UNIVERSITÉ DE NANTES

Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation

Jing Li, Rafal K.Mantiuk, Junle Wang,
Suiyi Ling, Patrick Le Callet

NIPS

Preference aggregation

- Application:

  Recommendation system

  Social networks

  Sports race, chess

  Online games

- Objective:

  Infer the underlying rating or ranking of the test candidates according to annotator’s label.

Preference aggregation

- Sometimes discovering the rating (true score) is more important

- Game players matching system
- e.g., MSR's TrueSkill system

- Friends-making website
- e.g., Facebook, Meetup

- Subjective image/video quality assessment

Pairwise comparison

- Advantage:
  - “human response to comparison questions is more stable in the sense that it is not easily affected by irrelevant alternatives” [Ailon,NIPS2009]
- Drawback:
  - $O(n^2)$ time complexity [ITU-R BT.500]
- Solutions:
  - Optimization on parameter estimation (deal with sparse data)
  - Novel model
  - **Pairwise sampling**

Outline

- The state of the art pairwise sampling strategy
- Proposed Methodology
- Experiment
- Results
- Conclusion

The state of the art

- Random sampling
  - Random Graph [Xu, TMM2012]
  - Subset Balanced design[Dykstra, 1960]
- Empirical sampling
  - Sorting based sampling [Silverstein, 1998]
  - Adaptive/Optimized Rectangular Design (ARD/ORD) [Li 2012][IEEEP3333.1.1][ITU-T P.915]
- Active sampling

Active sampling

- Active learning process
- Learn which pair could generate the maximum information gain (EIG)
- Bayesian theory (prior and posterior)

Active sampling

- [Pfeiffer, AAAI 2012]
  – Thurstone model + Bayesian framework
- [Chen,WSDM 2013] Crowd-BT
  – Bradley-Terry model + annotator's malicious behavior + Bayesian framework
- [Xu, AAAI 2018] Hodge-active
  – HodgeRank model + Bayesian framework

Drawbacks

- Sampling procedure is sequential
- Focusing on ranking aggregation, not accurate for rating
- Annotator’s unreliability is not considered
- High computational cost

The proposed method: Hybrid-MST

Preliminary

- n objects: $A_1, A_2, \ldots, A_n$
- True quality: $s = (s_1, s_2, \ldots, s_n)$
- Observed score: $r = (r_1, r_2, \ldots, r_n)$

$$r_i = s_i + \varepsilon_i$$

- Noise term: $\varepsilon_i \sim N(0, \sigma_i^2)$

In an observation:

If $r_i > r_j$, observer select A_i → $y_{ij} = 1$

If $r_i < r_j$, observer select A_j → $y_{ij} = 0$

Bradley-Terry model [Bradley1952]

The probability that Ai is preferred than Aj

$$Pr(A_i \succ A_j) \triangleq \pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}, \qquad \pi_i \geq 0, \qquad \sum_{i=1}^{t} \pi_i = 1$$

π_i is the merit of the object Ai

$$s_i = \log(\pi_i)$$

Thus, we obtain:

$$\pi_{ij} = \frac{e^{s_i}}{e^{s_i} + e^{s_j}} = \frac{1}{1 + e^{-(s_i - s_j)}}$$

Likelihood function:

$$L(\mathrm{s}|\mathrm{M}) = \prod_{i<j} \pi_{ij}^{m_{ij}} (1 - \pi_{ij})^{m_{ji}}$$

m_{ij} represents the total number of trial outcomes $A_i \succ A_j$

Using MLE:

$$\mathrm{s} \sim \mathcal{N}(\hat{\mathrm{s}}, \hat{\Sigma})$$

Active learning

- Gain information from the observations

$$\mathbf{s} \sim \mathcal{N}(\hat{\mathbf{s}}, \hat{\Sigma})$$

Multivariate Gaussian

- Utility function:
  - Fisher Information

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\middle|\theta\right]$$

  - Kullback-Leibler Divergence (KLD)

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right).$$

Active learning

- Gain information from the observations

$$\mathbf{s} \sim \mathcal{N}(\hat{\mathbf{s}}, \hat{\Sigma})$$

Multivariate Gaussian

- A straightforward way: **Global** KLD

posterior prior Maybe singular

$$D_{\mathrm{KL}}(\mathcal{N}(\hat{s}^{ij}, \hat{\Sigma}^{ij}) \| \mathcal{N}(\hat{s}^{c}, \hat{\Sigma}^{c})) = \frac{1}{2}\Big[\mathrm{tr}\left((\hat{\Sigma}^{c})^{-1}\hat{\Sigma}^{ij}\right) + \left(\hat{s}^{c} - \hat{s}^{ij}\right)^{\top}(\hat{\Sigma}^{c})^{-1}(\hat{s}^{c} - \hat{s}^{ij}) - \log\left(\frac{|\hat{\Sigma}^{ij}|}{|\hat{\Sigma}^{c}|}\right) - n\Big]$$

Active learning

- Gain information from the observations

$$\mathbf{s} \sim \mathcal{N}(\hat{\mathbf{s}}, \hat{\Sigma})$$

- Our proposal: **Local** Gain

$$s_{ij} \sim \mathcal{N}(\hat{s_i} - \hat{s_j}, \sigma_{ij}^2)$$

$$\sigma_{ij}^2 = \hat{\Sigma}(i,i) + \hat{\Sigma}(j,j) - 2\hat{\Sigma}(i,j)$$

$$U_{ij} = \int \sum_{y_{ij}} log \left\{ \frac{p(s_{ij}|y_{ij})}{p(s_{ij})} \right\} p(s_{ij}|y_{ij}) p(y_{ij}) ds_{ij}$$

posterior

prior

Utility function:

$$U_{ij} = \int \sum_{y_{ij}} log\left\{\frac{p(s_{ij}|y_{ij})}{p(s_{ij})}\right\} p(s_{ij}|y_{ij})p(y_{ij})ds_{ij}$$

A tractable form:

$$U_{ij} = E(p_{ij}log(p_{ij})) + E(q_{ij}log(q_{ij})) - E(p_{ij})logE(p_{ij}) - E(q_{ij})logE(q_{ij})$$

$$E(p_{ij}log(p_{ij})) = \int p_{ij}log(p_{ij})p(s_{ij})ds_{ij} = \int \frac{1}{1+e^{-x}} log(\frac{1}{1+e^{-x}}) \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x-(\hat{s_i}-\hat{s_j}))^2}{2\sigma_{ij}^2}} dx$$

With Gaussian-Hermite quadrature

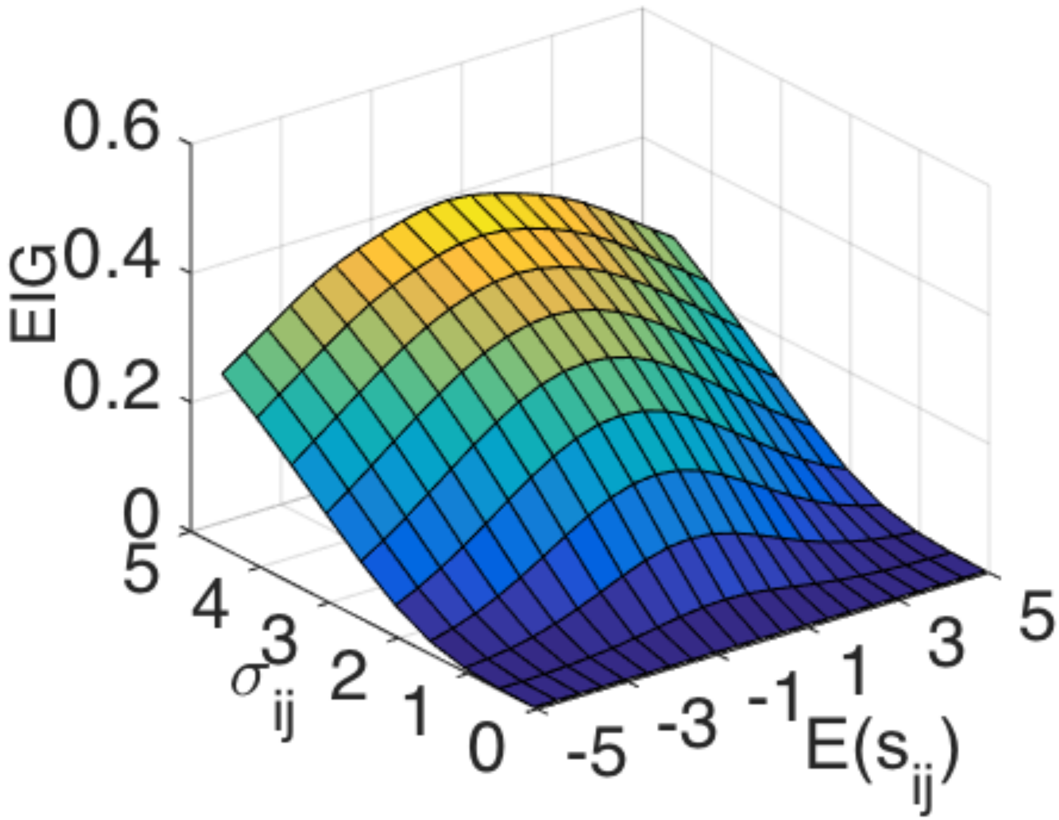
$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\,dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_i)]^2}.$$

In our model, *n*=30
Reduce the computational complexity!

Note that this *n* is sample points in Gaussian-Hermite quadrature, which is different from the number of test objects

Relationship between MLE estimates and EIG



The pairs which have similar scores or the score differences have higher uncertainties would generate more information

Pair selection strategy

- Global maximum (GM) method

$$\{A_i, A_j\} = argmax_{i \neq j} U_{ij}$$

Traditional method

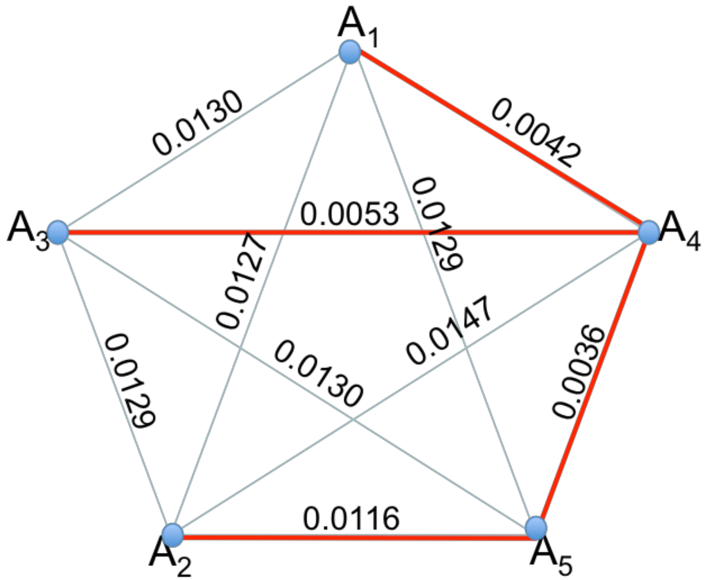
Pair selection strategy

- Global maximum (GM) method

$$\{A_i, A_j\} = argmax_{i \neq j} U_{ij}$$

Traditional method

- Minimum Spanning Tree (MST) method



Test objects as the vertices
EIG as the edges

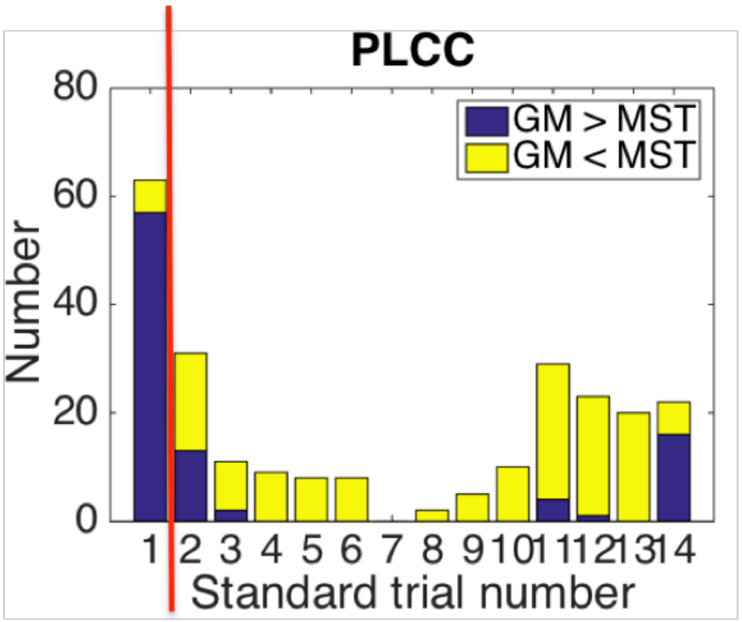
- n-1 edges
- All the vertices are connected
- Unique

Determination of strategy

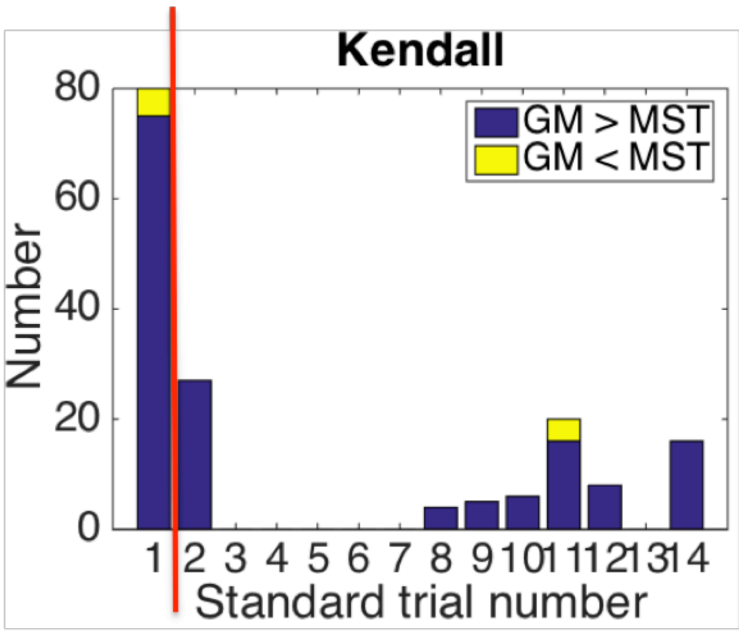
- When to use GM? When to use MST?
- Monte Carlo simulation
  - Number of test stimuli: 10, 16, 20, 40
  - True score ~ Uniform (1,5)
  - Noise ~ N(0, sigma^2), sigma~ Uniform (0,0.7)
  - Annotator's error: 10%, 20%, 30%, 40%
  - 100 repetitions
- Evaluation:

  PLCC, Kendall + Student's t-test

Hybrid strategy



Threshold



Threshold

1 standard trial number = n(n-1)/2 comparisons

$$\{A_i, A_j\} = \begin{cases} argmax_{i \neq j} U_{ij} & \text{if} \sum_{i,j} m_{ij} \leq \frac{n(n-1)}{2} \\ E_{mst} & \text{otherwise} \end{cases}$$

The whole Hybrid-MST procedure

According to current observations:

1. Calculating EIG for all pairs
2. If total comparison number < 1 standard number:

   →select pair using global Maximum

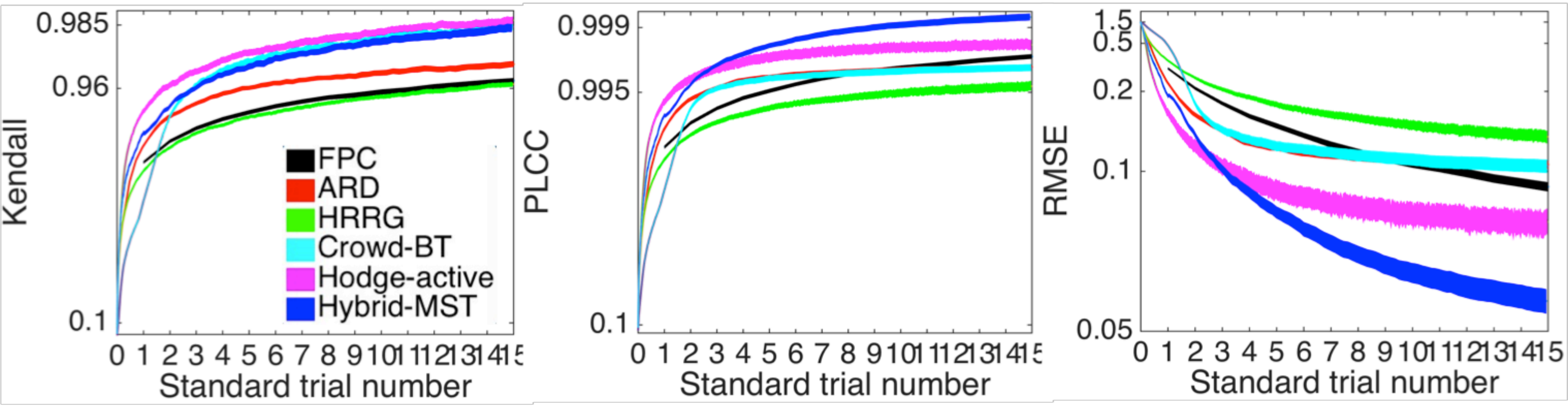
   Otherwise:

   →select pairs using MST

3. Run pairwise comparison

Experimental results

- Simulated data:
  - 60 stimuli ~Uniform[1,5]+$N(0,0.7^2)$
  - Observation error: 10,20,30,40%



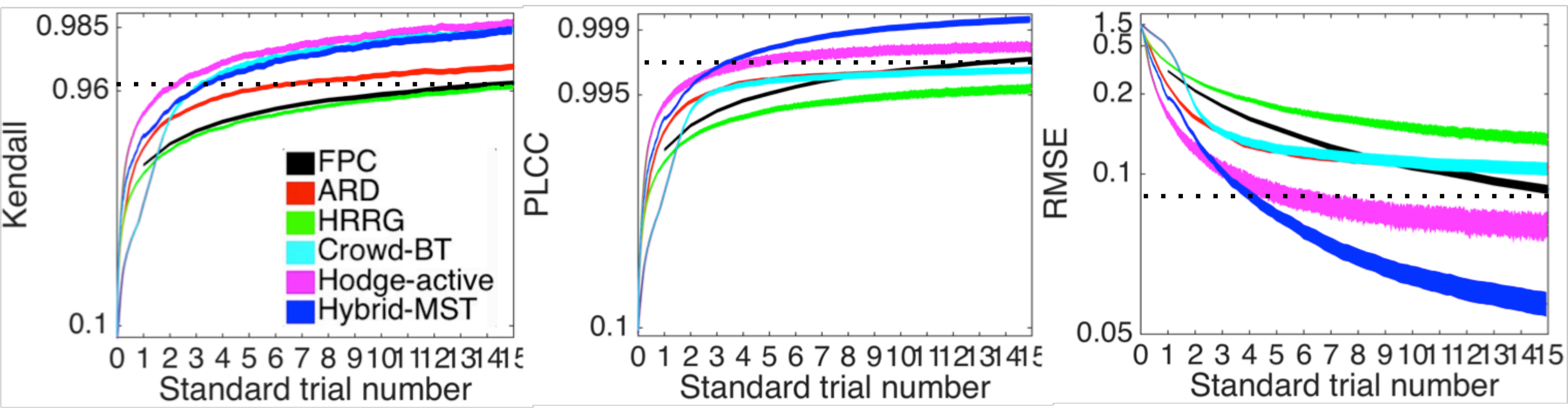
For better visualization, Kendall and PLCC are rescaled using Fisher transformation
RMSE is rescaled by y'=-1/y

To achieve the same accuracy with FPC of 15 annotators

Saving budget

$$\left(1 - \frac{D}{\frac{n(n-1)}{2} \times 15}\right) \times 100\%$$

	Kendall	PLCC	RMSE
Hybrid-MST	77.11%	74.89%	74.89%
Hodge-active	84.57%	68.61%	71.65%
Crowd-BT	78.43%	-	-

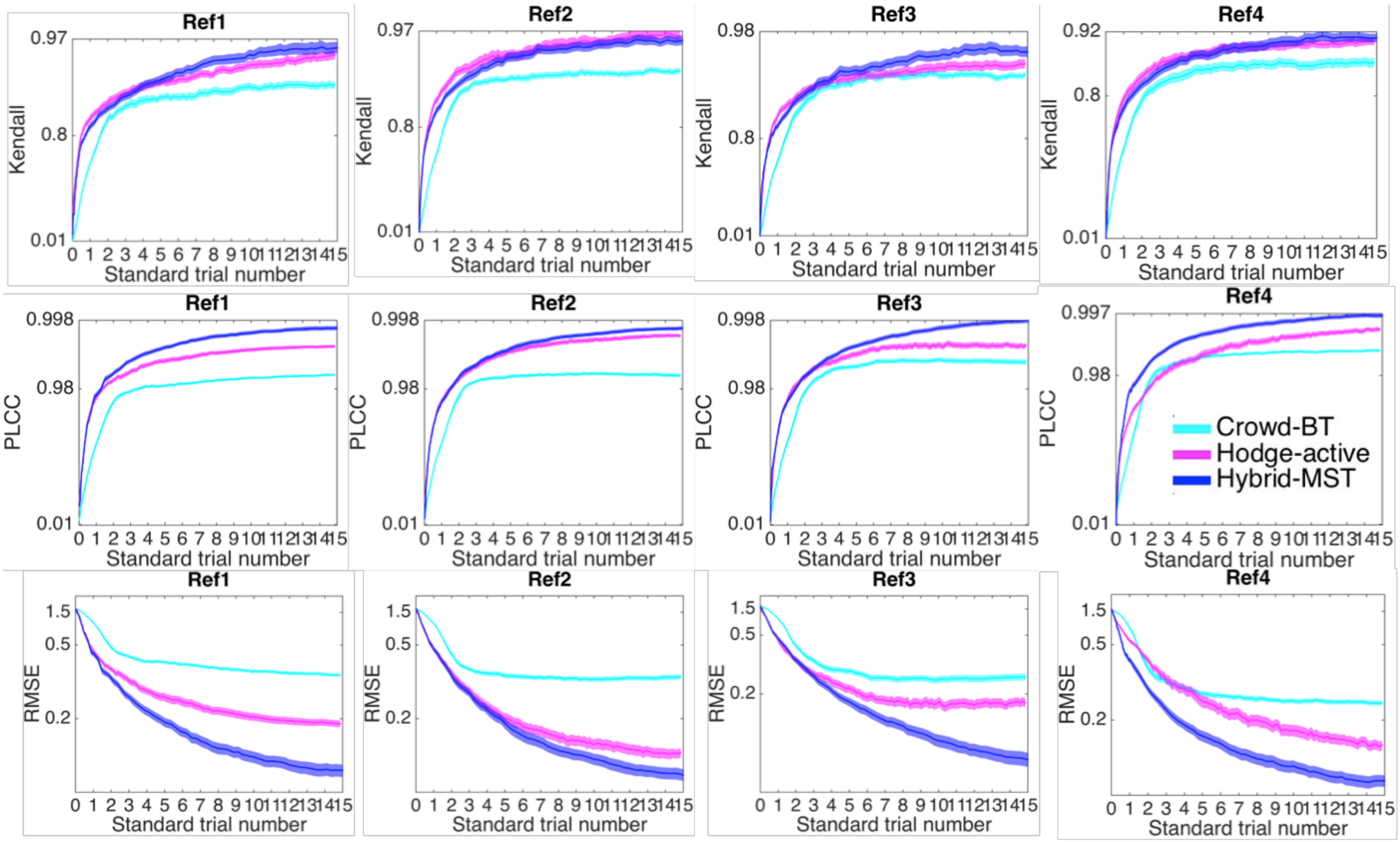


For better visualization, Kendall and PLCC are rescaled using Fisher transformation
RMSE is rescaled by y’=-1/y

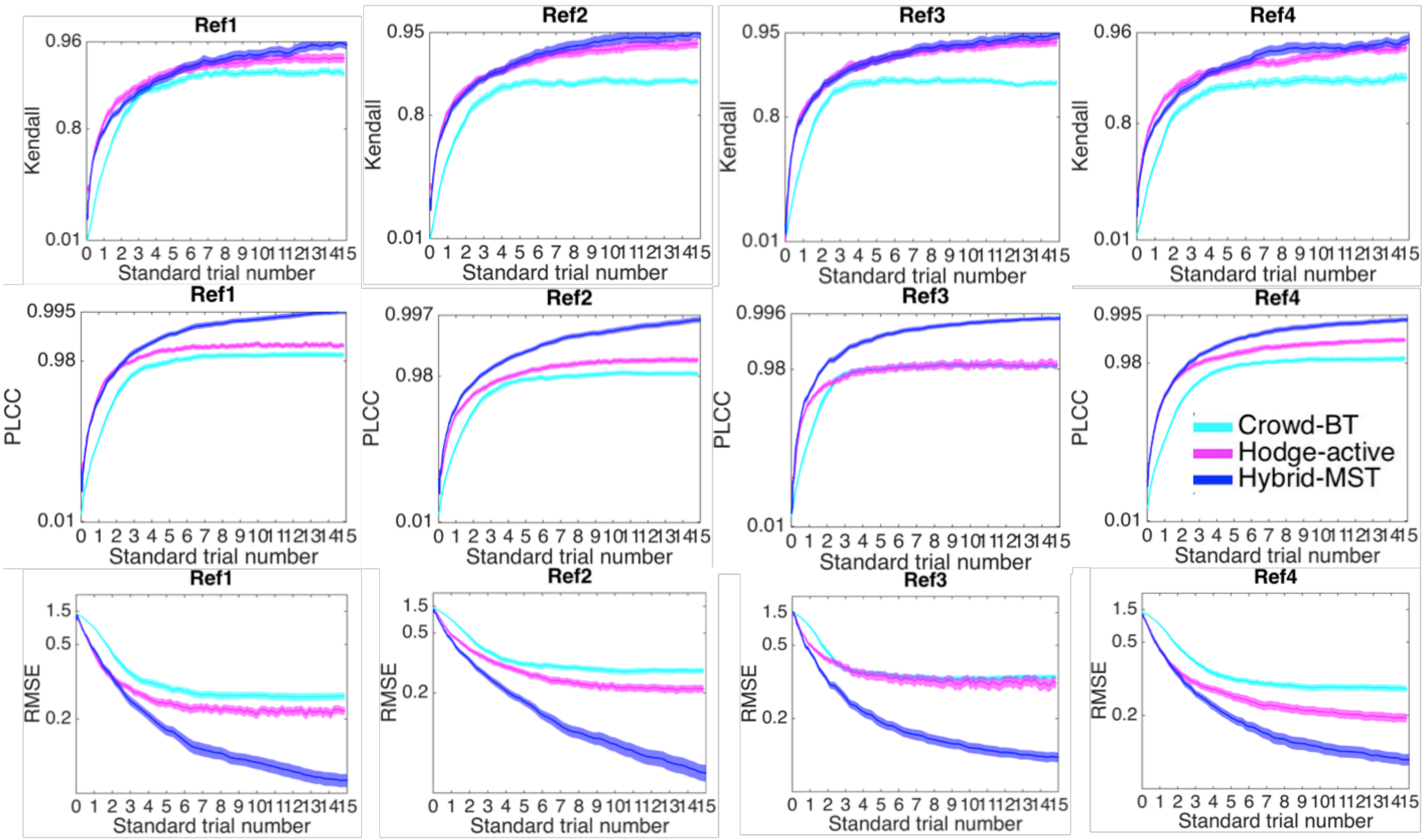
Real-world data

- Image Quality Assessment (IQA) dataset [Xu2012TMM]
  - 43266 pairwise comparison data,
  - 15 references from LIVE2008 and IVC2005,
  - 16 distortions
  - 328 annotators from internet
- Video Quality Assessment (VQA) dataset [Xu2011ACMMM]
  - 38400 pairwise comparison data
  - 10 references from LIVE database
  - 16 distortions
  - 209 annotators

Experimental results: IQA dataset



Experimental results: VQA dataset



Time complexity

Table 1: Runtime comparison on simulated data (ms/pair)

n	FPC	ARD	HRRG	Crowd-BT	Hodge-active	Hybrid-MST	
						GM	MST
10	0.11	1.24	0.38	85.69	0.34	48.72	6.16
20	0.10	0.62	0.34	188.56	0.22	153.61	8.97
100	0.10	0.16	0.65	3033.02	0.65	3007.08	30.04

FPC, ARD, HRRG, Hodge-active are the fastest

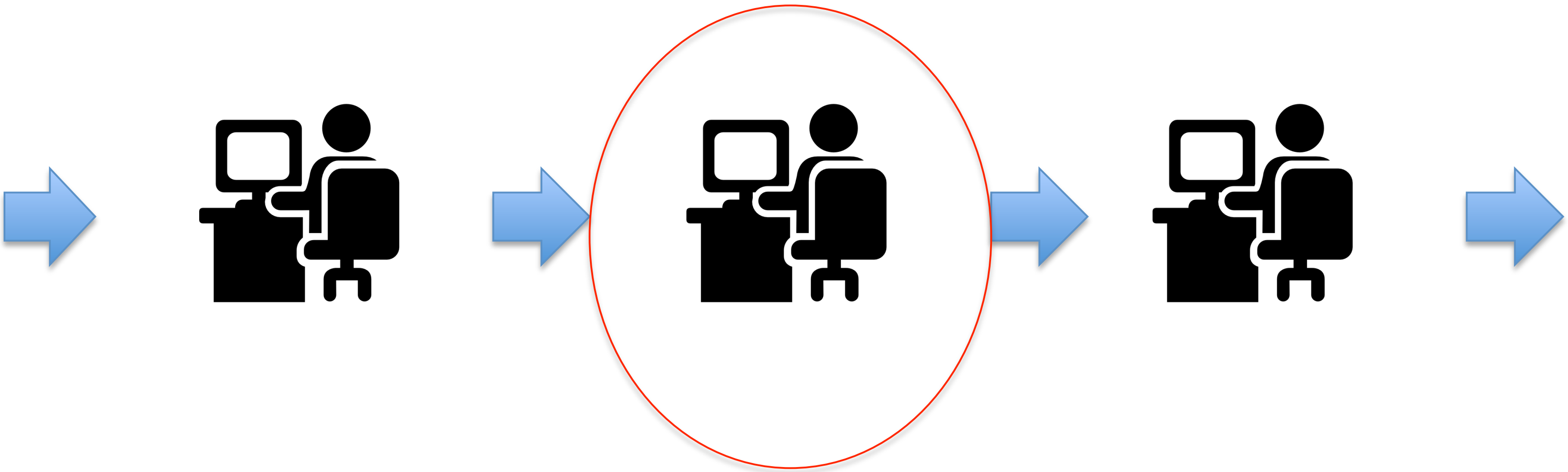
In learning based method:

- Hodge-active is faster than Crowd-BT and Hybrid-MST
- Hybrid-MST in GM mode is a little bit faster than Crowd-BT
- Hybrid-MST in MST mode is n times faster than Crowd-BT

In most cases, Hybrid-MST is in MST mode…

Considering crowd sourcing

Sequential sampling method: Hodge-active, Crowd-BT



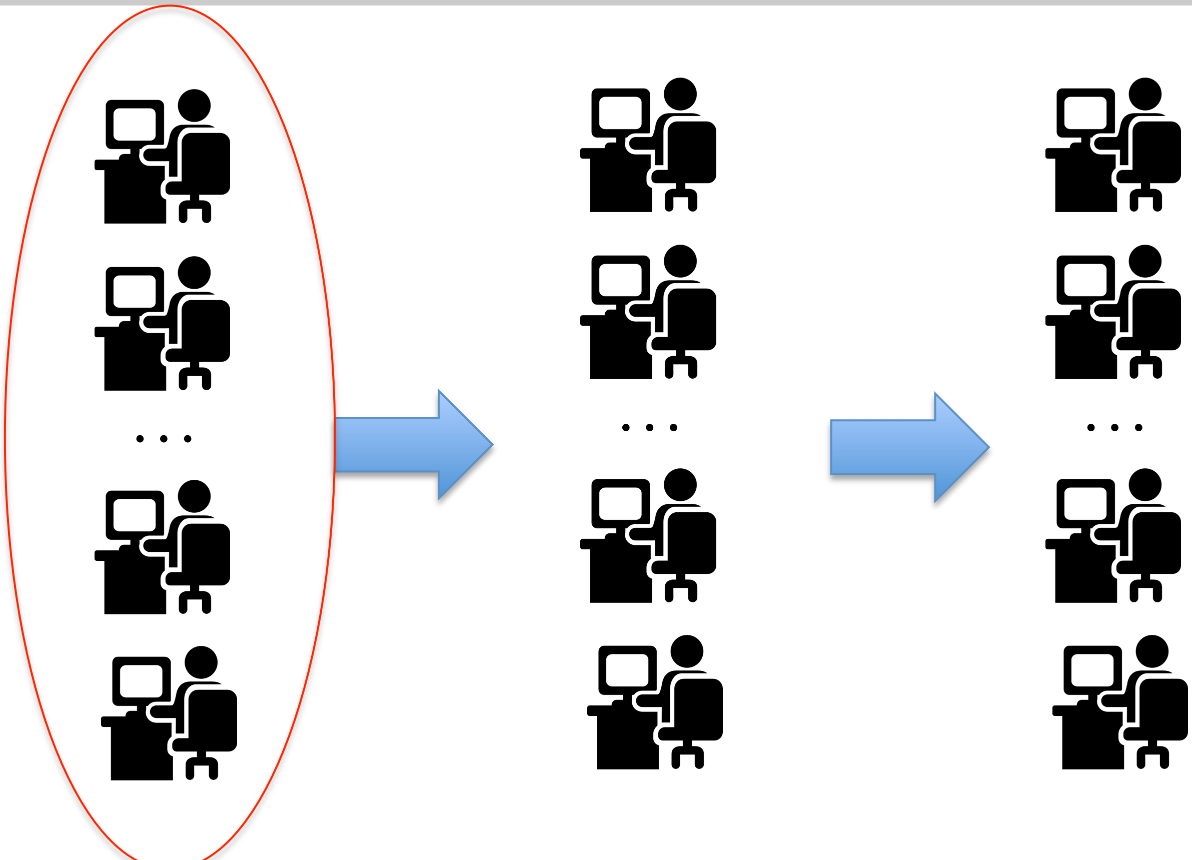
The next pair can only be determined when the previous voting is finished.

To finish **one** pairwise comparison procedure, T1+T2+T2 seconds are required:
T1: presentation time (e.g. 10 seconds)
T2: annotator voting time (e.g., 5 seconds)
T3: sampling algorithm runtime (according to the used algorithm)

Considering crowd-sourcing

Batch sampling method: Hybrid-MST (MST mode)

n-1 pairs



To finish **n-1** pairwise comparison procedure: T1+T2+T3 seconds

Time cost in real application

Table 2: Time cost (seconds) of comparing $n-1$ pairs in a typical VQA pair comparison experiment $(T1+T2+T3)$

n	Crowd-BT	Hodge-active	Hybrid-MST		
			GM	MST(ideal case)	MST (the worst case)
10	135.8	135.0	135.4	15.1	135.1
20	288.6	285.0	287.8	15.2	285.2
100	1782.0	1485.1	1782.0	17.9	1487.9

For MST:

- ❑ The worst case → the annotators work one after the other
- ❑ The ideal case → the annotators work at the same time

The proposed Hybrid-MST is more applicable in Crowd sourcing

Conclusion

- The contribution of our work:
  - ✓local information gain → faster computation
  - ✓Hybrid sampling strategy → reliable results
  - ✓MST → robustness to observation errors
  - ✓Batch mode → applicable in crowd sourcing

Conclusion

- Using Hodge-active [Xu,AAAI2018] when:
  - the test budget is small (< 2 standard trial numbers, i.e., 2n(n-1)/2) and the objective is for ranking aggregation
- Using Hybrid-MST when:
  - for rating aggregation
  - Test budget is large and for ranking aggregation
  - Small time budget

Beyond this…

UNIVERSITÉ DE NANTES

Thank you so much!

Paper is accepted by NIPS 2018
Code is available in github:
https://github.com/jingnantes/hybrid-mst
Paper is available in arXiv:
http://arxiv.org/pdf/1810.08851