# A covariate-adaptive test for replicability across multiple studies with false discovery rate control

Dennis Leung

University of Melbourne

Nov 28, 2025

# Replicability crisis

**Scientists replicated 100 recent psychology experiments. More than half of them failed.**

by **Julia Belluz**
Aug 28, 2015, 4:00 AM GMT+10

# Where our work fits into research on enhancing replicability

- There's been progress in "meta-research". These endeavors primarily focus on transparency, ethics, reproducible computing practices, etc.
- We focus on the statistical methodology for replicability analysis:
- Suppose we do have multiple *reliable* and independent studies of data on the efficacy of a new drug. Based on these data, how do we *test whether the drug is effective in at least a good portion of, if not all, studies* ?

# Background: Conjunction null hypothesis

- Let $[\ell] \equiv \{1, \ldots, \ell\}$ for any natural number $\ell \in \mathbb{N}$
- Suppose $\mu_1, \ldots, \mu_n$ are the effects of the same underlying phenomenon in $n$ different studies (e.g. effectiveness of the new drug on $n$ different populations).
- $\mathcal{A} \subset \mathbb{R}$ is *null region*; the phenomenon is deemed non-existent in study $j$ if the null hypothesis
$$H_j : \mu_j \in \mathcal{A}$$
is true. (e.g. if $\mathcal{A} = (-\infty, 0]$, then the drug is only effective when $\mu_j > 0$.)
- Rigorously, we can test the conjunction null hypothesis that
$$|\{i : \mu_i \notin \mathcal{A}\}| \leq n - 1;$$
rejecting this means the effect exists consistently in all studies.
- More generally, the analyst can pre-specify a *replicability level* $u \in \{1, \ldots, n\}$, and test the partial conjunction (PC) null hypothesis (Benjamini and Heller, 2008)
$$H^{u/[n]} : |\{i : \mu_i \notin \mathcal{A}\}| \leq u - 1,$$
and declare the phenomenon $u$ out of $n$ replicable if $H^{u/n}$ can be rejected.

# Testing a PC null hypothesis

- Suppose $p_1, \ldots, p_n$ are independent $p$-values for their respective base nulls $H_1, \ldots, H_n$.

- Ordering them as $p_{(1)}, \ldots, p_{(n)}$, a $p$-value for $H^{u/[n]}$, also called a partial conjunction (PC) $p$-value, is typically formed as

$$p^{u/[n]} = f(p_{(u)}, \ldots, p_{(n)}),$$

where $f$ is a known $p$-value combining function, such as the Fisher function

$$f(p_{(u)}, \ldots, p_{(n)}) = 1 - F_{\chi^2_{2(n-u+1)}}\left(-2\sum_{j=u}^{n} \log(p_{(j)})\right),$$

where $F_{\chi^2_s}$ is the chi-squared CDF of $s$ degree.

- Easy to show that

$$P(p^{u/[n]} \leq t) \leq t \text{ for all } t \in [0,1] \text{ under } H^{u/[n]}.$$

Rejecting $H^{u/[n]}$ when $p^{u/[n]} \leq q$ controls Type I error under $q \in (0,1)$.

# Multiple testing of PC hypotheses

High-throughput experiments usually gives us many PC hypotheses to test:

**Example (Differential gene expression for autoimmune disorders)**

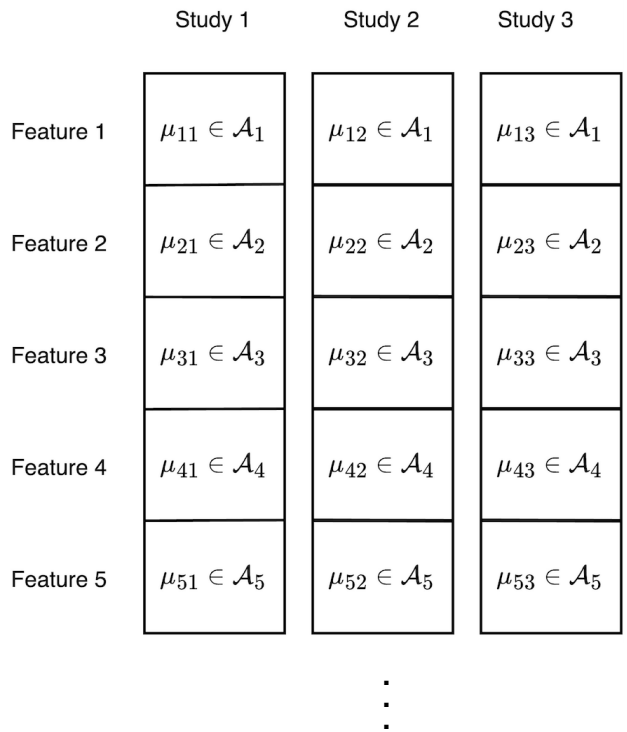
- Consider $n = 3$ independent mouse studies.
- Each study examines the same set of $m = 6,587$ genes in healthy and autoimmune medullary thymic epithelial cells (mTECs).
- For each $(i, j) \in [m] \times [n]$, $\mu_{ij} \in \mathbb{R}$ is the mean difference in expression level of gene $i$ between healthy and autoimmune mice in study $j$.
- If $\mu_{ij} \neq 0$ (i.e. $\mathcal{A}_i = 0$), then gene $i$ is deemed a potential marker for autoimmunity, as its expression differs between healthy and autoimmune mice on average.

# Multiple testing of PC hypotheses

- Let $\mathcal{A}_i \subseteq \mathbb{R}$ be the *null region* for feature $i$. We have the base *null hypotheses*

$$H_{ij} : \mu_{ij} \in \mathcal{A}_i \text{ for } (i,j) \in [m] \times [n].$$

- Visualization:



# Controlling the false discovery rate

- We aim to control the *false discovery rate (FDR)* when testing the PC nulls

$$H_1^{u/[n]}, H_2^{u/[n]}, \ldots, H_m^{u/[n]}.$$

- Suppose $\widehat{\mathcal{R}} \subseteq [m]$ is a data-driven set of rejected PC nulls; the FDR for $\widehat{\mathcal{R}}$ is

$$\mathrm{FDR}_{\mathrm{rep}} = \mathrm{FDR}_{\mathrm{rep}}(\widehat{\mathcal{R}}) \equiv \mathbb{E}\left[\frac{\sum_{i\in[m]} I\{i \in \widehat{\mathcal{R}}\} \times I\{H_i^{u/[n]} \text{ is true}\}}{\max(1, \sum_{i\in[m]} I\{i \in \widehat{\mathcal{R}}\})}\right].$$

- Standard protocol: applying the BH-procedure (Benjamini and Hochberg, 1995) to the PC $p$-values

$$p_1^{u/[n]}, \ldots, p_m^{u/[n]}.$$

  But it can be extremely underpowered, especially multiplicity has to be corrected for.

# Low power when $u = n$

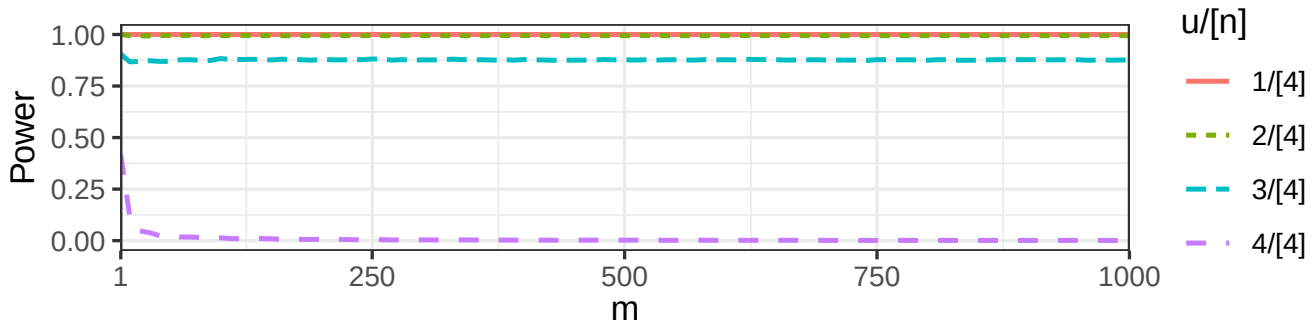



Figure: Power of the BH procedure (with FDR target $q = 0.05$) applied to $m = 1, 10, 20, \cdots, 1000$ PC $p$-values under replicability levels $u = 1, 2, 3, 4$, based on a simulation experiment with a total of $n = 4$ studies. ALL base hypotheses are non-null in this setting.

The ParFilter is our FDR-controlling method for simultaneously testing PC null hypotheses with power via *partitioning* and *filtering*.

# The ParFilter I

With a target FDR level $q$, a simplified version of ParFilter operates by:

1. *Partitioning* the $n$ studies into $K$ groups, i.e. for a chosen $K \in [u]$, let

$$\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K \subseteq [n]$$

be disjoint subsets partitioning $[n]$, and let $w_1, \ldots, w_K \in (0, 1]$ be some *local error weights* such that

$$\sum_{\ell=1}^{K} w_\ell = 1.$$

2. For each $i \in [m]$ and $k \in [K]$, define $u_{ik}$ as a *local replicability level* that satisfies

$$u_{ik} \leq |\mathcal{G}_k| \quad \text{for all } k \in [K] \qquad \text{and} \qquad \sum_{k \in [K]} u_{ik} = u.$$

# The ParFilter II

3. Define the *local* PC null hypothesis

$$H_i^{u_{ik}/\mathcal{G}_k} : |\{j \in \mathcal{G}_k : \mu_{ij} \notin \mathcal{A}_i\}| \leq u_{ik} - 1,$$

and form a *local* PC $p$-value

$$p_i^{u_{ik}/\mathcal{G}_k} \equiv f_{ik}\Big((p_{ij})_{j\in\mathcal{G}_k}\Big).$$

4. The ParFilter then considers a *candidate* rejection set

$$\mathcal{R}(\mathbf{t}) \equiv \bigcap_{k\in[K]} \left\{ i \in \mathcal{S}_k : p_i^{u_{ik}/\mathcal{G}_k} \leq \nu_{ik} \cdot t_k \right\}. \tag{1}$$

where

- $\mathbf{t} = (t_1, \ldots, t_K) \in [0, \infty)^K$ is a vector of thresholds.
- $\mathcal{S}_k \subseteq [m]$ is a *selected set* depending on $\{p_{ij}\}_{j\notin\mathcal{G}_k}$ ($p$-values outside of group $k$). Example:

$$\mathcal{S}_k = \bigcap_{\ell\in[K]\setminus\{k\}} \left\{ i \in [m] : p_i^{u_{i\ell}/\mathcal{G}_\ell} \leq w_\ell \cdot q \right\} \quad \text{for each } k \in [K]. \tag{2}$$

- $\nu_{1k}, \ldots, \nu_{mk} \in [0, \infty)$ are *local PC weights* that satisfies $\sum_{\ell\in\mathcal{S}_k} \nu_{\ell k} = |\mathcal{S}_k|$

# The ParFilter III

5. Consider the set of threshold vectors

$$\mathcal{T} \equiv \left\{ \mathbf{t} = (t_1, \cdots, t_K) \in [0, \infty)^K : \widehat{\mathrm{FDP}}_k(\mathbf{t}) \leq w_k \cdot q \text{ for all } k \in [K] \right\},$$

where

$$\widehat{\mathrm{FDP}}_k(\mathbf{t}) \equiv \frac{|\mathcal{S}_k| \cdot t_k}{|\mathcal{R}(\mathbf{t})| \vee 1}$$

conservatively estimates the groupwise false discovery proportion

$$\mathrm{FDP}_k(\mathbf{t}) \equiv \frac{\sum_{i \in [m]} I\left\{ i \in \mathcal{R}(\mathbf{t}) \right\} \times I\{ H_i^{u_{ik}/\mathcal{G}_k} \text{ is true} \}}{|\mathcal{R}(\mathbf{t})| \vee 1}.$$

6. Compute a data-dependent threshold vector $\hat{\mathbf{t}} = (\hat{t}_1, \cdots, \hat{t}_K)$ such that

$$\mathbf{t} \preccurlyeq \hat{\mathbf{t}} \text{ for all } \mathbf{t} \in \mathcal{T},$$

plug this into (1) and reach a final rejection set $\mathcal{R}(\hat{\mathbf{t}})$. It has the property

$$\mathrm{FDR}_{\mathrm{rep}}(\mathcal{R}(\hat{\mathbf{t}})) \leq q$$

under "standard assumptions".

The gist of the algorithm:

- When a feature $i$ is $H_i^{u_{ik}/\mathcal{G}_k}$ replicable for all group $k \in [K]$, then it is $u/[n]$ replicable.

- When the groupwise false discovery rate $\mathbb{E}[\mathrm{FDP}_k(\mathbf{t})]$ is under $w_i \cdot q$, then the overall $\mathrm{FDR}_{\mathrm{rep}}(\mathcal{R}(\mathbf{t}))$ is under $q$.

- The selection in (2) borrows information between different groups to filter out features that likely won't be $u/[n]$ replicable, so multiplicity in each group $k$ is cut down from $m$ to $|\mathcal{S}_k|$.

# Side information to further boost power

- There may be *side information* in the form of a valid *covariate* $x_{ij}$ that is also informative for testing $H_{ij}$.
- For instance, in the example of autoimmune disorders, $x_{ij}$ can be taken as the differential expression of gene $i$ in cells from a different part of the thymus other than the medulla, such as the cortex.
- These covariates can be used to train better local PC weights $\nu_{1k}, \ldots, \nu_{mk}$, to ultimately promote the rejection of the non-null $H_i^{u/[n]}$'s.

# Results for our applied example of autoimmune disorders

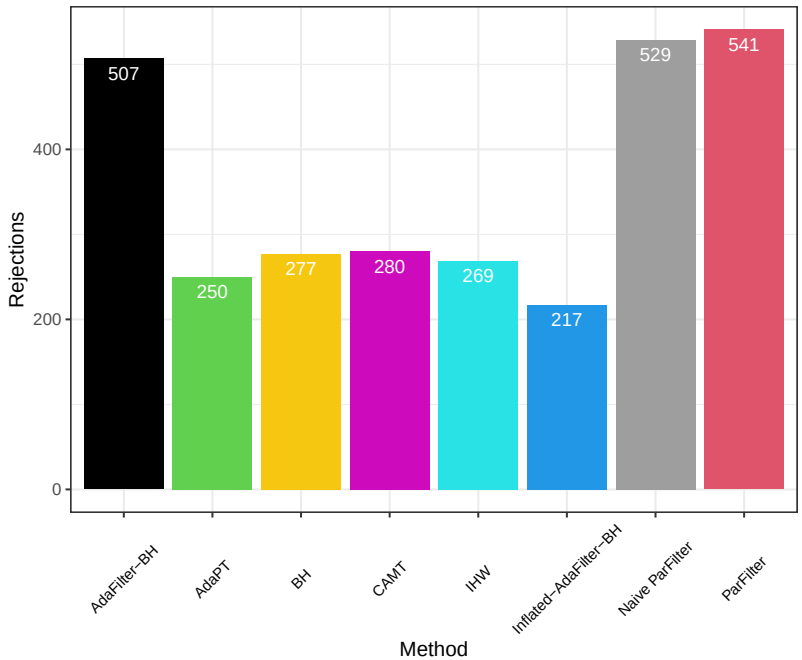



Figure: Rejection results for 3/[3] replicability across compared methods.

| Gene | Stouffer GBHPC $p$-value ($p_i^{3/[3]}$) |
|---|---|
| Mknk2 | 0.01260681 |
| Mreg | 0.01266433 |
| Ecscr | 0.01278160 |
| Jarid2 | 0.01286667 |
| Ncl | 0.01313040 |
| Nhsl1 | 0.01320058 |
| Bcl2l2 | 0.01328083 |
| Rell1 | 0.01344367 |
| Fgfbp1 | 0.01369939 |
| Antxr1 | 0.01378867 |
| Dkc1 | 0.01389120 |
| Hspg2 | 0.01389120 |
| Tnfrsf11a | 0.01485068 |

Table: Thirteen genes identified as 3/[3] replicated by ParFilter but not by other methods at $q = 0.05$.

# Future Work

- Undergoing revision.
- Extension to incorporate *e*-values (Ramdas and Wang, 2024) to more powerfully handle dependence across features.

# References


Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. Biometrics, 64(4):1215–1222.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300.

Ramdas, A. and Wang, R. (2024). Hypothesis testing with e-values. arXiv preprint arXiv:2410.23614.