Scoring and ranking strategies to benchmark cell type deconvolution pipelines

Vadim Bertrand Elise Amblard Magali Richard

Univ. Grenoble Alpes, CNRS, UMR 5525, TIMC, Grenoble, France

Corresponding author: vadim.bertrand@univ-grenoble-alpes.fr

Relevance of cell type deconvolution for cancer treatments

- **Bulk omics data** are nowadays a critical source of information for **cancer patients classification**, allowing finer **diagnostics** and **treatments**.
- However, current classifications could be improved, especially by taking into account the **tumor heterogeneity** through its **cell type proportion**.

Cell type deconvolution

$$\underset{F\times S}{D} = \underset{F\times K}{T} \times \underset{K\times S}{A}$$

with D being the **bulk matrix** of S samples and F features, T the **reference profiles matrix** of the F features for K cell types and A the **proportion matrix** of the K cell types in the S samples.

Constructing robust and comprehensive benchmarks

- Although many **deconvolution algorithms** have been proposed, there is **no consensus framework** for estimating cell type proportions from bulk samples and **new deconvolution methods are still being evaluated**.
- To compare these deconvolution approaches, a **robust scoring and ranking strategy** is needed, along with **comprehensive benchmarks**.

Desired properties [1]

Theoretical criteria

Several interesting theoretical criteria exist to evaluate the **resilience** of a ranking strategy **against judge or candidate perturbation**.

But they are quite hard to test in practice!

Gibbard's theorem states that no ranking process can satisfy all the desired theoretical properties. [2]

Empirical criteria

- Average rank of the winner: normalized averaged rank across the judges.
- Condorcet rate: the rate of ranking the existing Condorcet winner first.
- Generalization: similarity between a new judge ranking and the ranking of the original set of judges.

Our benchmark setting

- $N = 3$ **simulated** different datasets.
- $M = 10$ proportion matrices $A_{n,m}$ per dataset.

Bulk matrix simulation

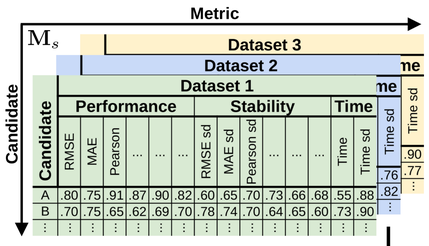
$$D_{n,m} = T_n \times A_{n,m} + \epsilon_{n,m}$$

where T_n is an **in-vitro** reference matrix, $A_{n,m}$ is sampled from a **Dirichlet** distribution to simulate the **biological noise** and $\epsilon_{n,m}$ is the **technical noise**.

- 182 deconvolution configurations (candidates) are evaluated.
- 3 metric categories: computational cost, performance and stability.
- 14 atomic metrics (judges) computed in total.
- Metric tensor $\mathbf{M}_m$ of size 182 (candidates) × 14 (judges) × 3 (datasets).

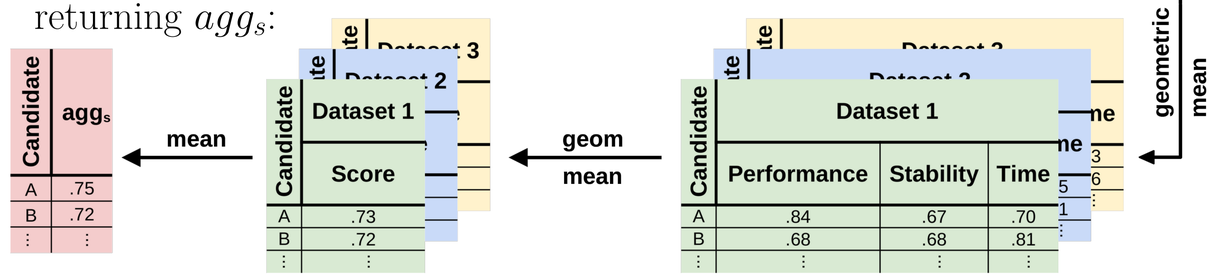
Scoring and ranking strategy

1. Per dataset scores **normalisation** s.t. they lie in $[0, 1]$ and 1 is best, yielding the tensor $\mathbf{M}_s$.



$\mathbf{M}_t$ and $\mathbf{M}_r$ are derived from $\mathbf{M}_s$ by computing the TOPSIS similarity [3] and ranks.

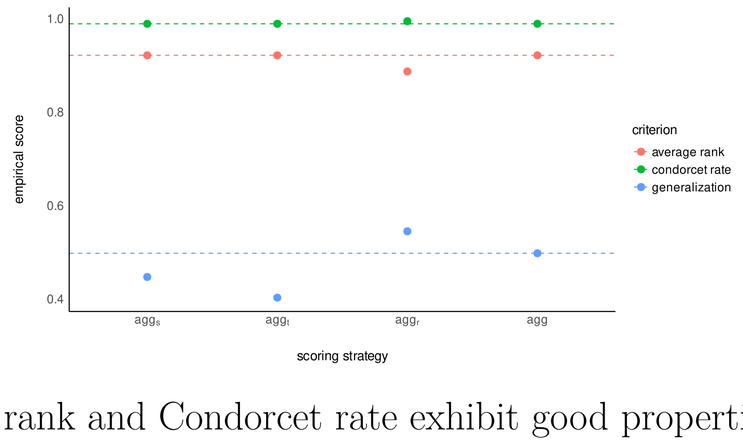
2. **Aggregation** per category, then across categories and datasets, returning agg_s:



Repeated with $\mathbf{M}_t$ and $\mathbf{M}_r$ as inputs, giving agg_t and agg_r.

3. **Final averaging** step $agg = (agg_s + agg_t + agg_r)/3$

Desired properties evaluation



- The average rank and Condorcet rate exhibit good properties of the proposed candidate winner.
- The aggregation of different categories of scores could explain the contrasted generalization performances.
- Our final strategy agg seems to benefit from all the 3 intermediate aggregations agg_s, agg_t, agg_r.

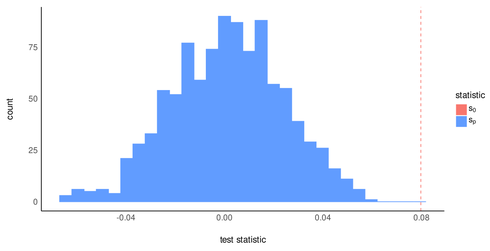
Statistically significant performance improvement

Non-parametric pair-wise permutation test for the final scores difference:

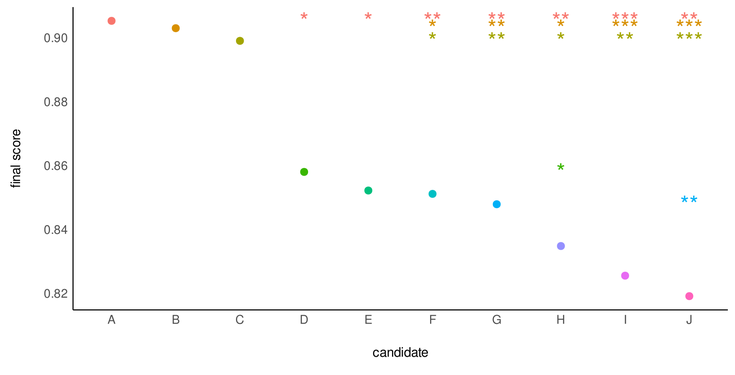
1. Compute the **observed** test statistic $s_0 = agg_0^A - agg_0^B$.
2. Repeat P times:
   - **Permute** "paired" metrics of the two candidates with a probability $p = 0.5$.
   - Compute agg_p^A, agg_p^B and s_p.
3. Compute the ***p*-value** associated with the test:

$$p - value = \frac{1}{P+1}(1 + \sum_{p=1}^{P} \mathbb{1}_{s_p \geq s_0})$$

p	Candidate	Dataset 1 metric 1	Dataset 1 ...	Dataset 1 metric 14	Dataset 2 metric 1	Dataset 2 ...	Dataset 2 metric 14	Dataset 3 metric 1	Dataset 3 ...	Dataset 3 metric 14	agg_p	s_p
0 (Obs.)	A	.80	...	.88	.76	...	.76	.78	...	.90	.75	.03
	B	.70	...	.90	.74	...	.82	.71	...	.77	.72	
1	A	.70	...	.88	.76	...	.76	.78	...	.77	.74	.01
	B	.80	...	.90	.74	...	.82	.71	...	.90	.73	
1000	A	.80	...	.90	.76	...	.76	.71	...	.90	.74	.03
	B	.70	...	.88	.74	...	.82	.78	...	.77	.71	



Final scores and significance levels of performance improvement:



References

[1] Adrien Pavao, Michael Vaccaro, and Isabelle Guyon. "Judging competitions and benchmarks: a candidate election approach". In: *ESANN 2021*. Oct. 2021.

[2] Allan Gibbard. "Manipulation of Voting Schemes: A General Result". In: *Econometrica* 41.4 (1973), pp. 587–601.

[3] Ching-Lai Hwang and Kwangsun Yoon. *Multiple Attribute Decision Making: Methods and Applications*. Springer Berlin, Heidelberg, 1981, pp. 69–70.