Probing differential expression patterns efficiently & robustly through adaptive linear multi-rank two-sample tests.

Đan Daniel Erdmann-Pham

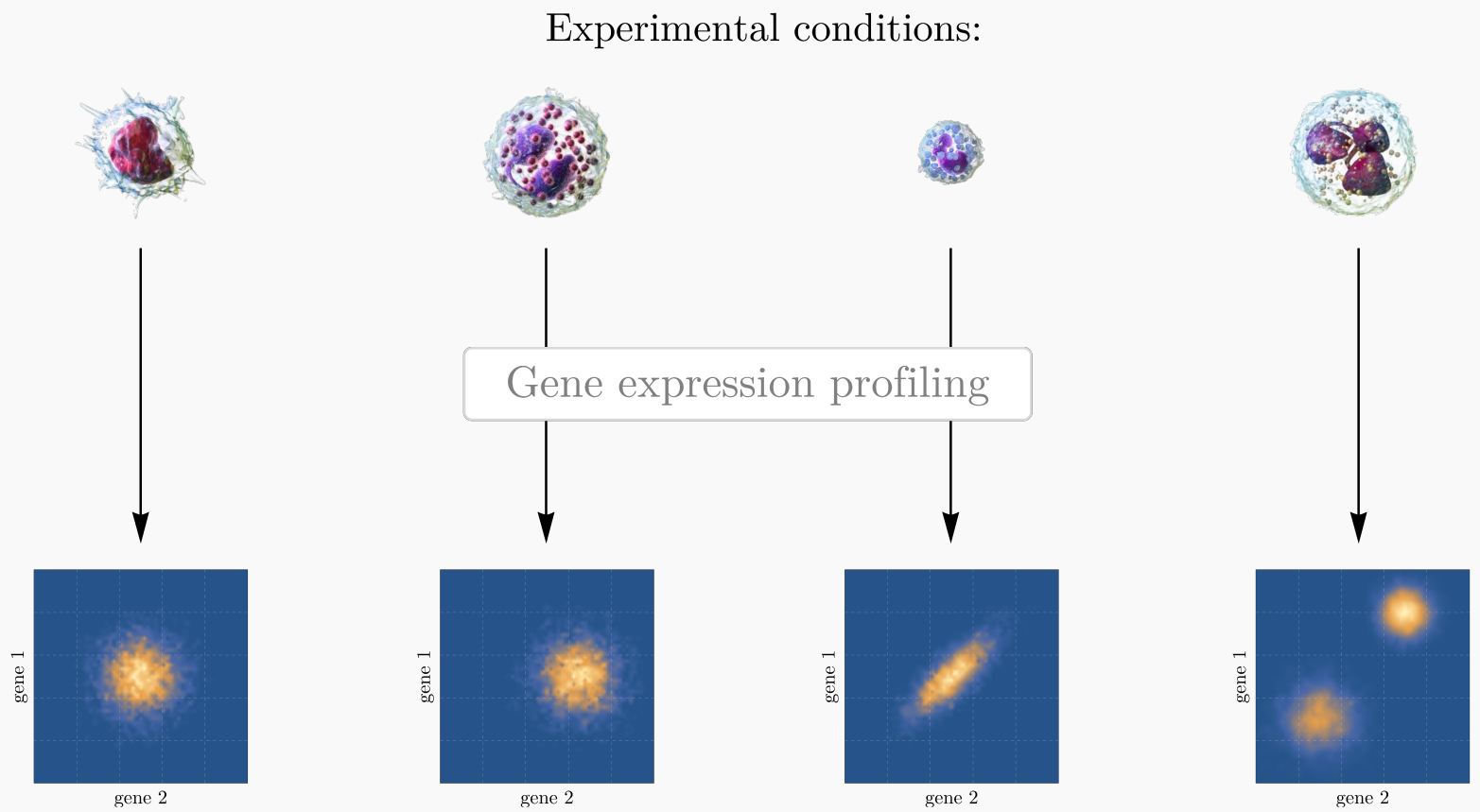
Stanford University

May 18, 2023

Workshop in Biostatistics

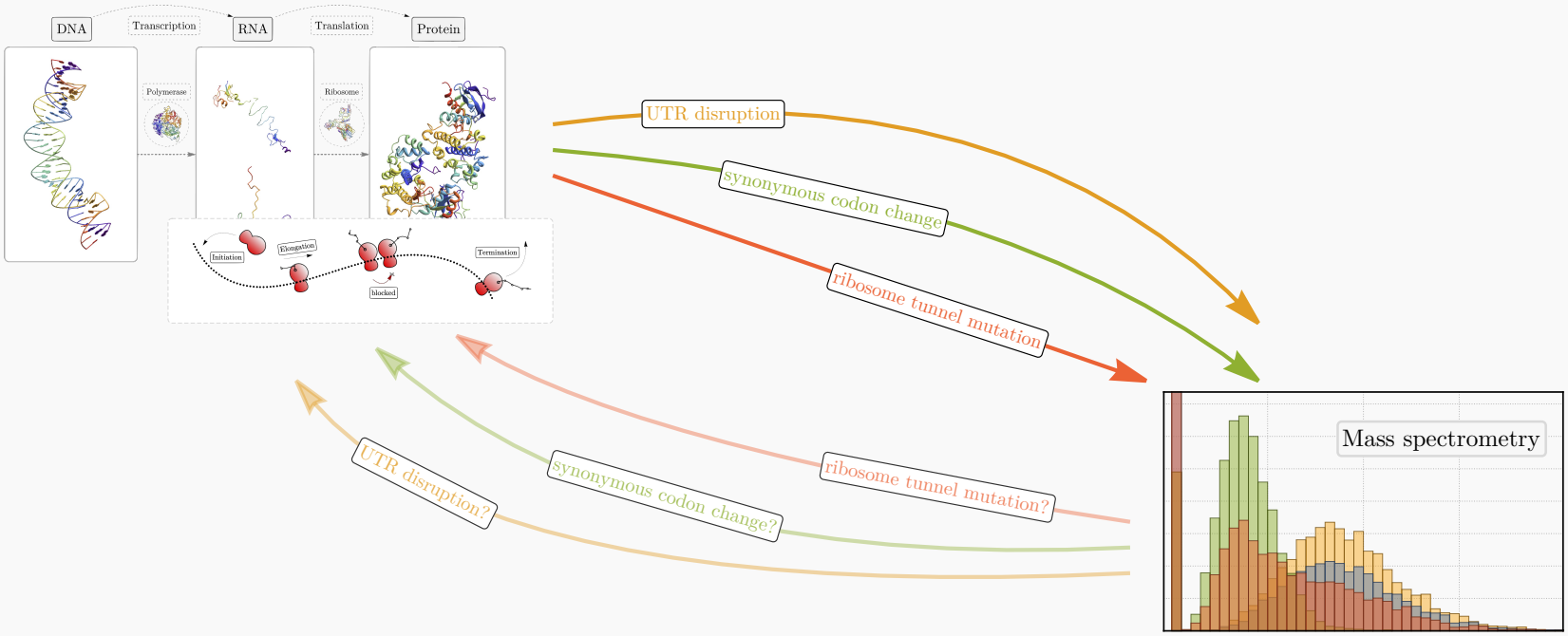
Motivation: Differential Gene Expression

Detecting expression heterogeneity



Broadly, differential expression analysis seeks to detect changes in expression patterns across experimental conditions.

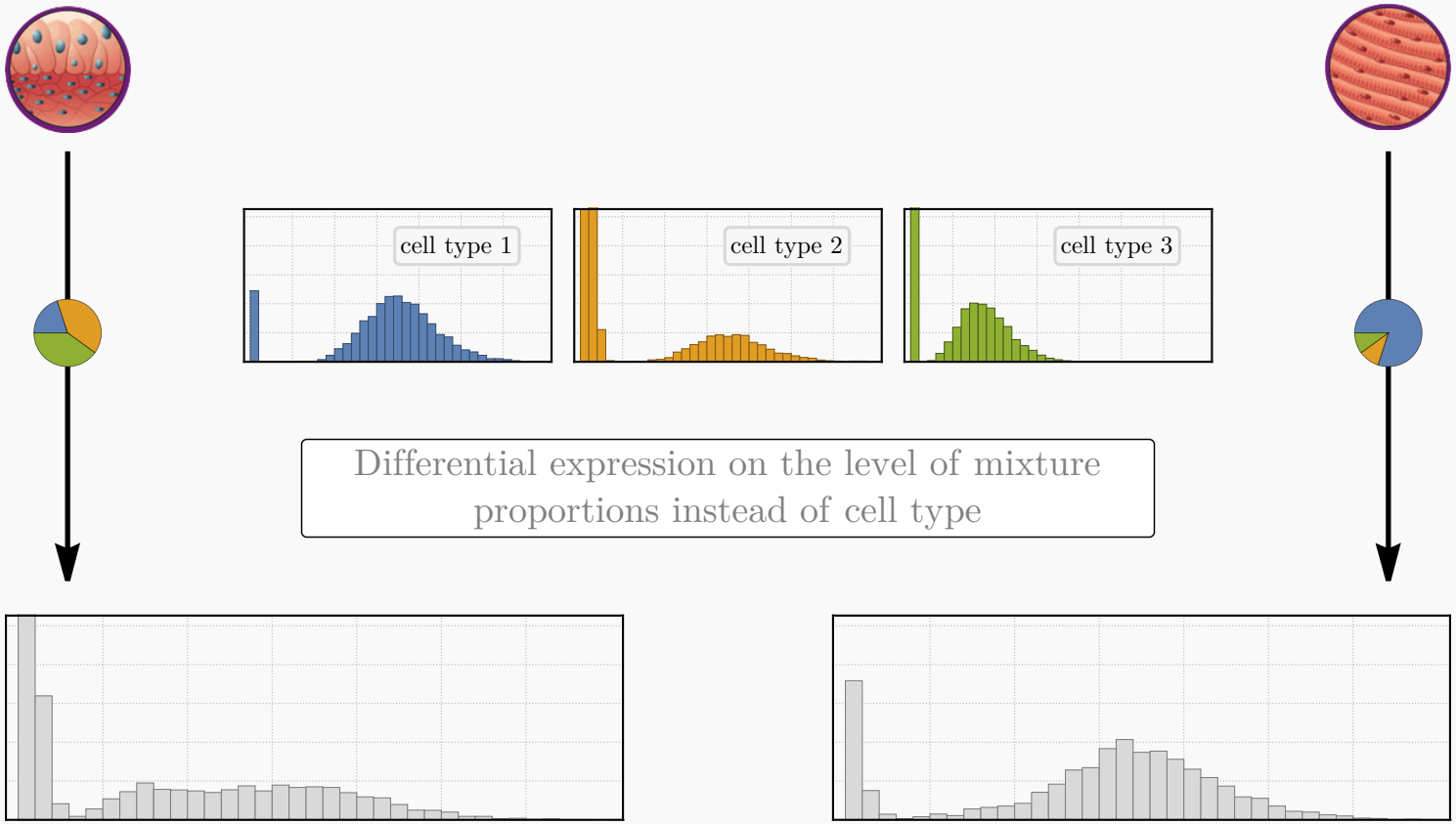
Two examples: single-cell proteomics & bulk RNA-Seq



Can we detect structural variation in a way that

- harnesses any existing model formations,
- is robust to model misspecification,
- allows for model selection,
- and is computationally feasible?

Two examples: single-cell proteomics & bulk RNA-Seq



Can we construct a differential expression test that is

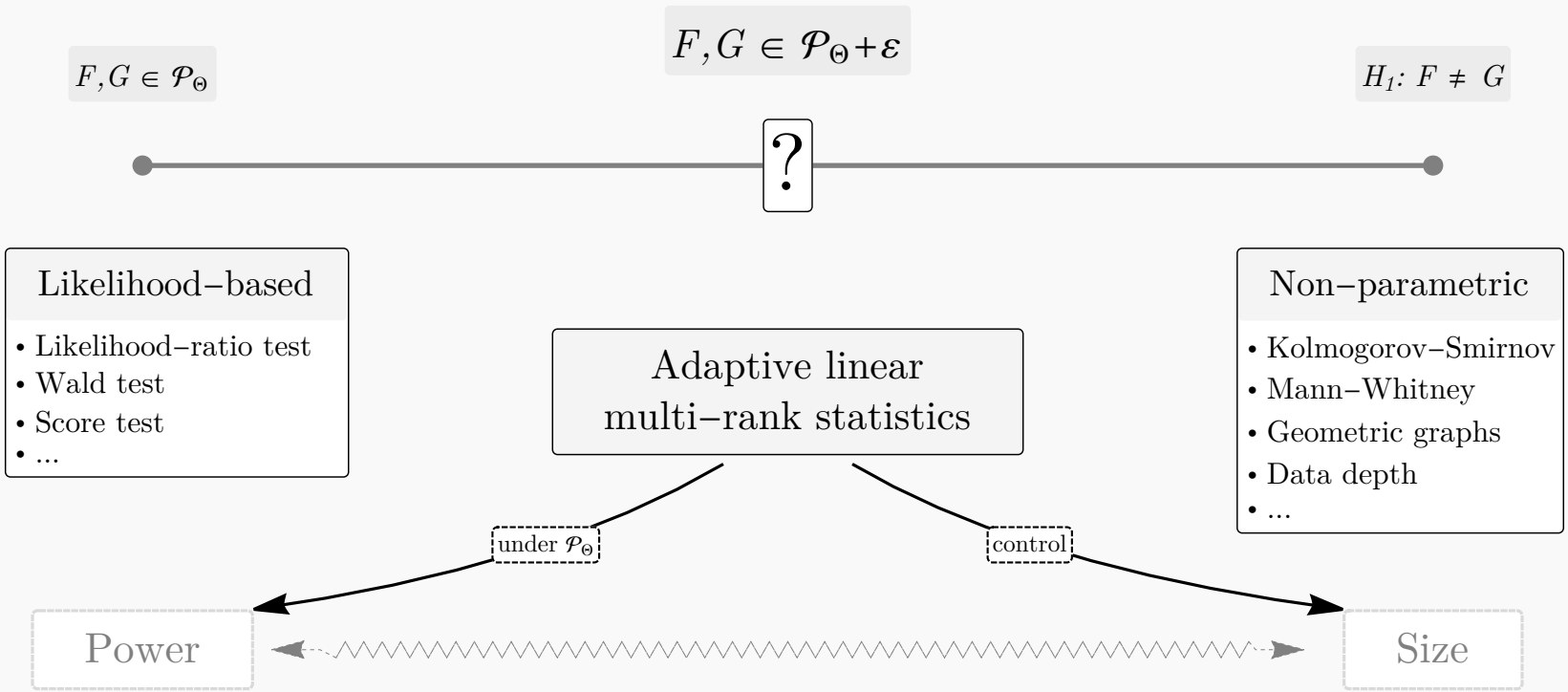
- sensitive to variation along desired directions,
- insensitive to variation along nuisance directions,
- and robust against model misspecification?

Statistical Formulation: K-sample tests

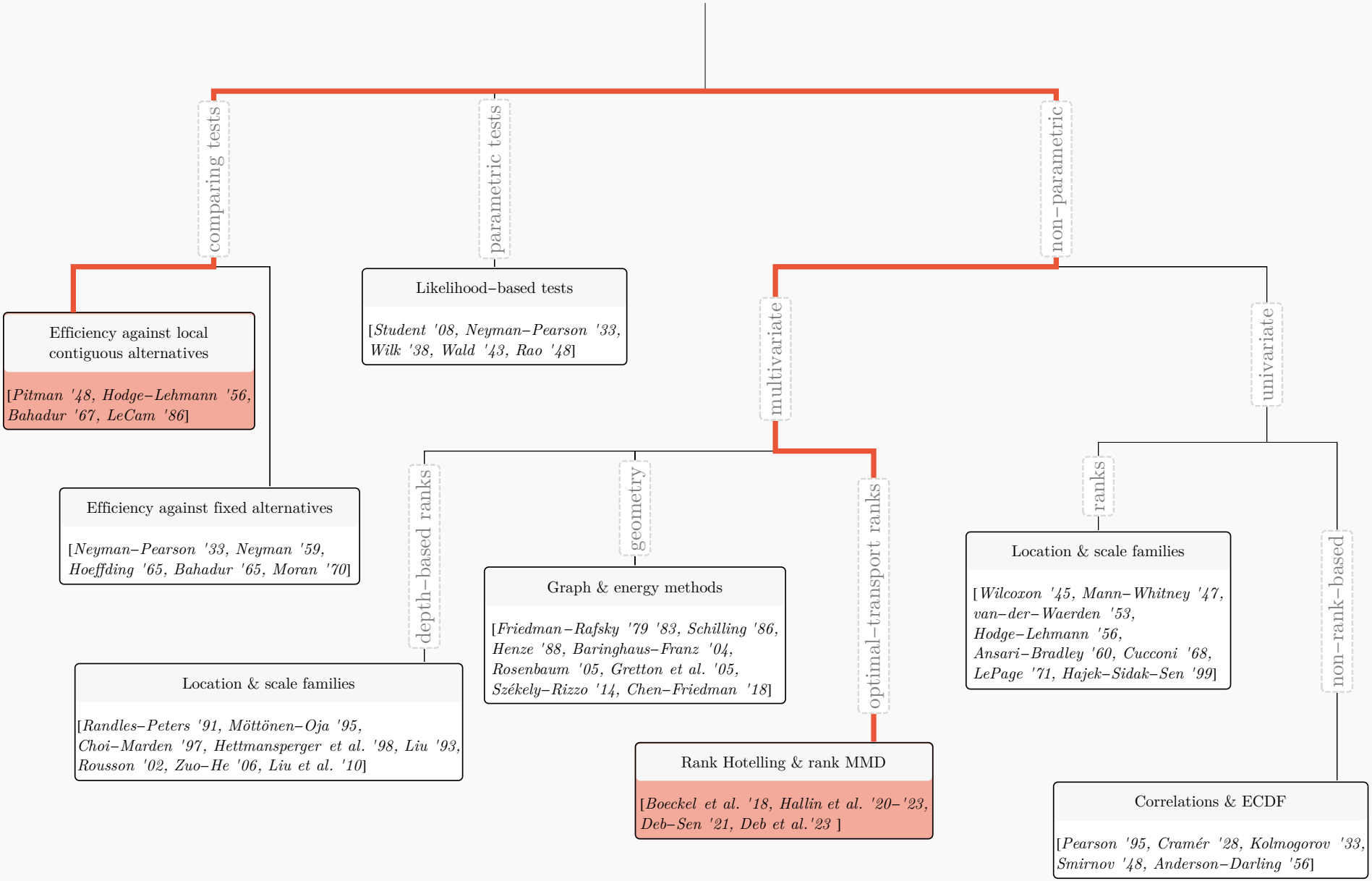
K-sample tests: $K = 2$

Given $\mathcal{X}_n = \{X_1, ..., X_n\} \sim F^{\otimes n}$ and $\mathcal{Y}_m = \{Y_1, ..., Y_m\} \sim G^{\otimes m}$, a two sample test probes

$$\mathrm{H}_0 : \{F = G\} \qquad \text{against} \qquad \mathrm{H}_1 \subseteq \{F \neq G\}.$$



Two-sample testing thus far

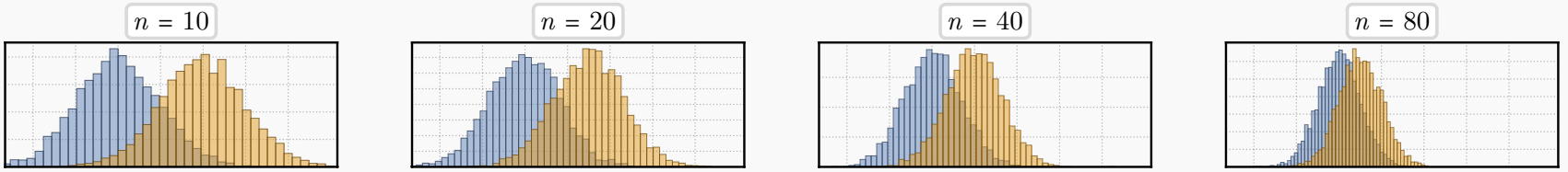


Univariate case: logistics of Mann-Whitney's U

Comparing tests: Pitman efficiency

Most reasonable tests are consistent against fixed alternatives and converge exponentially fast, rendering their comparison difficult.

Solution: Compare tests on alternatives approaching H_0 at rate $N^{-1/2}$ ($N = n + m$ and $\min\{n, m\} \to \infty$).



(Pitman) Asymptotic Relative efficiency

Informally, a test S_1 has Pitman efficiency η with respect to a test S_2 and alternatives H_n if for a given size α it requires η^{-1} as many samples as S_2 to achieve the same power.

Case study: Mann-Whitney is surprisingly efficient

Hodges and Lehmann showed that the Pitman efficiency of Mann-Whitney's U against Student's t is *lower bounded* by $108/125 \approx 0.864$ over all location families

$$\mathrm{H}_1 : G(x) = F(x + \theta N^{-1/2}).$$

Moreover, the Gaussian-score-transformed Mann-Whitney test (also known as van-der-Waerden test) is never less efficient than Student's t in the same setting.

That is, from an asymptotic, local perspective, van-der-Waerden tests should always be preferred over t tests.

Mann-Whitney: ingredients for efficiency

Reminder: Mann-Whitney's U is given by $\overline{\mathcal{R}}_m - \overline{\mathcal{R}}_n$, where $\mathcal{R}_n, \mathcal{R}_m$ are the ranks of $\mathcal{X}_n, \mathcal{Y}_m$ in the pooled sample $\mathcal{Z}_N = \mathcal{X}_n \cup \mathcal{Y}_m$. This test statistic is equivalent to

$$T_{n,m}^{\mathrm{id}} = N^{-1} \sum_{k=1}^{m} r_k = N^{-1} \sum_{k=1}^{m} \mathrm{id}\left(r_k\right),$$

where r_k is the normalized rank of Y_k in $\mathcal{Z}_N$. The van-der-Waerden test massages this into

$$T_{n,m}^{\Phi^{-1}} = N^{-1} \sum_{k=1}^{m} \Phi^{-1}\left(r_k\right),$$

where Φ^{-1} is the quantile function of a standard Gaussian.

Extracting efficiency ingredients: linear rank statistics

Both $T^{\mathrm{id}}_{n,m}$ and $T^{\Phi^{-1}}_{n,m}$ are examples of linear rank statistics:

Linear rank statistic

The linear rank statistic of weight $w : [0,1] \to \mathbb{R}$ associated with $\mathcal{X}_n, \mathcal{Y}_m$ is

$$T^{w}_{n,m} = N^{-1} \sum_{k=1}^{m} w(r_k) = (1-\alpha) \int w \circ H_{n,m} \, \mathrm{d}G_m,$$

where $\alpha = n/N$ and $H_{n,m} = \alpha F_n + (1-\alpha) G_m$.

Other examples of linear rank statistics

- Siegel-Tukey: $w(x) = |x - 1/2|$
- Mood: $w(x) = (x - 1/2)^2$
- Klotz: $w(x) = \Phi^{-2}(x)$

Question: How to choose w?

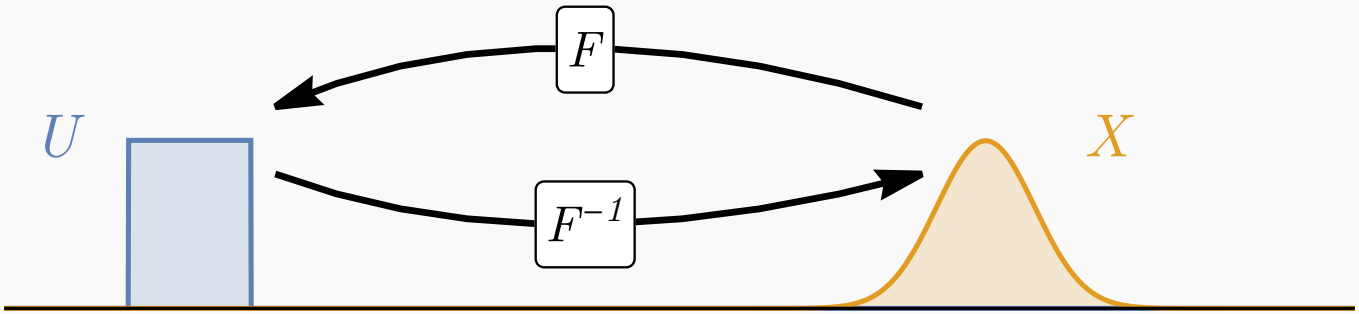
Multivariate case: Adaptive linear multi-rank statistics

Generalizing ranks from linear rank statistics

Question: What properties of (univariate) ranks are required for (univariate) linear rank statistics to work?

Answer contains essentially three parts:

- $r_k = F_n(X_k) \sim \text{Uniform}(\{1, 2, ..., N\}/N)$
- $F_n(X) \longrightarrow F(X) \sim \text{Uniform}([0, 1])$
- $F_n^{-1}(U) \longrightarrow F^{-1}(U) \sim F$ for $U \sim \text{Uniform}([0, 1])$

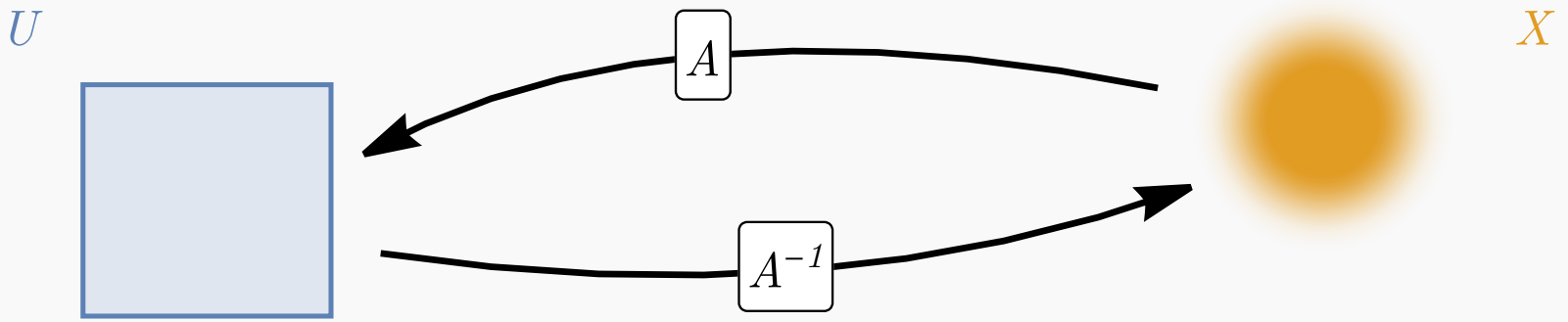


That is, F is a *transport map* sending the law of X to U.

Multivariate ranks as transport maps

This motivates the search for *multivariate transport maps* $A : \mathbb{R}^q \to \mathbb{R}^q$ such that for $U \sim \text{Uniform}(\square^q)$

$$A(X) \sim U \qquad A^{-1}(U) \sim X.$$



Generally, many such maps exist. A popular way to instantiate one is through *optimal transport*.

Multi-ranks via optimal transport

A population rank map A is given as the minimizer of $\mathbb{E}_X \|X - A(X)\|_p^p$ subject to $A(X) \sim U$. The empirical law of $\mathcal{X}_n$ gives rise to the empirical rank map A_n, where the constraint reads $A_n(\mathcal{X}_n) = \mathcal{A}_n$ for any $\mathcal{A}_n : \text{Uniform}(\mathcal{A}_n) \xrightarrow{w} \text{Uniform}(\square^q)$.

Two-sample testing based on multi-ranks

The notion of linear rank statistics then generalizes to multi-ranks in a straightforward manner:

$$T^w_{n,m} = N^{-1} \sum_{k=1}^{m} w \circ A_N(Y_k),$$

where $A_N : \mathcal{Z}_N \to \mathcal{A}_N$. The only existing two-sample test based on A_n targeting location families was proposed by Deb et al. ('23) and turns out to be equivalent to $(J : \square^q \to \mathbb{R}^q)$

$$S^J_{n,m} = \| N^{-1} \sum_{k=1}^{m} J \circ A_N(Y_k) \|_2^2 = \| T^J_{n,m} \|_2^2.$$

Linear multi-rank statistic

A linear multi rank-statistic is any statistic of the form $\| T^w_{n,m} \|_2^2$ for some $w : \square^q \to \mathbb{R}^\ell$.

Question: How to choose w?

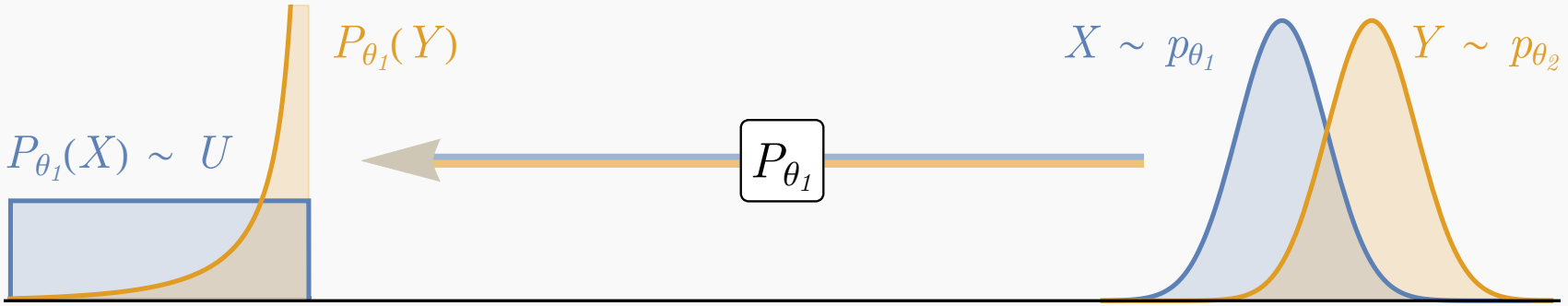
Connecting w and likelihood ratios

Given a family $\mathcal{P}_\Theta = \{p_\theta : \theta \in \Theta \subset \mathbb{R}^\ell\}$, the two-sample likelihood ratio statistic is given by

$$\lambda'_\Theta = \frac{\max_{\theta_1,\theta_2\in\Theta} p_{\theta_1}(\mathcal{X}_n)p_{\theta_2}(\mathcal{Y}_m)}{\max_{\theta\in\Theta} p_\theta(\mathcal{X}_n)p_\theta(\mathcal{Y}_m)}.$$

Pushing both p_{θ_1} and p_{θ_2} forward by the transport map $P_{\theta_1} : P_{\theta_1}\#p_{\theta_1} = \text{Uniform}(\square^q)$ gives the equivalent statistic

$$\lambda_\Theta = \max_{\theta_1,\theta_2\in\Theta} (P_{\theta_1}\#p_{\theta_2})(P_{\theta_1}(\mathcal{Y}_m)).$$



Observation: Under local alternatives, $A_N \approx P_{\theta_1} + O(N^{-1/2})$!

Connecting w and likelihood ratios

Locally around $(\theta_1, \theta_2) = (\theta^*, \theta^*)$, the log-likelihood ratio is governed by the behavior of the sample score

$$\sum_{k=1}^{m} z_{\theta^*}(Y_k) = \sum_{k=1}^{m} z_{\theta^*} \circ P_{\theta^*}^{-1} \circ P_{\theta^*}(Y_k) \approx \sum_{k=1}^{m} \underbrace{z_{\theta^*} \circ P_{\theta^*}^{-1}}_{w} \circ A_N(Y_k)$$

$$= \sum_{k=1}^{m} w \circ A_N(Y_k) = NT_{n,m}^{w}.$$

Theorem (Adaptive linear multi-ranks; informal)

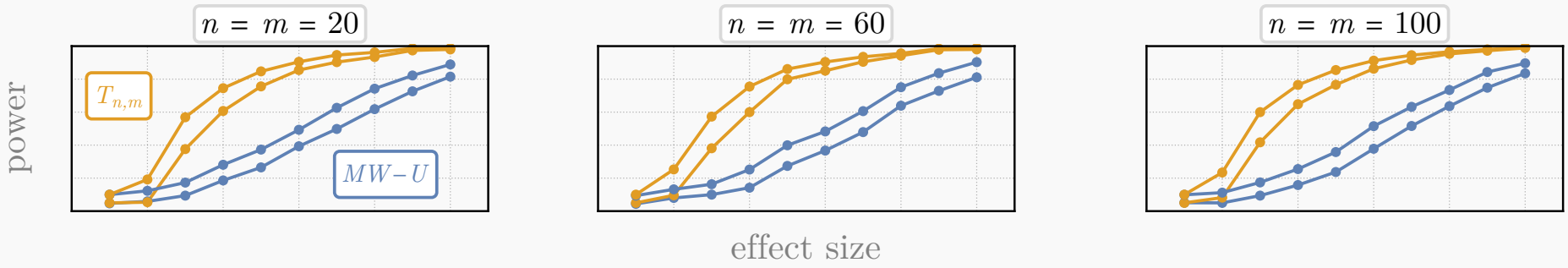
Set $\hat{w} = z_{\hat{\theta}} \circ P_{\hat{\theta}}^{-1}$ for any $\sqrt{N}$-consistent estimator $\hat{\theta}$ of θ based on $\mathcal{Z}_N$. Then under H_0, $(T_{n,m}^{\hat{w}} \mid \hat{\theta})$ is exactly distribution-free and converges to $\mathcal{N}(0, \alpha(1-\alpha)\int \hat{w} \otimes \hat{w})$. Moreover, under local contiguous alternatives in $\mathcal{P}_\Theta$ its Pitman efficiency relative to λ_Θ is 1.

Application: RNA-Seq Differential Expression Analysis

Benchmarking under zero-inflated NB model

Modeling considerations for RNA-Seq data

- data appear to follow negative binomial distributions more closely than other easily parametrized models, and so are modeled as such
- differential expression based on NB likelihood-ratios
- recent awareness [e.g., Li et al. '22] that popular differential expression packages are often miscalibrated (FDR/FWER inflation by a factor of 5)
- often suggested alternative: Mann-Whitney



Benchmarking under zero-inflated NB model

ε	n	m	GAMLSS	MDSeq	MOCHIS	diffVar	log-Bartlett
0	150	250	**0.803**	0.733	0.788	0.776	0.766
1	150	250	0.287	0.182	**0.747**	0.696	0.491
2	150	250	0.274	0.158	**0.724**	0.629	0.441
3	150	250	0.3	0.146	**0.688**	0.521	0.415
4	150	250	0.236	0.727	**0.755**	0.36	0.145
5	150	250	0.202	0.719	**0.731**	0.251	0.141
6	150	250	0.177	0.709	**0.726**	0.158	0.136
0	50	50	**0.59**	0.512	0.548	0.536	0.533
0	50	100	**0.658**	0.579	0.608	0.601	0.607
0	100	100	**0.738**	0.664	0.692	0.705	0.692
0	100	150	**0.765**	0.693	0.743	0.736	0.723
3	50	50	0.132	0.133	**0.43**	0.215	0.115
3	50	100	0.144	0.143	**0.49**	0.23	0.117
3	100	100	0.164	0.152	**0.59**	0.299	0.125
3	100	150	0.177	0.16	**0.636**	0.317	0.129

Additional features

Joint testing

Question: Many differential expression packages test separately for location & scale and adjust resulting p-values for multiple testing. Can this testing be performed jointly?

Solution: Given K weight functions $\{w_k : \square^q \to \mathbb{R}^{\ell_k}, j = 1, ..., K\}$, stack them to form weight $w = \sum_{k=1}^{K} e_k \otimes w_k : \square^q \to \mathbb{R}^{\ell}$, where $\ell = \sum_{k=1}^{K} \ell_k$, and perform tests on $T^w_{n,m}$.

Projecting out nuisance alternatives

Question: Can $T^w_{n,m}$ be designed to remain insensitive to batch effects, PCR amplification, etc.?

Solution: Given weight functions w_1 and w_2 sensitive to signal and noise alternatives, respectively, the weight $w = \Pi^{\perp}_{w_2} w_1$ remains powerful in the direction of w_1 while ignoring variation along w_2.

Additional features

$(K > 2)$-sample testing

Question: How is the previous discussion extended to the general setup involving K samples $\mathcal{X}_1, ..., \mathcal{X}_K$?

Answer: $T^w_{n,m}$ is naturally generalized to

$$T^w_{n_1,...,n_K} = N^{-1} \sum_{k=1}^{K} e_k \otimes \left(\sum_{j=1}^{n_k} w \circ A_N(X_{kj}) \right),$$

which enjoys all the previous power and robustness properties.

Qualitative weight constructions

Question: How should $T^w_{n,m}$ be used when a concrete candidate for $\mathcal{P}_\Theta$ doesn't exist?

Answer: Weight functions need not be derived from models, but can be assembled pure from qualitative observations.

Adaptive multi-ranks: summary

What happened?

- Starting from univariate linear rank statistics, $T^w_{n,m}$ generalizes them to
  - include multivariate sample spaces $\mathcal{X}_n, \mathcal{Y}_m \subset \mathbb{R}^q$
  - account for multivariate weight functions $w : \square^q \to \mathbb{R}^\ell$
- This is made possible through an extension of univariate ranks using transport maps.
- The resulting tests are as powerful as likelihood-ratios under correct model specification, yet remain well-calibrated (and comparably well-powered) in the absence thereof.
- Applications to RNA-Seq differential expression analysis show promise.

Thank you!