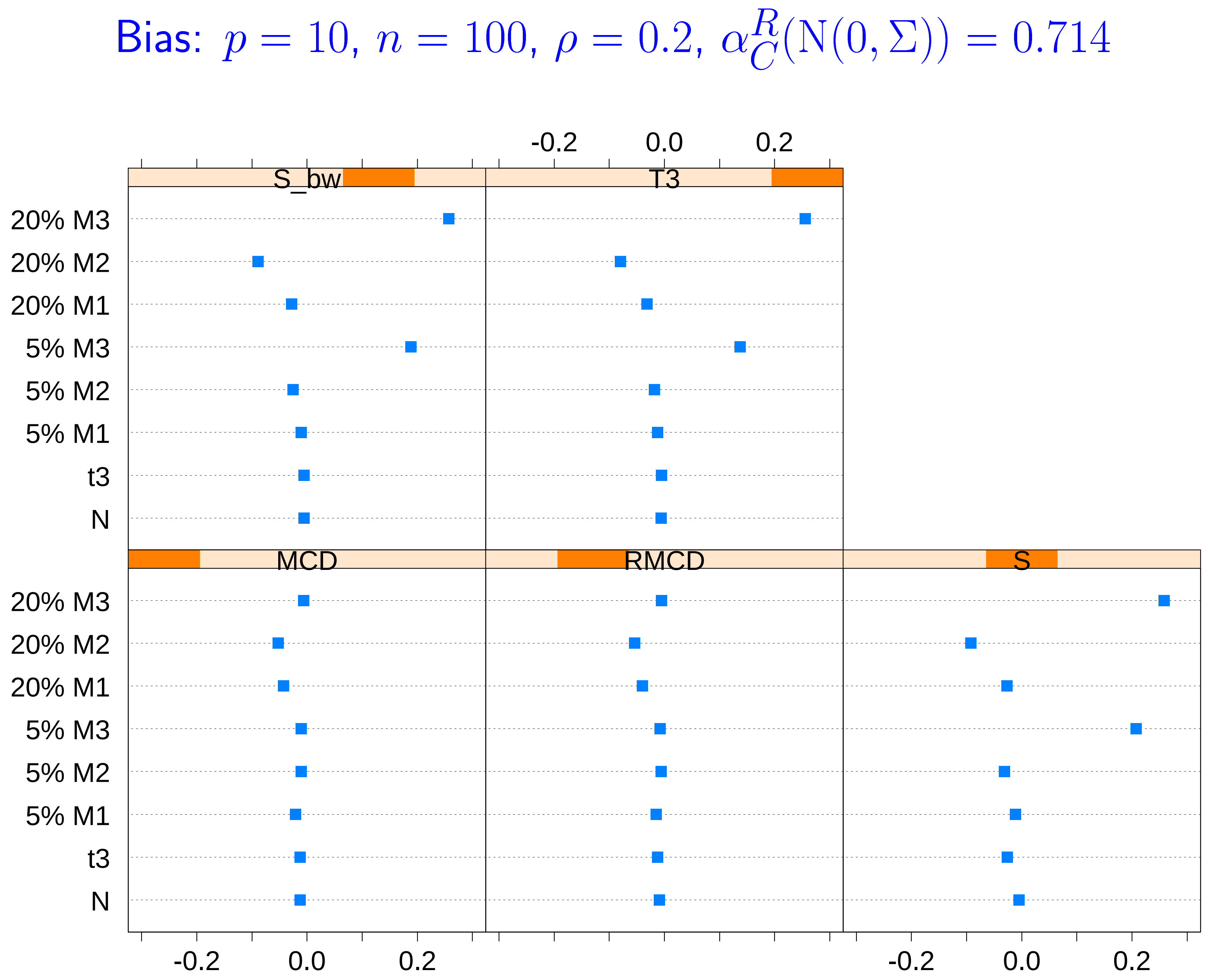


Bias: $p = 2$, $n = 100$, $\rho = 0.5$, $\alpha_C^R(\mathrm{N}(0,\Sigma)) = 0.667$



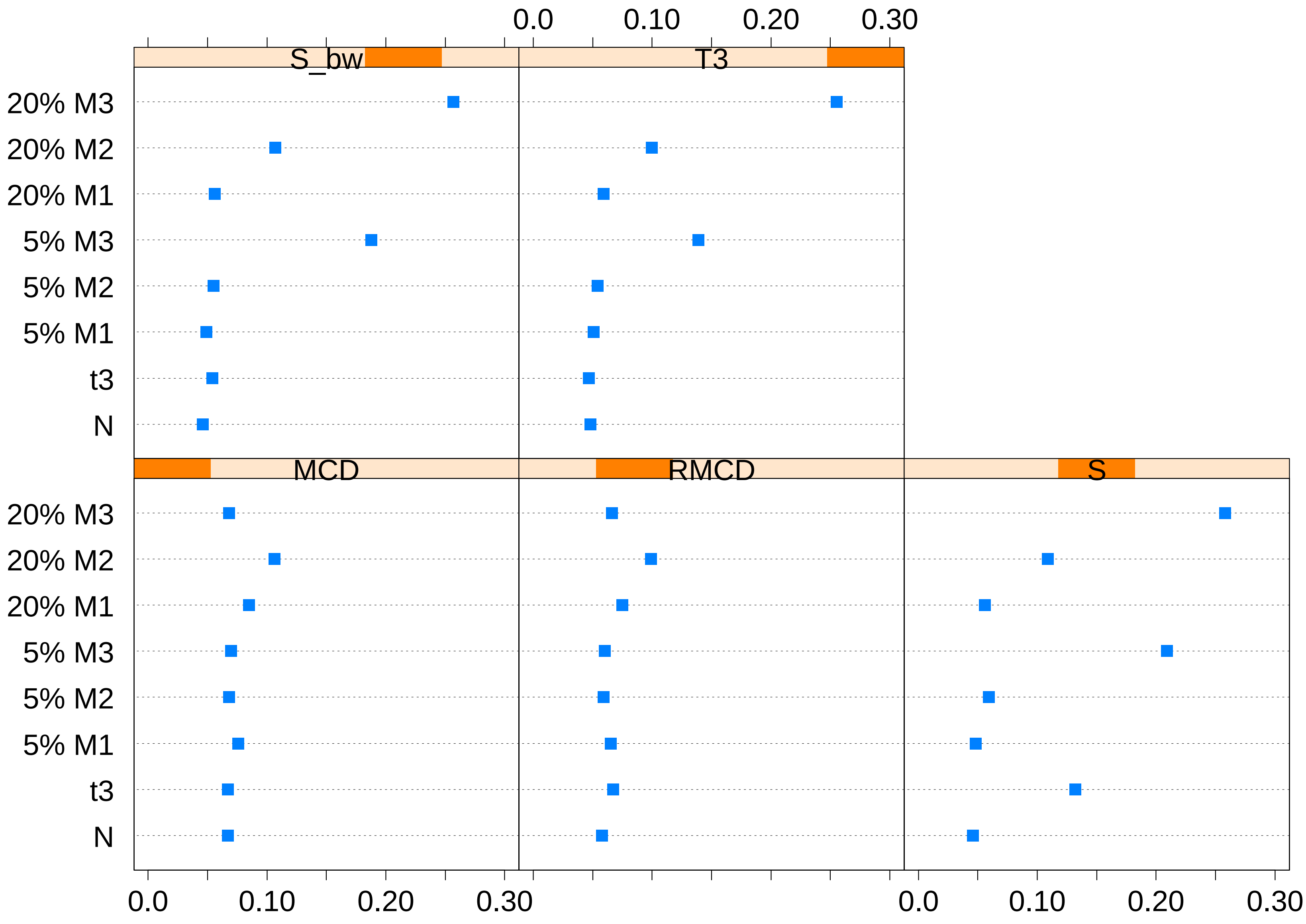
Square root of MSE: $p = 2,\ n = 100,\ \rho = 0.5,\ \alpha_C^R(\mathrm{N}(0,\Sigma)) = 0.667$





Bias: $p = 10$, $n = 100$, $\rho = 0.2$, $\alpha_C^R(\mathrm{N}(0,\Sigma)) = 0.714$
S_bw
T3
MCD
RMCD
S
20% M3
20% M2
20% M1
5% M3
5% M2
5% M1
t3
N
-0.2
0.0
0.2

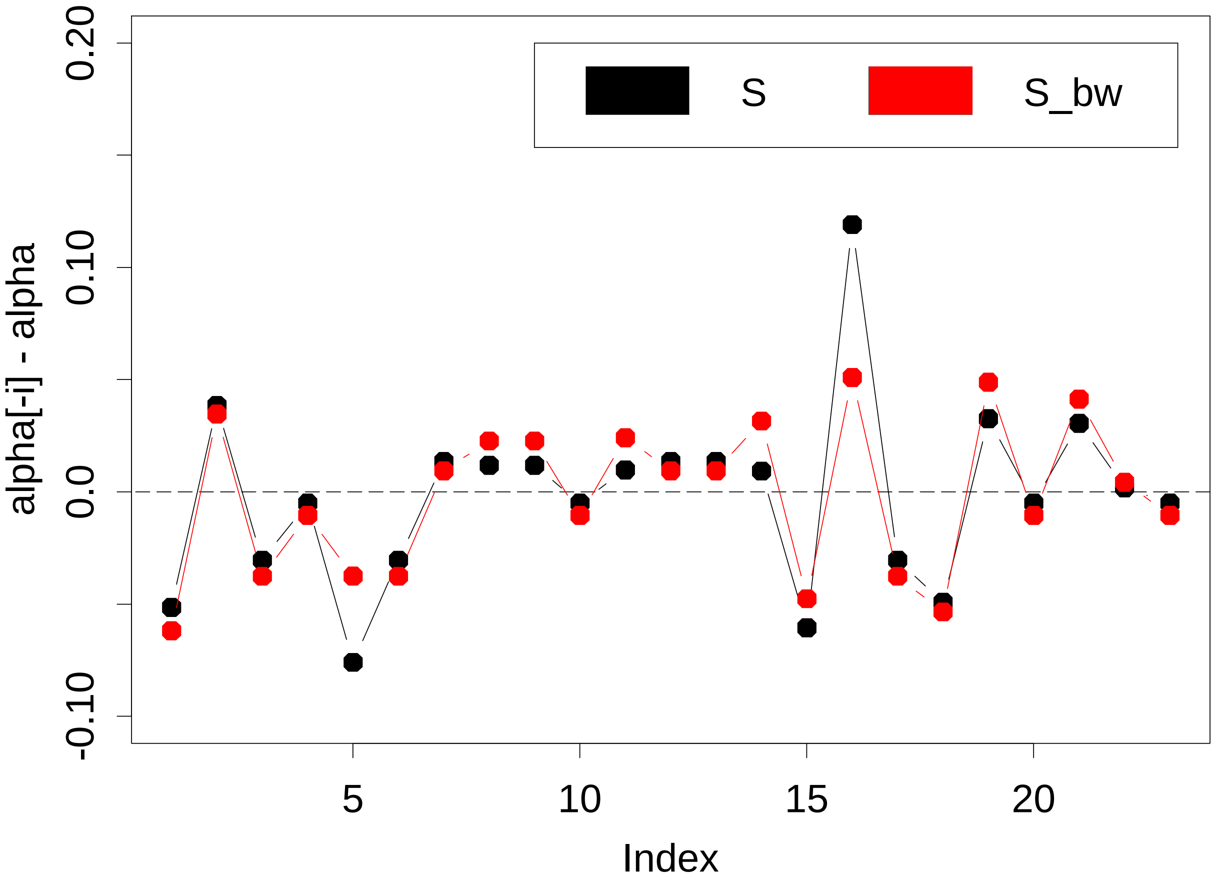
Square root of MSE: $p = 10$, $n = 100$, $\rho = 0.2$, $\alpha_C^R(\mathrm{N}(0,\Sigma)) = 0.714$



APPLICATION (cont.)

- a few seconds for $n \approx 800$ and $p = 14$
- often similar results, because many variables are ordinal

Sensitivity analysis: Bavaria ($n = 23, p = 3$)



CONCLUSION

- Classical Cronbach's alpha is not robust
- Robust versions based on robust estimators of the scatter matrix
- Using RMCD or $S_{bw}-$estimator:

  consistency, bounded IF, good results in simulations
- breakdown point of $\alpha_n^R(C_n) \leftrightarrow$ breakdown point of C_n ?
- asymptotical distribution, ... ?

REFERENCES

Butler, Davies, Jhun (1993). Asymptotics for the Minimum Covariance Determinant Estimator, *The Annals of Statistics*, 21, 1385–1400.

Cronbach (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334.

Croux, Haesbroeck (2000). "Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Function and Efficiencies. *Biometrika*, 87, 603–618.

Davies (1987). Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics*, 15, 1269–1292.

Kent, Tyler (1991). Redescending M-estimates of Multivariate Location and scatter. *The Annals of Statistics,* 19, 2102–2119.

Lopuhaä (1999). Asymptotics of Reweighted Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, 27, 1638–1665.

Pison, Van Aelst, Willems (2002). Small Sample Corrections for LTS and MCD. *Metrika*, to appear.

Rocke, Woodruff (1993). Computation of Robust Estimtes of Multivariate Location and Shape. *Statistica Neerlandica,* 47, 27–42.

Rousseeuw, Van Driessen (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 212–223.

Rousseeuw, van Zomeren (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85, 633–651.

Tukey (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.

Wilcox (1992). Robust Generalizations of Classical Test Reliability and Cronbach's Alpha. *British Journal of Mathematical and Statistical Psychology*, 45, 239–254.