A Supplementary Material

Here the proofs of the results are collected.

A.1 Proof of Proposition 1

We can generate $(X, Y) \sim F_p$ for $\rho \geqslant 0$ by

$$\begin{bmatrix} X \\ Y \end{bmatrix} = A \begin{bmatrix} U \\ V \\ W \end{bmatrix} \tag{A.1}$$

where U, V, W follow a symmetric unimodal distribution G and are i.i.d., and

$$A = \begin{bmatrix} \sqrt{1-\rho} & 0 & \sqrt{\rho} \\ 0 & \sqrt{1-\rho} & \sqrt{\rho} \end{bmatrix}.$$

For $G = N(0,1)$ the distribution of (A.1) equals (9). We now obtain $\xi(\rho) = E[\psi(u\sqrt{1-\rho} + w\sqrt{\rho})\psi(v\sqrt{1-\rho}+w\sqrt{\rho})]$. Since we are interested in $\rho \approx 0$, we can use the Taylor expansion (derived with $\delta = \sqrt{\rho}$) to obtain $\psi(u\sqrt{1-\rho}+w\sqrt{\rho}) = \psi(u) + w\sqrt{\rho}\psi'(u) + \frac{w^2\rho}{2}\psi''(u) + o(\rho)$ and similarly for the second factor, yielding 9 terms of which only one term remains, the others being $o(\rho)$ or zero since ψ is odd:

$$\begin{aligned}
\xi(\rho) = E\Bigg[&\psi(u)\left\{\psi(v) + w\sqrt{\rho}\psi'(v) + \frac{w^2\rho}{2}\psi''(v)\right\} \\
&+ w\sqrt{\rho}\psi'(u)\left\{\psi(v) + w\sqrt{\rho}\psi'(v) + \frac{w^2\rho}{2}\psi''(v)\right\} \\
&+ \frac{w^2\rho}{2}\psi''(u)\left\{\psi(v) + w\sqrt{\rho}\psi'(v) + \frac{w^2\rho}{2}\psi''(v)\right\}\Bigg] \\
=&\rho E\left[w^2\psi'(u)\psi'(v)\right] + o(\rho) \\
=&\rho E[\psi'(u)]E[\psi'(v)] + o(\rho)
\end{aligned}$$

Therefore $\xi'(0) = E[\psi'(u)]^2$ and we obtain $\mathrm{IF}((x,y), T, F_0) = \psi(x)\psi(y)/E[\psi']^2$.

A.2 Influence function for general ρ

We first consider the non Fisher-consistent functional $T_\psi = E[\psi(X)\psi(Y)]$. The raw influence function of T_ψ under the distribution F_ρ generated as in (A.1) is then

$$\mathrm{IF}_{raw}((x,y), T_\psi, F_\rho) = \psi(x)\psi(y) - E_{F_\rho}[\psi(X)\psi(Y)] \,.$$

Proof. Let $F_\epsilon = (1-\epsilon)F_\rho + \epsilon\Delta_{(x,y)}$. Then

$$T_\psi(F_\epsilon) = (1-\epsilon)E_{F_\rho}[\psi(X)\psi(Y)] + \epsilon E_{\Delta_{(x,y)}}[\psi(X)\psi(Y)] \; .$$

Differentiating with respect to ϵ at $\epsilon = 0$ yields $-E_{F_\rho}[\psi(X)\psi(Y)] + \psi(x)\psi(y)$. □

Now denote the finite sample version of T_ψ by $T_n = \frac{1}{n}\sum_{i=1}^n \psi(x_i)\psi(y_i)$. From the law of large numbers we have that T_n is strongly consistent for its functional value: $T_n \xrightarrow{a.s.} T_\psi(F_\rho)$ for $n \to \infty$. By the central limit theorem, we also have asymptotic normality of T_ψ:

$$\sqrt{n}(T_n - T_\psi(F_\rho)) \to N(0, V_{raw})$$

where the asymptotic variance V_{raw} is given by

$$\begin{aligned} V_{raw} &= E_\rho[\mathrm{IF}_{raw}((X,Y), T_\psi, F_\rho)^2] \\ &= E_\rho\left[(\psi(X)\psi(Y) - E_\rho[\psi(X)\psi(Y)])^2\right] \\ &= E_\rho\left[\psi(X)^2\psi(Y)^2\right] - E_\rho[\psi(X)\psi(Y)]^2 \quad . \end{aligned}$$

Now we switch to the Fisher-consistent functional $U_\psi(F) \coloneqq \xi^{-1}(T_\psi(F))$ given in (11). The general influence function defined in (12) then becomes

$$\begin{aligned} \mathrm{IF}((x,y), T_\psi, F_\rho) &\coloneqq \mathrm{IF}_{raw}((x,y), U_\psi, F_\rho) \\ &= \frac{\mathrm{IF}_{raw}((x,y), T_\psi, F)}{\xi'(\rho)} \\ &= \frac{\psi(x)\psi(y) - E_\rho[\psi(X)\psi(Y)]}{\xi'(\rho)} \end{aligned}$$

hence

$$\mathrm{IF}((x,y), T_\psi, F_\rho) = \frac{\psi(x)\psi(y) - C_\rho}{D_\rho} \tag{A.2}$$

where $C_\rho \coloneqq E_\rho[\psi(X)\psi(Y)]$ and $D_\rho \coloneqq \xi'(\rho)$ can be computed numerically to any given precision. For $\rho = 0$ this simplifies to the formula in Proposition 1. Note that the influence function has the same shape for all values of ρ (including $\rho = 0$), only the constants C_ρ and D_ρ differ which amounts to shifting and rescaling the IF along the vertical axis.

Now consider the estimator $T_n^* = \xi^{-1}(T_n)$ corresponding to the functional U_ψ. Since T_n is asymptotically normal, we can apply the delta method to establish the asymptotic normality of T_n^*. Using $(\xi^{-1}(x))' = 1/\xi'(\xi^{-1}(x))$ we obtain

$$\sqrt{n}(T_n^* - \rho) \to N(0, V)$$

where $V = V_{raw}/(\xi'(\rho))^2$ with V_{raw} as above. At $\rho = 0$ this corresponds to (14).

A.3 Relation with influence functions of rank correlations

At the model distribution F_0 of (9) the influence functions of the Quadrant and Spearman correlation (Croux and Dehon, 2010) and the normal scores (Boudt et al., 2012) correspond to those of certain ψ-product moments. This is not a coincidence, because if we write the rank transform as $g(x_i) = h(R_n(x_i))$ it tends to the function $\tilde{g}(x) = h(\Phi(x))$ when $n \to \infty$. If we put $\psi(x) := h(\Phi(x))$ we observe that (15) indeed holds, with $\mathrm{IF}(x, h, \Phi) = h(\Phi(x))/\int (h(\Phi))' d\Phi = \psi(x)/E[\psi']$.

For the quadrant correlation $h(u) = \mathrm{sign}(u - 1/2)$ we get the IF of the median:

$$\mathrm{IF}(x, L_h, \Phi) = \frac{\mathrm{sign}(x)}{2\Phi'(0)} = \sqrt{\frac{\pi}{2}}\,\mathrm{sign}(x)$$

and so $\gamma^* = \pi/2$ and $\mathrm{eff} = 4/\pi^2$.

For the normal scores rank correlation we have $h(u) = \Phi^{-1}(u)$ hence $\mathrm{IF}(x, L_h, \Phi) = x$ which is the influence function of the mean and thus unbounded, yielding $\gamma^* = \infty$ and $\mathrm{eff} = 1$. The truncated normal scores $h(u) = \Phi^{-1}\left([u]_\alpha^{1-\alpha}\right) = [\Phi^{-1}(u)]_{-b}^{b}$ where $\alpha = \Phi(-b)$ yields $\mathrm{IF}(x, L_h, \Phi) = \psi_b(x)/E[\psi_b']$, which is the influence function of Huber's ψ_b function.

For the Spearman correlation ($h(u) = u - 1/2$) we obtain

$$\mathrm{IF}(x, L_h, \Phi) = \frac{\Phi(x) - 1/2}{E[(\Phi')^2]} = 2\sqrt{\pi}\left(\Phi(x) - \frac{1}{2}\right)$$

which is also the influence function of the Hodges-Lehmann estimator and the Mann-Whitney and Wilcoxon tests (Hampel et al., 1986). It yields $\gamma^* = \pi$ and $\mathrm{eff} = 9/\pi^2$.

A.4 Proof of Proposition 2 and Corollary 1

Proof of Proposition 2. We give the proof for the maximum upward bias (the result for the maximum downward bias then follows by replacing Y by $-Y$). The uncontaminated distribution of (X, Y) is $F = F_\rho$ from (A.1). Since $\psi(X)$ and $\psi(Y)$ have the same distribution and ψ is odd and bounded we find $E_F[\psi(X)] = E_F[\psi(Y)] = 0$ and $E_F[\psi(X)^2] = E_F[\psi(Y)^2]$. Now consider the contaminated distribution $G = (1-\varepsilon)F_\rho + \varepsilon H$ where H is any distribution. At G we obtain

$$\mathrm{Cor}_G(\psi(X), \psi(Y)) = \frac{E_G[(\psi(X) - E_G[\psi(X)])(\psi(Y) - E_G[\psi(Y)])]}{\sqrt{E_G[(\psi(X) - E_G[\psi(X)]^2)]E_G[(\psi(Y) - E_G[\psi(Y)])^2]}}$$

which works out to be

$$\frac{(1-\varepsilon)\operatorname{Cov}_F(U,V)+\varepsilon E_H[UV]-\varepsilon^2 E_H[U]E_H[V]}{\sqrt{((1-\varepsilon)V_F+\varepsilon E_H[U^2]-\varepsilon^2 E_H[U]^2)((1-\varepsilon)V_F+\varepsilon E_H[V^2]-\varepsilon^2 E_H[V]^2)}} \tag{A.3}$$

where we denote $U := \psi(X)$ and $V := \psi(Y)$ to save space, as well as $V_F := \operatorname{Var}_F(U) = E_F[\psi(X)^2] = E_F[\psi(Y)^2] = \operatorname{Var}_F(V)$.

We will show the proof for $\rho = 0$ which implies that U and V are independent hence $\operatorname{Cov}_F(U,V) = 0$ as this reduces the notation, but the proof remains valid if the term $(1-\varepsilon)\operatorname{Cov}_F(U,V) = (1-\varepsilon)V_F T_\psi(F)$ is kept. The proof consists of two parts. We first show that the contaminated correlation (A.3) is bounded from above by

$$C(\varepsilon) := \frac{\varepsilon M^2}{(1-\varepsilon)V_F+\varepsilon M^2} \tag{A.4}$$

and then we provide a sequence of contaminating distributions H_n for which (A.3) tends to this upper bound.

1. Suppose first that $E_H[U]E_H[V] \leqslant 0$. Then we have for the numerator of (A.3):

$$\begin{aligned} E_H[UV]-\varepsilon E_H[U]E_H[V] &\leqslant E_H[UV]-E_H[U]E_H[V] \\ &\leqslant \sqrt{(E_H[U^2]-E_H[U]^2)(E_H[V^2]-E_H[V]^2)}\ . \end{aligned}$$

Now consider the denominator of (A.3) and note that

$$\begin{aligned} &\sqrt{((1-\varepsilon)V_F+\varepsilon(E_H[U^2]-\varepsilon E_H[U]^2))((1-\varepsilon)V_F+\varepsilon(E_H[V^2]-\varepsilon E_H[V]^2))} \geqslant \\ &\qquad \sqrt{((1-\varepsilon)V_F+\varepsilon(E_H[U^2]-E_H[U]^2))((1-\varepsilon)V_F+\varepsilon(E_H[V^2]-E_H[V]^2))} \end{aligned}$$

because $E_H[U^2]-E_H[U]^2 \geqslant 0$, $E_H[U^2] \geqslant 0$, $E_H[U]^2 \geqslant 0$ and $0 \leqslant \varepsilon \leqslant 1$. Therefore, we can bound (A.3) from above by

$$\frac{\varepsilon\sqrt{(E_H[U^2]-E_H[U]^2)(E_H[V^2]-E_H[V]^2)}}{\sqrt{((1-\varepsilon)V_F+\varepsilon(E_H[U^2]-E_H[U]^2))((1-\varepsilon)V_F+\varepsilon(E_H[V^2]-E_H[V]^2))}}$$

and this quantity is maximal when $(E_H[U^2]-E_H[U]^2)$ and $(E_H[V^2]-E_H[V]^2)$ are as large as possible. Their supremum is in fact M^2. Therefore, (A.3) is less than or equal to (A.4).

2. Suppose now that $E_H[U]E_H[V] > 0$. We will first show that the numerator is bounded as follows:

$$E_H[UV]-\varepsilon E_H[U]E_H[V] \leqslant \sqrt{(E_H[U^2]-\varepsilon E_H[U]^2)(E_H[V^2]-\varepsilon E_H[V]^2)}\ . \tag{A.5}$$

By squaring both sides we find that this is equivalent to showing

$$\begin{aligned} & E_H[UV]^2 - 2\varepsilon E_H[U]E_H[V]E_H[UV] \\ \leqslant\ & E_H[U^2]E_H[V^2] - \varepsilon(E_H[U^2]E_H[V]^2 + E_H[U]^2E_H[V^2]) \end{aligned}$$

which is equivalent to

$$E_H[U^2]E_H[V^2]-E_H[UV]^2+\varepsilon(2E_H[U]E_H[V]E_H[UV]-E_H[U^2]E_H[V]^2-E_H[U]^2E_H[V^2]) \geqslant 0. \tag{A.6}$$

We know that (A.5) holds for $\varepsilon = 1$ as it is equivalent to $\mathrm{Cov}_H(U,V) \leqslant \sqrt{\mathrm{Var}_H(U)\,\mathrm{Var}_H(V)}$ so (A.6) is true in that case.

The general version of (A.6) with $\varepsilon \leqslant 1$ equals the LHS for $\varepsilon = 1$, plus $(1-\varepsilon)$ times

$$E_H[U]^2E_H[V^2] - 2E_H[U]E_H[V]E_H[UV] + E_H[U^2]E_H[V]^2 \ . \tag{A.7}$$

Therefore, it would suffice to prove that (A.7) is nonnegative. We know that $|E_H[UV]| \leqslant \sqrt{E_H[U^2]E_H[V^2]}$ by Cauchy-Schwarz. Since $E_H[U]E_H[V] > 0$ we obtain

$$\begin{aligned} & E_H[U]^2E_H[V^2] - 2E_H[U]E_H[V]E_H[UV] + E_H[U^2]E_H[V]^2 \\ \geqslant\ & E_H[U]^2E_H[V^2] - 2E_H[U]E_H[V]\sqrt{E_H[U^2]E_H[V^2]} + E_H[U^2]E_H[V]^2 \\ =\ & \left(E_H[U]\sqrt{E_H[V^2]} - E_H[V]\sqrt{E_H[U^2]}\right)^2 \geqslant 0 \ . \end{aligned}$$

Now that we have shown (A.5) we can proceed as in part 1, since (A.3) is bounded from above by

$$\frac{\varepsilon\sqrt{(E_H[U^2]-\varepsilon E_H[U]^2)(E_H[V^2]-\varepsilon E_H[V]^2)}}{\sqrt{((1-\varepsilon)V_F+\varepsilon(E_H[U^2]-\varepsilon E_H[U]^2))((1-\varepsilon)V_F+\varepsilon(E_H[V^2]-\varepsilon E_H[V]^2))}}$$

and this quantity is maximal when $(E_H[U^2]-\varepsilon E_H[U]^2)$ and $(E_H[V^2]-\varepsilon E_H[V]^2)$ are as large as possible. Their supremum is again M^2, so (A.3) is less than or equal to (A.4).

3. Now all that is left to show is that the upper bound (A.4) is sharp. Let $(k_n)_{n\in\mathbb{N}}$ be a sequence such that $\lim_{n\to\infty}\psi(k_n) = \sup_x|\psi(x)| = M$ and consider the sequence of 'worst-placed' contaminating distributions

$$H_n = \frac{1}{2}\Delta_{(k_n,k_n)} + \frac{1}{2}\Delta_{(-k_n,-k_n)} \ . \tag{A.8}$$

For the numerator of (A.3) we have $\lim_{n\to\infty} \varepsilon E_{H_n}[UV] - \varepsilon^2 E_{H_n}[U]E_{H_n}[V] = \varepsilon M^2$ since $E_{H_n}[U] = 0 = E_{H_n}[V]$, and for the denominator we obtain analogously

$$\lim_{n\to\infty} \sqrt{((1-\varepsilon)V_F + \varepsilon E_{H_n}[U^2])((1-\varepsilon)V_F + \varepsilon E_{H_n}[V^2])} \;=\; (1-\varepsilon)V_F + \varepsilon M^2$$

so we reach the upper bound (A.4). The proof for the maximum downward bias is entirely similar, and there the worst placed contaminating distributions are of the form $H_n = \frac{1}{2}\Delta_{(k_n,-k_n)} + \frac{1}{2}\Delta_{(-k_n,k_n)}$. QED.

Proof of Corollary 1. For the breakdown value we start from $F = F_1$, that is $\rho = 1$ and $X = Y$, so $\mathrm{Cov}_F(\psi(X), \psi(Y)) = \mathrm{Var}_F(\psi(X))$ hence $T_\psi(F) = 1$. From Proposition 2 we know that

$$\inf_{G\in\mathcal{F}_\varepsilon} T_\psi(G) = \frac{(1-\varepsilon)\,\mathrm{Var}_F(\psi(X))\,T_\psi(F) - \varepsilon M^2}{(1-\varepsilon)\,\mathrm{Var}_F(\psi(X)) + \varepsilon M^2} \quad .$$

For this to be nonpositive the numerator has to be, i.e. $(1-\varepsilon)\,\mathrm{Var}_F(\psi(X)) - \varepsilon M^2 \leqslant 0$. The smallest ε for which this holds is indeed $\mathrm{Var}_F(\psi(X))/(\mathrm{Var}_F(\psi(X)) + M^2)$. QED.

Note that we can rewrite the breakdown value as $\varepsilon^* = 1 - (E_F[(\psi/M)^2] + 1)^{-1}$ so it is a strictly increasing function of $E_F[(\psi/M)^2]$. This implies that the maximizer of the breakdown value is $\psi(x) = \mathrm{sign}(x)$ which maximizes $E_F[(\psi/M)^2] = 1$, hence $\varepsilon^* = 0.5$ (this yields the quadrant correlation). Interestingly, the breakdown value of the scale M-estimator S defined by $\mathrm{ave}_i\, \rho(x_i/S) = E_F[\rho]$ where $\rho(z) := \psi^2(z)$ is also determined by the ratio $E_F[\rho]/M^2 = E_F[(\psi/M)^2]$, see e.g. Maronna et al. (2006).

A.5 Relation with breakdown values of rank correlations

The breakdown values of the rank correlations in Table 2 were derived by Capéraà and Garralda (1997) and Boudt et al. (2012), but not for the ε-contamination model (16). Instead they used *replacement contamination*, which means you can take out a certain fraction of the observations and replace them by arbitrary points. In fact ε-contamination is a special case of this, which corresponds to replacing a mass ε distributed exactly like the original distribution F, whereas in general one could replace an arbitrary part of F. Therefore the breakdown value for replacement is always less than or equal to that for

ε-contamination. However, in many situations the result turns out to be the same, as is the case here.

For rank correlations in the replacement model, Capéraà and Garralda (1997) and Boudt et al. (2012) showed that given a sorted sample $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_1 < \cdots < x_n$ and $x_i = y_i$ for all $i \in \{1, \ldots, n\}$, the worst possible bias is reached by replacing the highest and the lowest y_i by values beyond the other end of the range.

We can in fact obtain the same type of configuration through the ε-contamination model. Let us start from perfectly correlated data, that is $x_i = y_i$ for all $i \in \{1, \ldots, n\}$. Then choose a sequence of contaminating distributions $H_n = \frac{1}{2}\Delta_{(-k_n,k_n)} + \frac{1}{2}\Delta_{(k_n,-k_n)}$ in which the k_n are positive and tend to infinity, so the horizontal and vertical coordinates of the outliers move outside the range of the original data values. The resulting rank pairs then have the same configuration as was constructed for breakdown under replacement. Therefore the ε-contamination breakdown values of rank correlations equal those under replacement.

A.6 Construction of the optimal transformation

Theorem 3.1 in (Hampel et al., 1981) says that for any $0 < c < \infty$ and large enough $k > 0$ there exist positive constants $0 < b < c$, A and B such that $\tilde{\psi}$ defined by

$$\tilde{\psi}(z) = \begin{cases} z & \text{if } 0 \leqslant |z| \leqslant b \\ \sqrt{A(k-1)}\tanh\left(\frac{B}{2}\sqrt{\frac{k-1}{A}}(c - |z|)\right)\text{sign}(z) & \text{if } b \leqslant |z| \leqslant c \\ 0 & \text{if } c \leqslant |z| \end{cases} \tag{A.9}$$

satisfies

$$b = \sqrt{A(k-1)}\tanh\left(\frac{1}{2}\sqrt{\frac{(k-1)B^2}{A}}(c-b)\right) \quad ,$$

$A = \int_{-c}^{c} \tilde{\psi}(x)^2 d\Phi(x)$, $B = \int_{-c}^{c} \tilde{\psi}'(x) d\Phi(x)$ and $\kappa^*(\tilde{\psi}) = k$. Theorem 4.1 then says that this function $\tilde{\psi}$ minimizes the asymptotic variance among all odd functions ψ satisfying (21) subject to $\kappa^*(\psi) \leqslant k$, and that this optimal solution is unique (upto a positive nonzero factor). It can be verified that for a given value of c there is a strictly monotone relation between k and b, so we have decided to parametrize $\tilde{\psi}$ by the easily interpretable tuning constants b and c. A short R-script is available that for any b and c derives the other

constants A, B and k, in turn yielding $q_1 = \sqrt{A(k-1)}$ and $q_2 = (B/2)\sqrt{(k-1)/A}$. For instance, for $b = 1.5$ and $c = 4$ we obtain $A = 0.7532528$, $B = 0.8430849$ and $k = 4.1517212$ hence $q_1 = 1.540793$ and $q_2 = 0.8622731$, yielding the gross-error-sensitivity $(b/B)^2 = 3.16$ and the efficiency $(B^2/A)^2 = 0.890$.

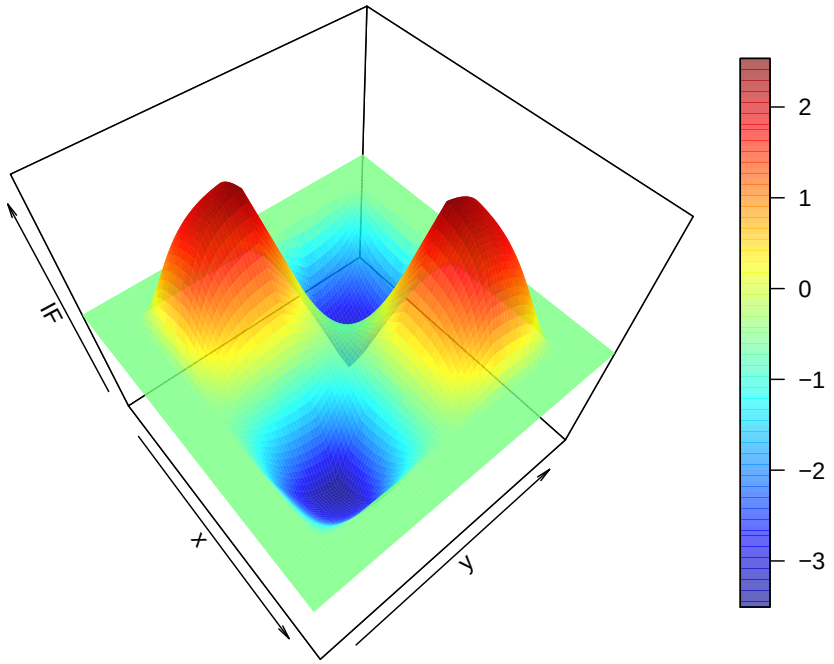

Figure 13: Influence function of T_ψ at F_ρ for $\rho = 0.5$.

Figure 13 shows the influence function (A.2) at $\rho = 0.5$ for the psi-function $\psi_{b,c}$ of (22). The influence function has the same shape at other values of ρ, up to shifting and rescaling the surface along the vertical axis, as shown in Section A.2.

A.7 Proof of Propositions 3 and 4

Proof of Proposition 3. It is assumed that (X, Y) follows a bivariate Gaussian distribution. Due to the invariance properties of correlation, we can assume w.l.o.g. that the distribution is F_ρ with center 0, unit variances and true correlation $-1 < \rho < 1$. The assumption that $\mathrm{Cor}(g_X(X), g_Y(Y)) = 0$ is equivalent to its numerator being zero, i.e.

$T(F_\rho) = E_\rho[\psi(X)\psi(Y)] = 0$. We need to show that this implies $\rho = 0$, from which independence between the components follows.

We first show that $\rho > 0$ implies that $T(F_\rho) = E_\rho[\psi(X)\psi(Y)] > 0$. Denote $A = \{(x,y) \in \mathbb{R}^2;\ xy > 0\}$ and $B = \{(x,y) \in \mathbb{R}^2;\ xy < 0\}$. We then have:

$$\begin{aligned}
E_\rho[\psi(X)\psi(Y)] &= \int_{\mathbb{R}^2} \psi(x)\psi(y) f_\rho(x,y)dxdy \\
&= \int_A \psi(x)\psi(y) f_\rho(x,y)dxdy + \int_B \psi(x)\psi(y) f_\rho(x,y)dxdy \\
&= \int_A \psi(x)\psi(y) f_\rho(x,y)dxdy + \int_A \psi(x)\psi(-y) f_\rho(x,-y)dxdy \\
&= \int_A \psi(x)\psi(y) f_\rho(x,y)dxdy - \int_A \psi(x)\psi(y) f_\rho(x,-y)dxdy \\
&= \int_A \psi(x)\psi(y) \left\{f_\rho(x,y) - f_\rho(x,-y)\right\} dxdy\ .
\end{aligned}$$

In the third equality we have changed the integration variables from (x,y) to $(x,-y)$. This transformation has Jacobian 1 and maps B to A. In the fourth equality we have used that ψ is odd so $\psi(-y) = -\psi(y)$. Now note that $f_\rho(x,y) > f_\rho(x,-y)$ for all $(x,y) \in A$ since $\rho > 0$. We conclude that $T(F_\rho) > 0$. The proof that $T(F_\rho) < 0$ for $\rho < 0$ follows by symmetry. Therefore, $T(F_\rho) = 0$ implies $\rho = 0$.

Proof of Proposition 4.

(i) From (23) and equivariance it follows that $\hat{\mu}_Y = \alpha + \beta\hat{\mu}_X$ and $\hat{\sigma}_Y = \beta\hat{\sigma}_X$ hence $g_Y(y_i) = (y_i - \hat{\mu}_Y)/\hat{\sigma}_Y = (x_i - \hat{\mu}_X)/\hat{\sigma}_X = g_X(x_i)$ for all i.

(ii) From $\mathrm{Cor}(g_X(x_i), g_Y(y_i)) = 1$ and $\mathrm{ave}_i(g_X(x_i)) = 0$ and $\mathrm{ave}_i(g_Y(y_i)) = 0$ it follows that there is a constant $\gamma > 0$ such that $g_Y(y_i) = \gamma g_X(x_i)$ for all i. For the i for which $|x_i - \hat{\mu}_X|/\hat{\sigma}_X \leqslant b$ and $|y_i - \hat{\mu}_Y|/\hat{\sigma}_Y \leqslant b$ it holds that $g_Y(y_i) = (y_i - \hat{\mu}_Y)/\hat{\sigma}_Y$ and $g_X(x_i) = (x_i - \hat{\mu}_X)/\hat{\sigma}_X$ hence $(y_i - \hat{\mu}_Y)/\hat{\sigma}_Y = \gamma(x_i - \hat{\mu}_X)/\hat{\sigma}_X$ which implies (23) with $\alpha = \hat{\mu}_Y - \gamma\hat{\mu}_X\hat{\sigma}_Y/\hat{\sigma}_X$ and $\beta = \gamma\hat{\sigma}_Y/\hat{\sigma}_X$.

A.8 Illustration of anomaly detection based on robust location and scatter

To visualize things we consider a small bivariate data set, about the star cluster CYG OB1 consisting of 47 stars in the direction of Cygnus. Their Hertzsprung-Russell diagram is a

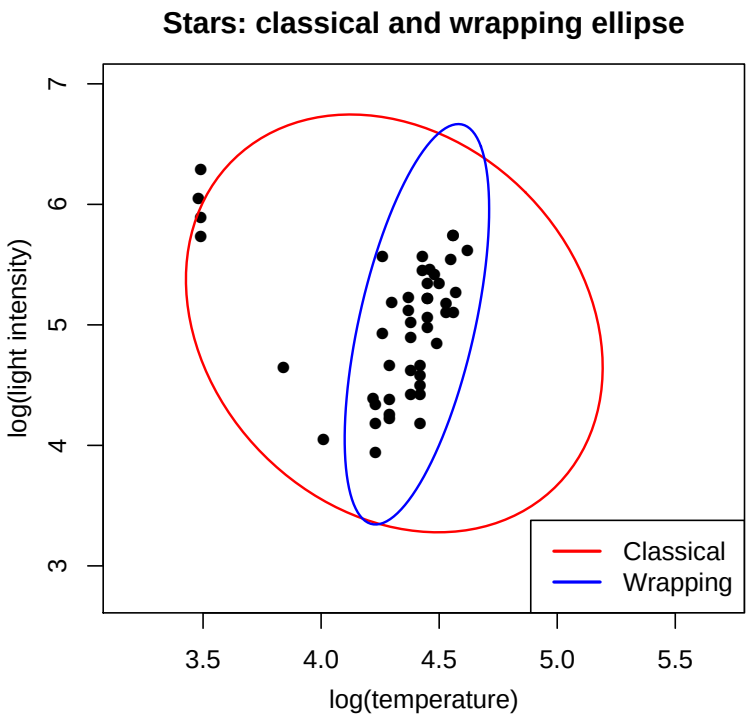

Figure 14: Plot of the 47 stars with their classical tolerance ellipse (red) and the one based on wrapped covariance (blue).

plot of the logarithm of each star's light intensity versus the logarithm of its temperature. The data can be found on page 27 of (Rousseeuw and Leroy, 1987) and is plotted in Figure 14. We see that the majority of the stars (the so-called main sequence stars) follows a certain upward trend, whereas there are four anomalous stars in the upper left corner. These are red giant stars. In this data set the anomalies are measured correctly, but they belong to a different population.

The classical correlation between the variables is -0.21 which would indicate a negative relation. However, this decreasing trend is caused by the four outliers, and without them the trend would be increasing. Indeed, the wrapped correlation is 0.57 indicating a positive relation. Figure 14 shows the 99% tolerance ellipse derived from the classical mean and covariance matrix, in red. The four outliers have pulled the ellipse toward them, making them lie on its boundary. In contrast, the tolerance ellipse from the wrapped mean and covariance (in blue) fits the majority of the stars, leaving aside the four outliers.

Of course, in higher dimensions we can no longer plot the data points or draw the tolerance ellipsoids. But in that case we can still look at the classical Mahalanobis distance

of each case $\boldsymbol{x}_i$ given by

$$\mathrm{MD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})} \ , \tag{A.10}$$

in which $\hat{\boldsymbol{\mu}}$ is the arithmetic mean and $\hat{\boldsymbol{\Sigma}}$ the empirical covariance matrix. The left panel of Figure 15 plots $\mathrm{MD}(\boldsymbol{x}_i)$ versus the case number i. In this plot the four giant stars lie close to the cutoff value $\sqrt{\chi^2_{d,0.99}}$ for dimension $d = 2$. But they are easily detected in the right hand panel, which plots the robust distances given by (A.10) where this time $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the location and scatter matrix obtained from the wrapped data. These robust estimates have thus allowed us to detect the anomalies.

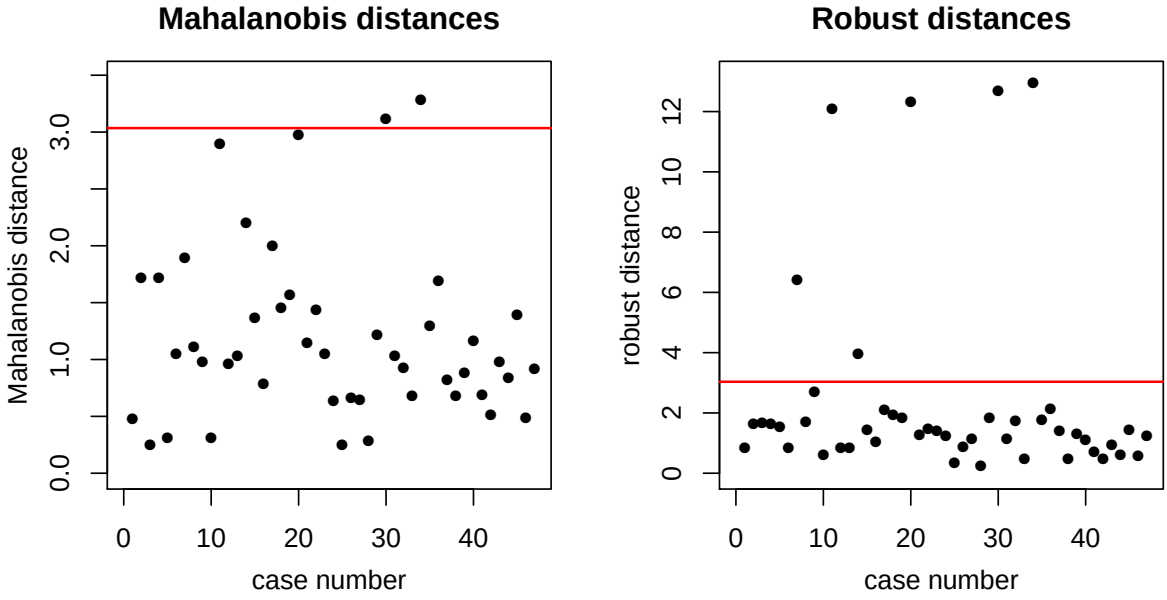

Figure 15: Classical distances of the stars (left) and their robust distances based on wrapped location and covariance (right).

A.9 Distance correlation after transformation

The distance correlation dCor between random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined by the Pearson correlation between the doubly centered interpoint distances of $\boldsymbol{X}$ and those of $\boldsymbol{Y}$ (Székely et al., 2007). It always lies between 0 and 1. Interestingly, $\mathrm{dCor}(\boldsymbol{X}, \boldsymbol{Y})$ can also be written in terms of the characteristic functions of the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ and the marginal distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Using this result Székely et al. (2007) prove that

$\mathrm{dCor}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ implies that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, which is not true for the plain Pearson correlation (except for multivariate Gaussian data).

The population $\mathrm{dCor}(\boldsymbol{X}, \boldsymbol{Y})$ is estimated by its finite-sample version $\mathrm{dCor}(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ which is a test statistic for dependence. Unfortunately this statistic is very sensitive to outliers. To illustrate this we first generate $n = 100,000$ data points from the standard bivariate Gaussian distribution, which has $\mathrm{dCor}(\boldsymbol{X}, \boldsymbol{Y}) = 0$, and replace a single observation by an outlier in the point (a, a). The left panel of Figure 16 shows $\mathrm{dCor}(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ as a function of a. For this we used the fast algorithm of Huo and Székely (2007) as implemented in the function *dcor2d* in the R package *energy*, which can handle such a large sample size n. For $a = 0$ we obtain $\mathrm{dCor}(\boldsymbol{X}_n, \boldsymbol{Y}_n) \approx 0$ but by letting a increase we can bring the result close to 1, even though the remaining $99,999$ points were generated independently.

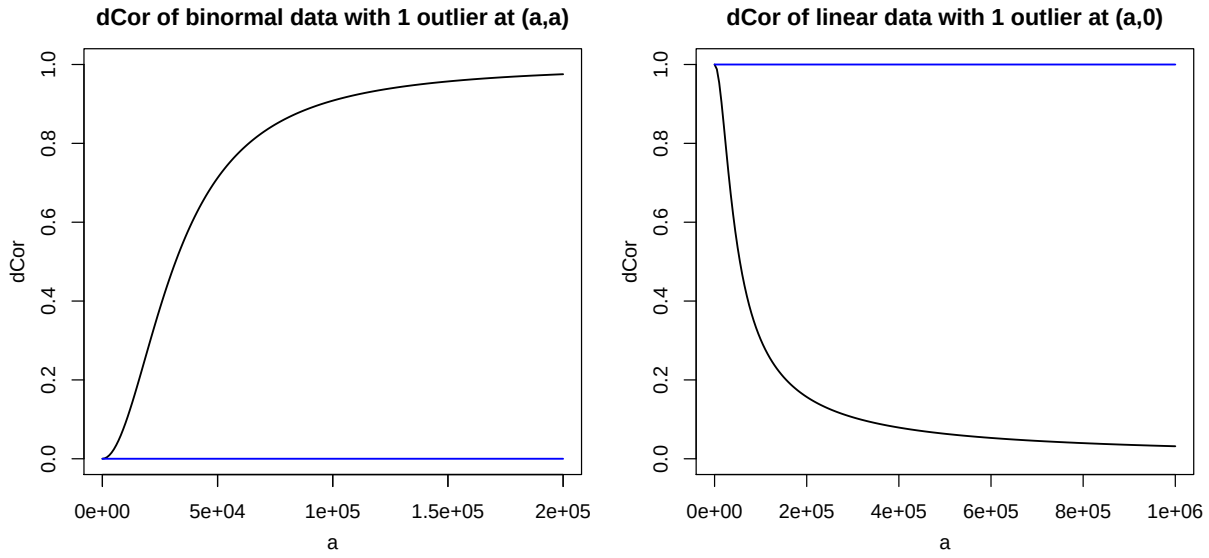

Figure 16: Left panel: distance correlation (black curve) and its robust version (blue curve) of a data set with $99,999$ standard Gaussian data points and one outlier at (a, a) versus a. Right panel: distance correlation of data with $99,999$ data points (x_i, x_i) with standard Gaussian x_i and one outlier at $(a, 0)$.

We can also do the opposite, by starting from a perfectly dependent setting. For this we generate $\boldsymbol{X}_n$ from the univariate standard Gaussian distribution, and take $\boldsymbol{Y}_n := \boldsymbol{X}_n$ so that $\mathrm{dCor}(\boldsymbol{X}_n, \boldsymbol{Y}_n) = 1$. Then we replace a single observation by an outlier in the point $(a, 0)$. In the right panel of Figure 16 we now see that we can bring $\mathrm{dCor}(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ close to

0 by this single outlier out of $100,000$ data points.

We now apply our methodology of first transforming the individual variables. For this we use the function g of (25) where $\hat{\mu}_j$ is the sample median and $\hat{\sigma}_j$ is the median absolute deviation. For the ψ-function we use the sigmoid $\psi(z) = \tanh(z)$. After this transformation we compute the distance correlation. This combined method no longer requires the first moments of the original variables to exist because ψ is bounded, and its population version is again zero if and only if the original $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, since ψ is invertible. The blue lines in Figure 16 are the result of applying the combined method, which by construction is insensitive to the outlier.

The robustness of the proposed method can help even when no outliers are added but distributions are long-tailed, as illustrated in Figure 8.

A.10 Simulation with cellwise outliers

This section repeats the simulation in Section 4 for cellwise outliers. The clean data are exactly the same, but now we randomly select data cells and replace them by outliers following the distribution $N(k, 0.01^2)$ when they occur in the x-coordinate and $N(-k, 0.01^2)$ when they occur in the y-coordinate. The simulation was run for 10%, 20% and 30% of cellwise outliers, but the patterns were similar across contamination levels.

Figure 17 shows the MSE of the same transformation-based correlation measures as in Figure 4, with 10% of cellwise outliers for $k = 3$ and $k = 5$. Within this class Pearson again has the worst MSE, followed by normal scores. The quadrant correlation is next, and does not look as good here as for rowwise outliers. Wrapping has the lowest MSE, and again outperforms Spearman, sigmoid and Huber because it moves the outlying cells to the central part of their variable.

Figure 18 compares wrapping to the correlation measures in Figure 7 in the presence of these cellwise outliers. Also here the SSCM has the largest bias, especially in $d = 10$ dimensions, followed by Kendall's tau. Wrapping does well but not as well as MCD and GK when $k = 3$, and their performance is similar for $k = 5$. But in higher dimensions wrapping still has the redeeming feature that it yields a PSD correlation matrix unlike the GK method, whereas the MCD suffers from the propagation of cellwise outliers and a high computation time.

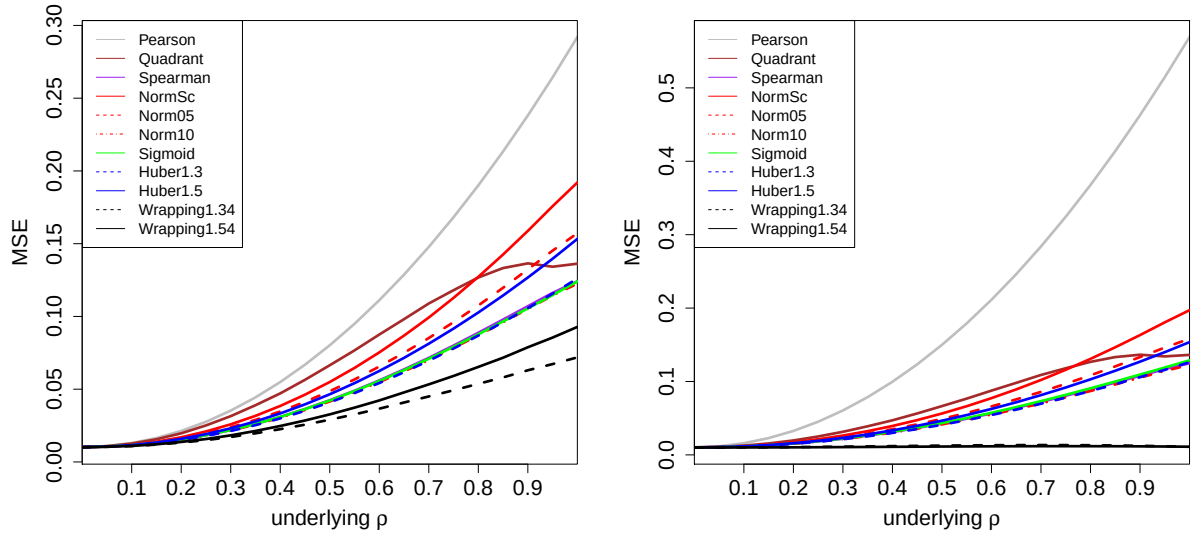

Figure 17: MSE of the correlation measures in Figure 4 with 10% of cellwise outliers placed with $k = 3$ (left) and $k = 5$ (right).

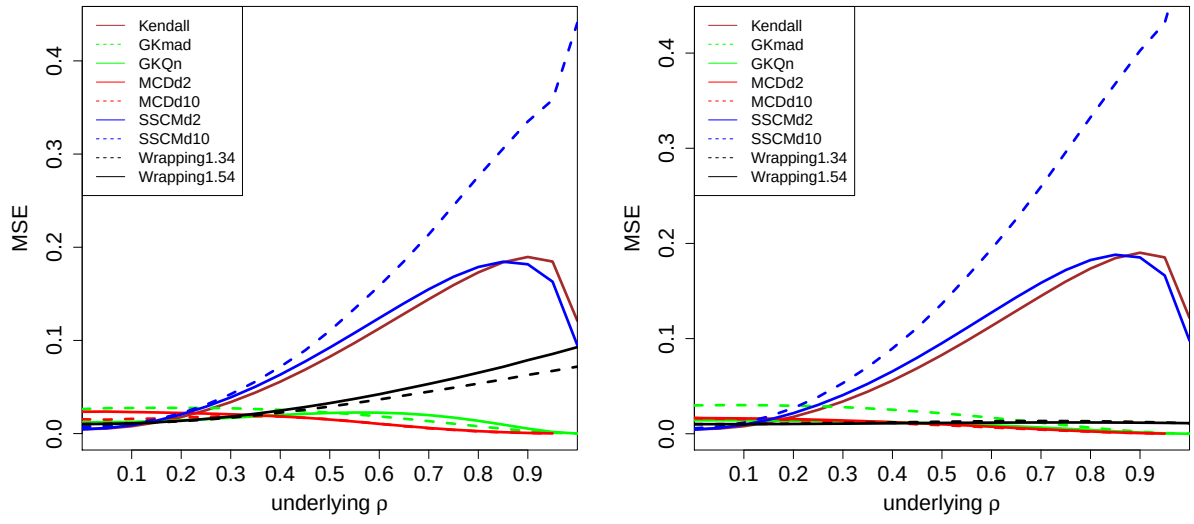

Figure 18: MSE of the correlation measures in Figure 6 with 10% of cellwise outliers placed with $k = 3$ (left) and $k = 5$ (right).