SUPPLEMENTARY MATERIAL TO "LOCAL ROBUST ESTIMATION OF THE PICKANDS DEPENDENCE FUNCTION"

By Mikael Escobar-Bach[†], Yuri Goegebeur[†] and Armelle Guillou[‡]

University of Southern Denmark[†] and Université de Strasbourg et CNRS[‡]

In this supplement we provide additional simulation results for the main work Escobar et al. (2017).

1. Asymptotic standard deviation from Theorem 2.4.

Figure S1 shows the asymptotic standard deviation of $\sqrt{nh^p}(\widehat{A}_{\alpha,n}(t|x_0) - A_0(t|x_0))$ derived in Theorem 2.4, apart from the factor $\|K\|_2 A_0(t|x_0)/\sqrt{f(x_0)}$, as a function of α. As is clear, and as expected, the asymptotic standard deviation is increasing in α, indicating that within the class of MDPDE, the maximum likelihood estimator is the efficient choice. The maximum likelihood estimator is though not robust, while the MDPDE is, as illustrated in the simulations.

2. Additional simulation results for the two types of contamination included in the main text.

In Figures S2 and S3 we show the scatterplots of datasets generated under the first and second type of contamination, respectively. In these plots, the non-contaminated sample is represented as circles whereas the contaminated pairs are represented as crosses. Here, ε is set to the value 0.1. These scatterplots are obtained before the empirical transformation of the margins into unit exponential distributions on the left panel and after this transformation on the right panel.

Figures S4 till S6 represent the boxplots of the estimator $\breve{A}_{\alpha,n}(.|x)$ based on 200 samples for three positions: $x = 0.1, 0.3$ and 0.5, respectively, in case $n = 1000$ for the first type of contamination (contamination on the axes). Four values of α have been reported: 0, 0.1, 0.5, and 1 corresponding to the different rows on each figure, whereas three different percentages of contamination have been used: from the left to the right of each figure we show $0\%, 10\%$ and 20%. Then, Figures S7 till S9 have been constructed similarly, but this time for a sample size $n = 5000$. Figures S10 till S15 are also similar but for the second type of contamination (contamination on the diagonal) and for $x = 0.5, 0.7$ and 0.9. Based on these figures, we can draw the following conclusions:

- in case of uncontaminated datasets, the best estimator is achieved with $\alpha = 0$, although there are no big differences. This is not surprising since when $\alpha = 0$, one recovers the maximum likelihood estimator which is well-known to be efficient, but not robust;
- increasing the sample size decreases the variability;
- in case of contamination, a larger value of α (0.5 or 1) is needed;

*This work was supported by a research grant (VKR023480) from VILLUM FONDEN and an international project for scientific cooperation (PICS-6416). Computation/simulation for the work described in this paper was supported by the DeIC National HPC Centre, SDU.

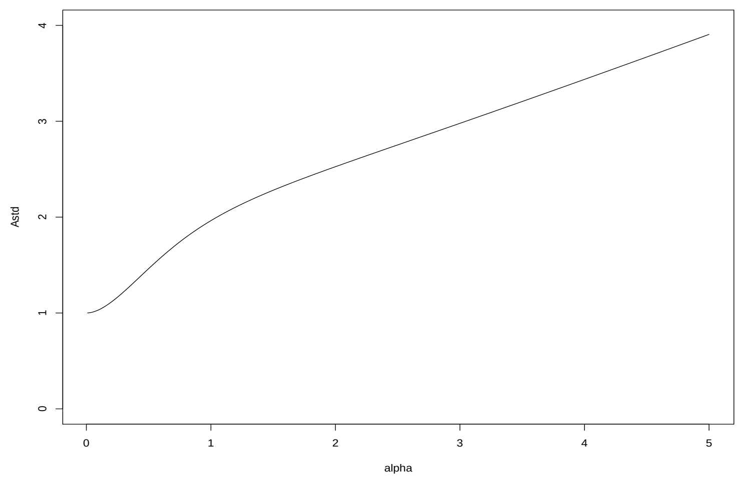

Fig S1. *Asymptotic standard deviation of* $\sqrt{nh^p}(\widehat{A}_{\alpha,n}(t|x_0) - A_0(t|x_0))$ *as a function of* α.

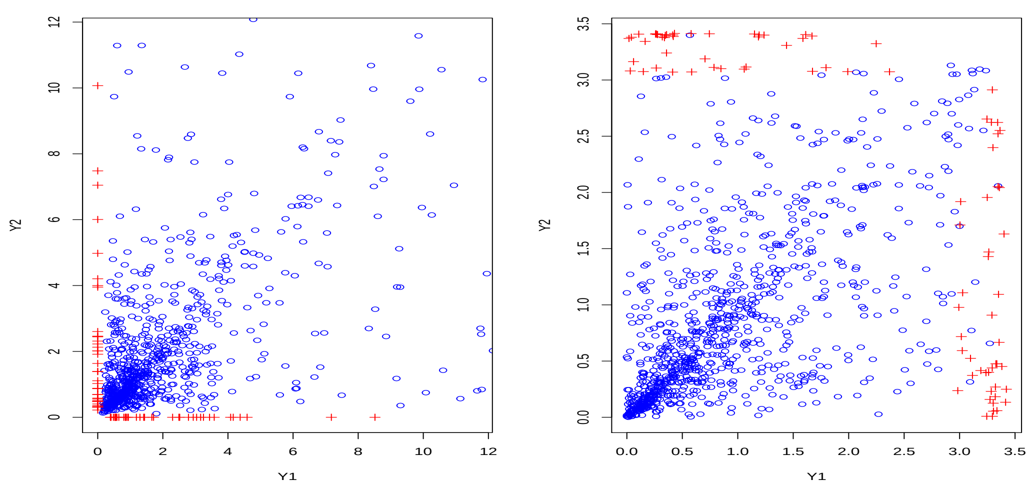

Fig S2. *First type of contamination: on the left the original data and on the right the data after transformation into (approximate) unit exponentials. The non-contaminated observations are represented as circles whereas the contaminated pairs are represented as crosses. Here* ϵ *is set to the value 0.1.*

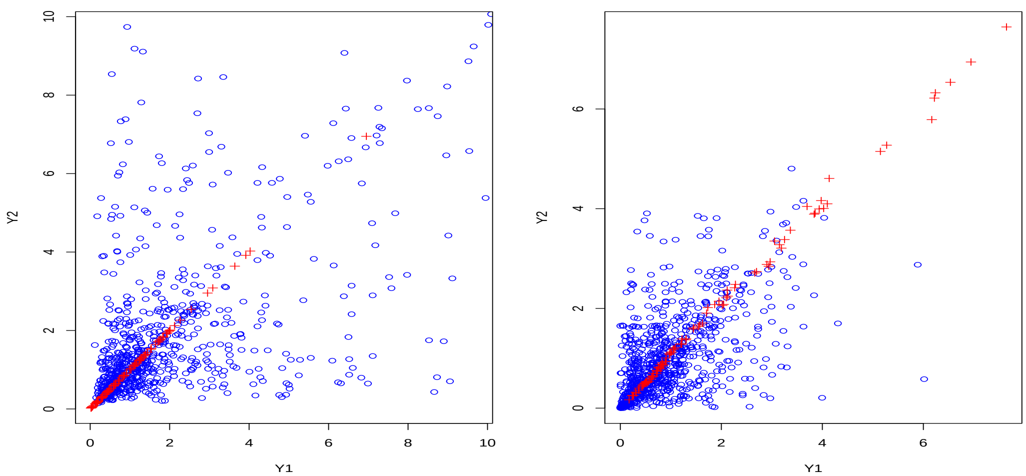

FIG S3. *Second type of contamination: on the left the original data and on the right the data after transformation into (approximate) unit exponentials. The non-contaminated observations are represented as circles whereas the contaminated pairs are represented as crosses. Here ϵ is set to the value 0.1.*

- as expected, increasing the percentage of contamination negatively affects the estimator, whatever the type of contamination and value of α.

These conclusions are in line with those provided by the MISE-plots or the empirical coverage probabilities, as reported in the main text.

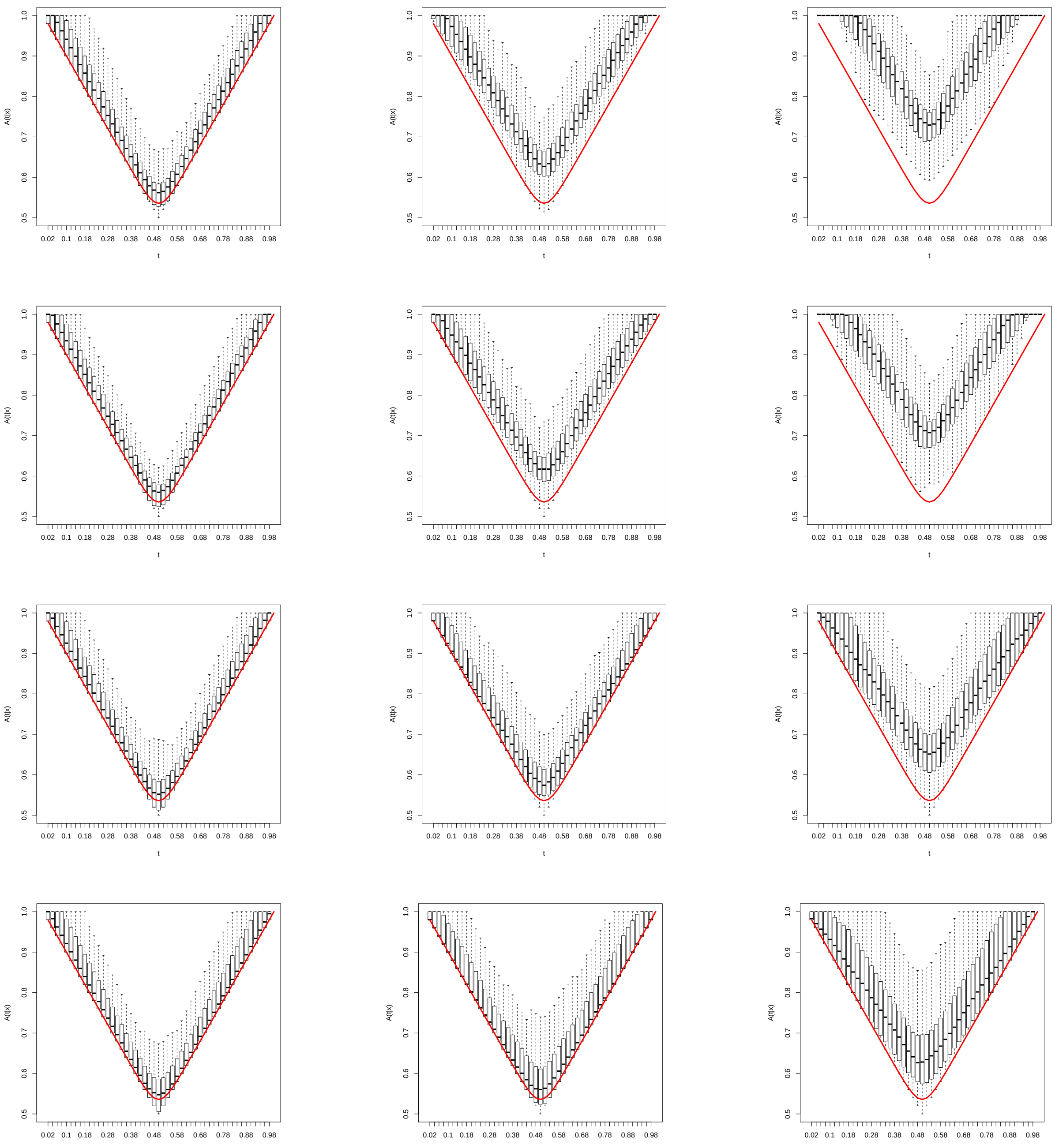

Fig S4. *First type of contamination: Estimation of* $A_0(.|0.1)$ *(full line) for the logistic distribution with* $n = 1000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

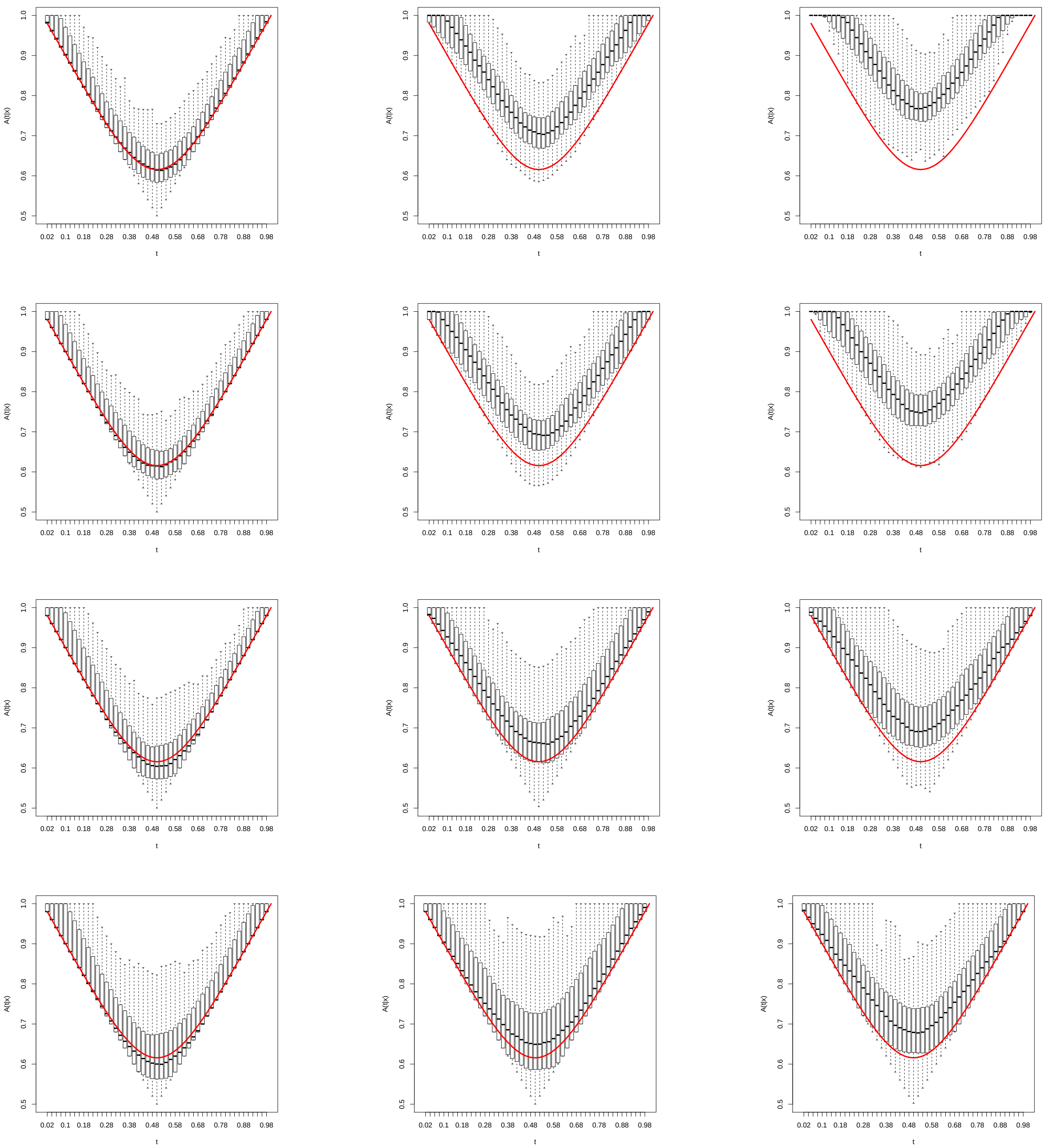

FIG S5. *First type of contamination: Estimation of* $A_0(.|0.3)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

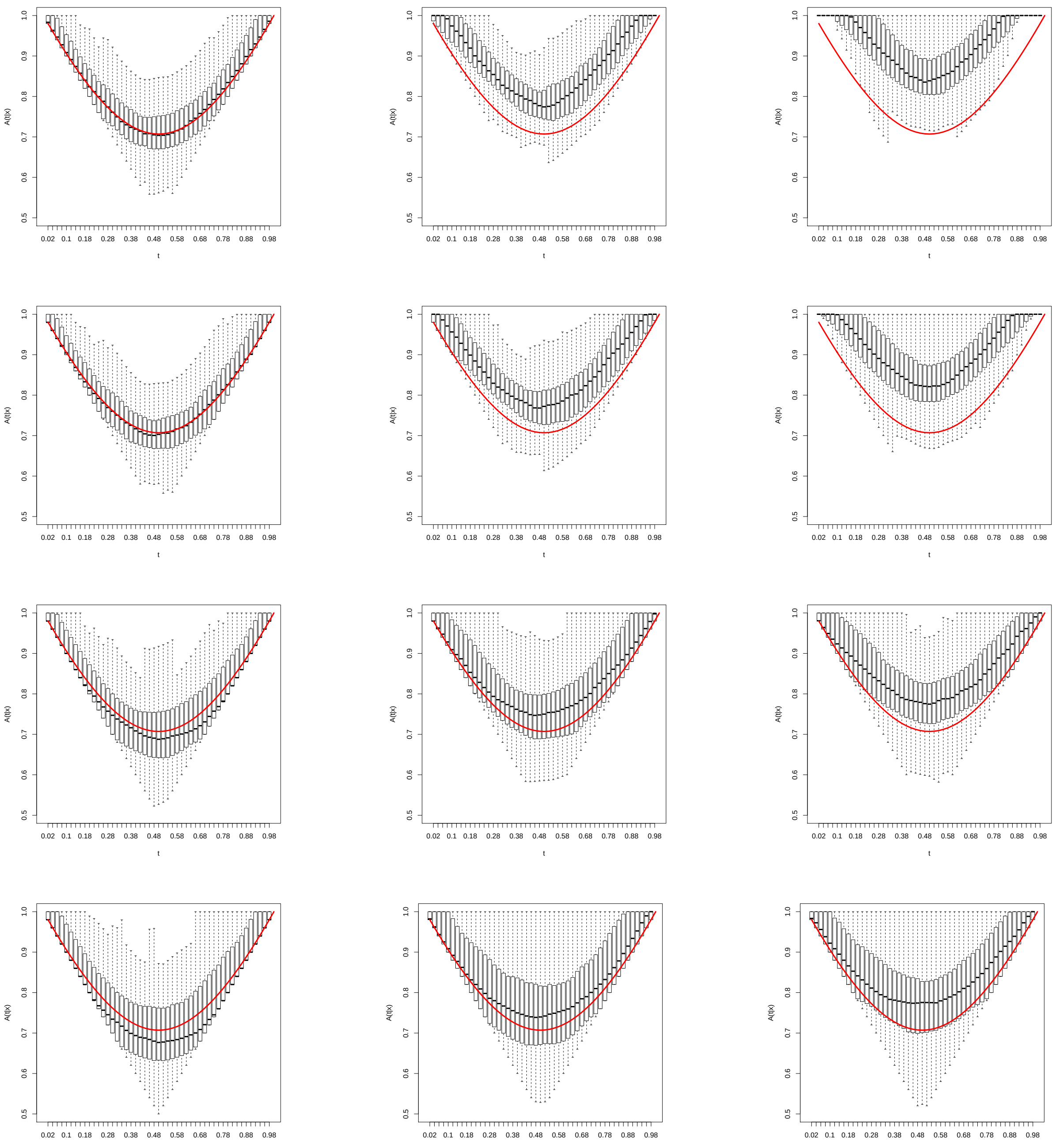

Fig S6. *First type of contamination: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

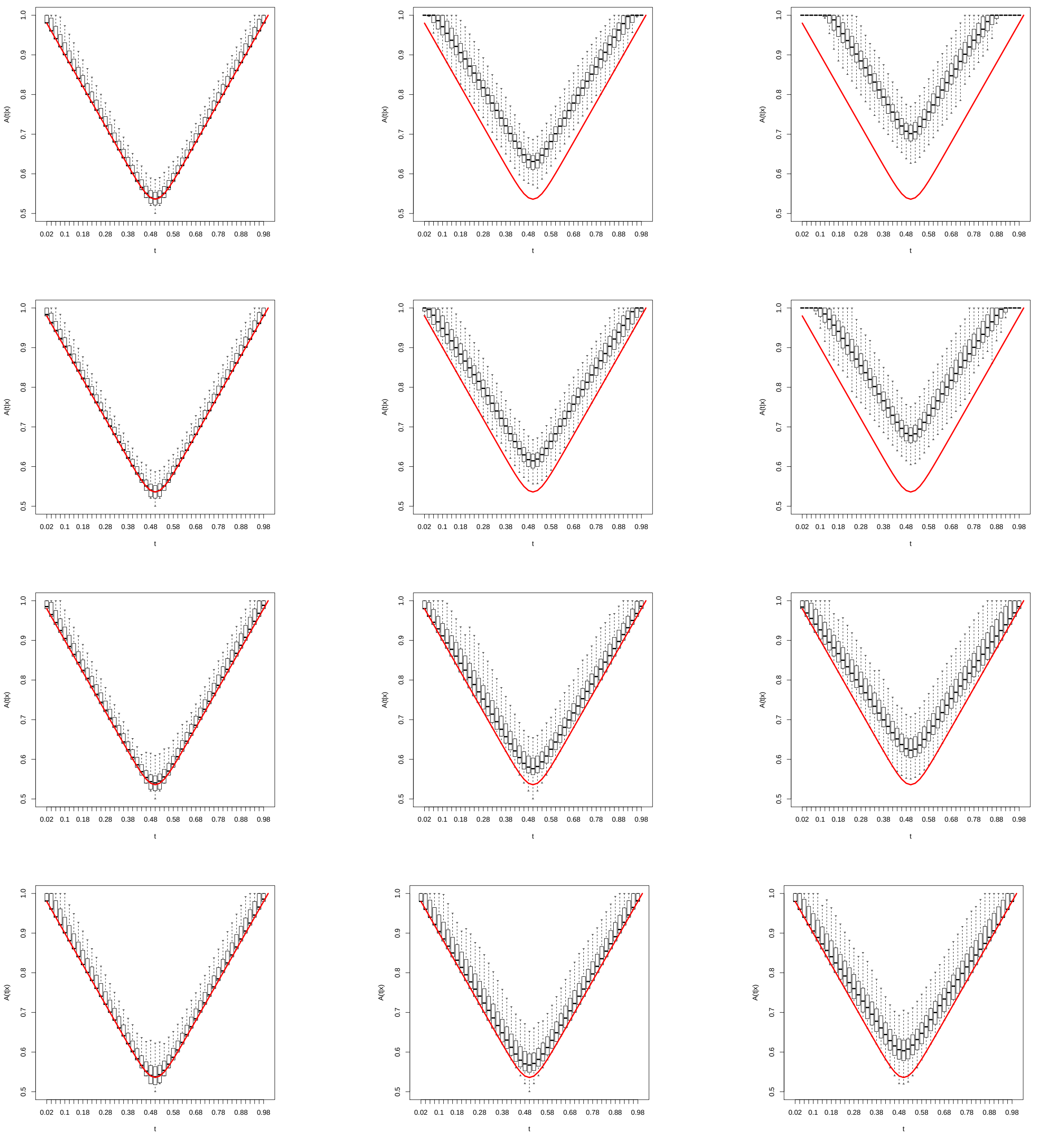

FIG S7. *First type of contamination: Estimation of* $A_0(.|0.1)$ *(full line) for the logistic distribution with* $n = 5000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

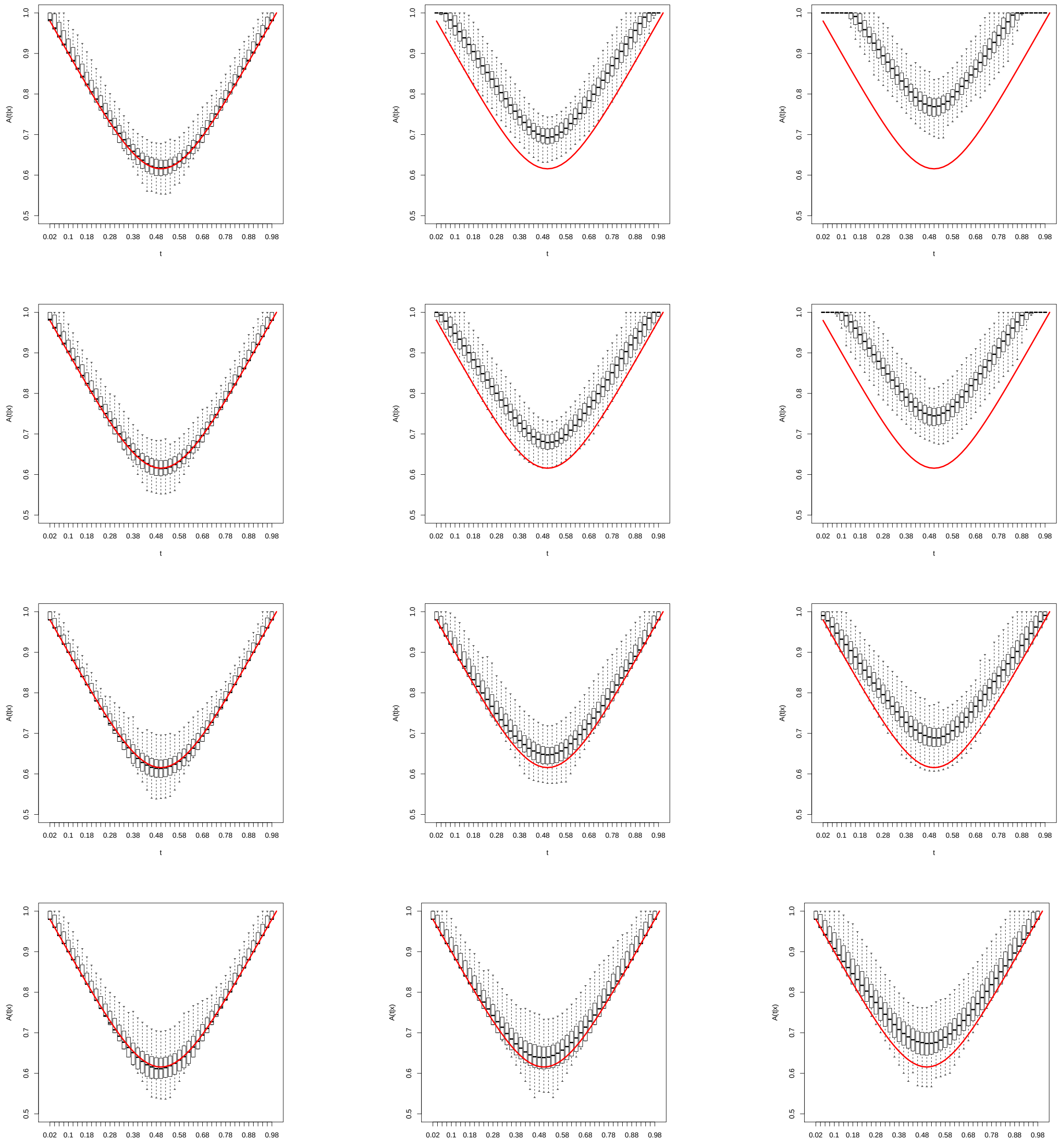

Fig S8. *First type of contamination: Estimation of* $A_0(.|0.3)$ *(full line) for the logistic distribution with* $n = 5000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

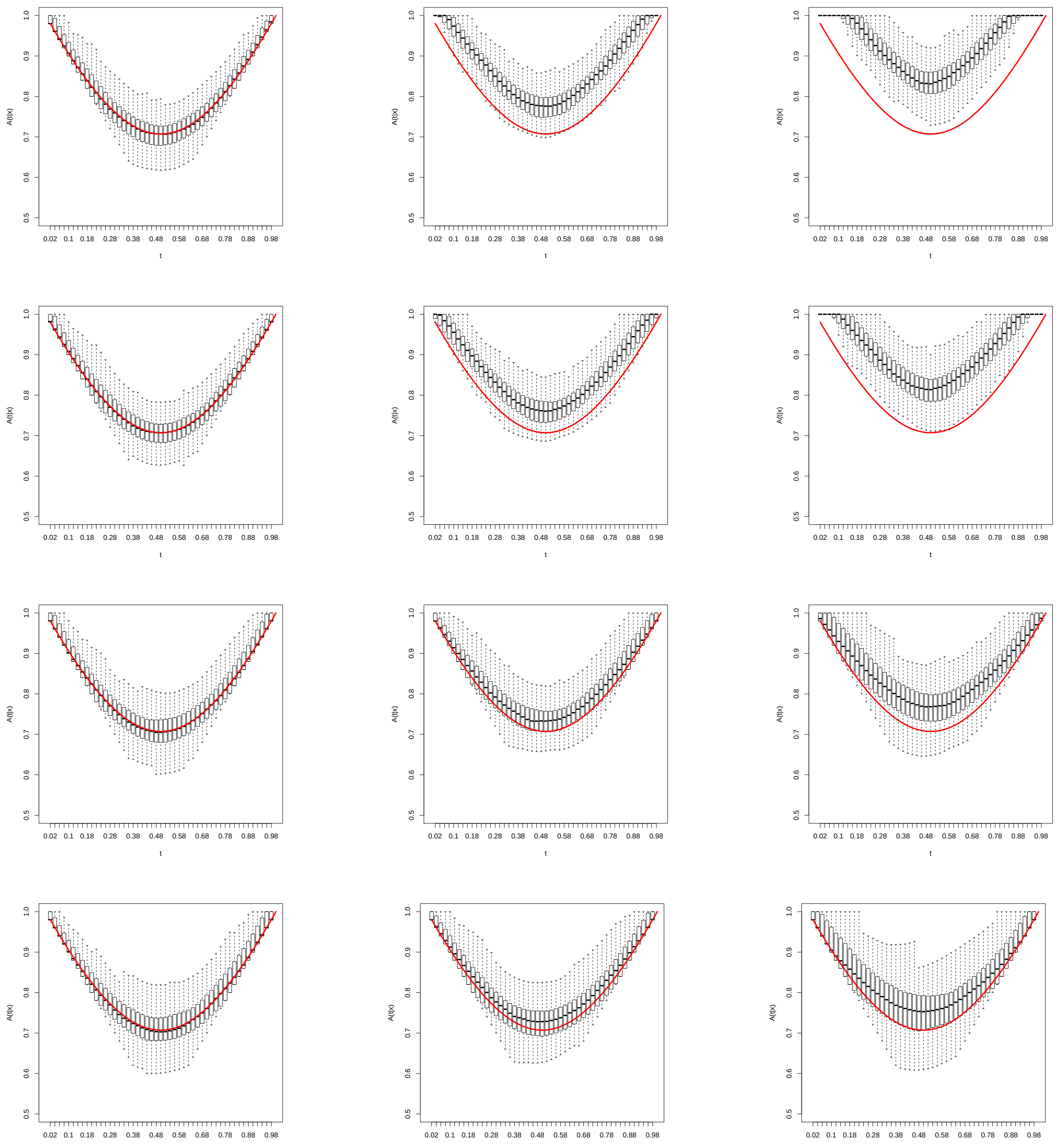

FIG S9. *First type of contamination: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 5000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

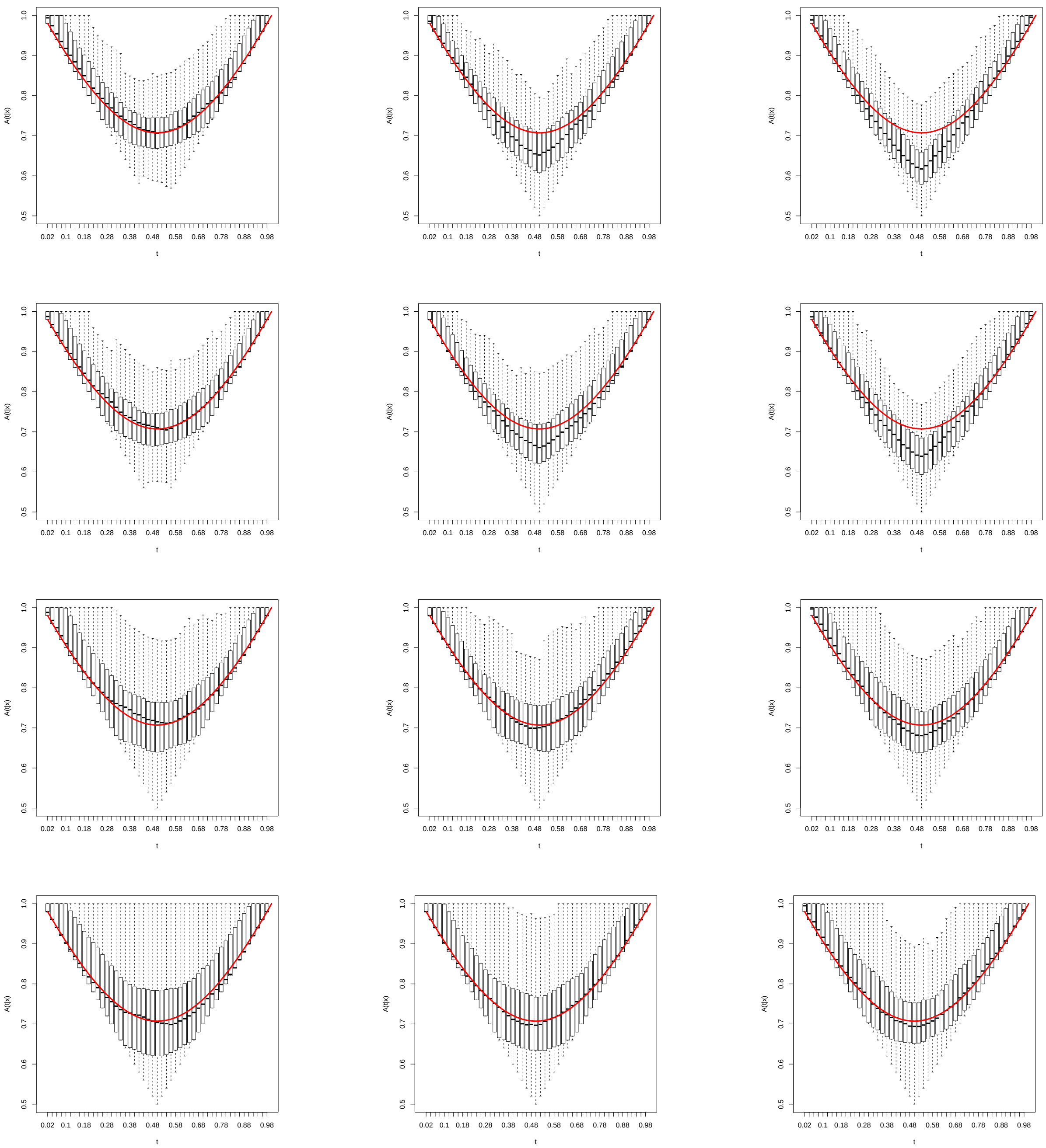

FIG S10. *Second type of contamination: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 1000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

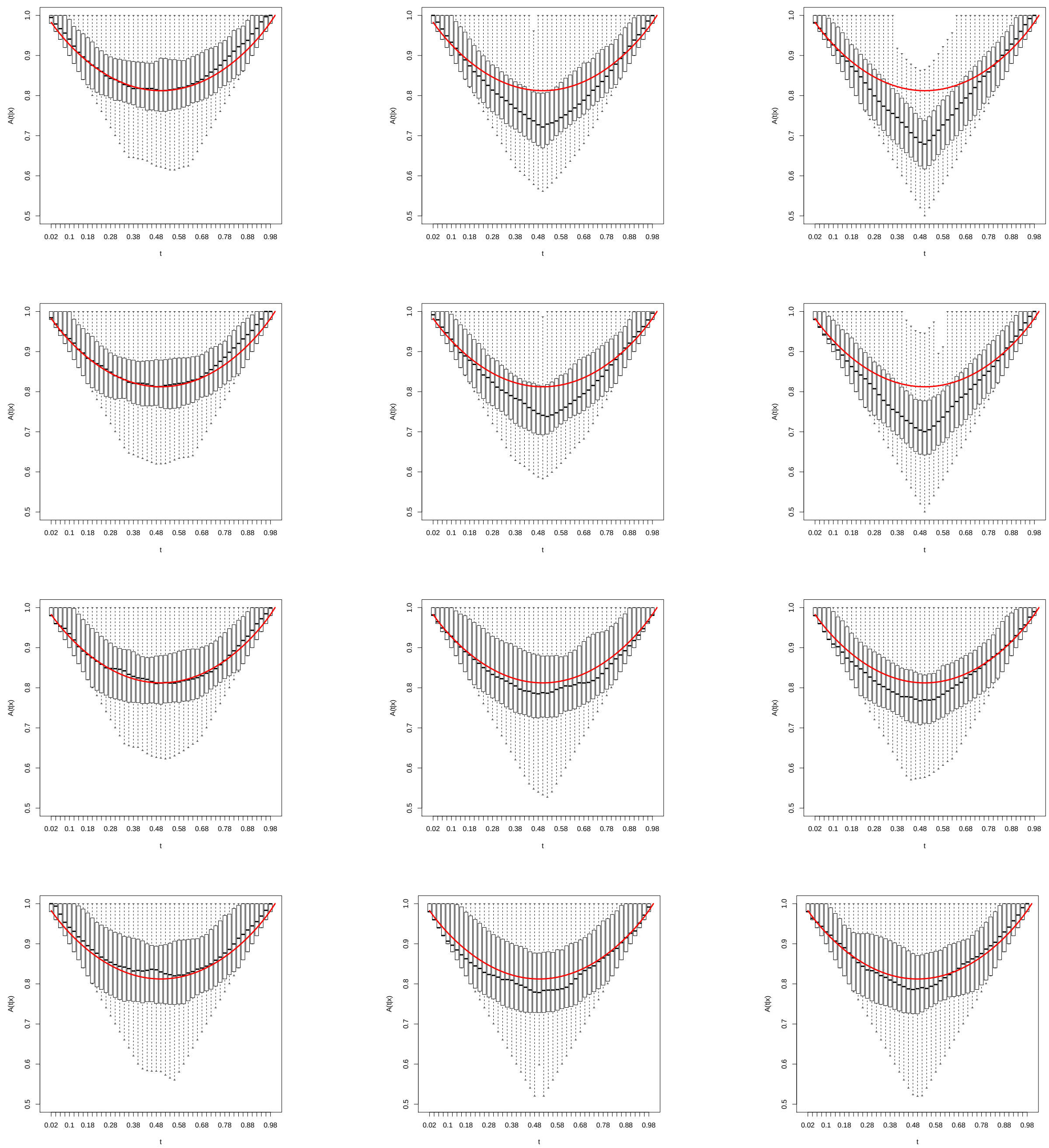

FIG S11. *Second type of contamination: Estimation of* $A_0(.|0.7)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

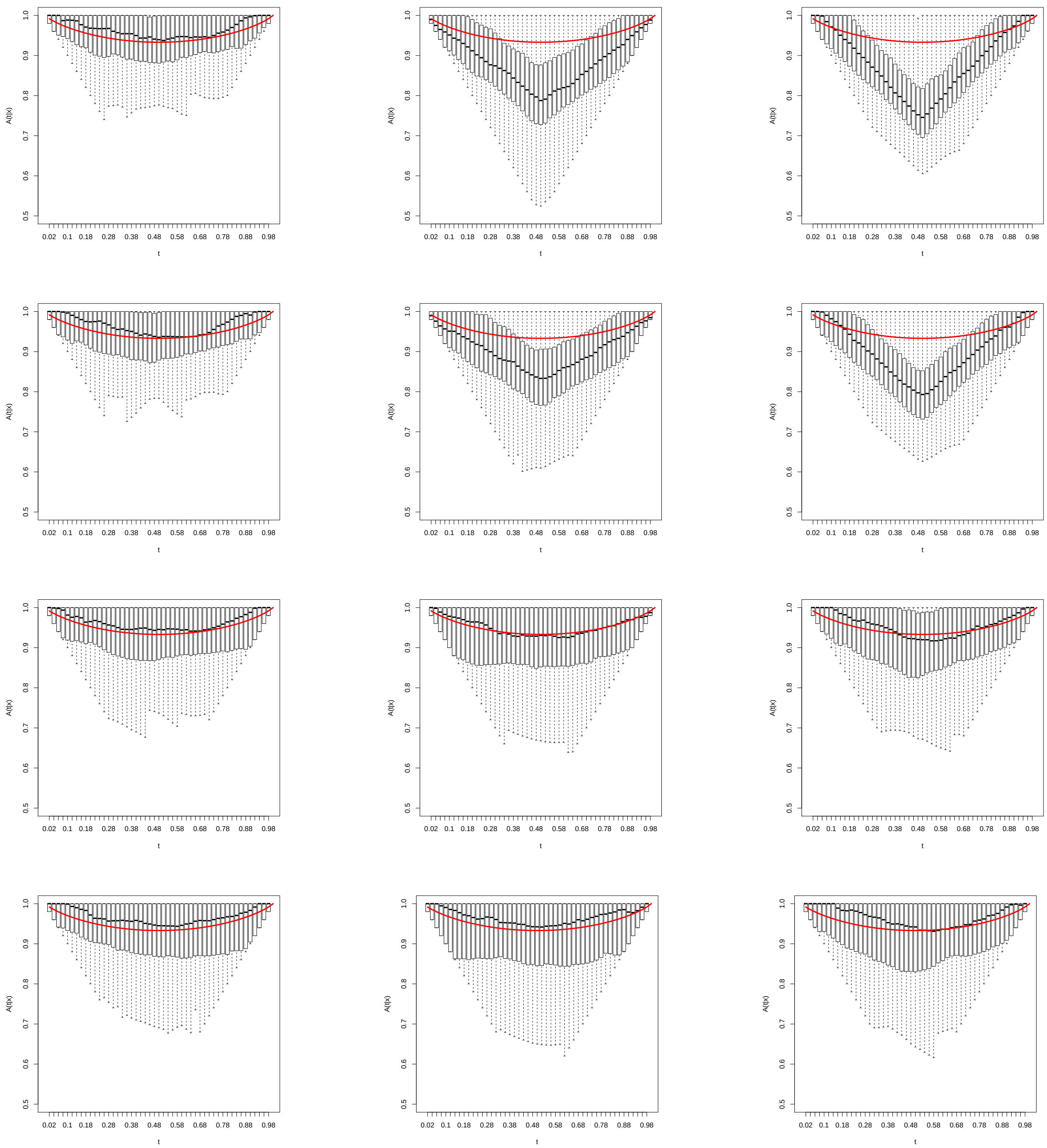

Fig S12. *Second type of contamination: Estimation of* $A_0(.|0.9)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

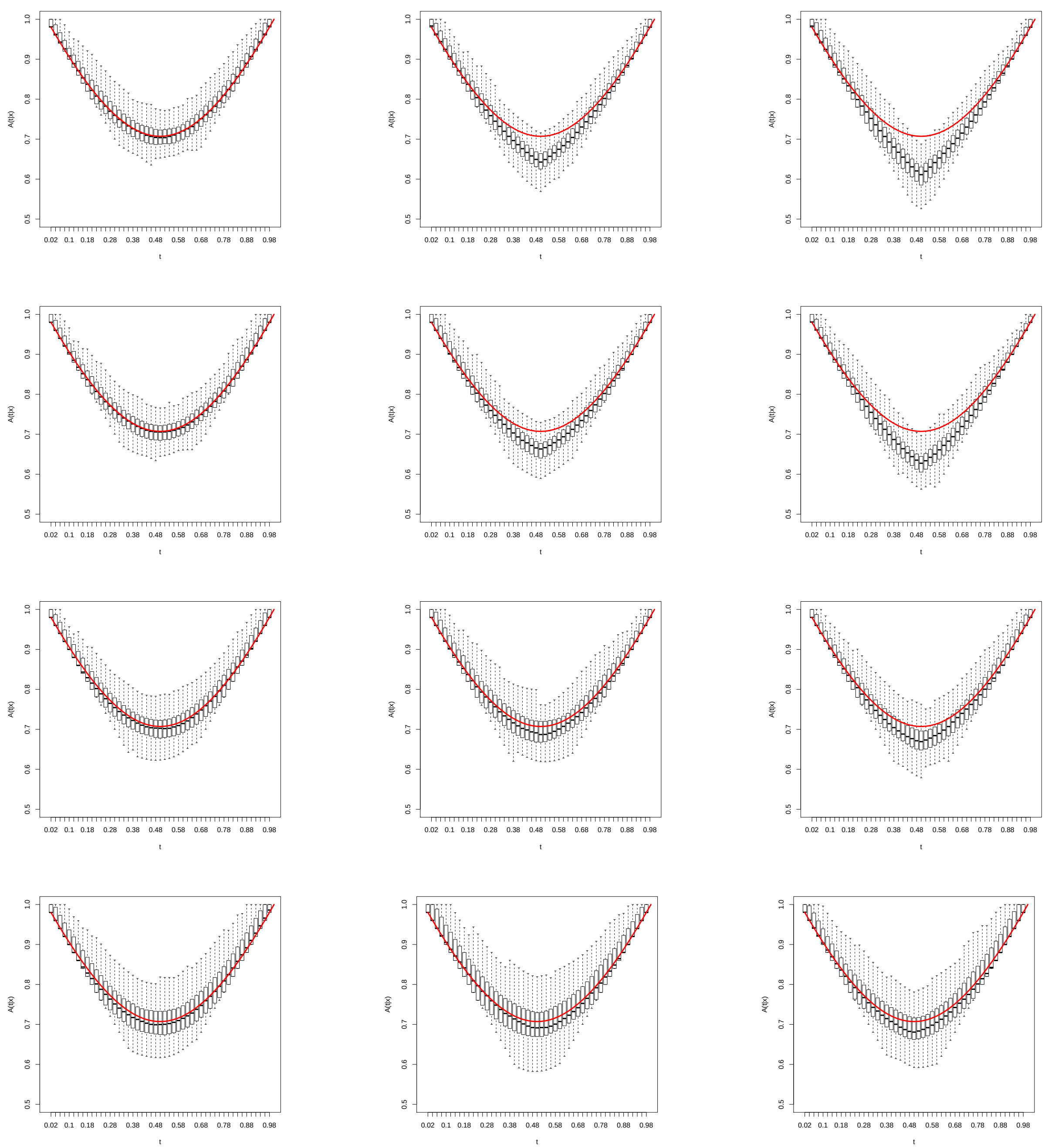

FIG S13. *Second type of contamination: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 5000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* 0%, 10% *and* 20% *of contamination.*

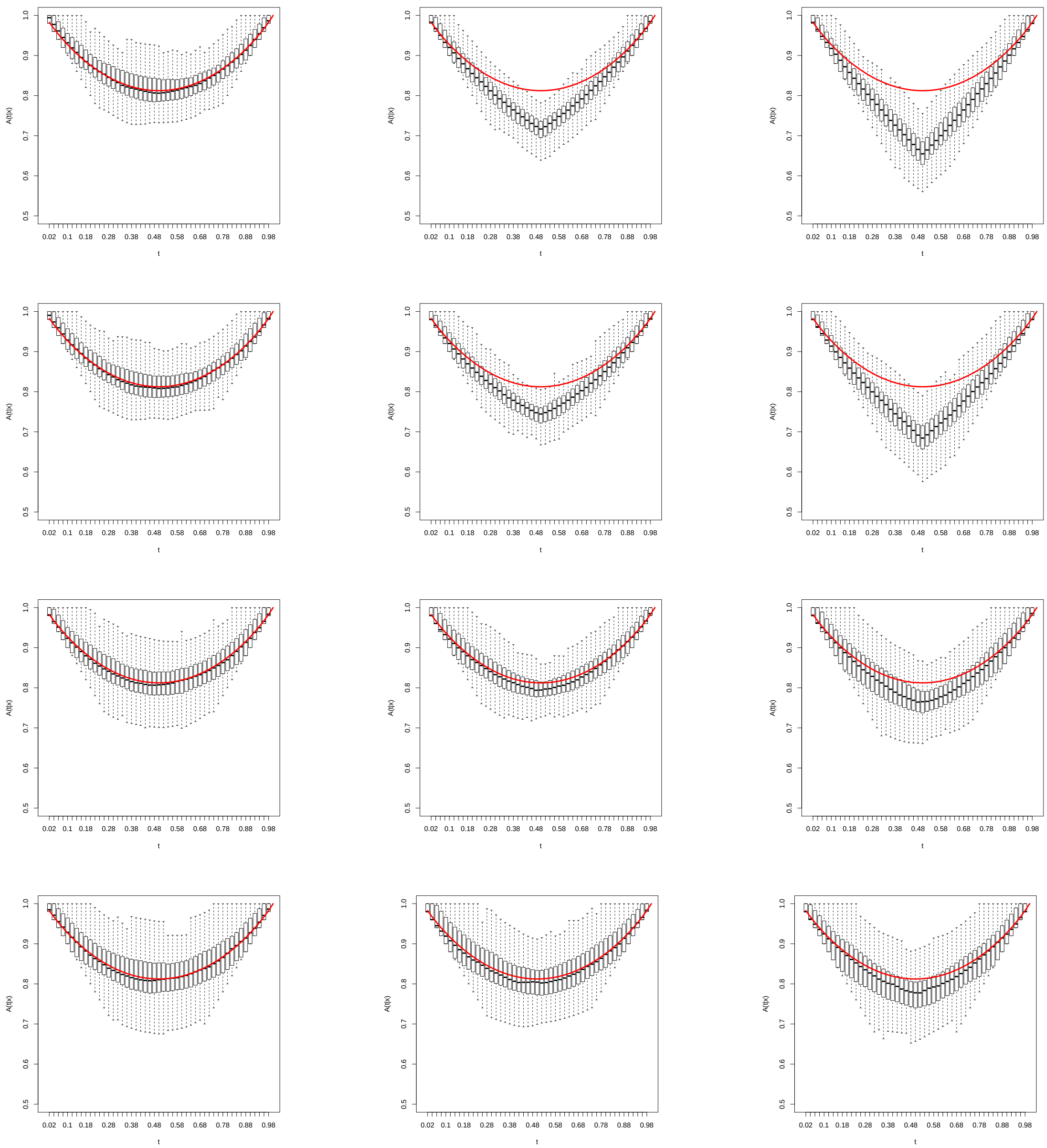

Fig S14. *Second type of contamination: Estimation of* $A_0(.|0.7)$ *(full line) for the logistic distribution with* $n = 5000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

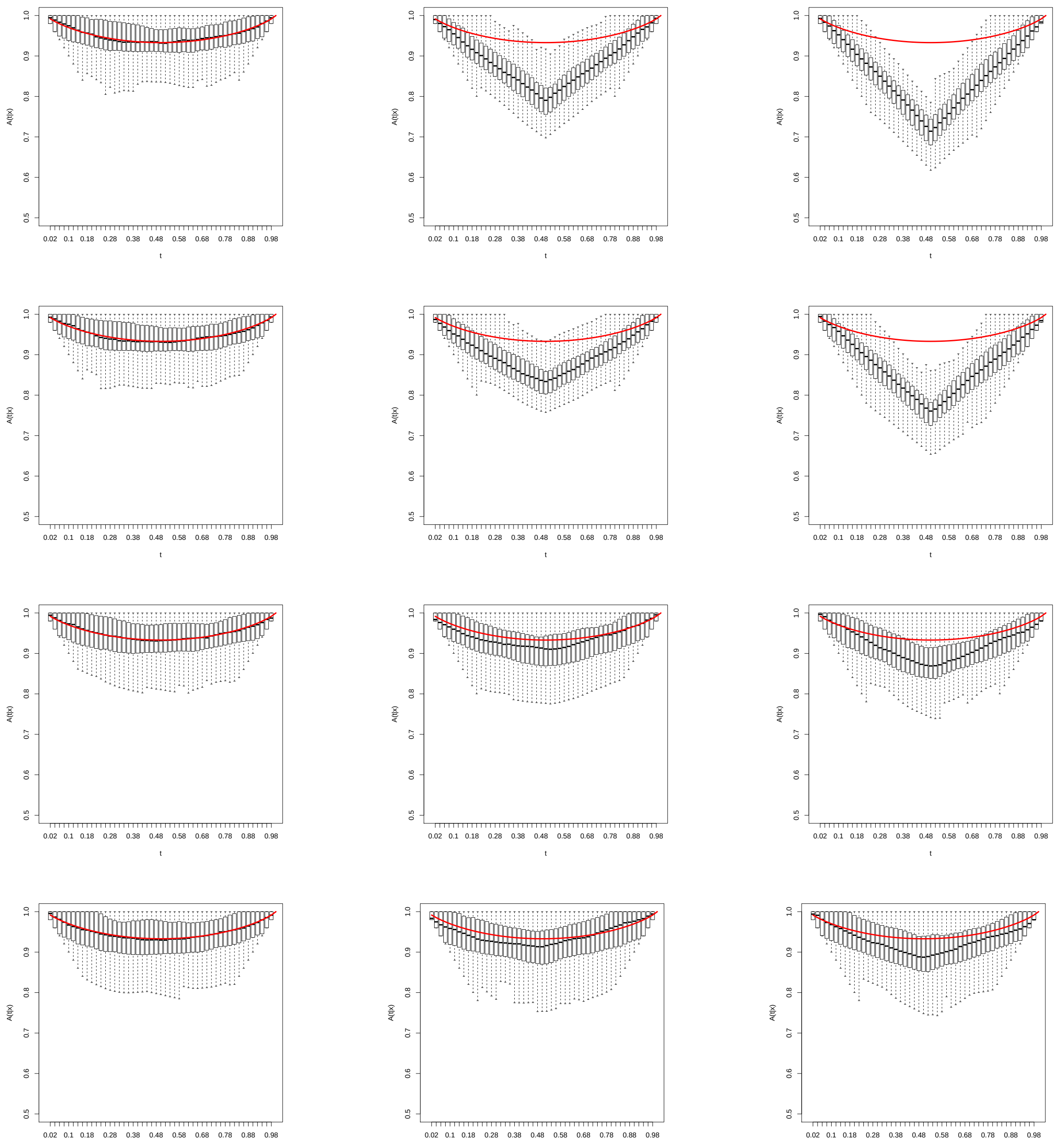

Fig S15. *Second type of contamination: Estimation of* $A_0(.|0.9)$ *(full line) for the logistic distribution with* $n = 5000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

3. An additional example of contamination (angular contamination). As in the main text, we consider a mixture model F_ε, where F_ℓ is again the logistic distribution, but this time, given $X = x$, F_c is the distribution function of the bivariate random vector

$$Y = \begin{cases} V\,(\cos(\theta), \sin(\theta))\,, & \text{with probability } 1/2 \\ V\,(\sin(\theta), \cos(\theta))\,, & \text{with probability } 1/2 \end{cases}$$

where V is a standard exponential random variable and $\theta \in [0, \pi/2]$. Here we use $\theta = \pi/8$ but varying this parameter changes obviously the positions of the contaminated points and thus influences the results. This mixture is illustrated in Figure S16 with the non-contaminated sample as circles and the contaminated pairs with crosses. Here ε is set to the value 0.1.

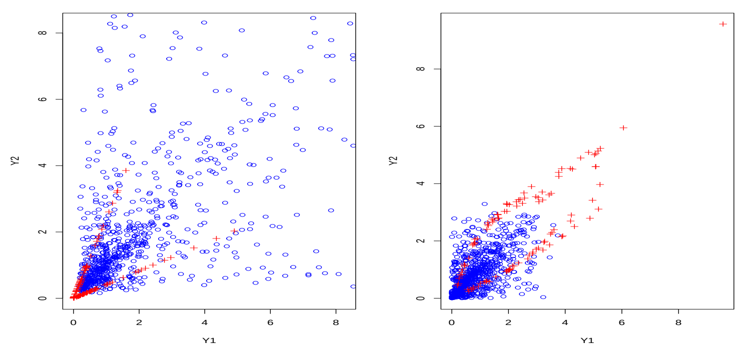

Fig S16. *Angular contamination: on the left the original data and on the right the data after transformation into (approximate) unit exponentials. The non-contaminated observations are represented as circles whereas the contaminated pairs are represented as crosses. Here ε is set to the value 0.1.*

Figure S17 represents the MISE$(\varepsilon, \alpha|x)$ as a function of ε for the angular contamination and the three positions: $x = 0.5, 0.7$ and 0.9. Again, the same conclusions can be done, i.e. for very small values of ε, all the estimators behave similarly, whatever α. On the contrary, for larger values of ε, it is crucial to increase α if one wants to keep a reasonable small value for the MISE when x is close to 1. These observations are clearer when $n = 5000$ since the MISE-curves exhibit less variability in that case. Similarly as Figures S4 till S15, the boxplots of our estimator $\breve{A}_{\alpha,n}(.|x)$ based on 200 samples for the three positions: $x = 0.5, 0.7$ and 0.9, respectively, are provided in Figures S18 till S20 for $n = 1000$ and in Figures S21 till S23 for $n = 5000$. Also the empirical coverage probabilities of 90% confidence intervals based on the limiting distribution are given in Table S1. Again all these results outline the performance of our robust estimator. It can handle a large percentage of contamination, and behaves quite similarly to the maximum likelihood estimator in case of no-contamination.

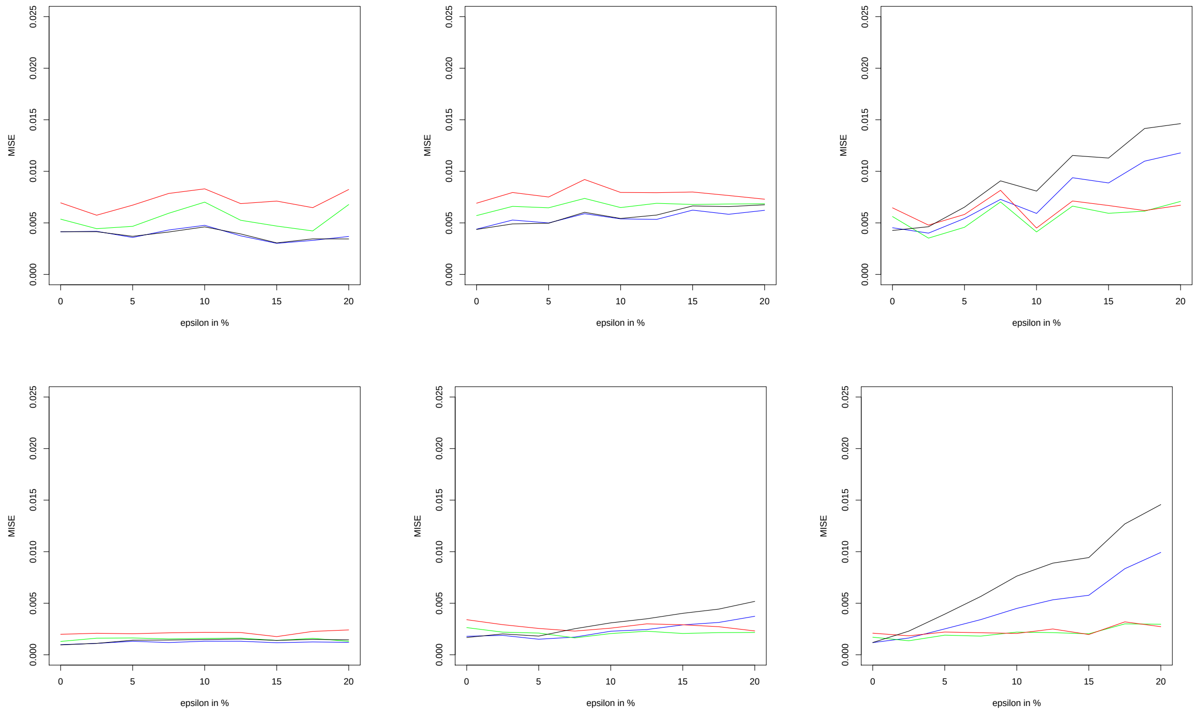

FIG S17. *Angular contamination: MISE*$(\varepsilon, \alpha|x)$ *as a function of* $\varepsilon \in \{0, 0.025, 0.05, ..., 0.2\}$ *with* $\alpha = 0$ *(black),* $\alpha = 0.1$ *(blue),* $\alpha = 0.5$ *(green) and* $\alpha = 1$ *(red) and* $x = 0.5, 0.7, 0.9$ *from the left to the right. Top row:* $n = 1000$ *and bottom row:* $n = 5000$.

Table S1

Angular contamination - coverage probabilities of 90% confidence intervals.

$x = 0.5$	α	$t = 0.3$				$t = 0.5$				$t = 0.7$			
		0	0.1	0.5	1	0	0.1	0.5	1	0	0.1	0.5	1
$n = 1000$	$\varepsilon = 0.0$	0.93	0.94	0.95	0.94	0.94	0.94	0.91	0.94	0.94	0.95	0.95	0.97
	$\varepsilon = 0.1$	0.94	0.94	0.93	0.94	0.90	0.92	0.91	0.92	0.94	0.95	0.94	0.94
	$\varepsilon = 0.2$	0.98	0.98	0.95	0.94	0.95	0.96	0.93	0.93	0.97	0.97	0.93	0.93
$n = 5000$	$\varepsilon = 0.0$	0.95	0.97	0.98	0.98	0.96	0.96	0.96	0.96	0.96	0.97	0.98	0.98
	$\varepsilon = 0.1$	0.91	0.97	0.96	0.97	0.78	0.84	0.94	0.94	0.89	0.97	0.97	0.96
	$\varepsilon = 0.2$	0.93	0.98	0.99	0.97	0.73	0.86	0.95	0.96	0.93	0.99	0.97	0.96
$x = 0.7$	α	$t = 0.3$				$t = 0.5$				$t = 0.7$			
		0	0.1	0.5	1	0	0.1	0.5	1	0	0.1	0.5	1
$n = 1000$	$\varepsilon = 0.0$	0.94	0.95	1.00	1.00	0.96	0.95	0.97	1.00	0.98	0.98	1.00	1.00
	$\varepsilon = 0.1$	0.98	0.98	1.00	1.00	0.86	0.90	0.97	0.99	0.97	0.98	1.00	1.00
	$\varepsilon = 0.2$	0.98	0.96	1.00	1.00	0.78	0.86	0.93	0.98	0.98	0.98	1.00	1.00
$n = 5000$	$\varepsilon = 0.0$	0.92	0.92	0.90	0.91	0.91	0.91	0.92	0.90	0.90	0.92	0.94	0.93
	$\varepsilon = 0.1$	0.72	0.84	0.98	0.98	0.48	0.73	0.94	0.96	0.76	0.84	0.93	0.94
	$\varepsilon = 0.2$	0.50	0.65	0.91	0.95	0.24	0.51	0.94	0.94	0.52	0.65	0.90	0.95
$x = 0.9$	α	$t = 0.3$				$t = 0.5$				$t = 0.7$			
		0	0.1	0.5	1	0	0.1	0.5	1	0	0.1	0.5	1
$n = 1000$	$\varepsilon = 0.0$	0.98	0.99	0.99	0.99	0.97	0.97	0.98	0.99	0.97	0.98	0.98	0.99
	$\varepsilon = 0.1$	0.90	0.94	1.00	0.99	0.80	0.90	0.99	0.99	0.91	0.95	0.99	1.00
	$\varepsilon = 0.2$	0.79	0.85	0.97	0.99	0.56	0.72	0.98	0.99	0.78	0.87	0.96	1.00
$n = 5000$	$\varepsilon = 0.0$	0.98	0.98	0.99	1.00	0.97	0.98	1.00	1.00	0.97	0.97	1.00	1.00
	$\varepsilon = 0.1$	0.45	0.67	0.95	0.98	0.20	0.53	0.96	0.99	0.39	0.62	0.96	0.99
	$\varepsilon = 0.2$	0.14	0.27	0.86	0.95	0.09	0.17	0.90	0.97	0.17	0.29	0.88	0.96

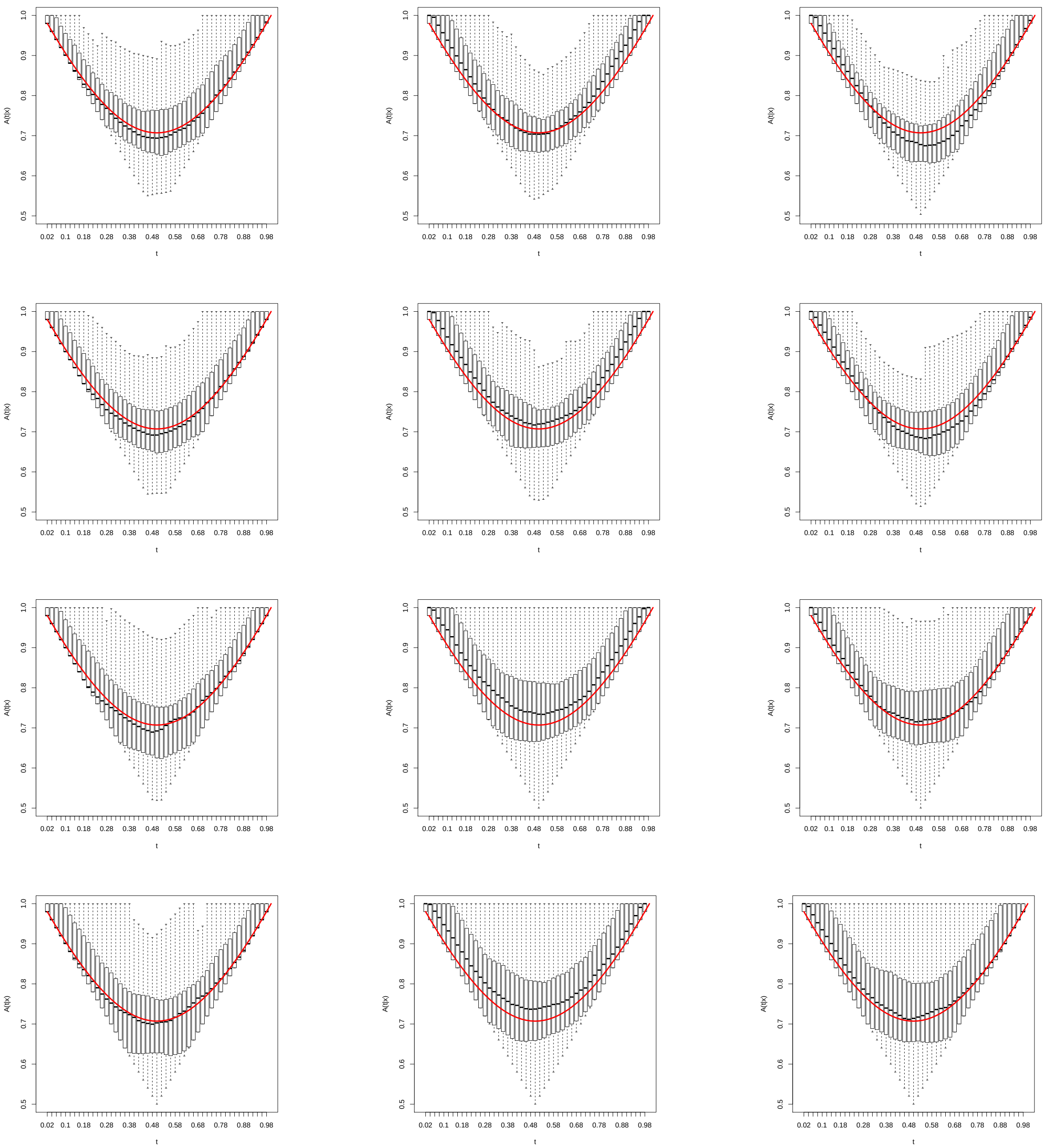

Fig S18. *Angular contamination: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 1000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

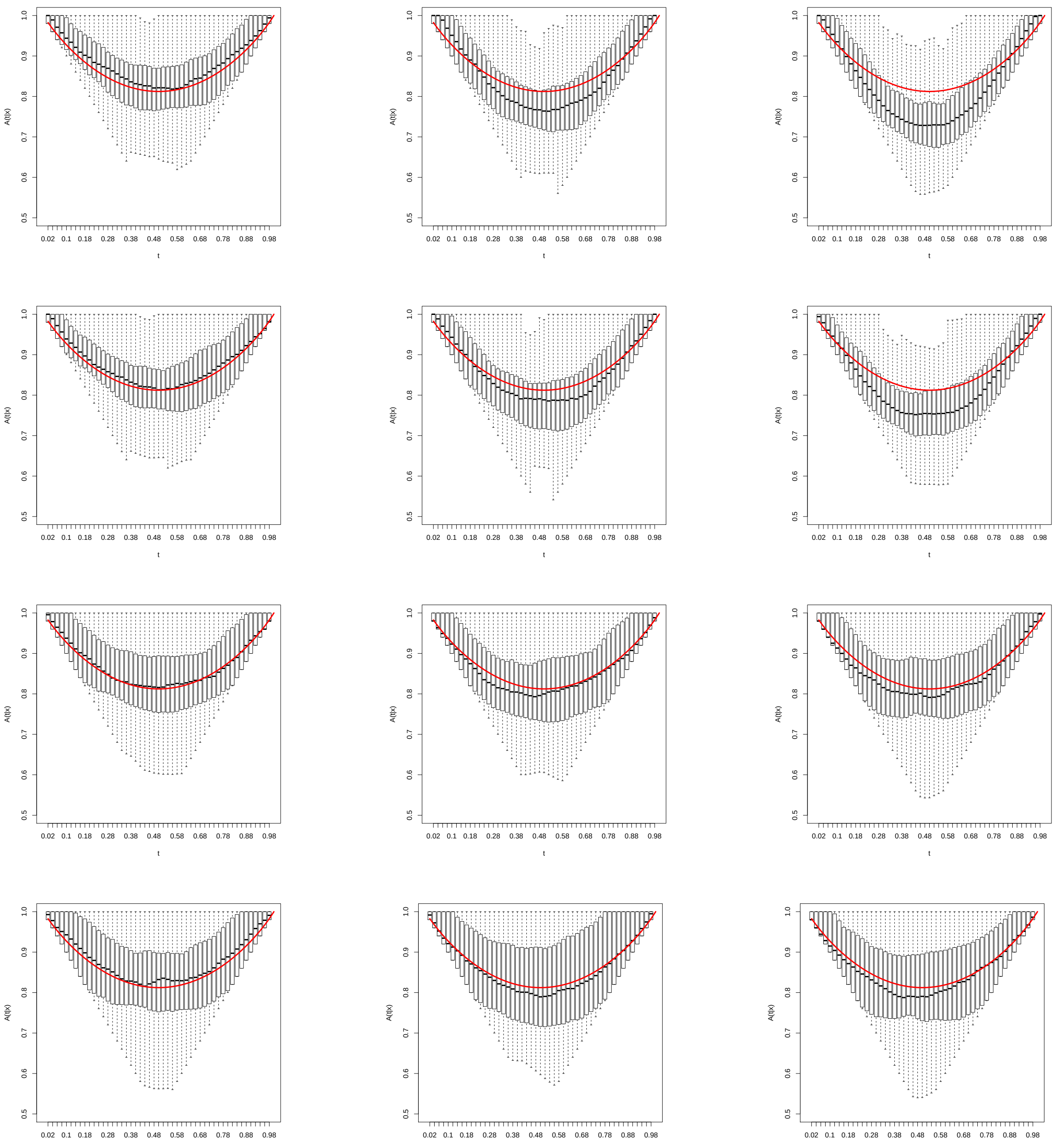

Fig S19. *Angular contamination: Estimation of* $A_0(.|0.7)$ *(full line) for the logistic distribution with* $n = 1000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

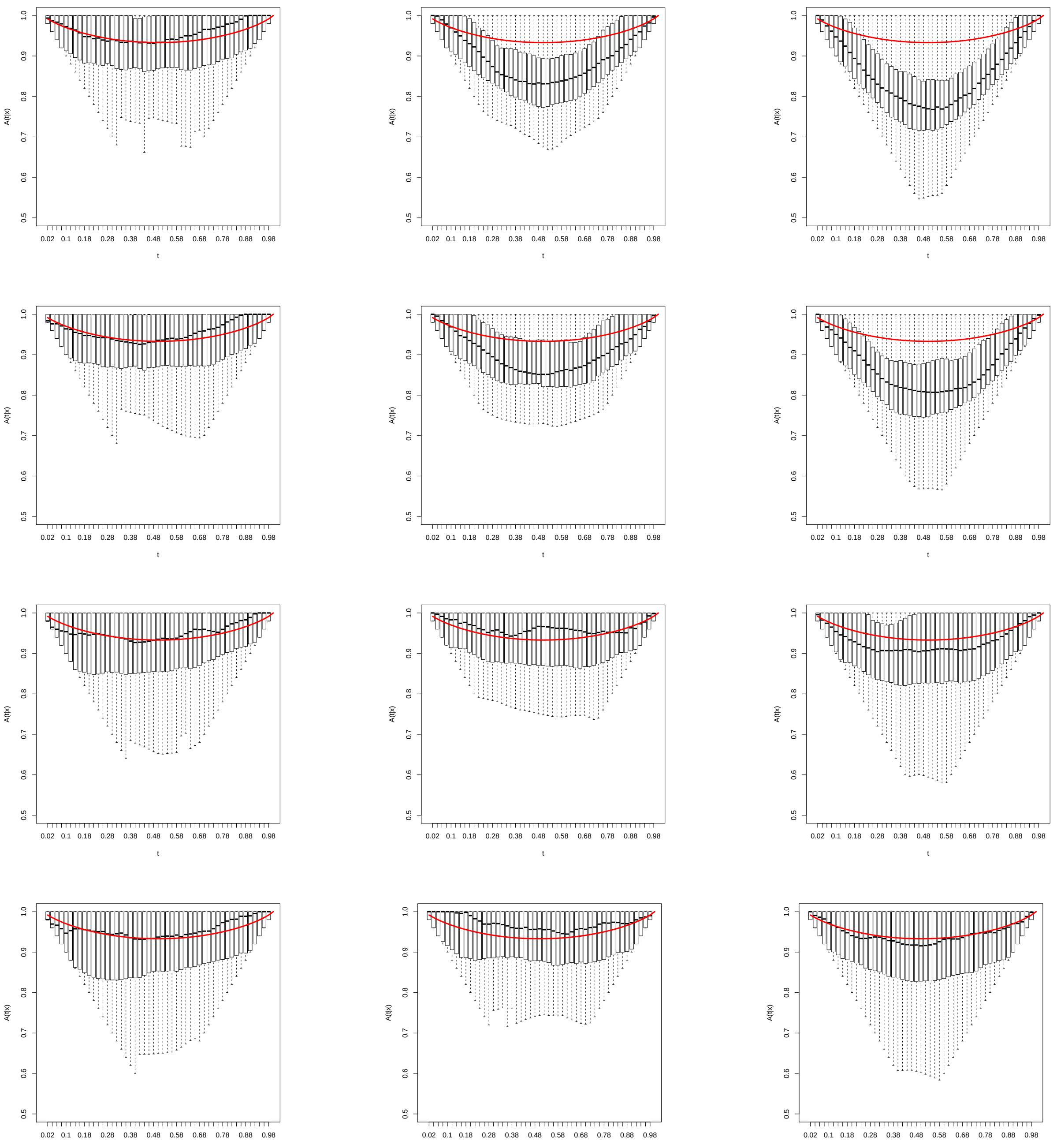

FIG S20. *Angular contamination: Estimation of* $A_0(.|0.9)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

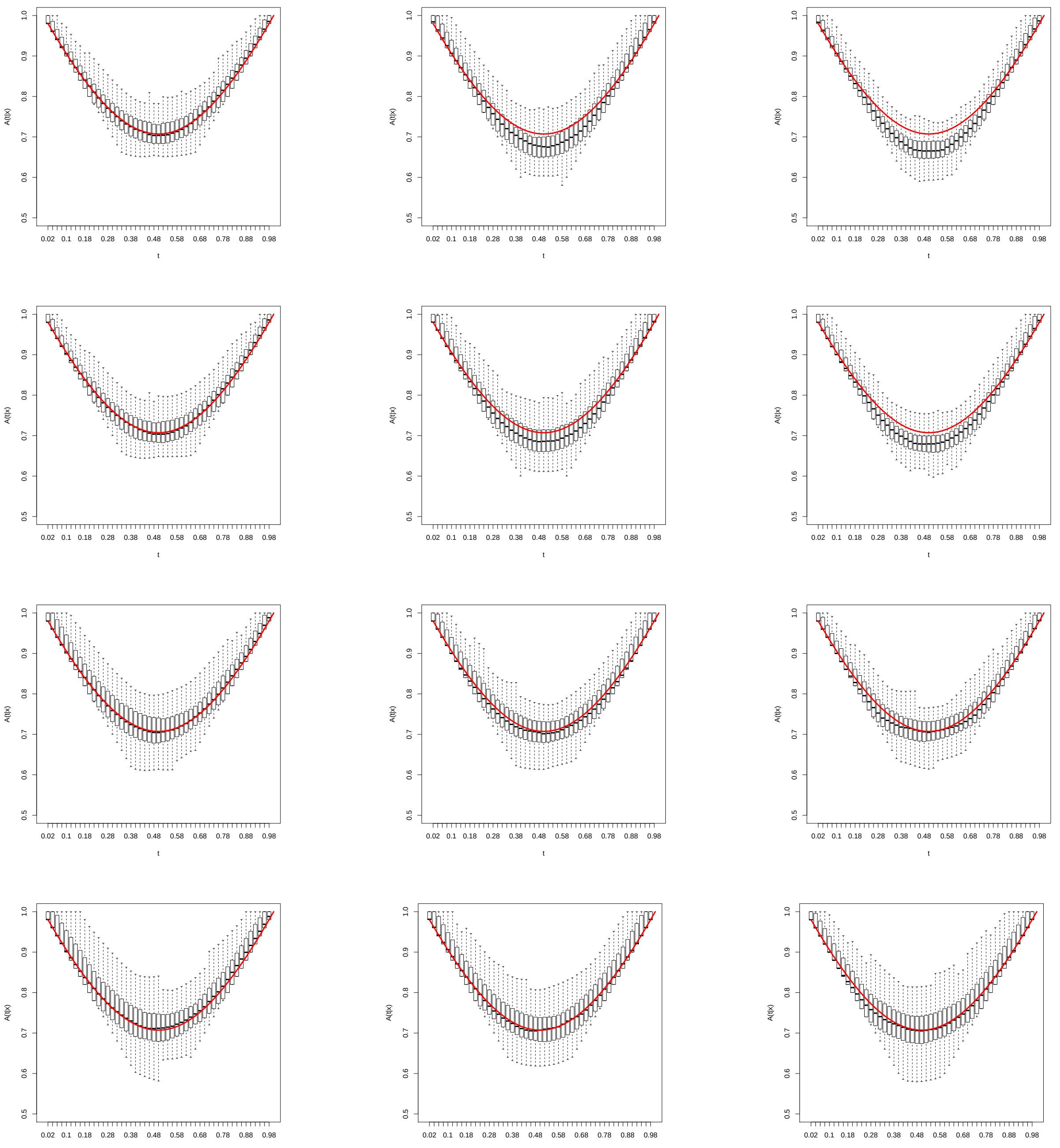

Fig S21. *Angular contamination: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 5000$*. From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

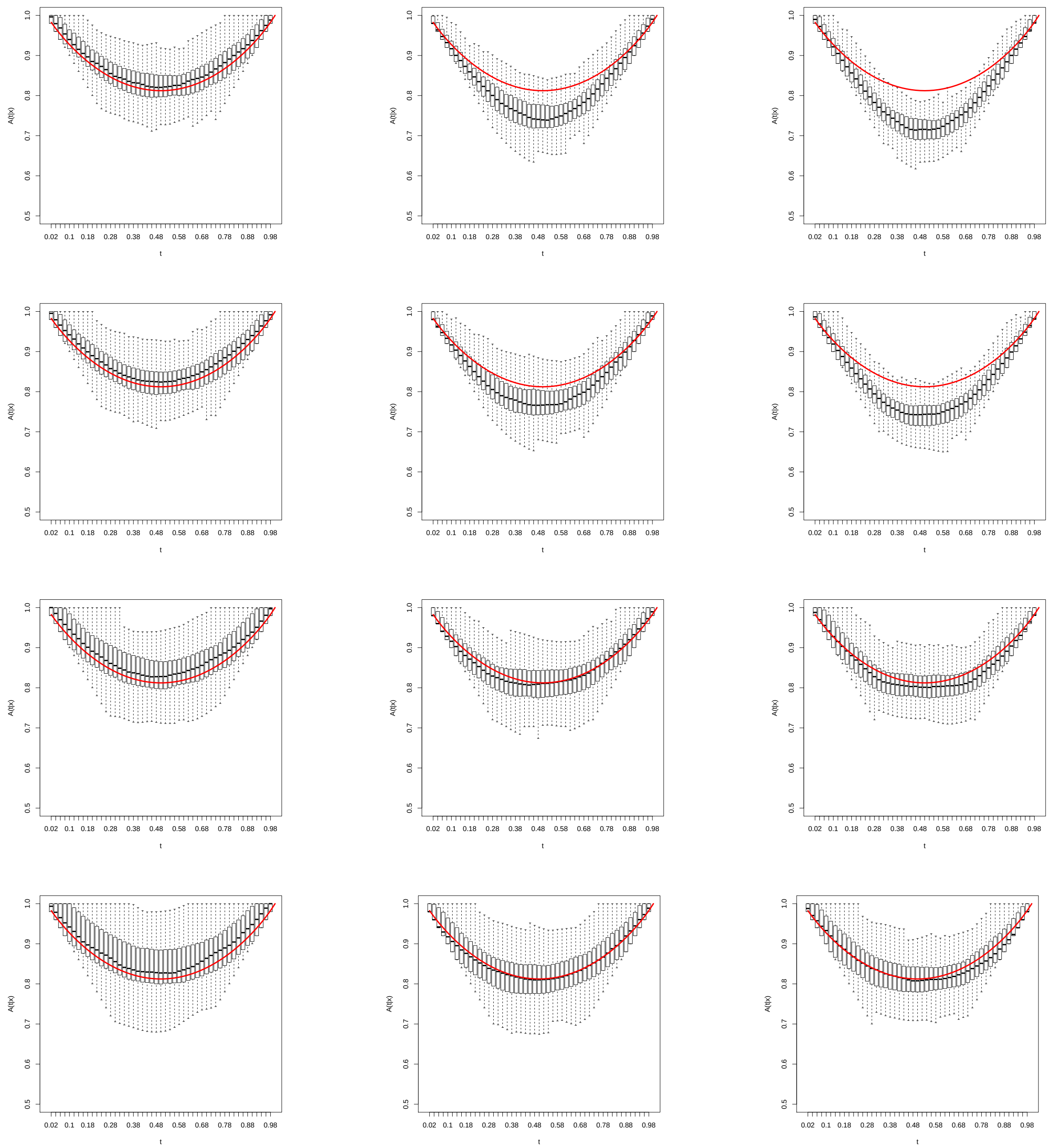

FIG S22. *Angular contamination: Estimation of* $A_0(.|0.7)$ *(full line) for the logistic distribution with* $n = 5000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

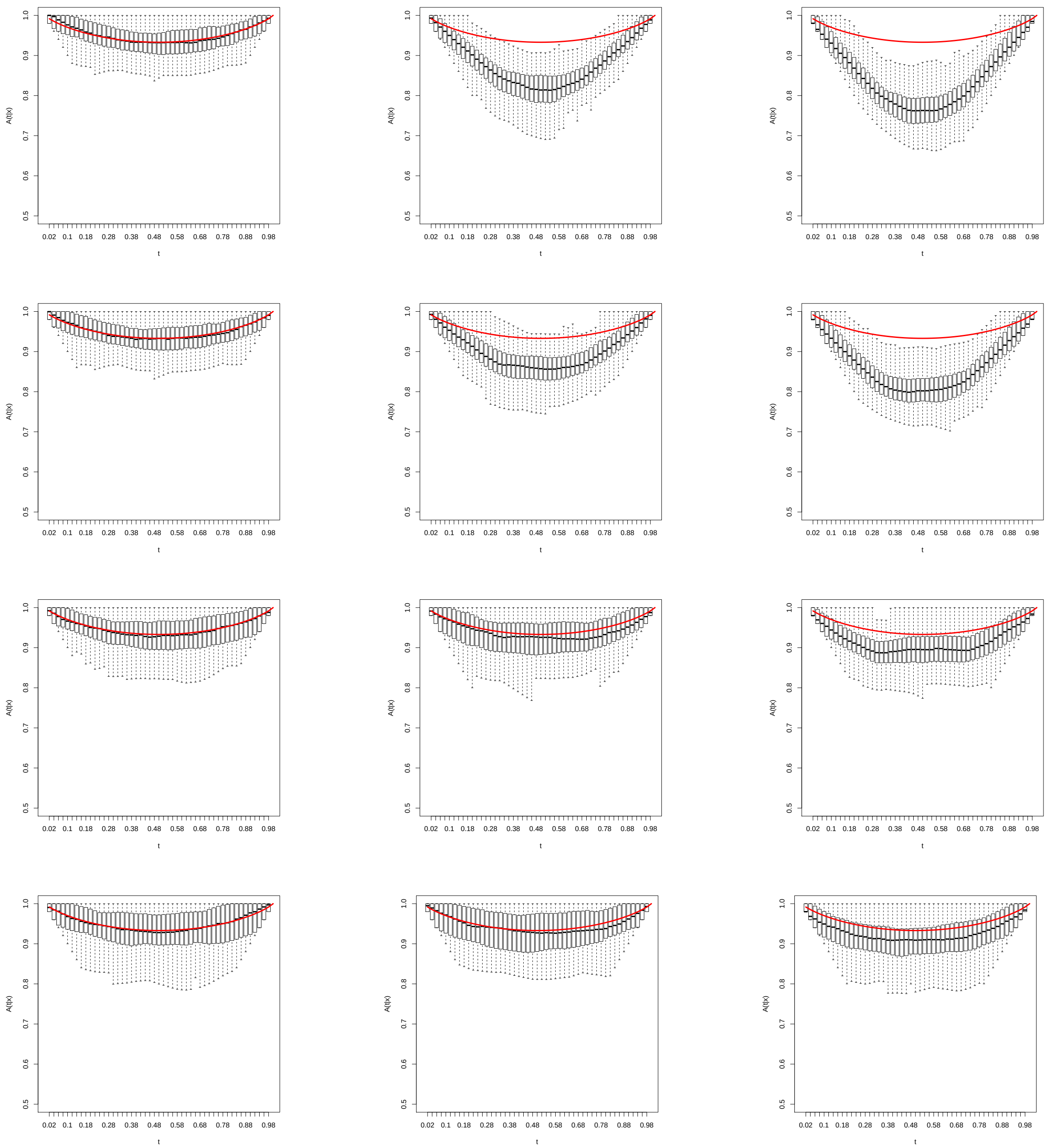

FIG S23. *Angular contamination: Estimation of* $A_0(.|0.9)$ *(full line) for the logistic distribution with* $n = 5000$. *From the top to the bottom:* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

4. Comparison with the Gardes and Girard (2015) estimator. Gardes and Girard (2015) have proposed an estimator of the tail copula, and hence of the Pickands dependence function, based on a kernel-type estimator of the multivariate conditional tail distribution. In this section, we compare this estimator with ours in case of contaminated and non-contaminated observations.

We compute their estimator using the algorithm described in their paper and especially with their grid of values for h, and thus, to be comparable, our estimator $\breve{A}_{\alpha,n}(.|x)$ is also computed with the same grid.

Note that the Gardes and Girard (2015) estimator is based on the original observations and due to its construction, a contaminated point corresponds to a point in the upper tail region. However, such points will not be considered as contaminated with our methodology since we transform the margins to the exponential scale, and thus we expect to see no effect of α in such a context.

As a final example, we consider a mixture model based again on the logistic distribution for F_ℓ, but this time, given $X = x$, F_c is the distribution of an independent bivariate random vector with unit Fréchet margins shifted by the vector $(7, 7)$. This example of contamination is illustrated in Figure S24 with the non-contaminated sample as circles and the contaminated pairs with crosses. Here ε is set to the value 0.1.

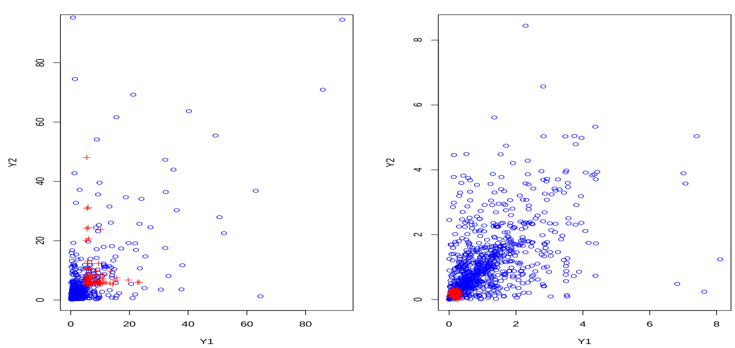

Fig S24. *Independent contamination - Shifted: on the left the original data and on the right the data after transformation into (approximate) unit exponentials. The non-contaminated observations are represented as circles whereas the contaminated pairs are represented as crosses. Here ε is set to the value 0.1.*

Figure S25 represents the MISE-plots as a function of ε for this contamination example, with the three positions: $x = 0.1, 0.3$ and 0.5. Clearly, the Gardes and Girard (2015) estimator is not robust with MISE-curves which increase drastically as ε increases, in particular when $x = 0.1$. This is expected since x close to 0 corresponds to complete dependence, while the contamination consists of independent pairs. Also, as expected the MISE-curves for $\breve{A}_{\alpha,n}(.|x)$ are flatted, without any effect of the value α. The boxplots of the Gardes and Girard (2015) estimator and $\breve{A}_{\alpha,n}(.|x)$ are given in Figures S26 till S28 based on 200 samples for the three positions $x = 0.1, 0.3$ and 0.5, respectively. As is clear, the Gardes and Girard (2015) estimator is not robust and has a high variability and bias when ε is getting large. On the contrary, $\breve{A}_{\alpha,n}(.|x)$ has very small variability and almost no bias, whatever α. Note also that in case of no-contamination ($\varepsilon = 0$), $\breve{A}_{\alpha,n}(.|x)$ performs well, with less variability than the Gardes and Girard (2015) estimator.

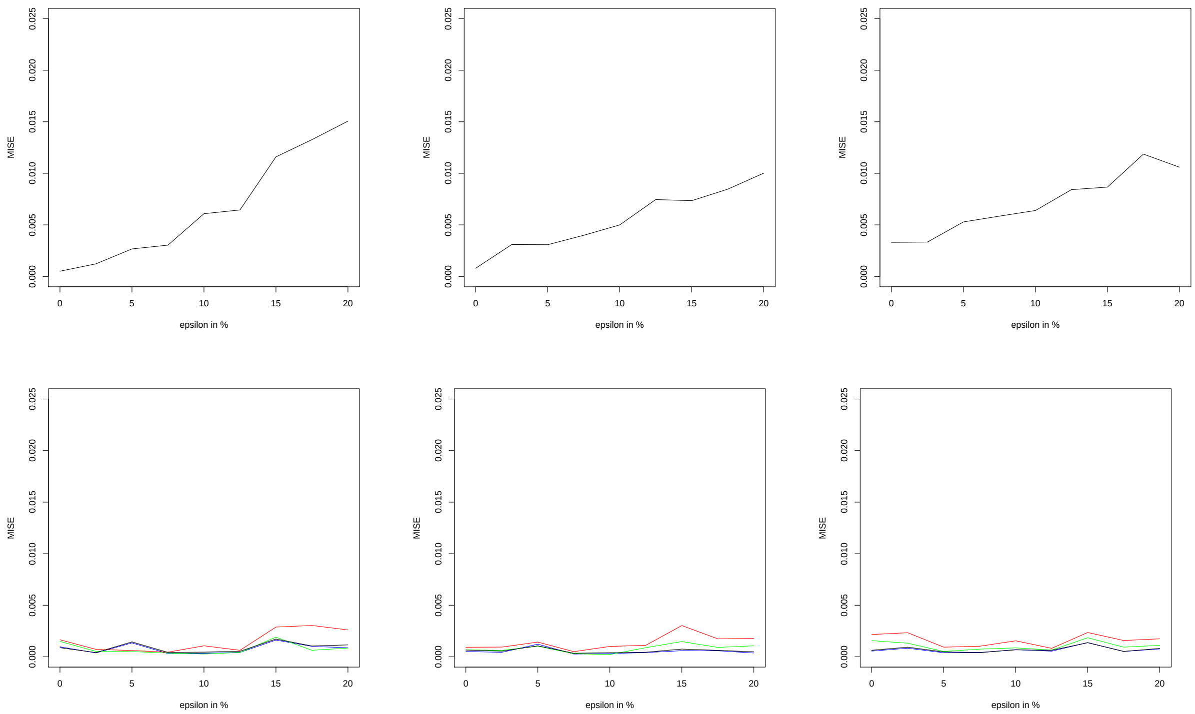

FIG S25. *Independent contamination - Shifted: MISE-plots as a function of* $\varepsilon \in \{0, 0.025, 0.05, ..., 0.2\}$*. Top row: Gardes and Girard (2015) estimator and bottom row: MDPD estimator with* $\alpha = 0$ *(black),* $\alpha = 0.1$ *(blue),* $\alpha = 0.5$ *(green) and* $\alpha = 1$ *(red), and* $x = 0.1, 0.3, 0.5$ *from the left to the right.*

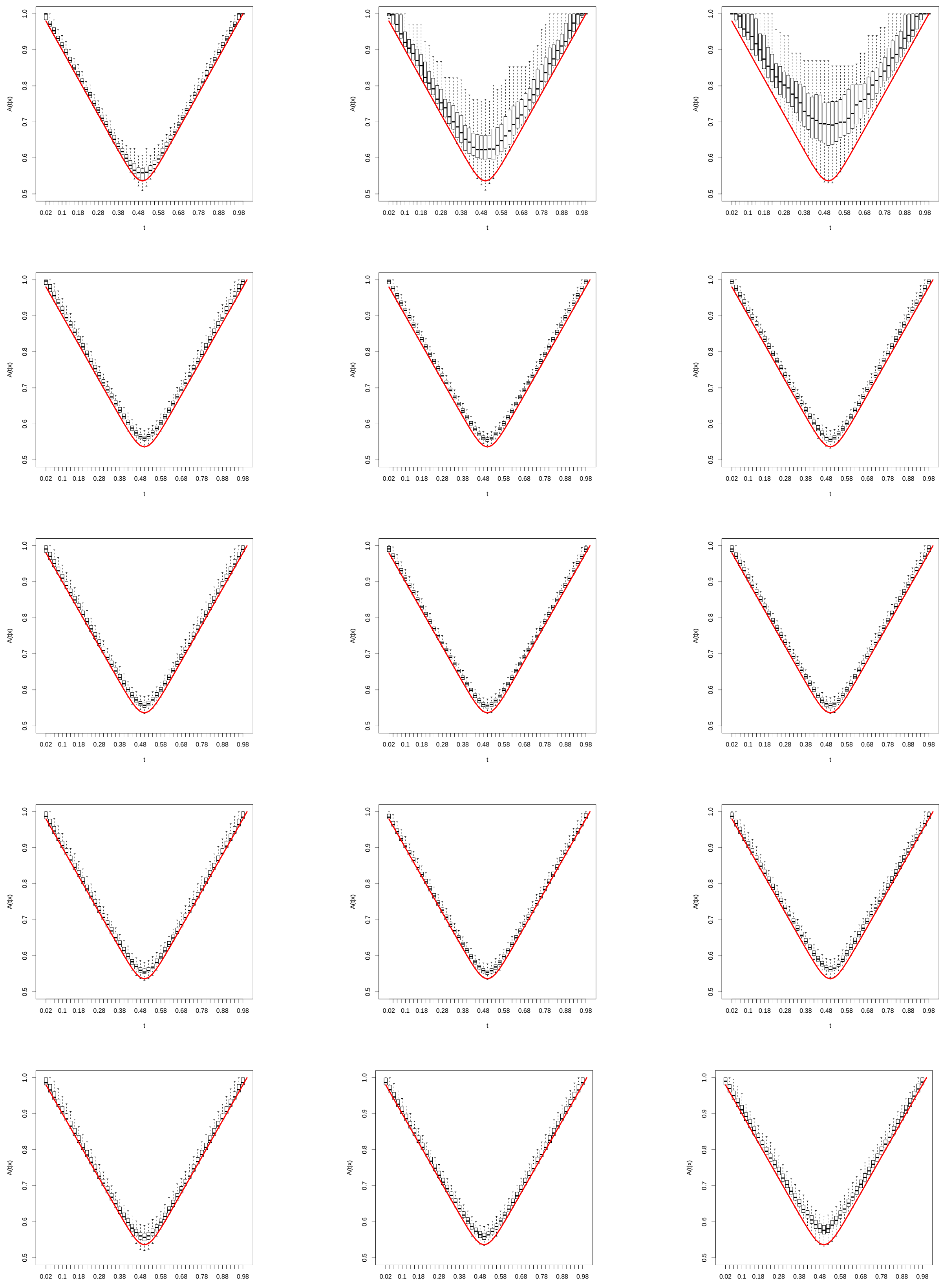

FIG S26. *Independent contamination - Shifted: Estimation of* $A_0(.|0.1)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom: Gardes and Girard (2015) estimator, MDPD estimator with* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

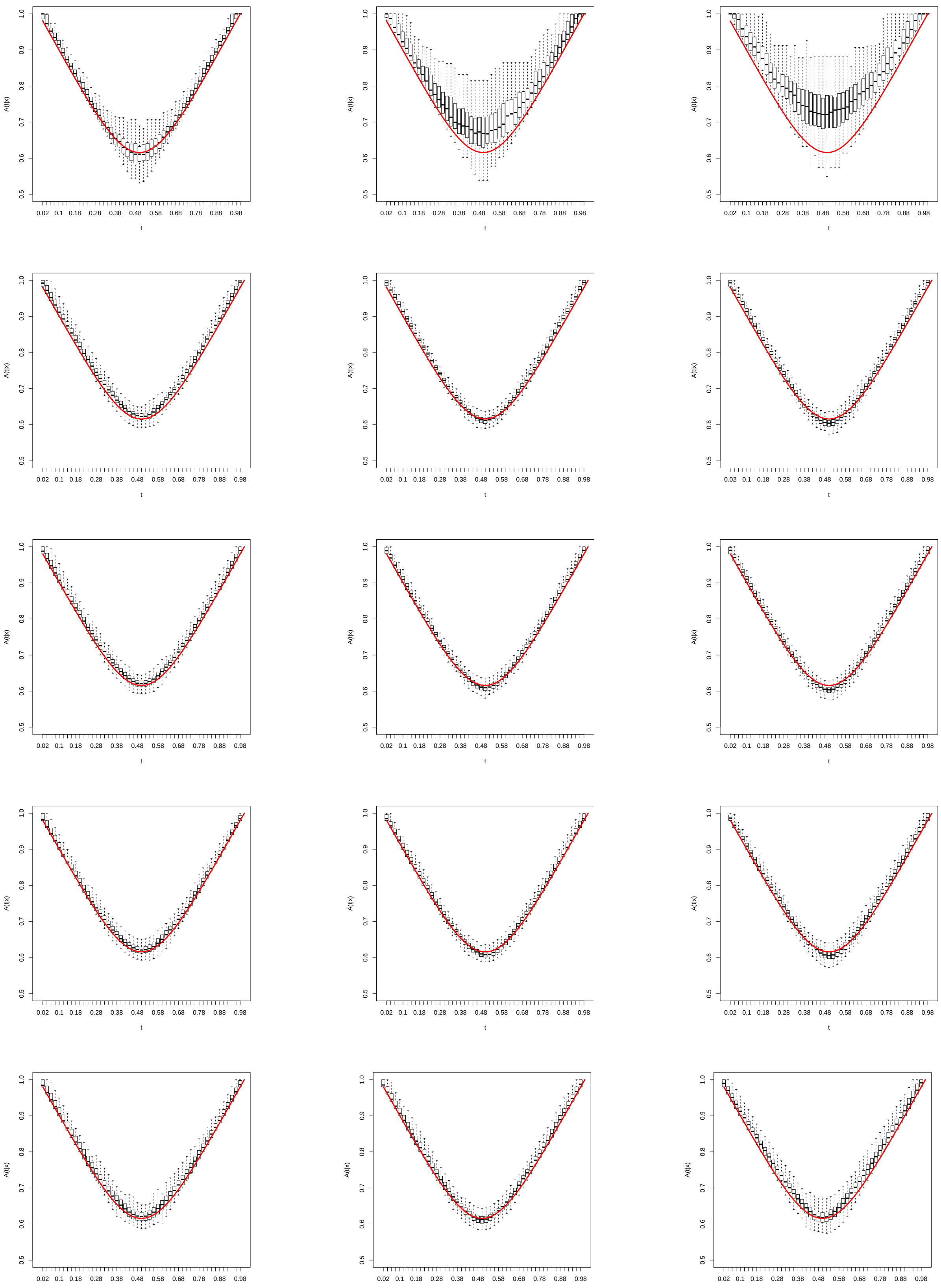

FIG S27. *Independent contamination - Shifted: Estimation of* $A_0(.|0.3)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom: Gardes and Girard (2015) estimator, MDPD estimator with* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

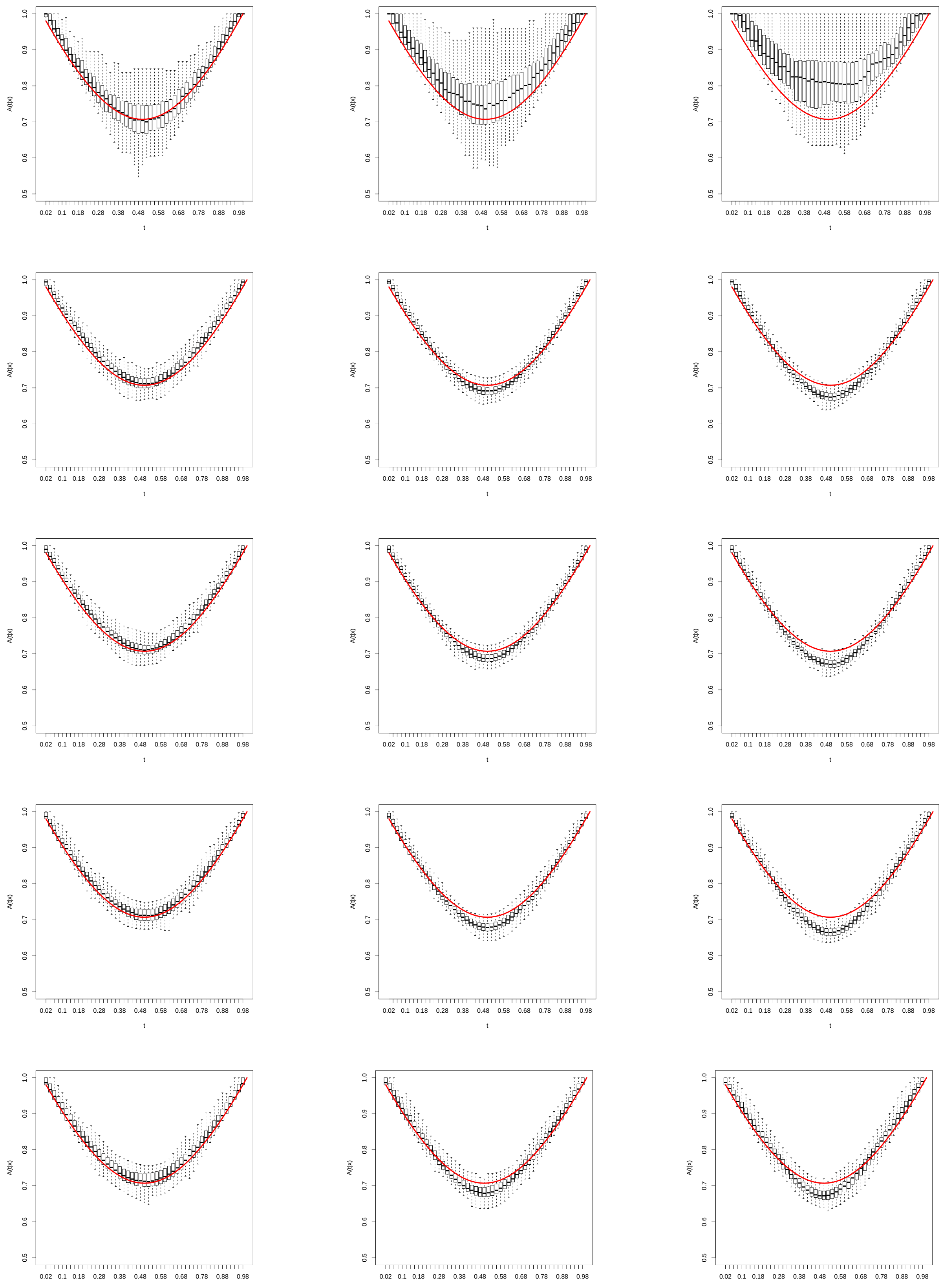

FIG S28. *Independent contamination - Shifted: Estimation of* $A_0(.|0.5)$ *(full line) for the logistic distribution with* $n = 1000$. *From the top to the bottom: Gardes and Girard (2015) estimator, MDPD estimator with* $\alpha = 0, 0.1, 0.5, 1$ *and from the left to the right:* $0\%, 10\%$ *and* 20% *of contamination.*

References.

Escobar-Bach, M., Goegebeur, Y. and Guillou, A. (2017). Local robust estimation of the Pickands dependence function. https://hal.archives-ouvertes.fr/hal-01340166 .

Gardes, L. and Girard, S. (2015). Nonparametric estimation of the conditional tail copula. *J. Multivariate Anal.* **137** 1–16. MR3332795

Department of Mathematics and Computer Science
University of Southern Denmark
Campusvej 55
5230 Odense M
Denmark
E-mail: escobar@imada.sdu.dk
yuri.goegebeur@imada.sdu.dk

Institut Recherche Mathématique Avancée, UMR 7501
Université de Strasbourg et CNRS
7 rue René Descartes
67084 Strasbourg cedex
France
E-mail: armelle.guillou@math.unistra.fr