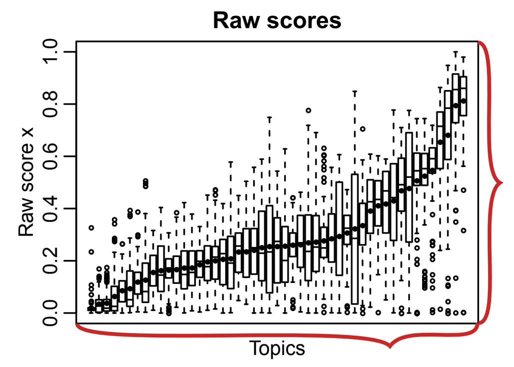
A New Perspective on Score Standardization

Julián Urbano, Harlley Lima and Alan Hanjalic

PROBLEM

- Very large **variability of effectiveness scores** within and between topics

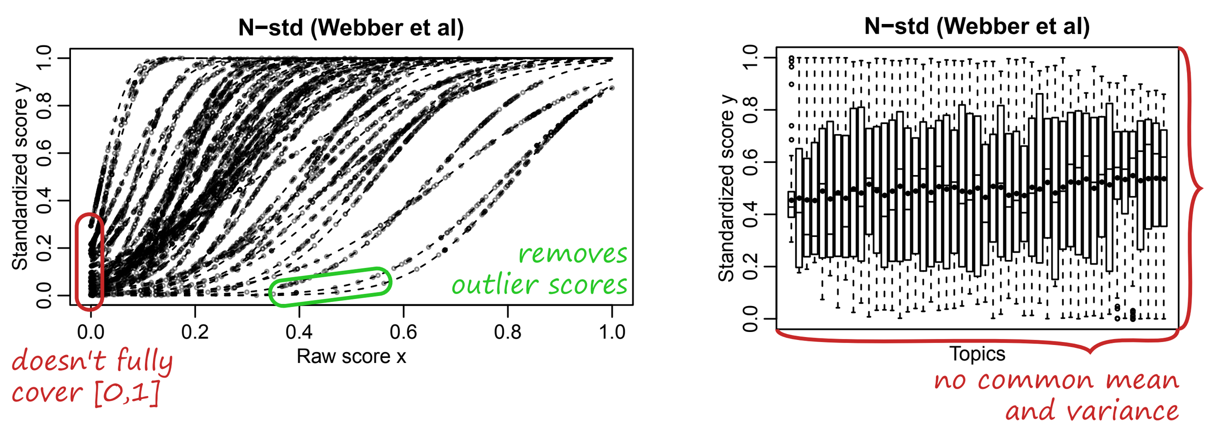


CONSEQUENCES

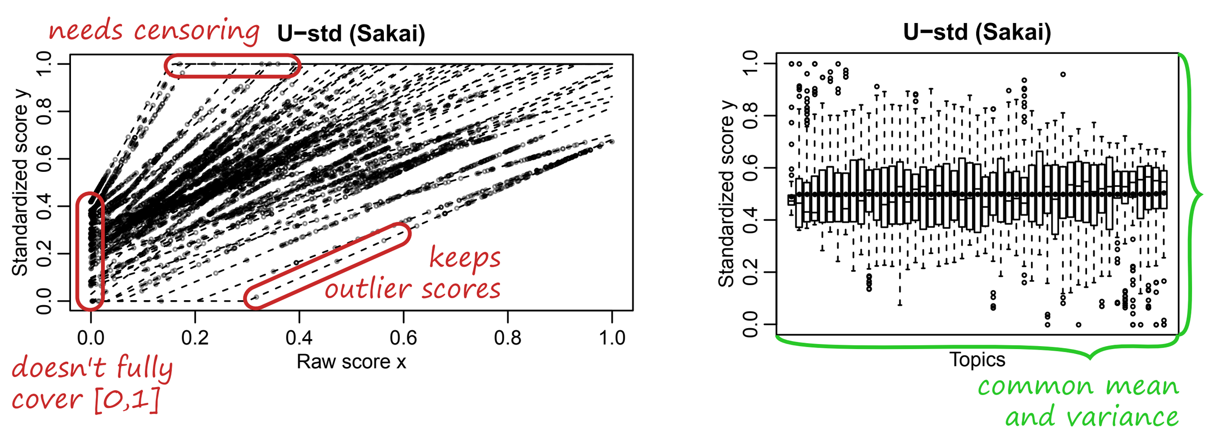
- **Within-collection** system comparisons are difficult: observed differences disproportionately due to a few topics
- **Between-collection**: very unstable, just impossible

SOLUTION?

- **Take topic difficulty into account**
- *Webber et al 2008*: 2-step standardization
  1. Compute z-score: $z=(x-\mu)/\sigma$, μ and σ per topic
  2. Nonlinear, **Gaussian** transform: $y=\Phi(z)$, so $y\in[0,1]$

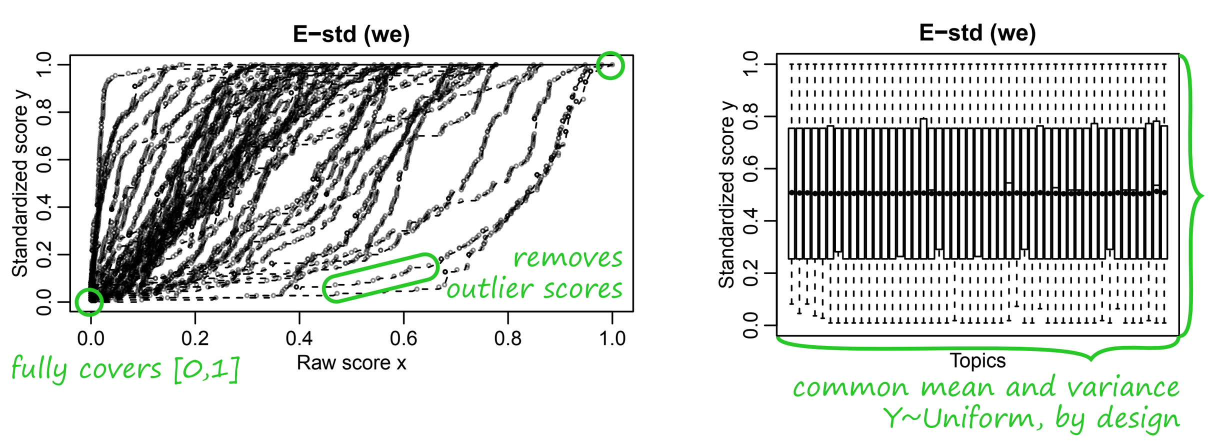


- *Sakai 2016*: 2-step standardization
  2. **Linear** transform: $y=Az+B$, $A=0.15$ and $B=0.5$



OUR PROPOSAL

- **Standardize with per-topic distributions: $y=F_X(x)=P(X\leq x)$**
- *"How does the system rank for the topic?"*
- From this perspective, it turns out that Webber et al. and Sakai are **special cases**, just assuming a specific F_X:
  - Webber et al: $X\sim Normal(\mu,\sigma^2)$
  - Sakai: $X\sim Uniform(\mu-\sigma B/A, \mu+\sigma(1-B)/A)$
- But **why assume anything**, and not just $X\sim ecdf(x_1,\dots,x_n)$?



Current score standardizations through gaussian and linear transformations are **special cases** of a standardization that **assumes specific distributions** of per-topic scores

The **empirical distribution** has **better properties,** seems to **work better,** and is more faithful to our **notion of "ranking"**

paper

poster

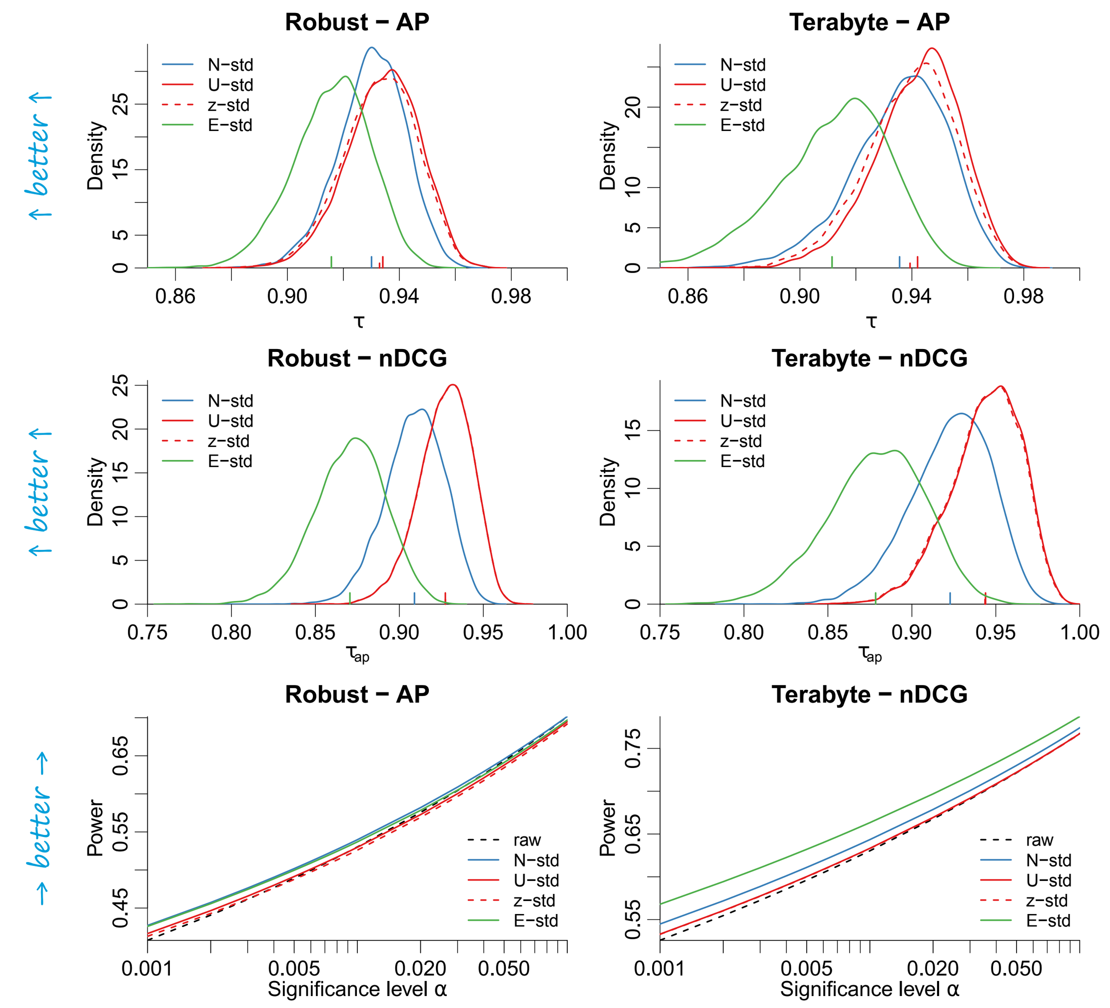
data & code

TU Delft

SIGIR 2019

WITHIN-COLLECTION COMPARISONS

- Repeat 10,000 times:
  - Randomly sample 50 topics and standardize
  - Compare the std. system rankings vs. raw (τ and τ_{ap})
  - Compare all pairs of systems (power)



BETWEEN-COLLECTION COMPARISONS

- Repeat 10,000 times:
  - Randomly sample 2 sets of 50 topics and standardize
  - Compare system rankings between sets (τ and τ_{ap})
  - Compare every system with itself (type I errors)
  - Compare all cross-collection pairs of systems (power)