Distance correlation: Discovering meta-analytic relationships between variables when other correlation coefficients fail

Research Synthesis, Dubrovnik: Methods in meta-analysis (29.05.2019)

Lukasz Stasielowicz & Reinhard Suck

- Osnabrück University
- lukasz.stasielowicz@uos.de

Foosball (table soccer)

Correlations in meta-analyses

- Usual main goal of a meta-analysis: Computing the mean correlation across studies (i.e. *r*)
  - Example: Is there some kind of dependence between personality constructs?

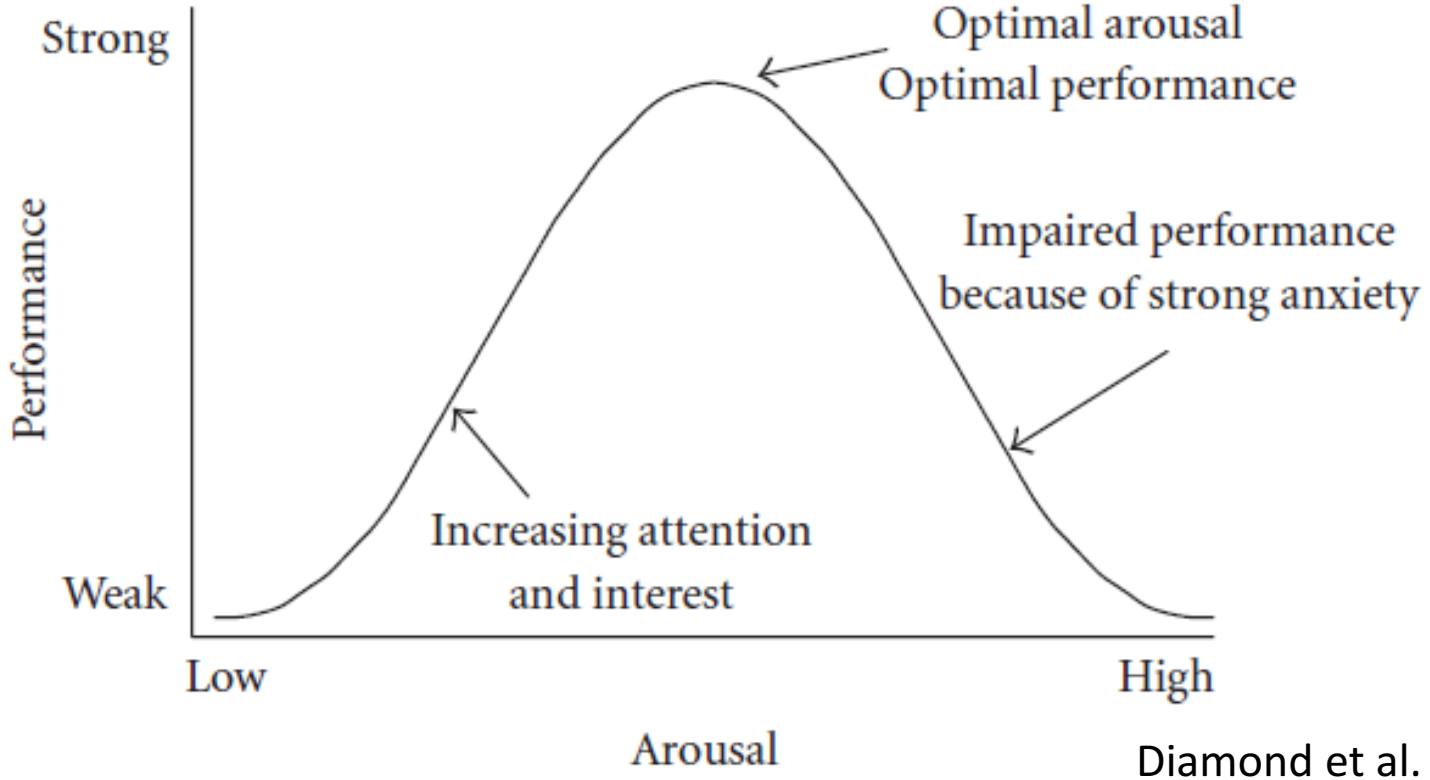
- Issue 1: Correlation ≠ causation

- Issue 2: *r* = 0 ≠ Lack of dependence
  - Crux of this presentation
  - *r* = 0 only means that there is no linear relationship
  - Risk of failing to identify nonlinear relationships, e.g. inverted-U

Nonlinear relationship: Example 1

➢ Yerkes–Dodson law

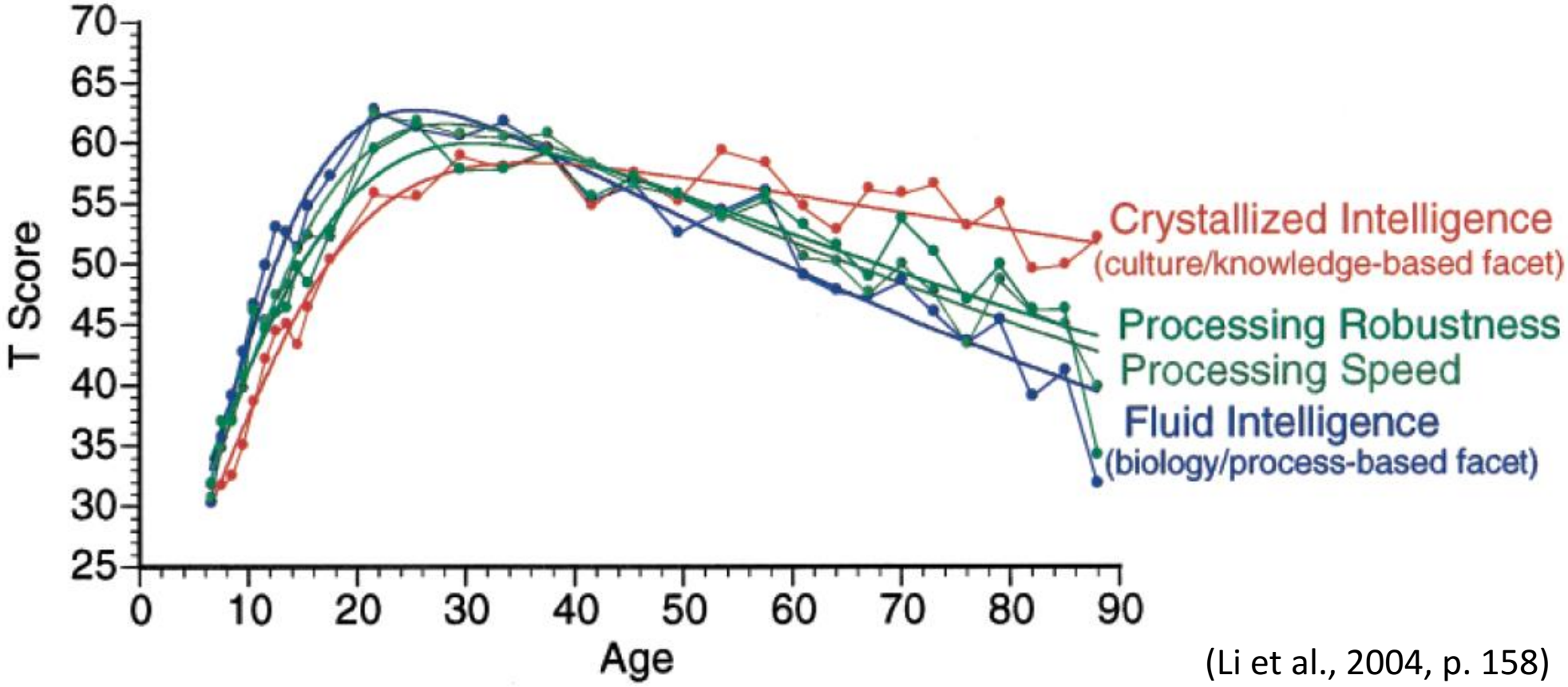
- Relationship between arousal and performance
- Nonlinear relationship (inverted-U relationship)



Diamond et al. (2007, p. 3)

Nonlinear relationship: Example 2

- Relationship between Age and cognitive abilities
  - Non-monotonic relationship
    - Increase + decrease of cognitive abilities



(Li et al., 2004, p. 158)

Nonlinear relationships in a meta-analysis

- *r* can fail in meta-analyses when dealing with nonlinear relationships

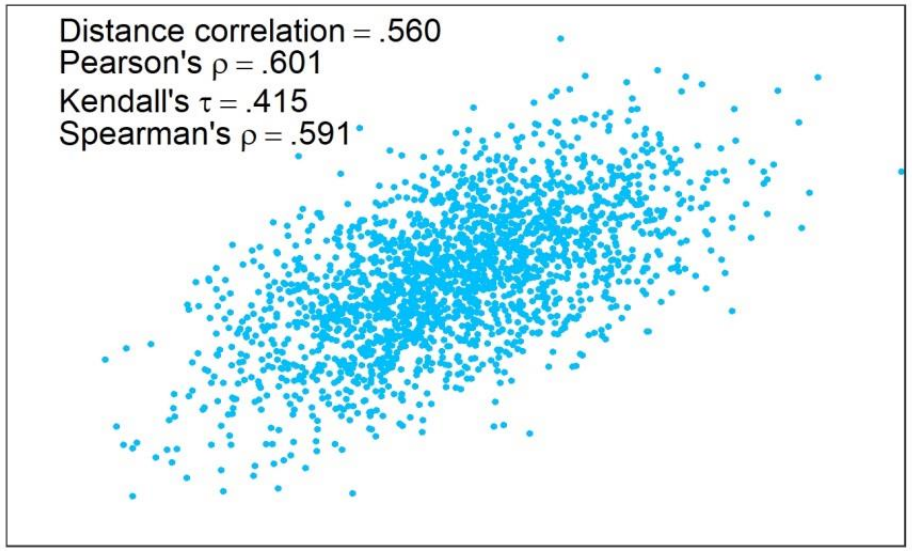
- What about other well-known effect sizes?
  - Spearman‘s rho, Kendall‘s tau etc. cannot detect non-monotonic relationships

- Distance correlation ($\mathcal{R}$) as a potential solution (Rizzo & Székely, 2016)
  - Different types of dependence can be assessed simultaneously
  - $\mathcal{R}_{Min} = 0$, $\mathcal{R}_{Max} = 1$
    - 0 means that there is no dependence

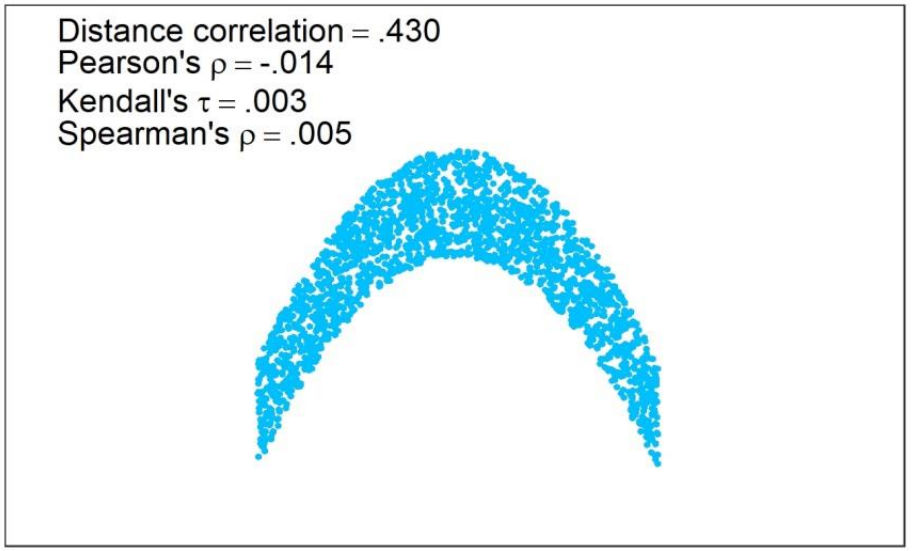
Assessing nonlinear relationships

- Comparison of four different coefficients
  - Distance correlation, Pearson's ρ, Kendall's τ, and Spearman's ρ

Linear relationship



Inverted-U relationship



Distance correlation

- Many applications of distance correlation
  - Exploratory data analysis (Székely & Rizzo, 2009)
  - Variable selection in regression models (Kong et al., 2015; Li et al., 2012; Yenigün & Rizzo, 2015)
  - Principal component analysis (Mishra, 2014)
  - Modelling autocorrelation in longitudinal studies (Edelmann et al., 2018; Zhou, 2012)
  - Measuring dependence between networks in brain imaging studies (Chen et al., 2019)
- Potentially relevant in the **meta-analytic context** (Székely et al., 2007)
  - *„Distance correlation can also be applied as an index of dependence; for example, in meta-analysis distance correlation would be a more generally applicable index than product-moment correlation"* (p. 2770)

Current study

- Goals of the present study
  - Testing the feasibility of using distance correlation in a meta-analysis
  - Comparing distance correlation to standard effect sizes

- Computing distance correlation
  - R package *energy*
  - Conceptual similarity to Pearson correlation: $\mathcal{R}(X, Y) := \frac{\mathcal{V}(X,Y)}{\sqrt{\mathcal{V}(X) \cdot \mathcal{V}(Y)}}$

    Distance Correlation = $\frac{Distance\ Covariance}{\sqrt{Distance\ Variance_X * Distance\ Variance_Y}}$
  - It is based on distances between individual values
    - i.e. X = cognitive abilities: Person 1 vs Person 2; Person 1 vs Person 3 etc.
    - i.e. Y = age: Person 1 vs Person 2; Person 1 vs Person 3 etc.

Computing distance correlation

- Distances for the X variable:

$$a_{km} = |X_k - X_m|$$

$$\bar{a}_{k.} = \frac{1}{n} \sum_{m=1}^{n} a_{km}$$

$$\bar{a}_{.m} = \frac{1}{n} \sum_{k=1}^{n} a_{km}$$

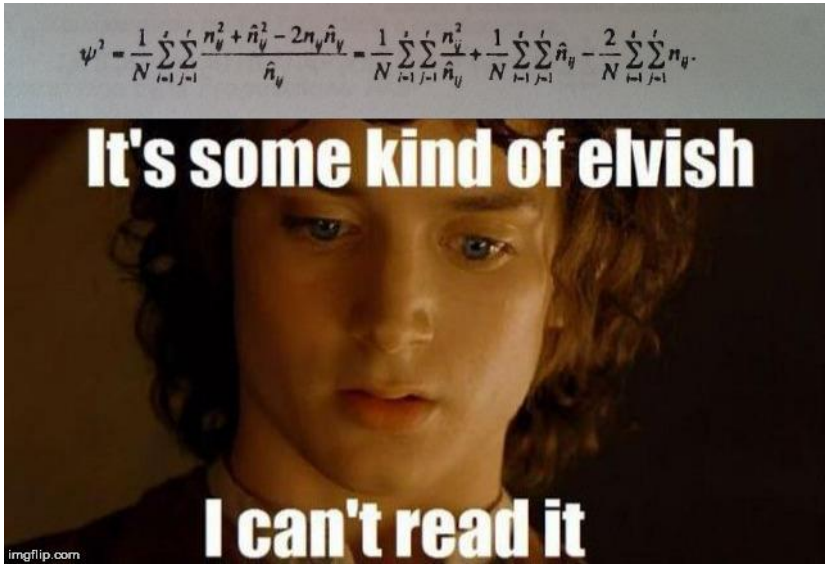
$$\bar{a}_{..} = \frac{1}{n^2} \sum_{k,m=1}^{n} a_{km}$$

$$A_{km} = a_{km} - \bar{a}_{k.} - \bar{a}_{.m} + \bar{a}_{..}$$

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X,X) = \frac{1}{n^2} \cdot \sum_{k,m=1}^{n} A_{km}^2$$ ⇒ Distance Variance

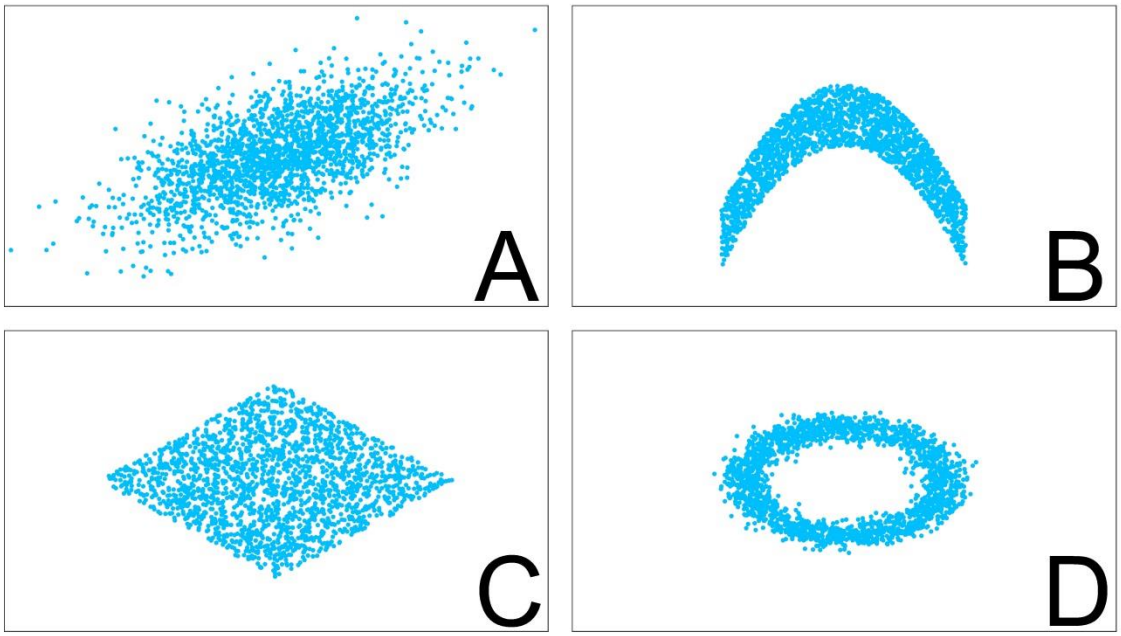
$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \cdot \sum_{k,m=1}^{n} A_{km} B_{km}$$ ⇒ Distance Covariance

- For the Y variable b values are computed



Current study

- 36 scenarios (4 x 3 x 3)
  - 4 different kinds of dependence (see figure)
  - Number of samples in the meta-analysis (*k*: 20, 50, 100)
  - Size of each sample (*N*: 50, 200, 1000)

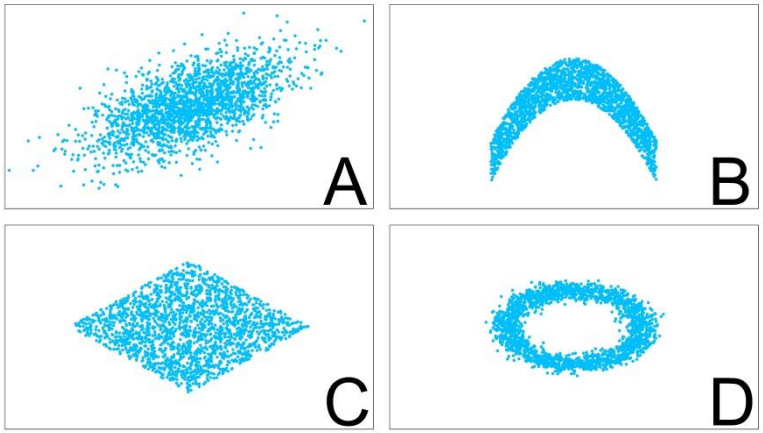


- For each sample the following effect sizes were computed: Kendall's tau (τ), Spearman's rho (ρ), Pearson correlation (r), distance correlation ($\mathcal{R}$), and unbiased distance correlation ($\mathcal{R}_U$) were computed
- Next the mean effect sizes were computed (180 in total)

Current study

- R packages: *energy*, *bootstrap, metafor*

- Meta-analytic model: Random-effects model



- Heterogeneity estimator: Restricted maximum likelihood (REML)
  - Good performance in simulation studies
    (Langan et al., 2017; Veroniki et al., 2016)

Distance correlation in meta-analysis

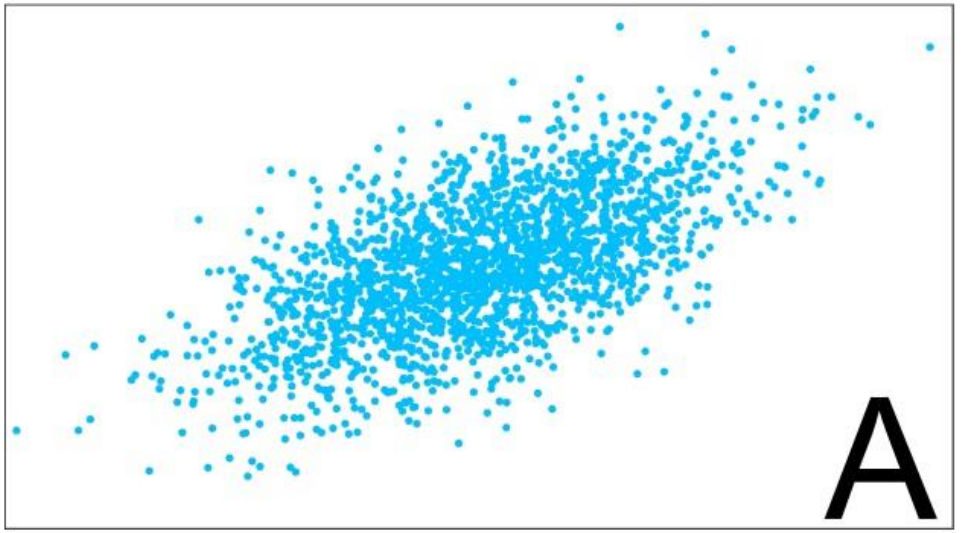
- Usually effect sizes are weighted (w_i) in a meta-analysis
  - They depend on the sampling variance (v_i)
  - Small samples → large variance → small weight

- Sampling variance for distance correlation
  - Jackknife method has been recommended (Székely & Rizzo, 2009)
    - Leave-one-out procedure
    - Compute distance correlation after „deleting“ one pair of observation (i.e. data for one person)
    - Compare mean correlation across leave-one-out subsets to the correlation of each subset

Image by HOerwin56 from Pixabay

Results (pattern A)

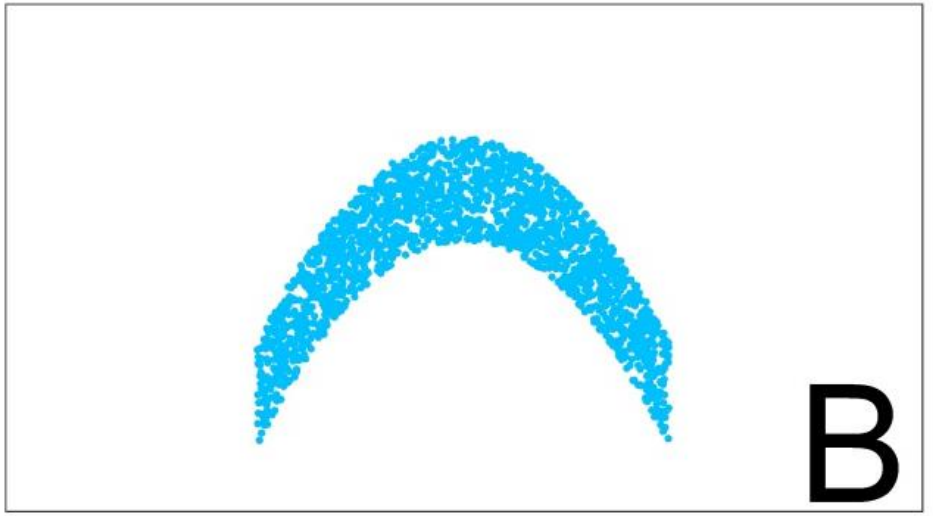
- Data sets were simulated based on a true Pearson correlation (*r*) of .60
- *r* performs best
- τ understimates the dependence
- Spearman‘s rho and distance correlations ($\mathcal{R}$) perform similarly (slight underestimation)
  - Interestingly distance correlations perform worse in large samples



		A				
k	*N*	τ	ρ	*r*	$\mathcal{R}$	$\mathcal{R}_U$
20	50	.397	.563	.592	.574	.539
	200	.410	.579	.605	.561	.551
	1000	.408	.579	.596	.550	.548
50	50	.402	.565	.612	.588	.556
	200	.411	.581	.606	.563	.553
	1000	.409	.581	.601	.552	.550
100	50	.408	.571	.621	.595	.563
	200	.407	.577	.601	.560	.550
	1000	.412	.585	.604	.556	.554

Results (pattern B)

- τ, r, and ρ fail to identify an inverted-U relationship
  - Values close to 0

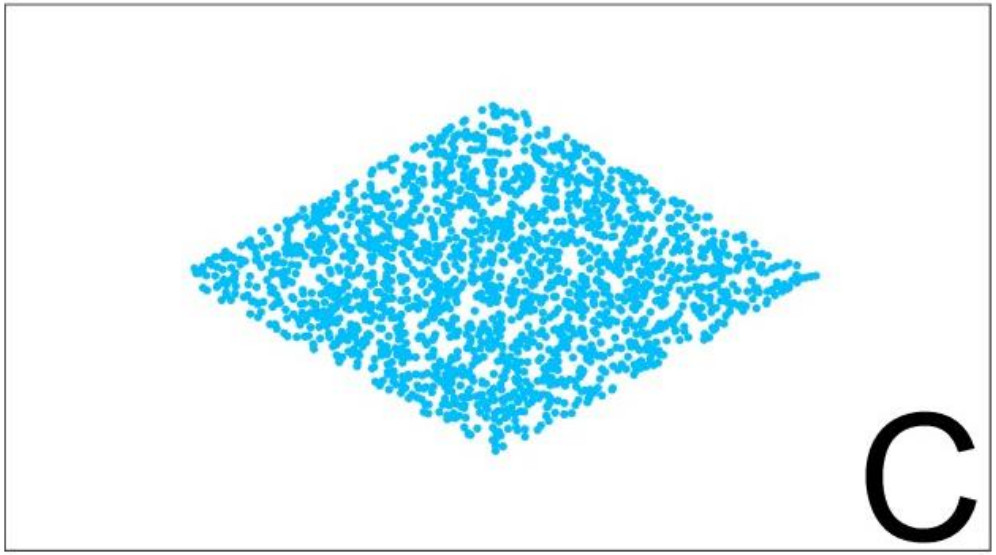


- Only distance correlations ($\mathcal{R}$) yield large values

		B				
k	N	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$
20	50	-.040	-.039	-.066	.586	.522
	200	-.017	-.024	-.056	.550	.533
	1000	-.006	-.005	.019	.540	.537
50	50	-.048	-.056	-.079	.590	.524
	200	-.011	-.015	-.014	.552	.535
	1000	.008	.010	.008	.540	.536
100	50	-.020	-.026	-.046	.584	.520
	200	-.006	-.005	.021	.548	.531
	1000	0	0	-.007	.541	.538

Results (pattern C)

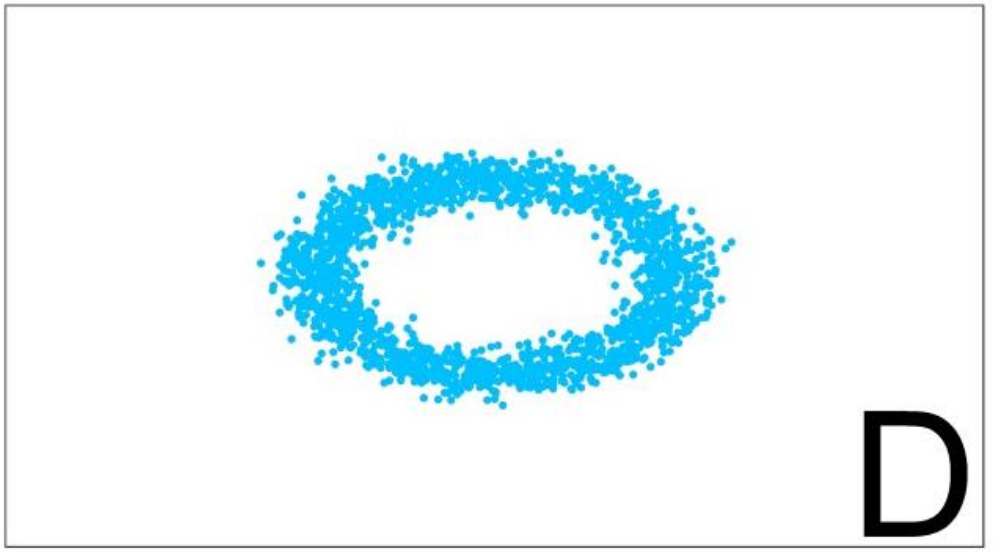
- τ, r, and ρ fail to identify the non-monotonic relationship
  - Values close to 0
- Only distance correlations ($\mathcal{R}$) yield values greater than 0
  - Unbiased estimator yielded negative values for some small samples (N = 50)



		C				
k	N	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$
20	50	.030	.053	.047	.248	.170[a]
	200	.005	.010	.004	.180	.147
	1000	-.004	-.006	-.005	.154	.146
50	50	.012	.018	.016	.242	.171[a]
	200	.007	.013	.010	.178	.145
	1000	.002	.003	.003	.152	.144
100	50	.011	.018	.016	.243	.172[a]
	200	-.004	-.006	-.005	.177	.143
	1000	0	-.001	-.001	.151	.143

Results (pattern D)

- τ, r, and ρ fail to identify the non-monotonic relationship
  - Values close to 0

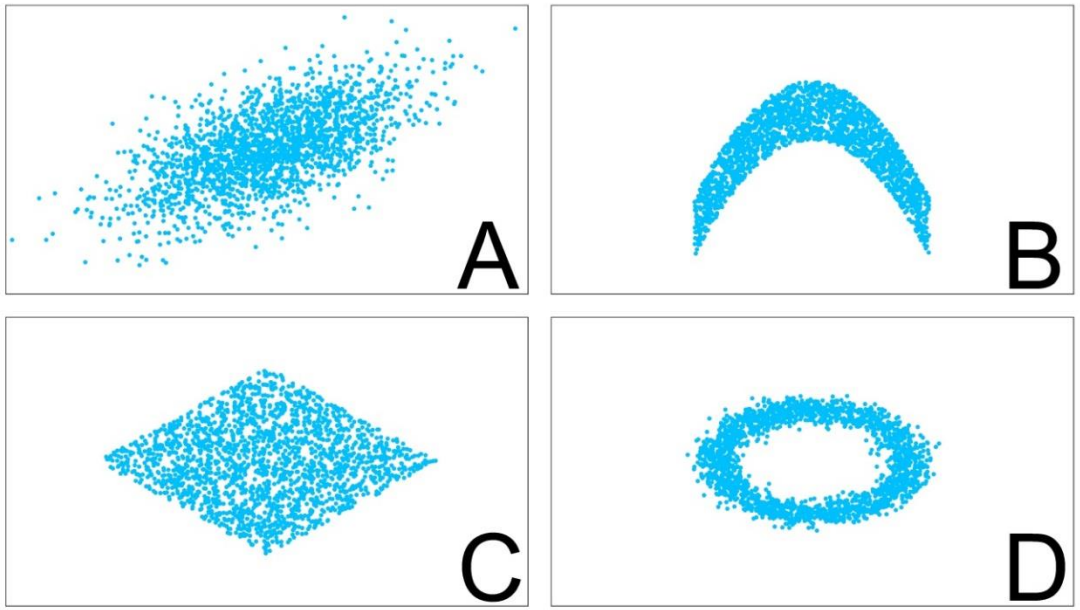


- Only distance correlations ($\mathcal{R}$) yield values greater than 0
  - Unbiased estimator yielded negative values for some small samples (N = 50)

		D				
k	N	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$
20	50	-.003	-.005	-.002	.239	.157[a]
	200	.006	.012	.013	.185	.157
	1000	-.002	-.006	-.007	.174	.168
50	50	.002	.008	.017	.241	.161[a]
	200	-.005	-.008	-.008	.189	.163
	1000	-.001	-.002	-.003	.172	.166
100	50	.004	.007	.009	.241	.168[a]
	200	-.001	-.003	-.003	.190	.163
	1000	0	-.002	-.002	.173	.167

Summary

- Only distance correlation was able to identify dependence across all 36 scenarios
- Use of distance correlation in a meta-analysis can be fruitful



- Recommendation: Preliminary check
  - No dependence? Use *r* (software available: metafor etc.)
  - Dependence: Check scatter plots for each sample
    - If the relationship is linear – use *r*
    - Nonlinear or nonmonotonic relationships – use distance correlation

Issues

- Interpretation: Does a value of .01 imply dependence?
  - Statistical tests exist (Székely & Rizzo, 2009; Székely et al., 2007)
    - Pitfalls of *p*-value (Amrhein, Greenland, & McShane, 2019)

- Unbiased estimator: Problems in small samples (negative values)
  - Common when dealing with unbiased statistics, i.e. $adjR^2$ in multiple regression etc. (Rizzo & Székely, 2016; Székely & Rizzo, 2013)
  - How to deal with this issue in a meta-analysis?
    - Set negative values to zero?
      - Requires adjusting the jackknife technique – setting distance correlations to 0
    - Delete them from the meta-analysis?

Issues

- Full data sets needed to compute distance correlation
  - It cannot be derived from summary statistics (*M*, *SD*, *t*, *p* etc.)
  - It cannot be derived from standard effect sizes (*r*, *d*, *OR* etc.)
  - Open Science to the rescue!
    - Willingness to share data is increasing
    - Many platforms available (osf, PsychArchives etc.)
    - Multi-lab studies (replications)
    - Peer Reviewers' Openness Initiative

Issues

- The same distance correlation value can correspond to different patterns across samples (i.e. linear, quadratic)
- Dealing with heterogeneity
  - Common heterogeneity statistics (I^2, Q, τ) may fail
    - Different patterns but the same distance correlation value
  - Failure of identifying moderators may lead to bad consequences, i.e.
    - Approval of interventions with side effects in certain groups
    - Rejection of promising interventions
  - Visual inspection of the data necessary
  - Changing the sign of distance correlation if plausible (i.e. U-relationship vs inverted-U relationship)
  - Subgroup analysis: Analyzing data sets with different patterns separately

Future research questions

- Conducting meta-analyses based on real data
- Benchmarks for interpreting $\mathcal{R}$ values
- Applying distance correlation to three-level meta-analytic models
- Bayesian distance correlation

- Comparing distance correlation to other new dependence measures
  - Maximal Information Coefficient (MIC), Total Information Coefficient (TIC), Heller Heller Gorfine measure (HHG) or Hoeffding's D (de Siqueira Santos et al., 2014; Kinney & Atwal, 2014; Reshef et al. 2018; Speed, 2011)
  - MIC and TIC seem to perform worse when dealing with linear patterns but are better when dealing with nonlinear patterns (Reshef et al., 2018).

References

- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, *567*, 305–307. http://doi.org/10.1038/d41586-019-00857-9
- Chen, H., Liu, K., Zhang, B., Zhang, J., Xue, X., Lin, Y., … Deng, Y. (2019). More optimal but less regulated dorsal and ventral visual networks in patients with major depressive disorder. *Journal of Psychiatric Research*, *110*, 172–178.
- de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., & Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in Bioinformatics*, *15*(6), 906–918. http://doi.org/10.1093/bib/bbt051
- Diamond, D. M., Campbell, A. M., Park, C. R., Halonen, J., & Zoladz, P. R. (2007). The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. *Neural plasticity, 2007*. https://doi.org/10.1155/2007/60803
- Edelmann, D., Fokianos, K., & Pitsillou, M. (2018). An updated literature review of distance correlation and its applications to time series. *International Statistical Review*. http://doi.org/10.1111/insr.12294

References

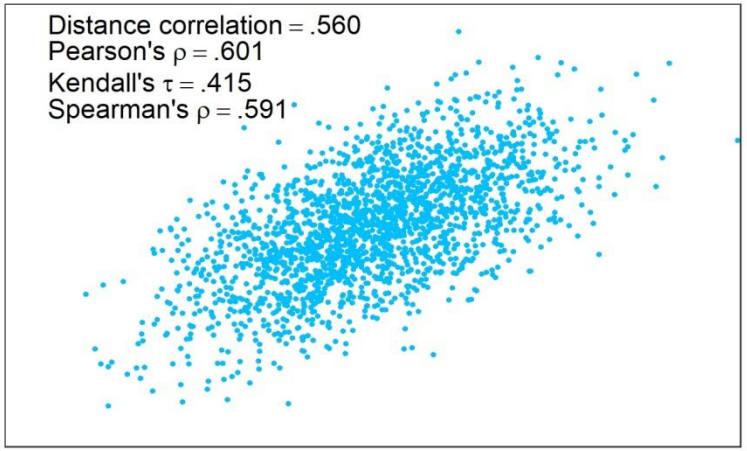
- Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *PNAS*, *111*(9), 3354–3359. http://doi.org/10.1073/pnas.1309933111
- Kong, J., Wang, S., & Wahba, G. (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine*, *34*(10), 1708–1720. http://doi.org/10.1002/sim.6441
- Langan, D., Higgins, J. P. T., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research Synthesis Methods*, *8*(2), 181–198. http://doi.org/10.1002/jrsm.1198
- Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, *107*(499), 1129–1139. http://doi.org/10.1080/01621459.2012.695654
- Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*(3), 155–163. http://doi.org/10.1111/j.0956-7976.2004.01503003.x

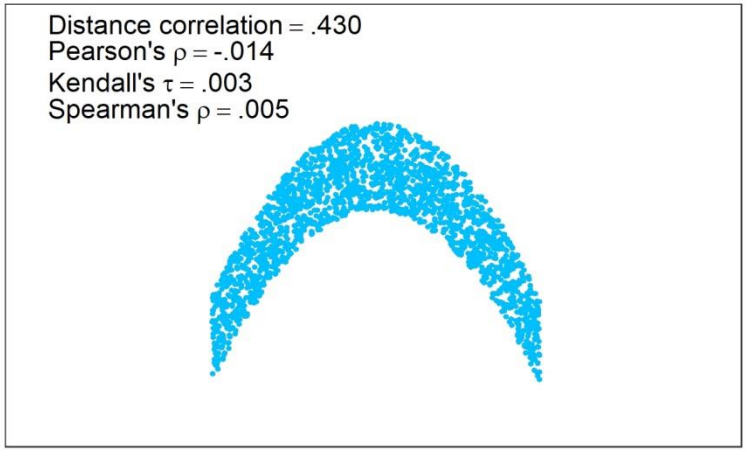
References

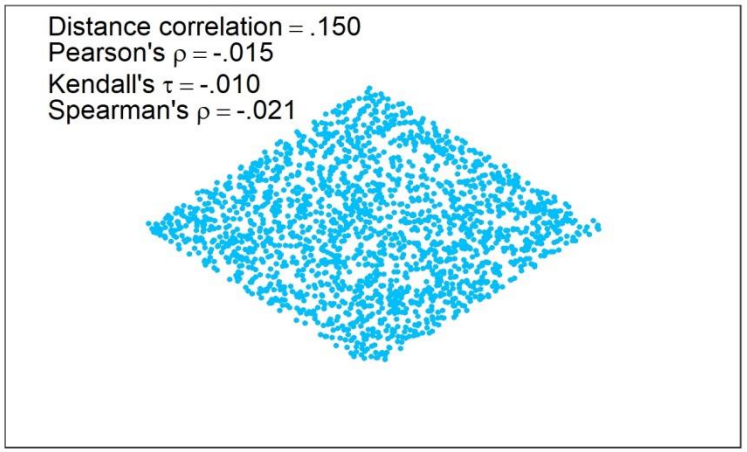
- Mishra, S. K. (2014). What happens if in the principal component analysis the Pearsonian is replaced by the Brownian coefficient of correlation? *SSRN*. http://doi.org/10.2139/ssrn.2443362
- Reshef, D. N., Reshef, Y. A., Sabeti, P. C., & Mitzenmacher, M. (2018). An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, *12*(1), 123–155. http://doi.org/10.1214/17-AOAS1093
- Rizzo, M. L., & Székely, G. J. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(1), 27–38. http://doi.org/10.1002/wics.1375
- Speed, T. (2011). A correlation for the 21st century. *Science*, *334*(6062), 1502–1503. http://doi.org/10.1126/science.1215894
- Székely, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics*, *3*(4), 1236–1265. http://doi.org/10.1214/09-AOAS312
- Székely, G. J., & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, *117*, 193–213. http://doi.org/10.1016/j.jmva.2013.02.012

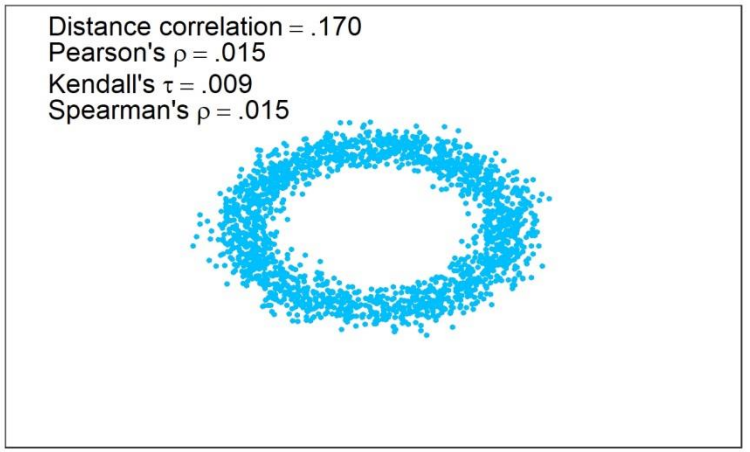
References

- Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, *35*(6), 2769–2794. http://doi.org/10.1214/009053607000000505
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., … Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. doi:10.1002/jrsm.1164
- Yenigün, C. D., & Rizzo, M. L. (2015). Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation*, *85*(8), 1692–1705. http://doi.org/10.1080/00949655.2014.895354
- Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, *33*(3), 438–457. http://doi.org/10.1111/j.1467-9892.2011.00780.x


Distance correlation = .560
Pearson's ρ = .601
Kendall's τ = .415
Spearman's ρ = .591


Distance correlation = .430
Pearson's ρ = -.014
Kendall's τ = .003
Spearman's ρ = .005


Distance correlation = .150
Pearson's ρ = -.015
Kendall's τ = -.010
Spearman's ρ = -.021


Distance correlation = .170
Pearson's ρ = .015
Kendall's τ = .009
Spearman's ρ = .015

Appendix: Unbiased vs standard estimator

- Unbiased estimator (Rizzo & Székely, 2016)

$$\tilde{A}_{i,j} = \begin{cases} a_{i,j} - \frac{1}{n-2}\sum_{i=1}^{n} a_{i,j} - \frac{1}{n-2}\sum_{j=1}^{n} a_{i,j} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^{n} a_{i,j}, & i \neq j; \\ 0, & i = j. \end{cases}$$

p. 33

- Standard estimator

$$A_{km} = a_{km} - \bar{a}_{k.} - \bar{a}_{.m} + \bar{a}_{..}$$

Appendix: All results

		A					B					C					D				
k	N	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$	τ	ρ	r	$\mathcal{R}$	$\mathcal{R}_U$
	50	.397	.563	.592	.574	.539	-.040	-.039	-.066	.586	.522	.030	.053	.047	.248	.170[a]	-.003	-.005	-.002	.239	.157[a]
20	200	.410	.579	.605	.561	.551	-.017	-.024	-.056	.550	.533	.005	.010	.004	.180	.147	.006	.012	.013	.185	.157
	1000	.408	.579	.596	.550	.548	-.006	-.005	.019	.540	.537	-.004	-.006	-.005	.154	.146	-.002	-.006	-.007	.174	.168
	50	.402	.565	.612	.588	.556	-.048	-.056	-.079	.590	.524	.012	.018	.016	.242	.171[a]	.002	.008	.017	.241	.161[a]
50	200	.411	.581	.606	.563	.553	-.011	-.015	-.014	.552	.535	.007	.013	.010	.178	.145	-.005	-.008	-.008	.189	.163
	1000	.409	.581	.601	.552	.550	.008	.010	.008	.540	.536	.002	.003	.003	.152	.144	-.001	-.002	-.003	.172	.166
	50	.408	.571	.621	.595	.563	-.020	-.026	-.046	.584	.520	.011	.018	.016	.243	.172[a]	.004	.007	.009	.241	.168[a]
100	200	.407	.577	.601	.560	.550	-.006	-.005	.021	.548	.531	-.004	-.006	-.005	.177	.143	-.001	-.003	-.003	.190	.163
	1000	.412	.585	.604	.556	.554	0	0	-.007	.541	.538	0	-.001	-.001	.151	.143	0	-.002	-.002	.173	.167