

Original software publication



FastRerandomize: Fast rerandomization using accelerated computing

Rebecca Goldstein^a, Connor T. Jerzak^{b,*}, Aniket Kamat^c, Fucheng Warren Zhu^d^a Robert F. Wagner School of Public Service, New York University, New York City, NY, USA^b Department of Government, The University of Texas at Austin, Austin, TX, USA^c Department of Statistics, University of California, Berkeley, CA, USA^d Departments of Statistics and Computer Science, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Keywords:

Rerandomization
Experimental design
Randomization tests
Covariate balance
GPU acceleration
R

ABSTRACT

We present *fastrerandomize*, an R package for fast, scalable rerandomization in experimental design. Rerandomization improves precision by discarding treatment assignments that fail a prespecified covariate-balance criterion, but existing implementations can become computationally prohibitive as the number of units or covariates grows. *fastrerandomize* introduces three complementary advances: (i) optional GPU/TPU acceleration to parallelize balance checks, (ii) memory-efficient key-only storage that avoids retaining full assignment matrices, and (iii) auto-vectorized, just-in-time compiled kernels for batched candidate generation and inference. This approach enables exact or Monte Carlo rerandomization at previously intractable scales, making it practical to adopt the tighter balance thresholds required in modern high-dimensional experiments while simultaneously quantifying the resulting gains in precision and power for a given covariate set. Our approach also supports randomization-based testing conditioned on acceptance. In controlled benchmarks, we observe order-of-magnitude speedups over baseline workflows, with larger gains as the sample size or dimensionality grows, translating into improved precision of causal estimates. Code: github.com/cjerzak/fastrerandomize-software. Interactive capsule: fastrerandomize.github.io/space.

Metadata

Code metadata.

Nr.	Code metadata description	Metadata
C1	Current code version	v0.1.0
C2	Permanent link to code/repository used for this code version	github.com/cjerzak/fastrerandomize-software
C3	Permanent link to Reproducible Capsule	fastrerandomize.github.io/space
C4	Legal Code License	GNU General Public License v3.0 (GPL-3)
C5	Code versioning system used	git
C6	Software code languages, tools, and services used	R; Python (JAX/XLA backend via <i>reticulate</i>); CUDA and METAL support for GPU
C7	Compilation requirements, operating environments & dependencies	R ≥ 4.3; Python ≥ 3.11; <i>reticulate</i> ; <i>jax</i> , <i>jaxlib</i> , <i>numpy</i> ; optional GPU drivers
C8	If available, link to developer documentation/manual	fastrerandomize.github.io
C9	Support email for questions	connor.jerzak@austin.utexas.edu

* Corresponding author.

Email address: connor.jerzak@austin.utexas.edu (C.T. Jerzak).

URL: <https://RebeccaSGoldstein.com> (R. Goldstein), <https://ConnorJerzak.com> (C.T. Jerzak), <https://github.com/aniketkamat> (A. Kamat), <https://WarrenZhu.com> (F.W. Zhu).

<https://doi.org/10.1016/j.softx.2026.102508>

Received 27 September 2025; Received in revised form 22 November 2025; Accepted 4 January 2026

Available online 17 January 2026

2352-7110/© 2026 Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Motivation and significance

Randomized experiments remain a cornerstone of empirical research across the social, biomedical, and other data-intensive sciences. While randomization guarantees unbiasedness, it can—especially in small samples or high-dimensional settings—produce treatment and control groups with appreciable covariate imbalance, inflating variance and reducing statistical power. Rerandomization addresses this challenge by repeatedly drawing candidate assignments and accepting only those that pass a prespecified balance criterion [1,2]. Conditioning subsequent inference on the accepted set maintains valid design-based guarantees while improving precision.

Despite its conceptual simplicity, rerandomization can be computationally demanding. As the number of units grows, the space of possible assignments explodes combinatorially, and stringent balance thresholds may imply very low acceptance probabilities. In contemporary applications, covariate sets often include hundreds or thousands of features from text, images, or networks, exacerbating both runtime and memory pressure [3–5]. Practitioners might either forgo rerandomization or rely on ad hoc workflows that are difficult to scale.

Open-source software has advanced the state of practice in random assignment and randomization-based inference (e.g., `RIttools`, `ri2`, `randomizeR`, `RATest`, `RCT2` [6–10]). A smaller ecosystem targets rerandomization directly, including tools for power analysis, threshold selection, or iterative assignment generation [11–14]. Yet two persistent gaps remain: (i) scalable generation and storage of large pools of candidate randomizations under tight balance thresholds, and (ii) efficient, design-respecting inference once those pools are constructed.

`fastrerandomize` addresses these gaps through three complementary contributions:

1. **Accelerated balance checks.** Batched, auto-vectorized kernels evaluate balance criteria across many candidates in parallel; GPU/TPU support further reduces latency.
2. **Key-only storage.** Instead of caching full assignment matrices, the backend retains compact pseudo-random keys sufficient to regenerate any accepted assignment on demand, greatly reducing memory requirements.
3. **Integrated design-based inference.** Exact or Monte Carlo generation integrates with randomization tests conditioned on acceptance, including optional fiducial intervals [15,16].

Together, these features extend rerandomization to experimental scales—large n , high-dimensional covariates, and tight balance thresholds—that were previously impractical, while maintaining the rigorous design-based guarantees emphasized in the theoretical literature [1,2]. The practical need for such acceleration is most acute in modern applications where covariates are themselves outputs of deep learning pipelines—text embeddings, satellite-image descriptors, or neural-network features—that can yield hundreds or thousands of

covariates per unit [3–5]. Achieving the level of balance needed to markedly improve precision and reduce p -hacking in high-dimensional, representation-rich experiments [17] requires much more selective rerandomization rules and correspondingly larger pools of candidate assignments. `fastrerandomize` is designed to make this high-dimensional regime practically accessible: its accelerated balance checks and key-only storage allow researchers to adopt stringent, theoretically motivated thresholds in settings where covariates are derived from contemporary machine-learning models, without abandoning the rerandomization framework or incurring prohibitive computational costs.

2. Software description

2.1. Architecture

`fastrerandomize` follows a two-layer design (Fig. 1):

- An R front-end provides a compact API for assignment generation and inference, handling data validation, user-facing options, and I/O.
- A JAX backend (managed via `reticulate`) compiles batched kernels for candidate generation, balance evaluation, and randomization tests. Where available, XLA (the Accelerated Linear Algebra compiler) compiles and dispatches these kernels to GPUs/TPUs; otherwise, optimized CPU kernels are used.

Two additional design choices underpin the scalability of randomization to previously computationally prohibitive scales:

Batched processing. Candidate assignments are generated and evaluated in batches to limit peak memory usage. Batching amortizes kernel launches and data transfer overhead while allowing auto-vectorized linear algebra to saturate available compute units.

Key-only storage. For Monte Carlo workflows, `fastrerandomize` stores pseudo-random number generator (PRNG) keys for accepted assignments (rather than full n -length vectors). When an assignment is needed—for plotting, export, or inference—the PRNG deterministically regenerates it from its key. If the key length is $L \ll n$, memory decreases by a factor of approximately n/L , enabling very large pools without overwhelming memory (Fig. 2).

2.2. Choosing stringency

Having described the batched, accelerator-aware pipeline, we now turn to the practical question that governs how the software is used: *given an observed imbalance distance, what does it imply for precision, and how stringent should rerandomization be?*

We begin with a simple working model, drawing on the foundations of [1,2]. Let $X \in \mathbb{R}^{n \times d}$ denote pre-treatment covariates that have been column-standardized and whitened so that $\Sigma_X \approx I_d$, and (by a multivariate CLT / Gaussian approximation for sample mean differences)

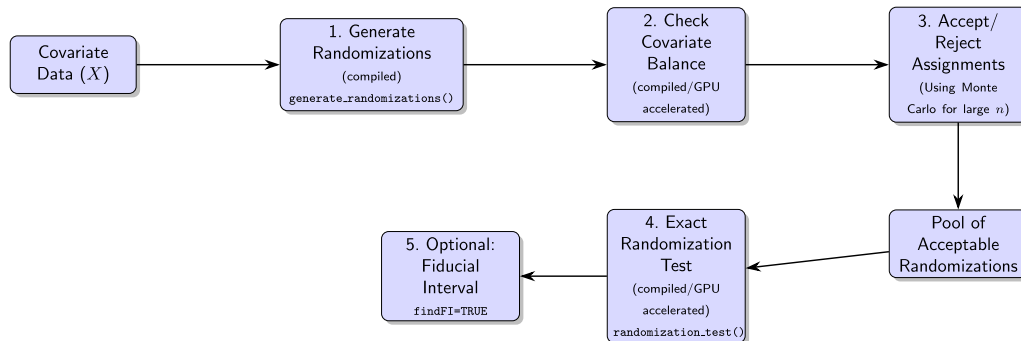


Fig. 1. Core workflow in `fastrerandomize`.

Batch Key Generation for Scalable Randomization

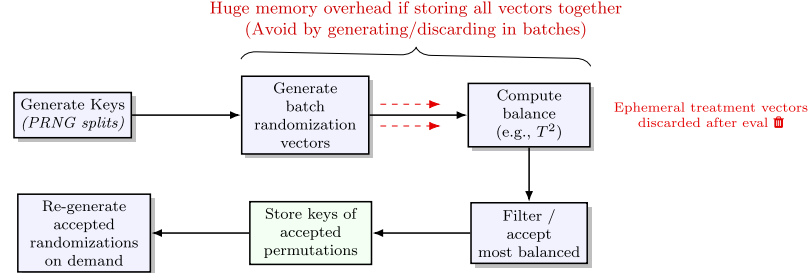


Fig. 2. Key-only storage. Ephemeral assignment vectors are generated, checked for balance, and discarded; only keys for accepted draws are retained. Assignments can then be regenerated on demand for analysis, reducing memory by $\approx n/L$.

Δ_X is approximately $\mathcal{N}\left(0, \left(\frac{1}{n_T} + \frac{1}{n_C}\right) I_d\right)$. For a realized assignment, the treated-control mean difference is $\Delta_X = \bar{X}_T - \bar{X}_C$, and a resulting multivariate imbalance measure is

$$M = \Delta_X^\top \Delta_X = \sum_{j=1}^d \text{SMD}_j^2,$$

reducing to the sum of squared standardized mean differences (SMDs). A larger M indicates greater imbalance; the goal is to understand how much that matters for precision and power.

Under a linear outcome model $Y_i(t) = \beta^\top X_i + \tau t + \epsilon_i$ with $\text{Var}(\epsilon_i) = \sigma^2$, define the share of outcome variability attributable to covariates:

$$R^2 \equiv \frac{\sigma_{\text{Prog}}^2}{\sigma_{\text{Prog}}^2 + \sigma^2}, \quad \sigma_{\text{Prog}}^2 \equiv \text{Var}(\beta^\top X_i).$$

A rotation-invariance argument (Appendix A) implies the following *typical-orientation* approximation for the difference-in-means RMSE targeting τ :

$$\text{RMSE}_{\hat{\tau}} \approx \sqrt{\sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C} \right) + \frac{R^2}{1-R^2} \frac{\sigma^2}{d} M},$$

where n_T and n_C denote the treated and control group sizes, respectively. Intuitively, the first term is irreducible residual variance; the second term is a data-dependent penalty from the observed imbalance M . The power-relevant signal-to-noise ratio $|\tau|/\text{RMSE}_{\hat{\tau}}$ therefore improves either by larger effects $|\tau|$ or by smaller M ; once this ratio is already large enough for the study's power target, tightening balance further will not change conclusions in a meaningful way.

Instead of setting a threshold on M , it is operationally simpler to specify an *acceptance probability* q (e.g., retaining the best 1% of candidates). Since $M \approx \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \chi_d^2$ for whitened covariates, this strategy effectively truncates this χ_d^2 distribution. Concretely, if $q \equiv \Pr(\text{accept}) = \Pr(M \leq a)$, then under the χ^2 approximation we have $q = \Pr(\chi_d^2 \leq c_q)$ with $c_q \equiv F_{\chi_d^2}^{-1}(q)$, and the implied threshold on M is $a_q = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) c_q$. Under this rule, the shrinkage factor can be written as $v_a(d) = \mathbb{E}[\chi_d^2 | \chi_d^2 \leq c_q]/d = \Pr(\chi_{d+2}^2 \leq c_q)/q$ (see Appendix A). Geometrically, this restricts Δ_X to a tighter hypersphere, shrinking the imbalance covariance by a scalar factor $v_a(d) \in (0, 1)$, which we compute in `diagnose_rerandomization()`. This shrinkage directly reduces the expected MSE:

$$\mathbb{E}[\text{MSE}(\hat{\tau}_{\text{DiM}}) | \text{accept}] = \left(\frac{1}{n_T} + \frac{1}{n_C} \right) \left(\sigma^2 + v_a(d) \sigma_{\text{Prog}}^2 \right),$$

$$\text{where } \sigma_{\text{Prog}}^2 = \frac{R^2}{1-R^2} \sigma^2.$$

Thus, increasing stringency (lowering q) leaves irreducible residual variance (σ^2) unchanged while systematically attenuating the error contribution from prognostic imbalance.

As a rule of thumb, at $q=0.01$ acceptance, the expected scaling term on the covariate-imbalance factor is $v_a(d) = 0.45$ at $d=30$, it is 0.66 at $d=100$, and it is 0.88 at $d=1000$ (larger values mean greater contribution to MSE from imbalance). In precision terms, this translates, at $n_T=n_C=500$ and $R^2=0.5$, to an RMSE reduction of about 15% at $d=30$, but only 3% at $d=1000$. Therefore, in very high- d settings, even more selective acceptance rates are necessary to yield the same precision gains as in small dimensionality settings [1]—further motivating the need for the accelerated computing approach taken here. Fig. 3 illustrates this compute-precision frontier, showing that while modest pools suffice for low dimensions, achieving comparable precision gains in high-dimensional settings requires stringent acceptance rates (low q) and therefore larger candidate pools.

Two recurring planning questions follow naturally. First, *does drawing more candidates improve balance?* No: balance is governed by q , and increasing draws mainly increases the number of accepted assignments and hence the Monte Carlo resolution. Holding $q=0.01$ fixed, going from 10^5 to 2×10^5 candidates simply doubles the accepted set (from 1000 to 2000 assignments) and halves the minimum attainable randomization-test p (from 0.0010 to 0.0005); balance itself is unchanged.

Second, *how should q be chosen?* A simple planning heuristic is that rerandomization at level q typically requires about $1/q$ candidate draws per accepted assignment. A power-focused approach simplifies q selection: specify (n_T, n_C) , an outcome scale σ , a plausible R^2 , and the effect size of interest $|\tau|$ with desired power and size. In an example high-dimensional case ($d=1000$, $n_T=n_C=500$, $\sigma=1$, $R^2=0.4$), targeting a two-sided $\alpha=0.05$ test with 80% power for $|\tau|=0.2$ implies a recommended acceptance probability of $q^* \approx 0.000012$, or about 81226 draws per accepted assignment. This inversion is implemented by `diagnose_rerandomization()`.

In sum, we recommend the following iteration cycle for selecting imbalance thresholds in a given scenario:

1. Start with a coarse q , compute $\text{RMSE}_{\hat{\tau}}$ from the observed M , using parameters from prior literature.
2. Tighten q if $|\tau|/\text{RMSE}$ is below the power target.
3. Stop when additional tightening results in minimal change.

2.3. Software functionalities

The package exposes four core functions. Further details are provided in Appendix B.

`build_backend()`. Creates a minimal conda environment (e.g., named "fastrerandomize") with Python and JAX dependencies. If a GPU or TPU device is present, the package selects an appropriate accelerated backend.

`generate_randomizations()`. Constructs pools of acceptable assignments under a specified balance criterion:

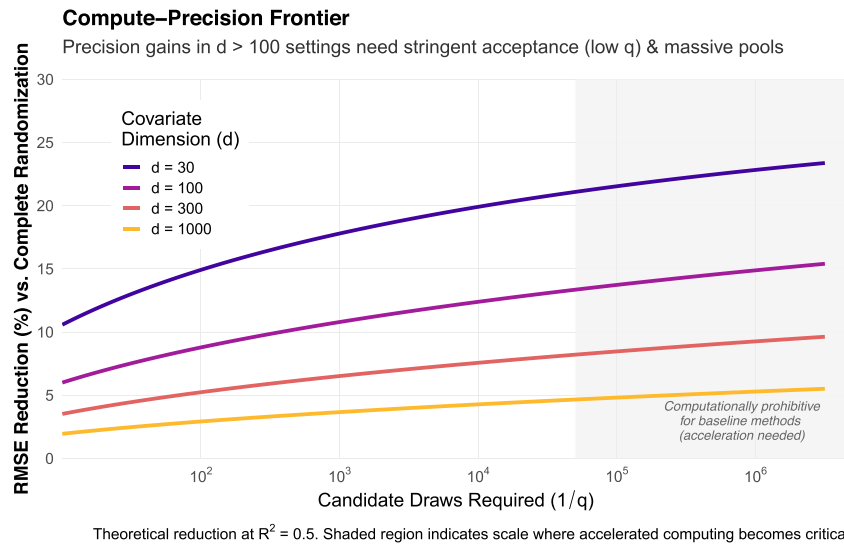


Fig. 3. The Compute-Precision Frontier. Theoretical RMSE reduction (%) versus complete randomization (assuming $R^2 = 0.5$) as a function of required candidate draws ($1/q$) across dimensions $d \in \{30, 100, 300, 1000\}$. Achieving meaningful precision gains in high-dimensional settings ($d > 100$) requires stringent acceptance criteria.

- **Exact enumeration** (`randomization_type = "exact"`): systematically iterates over all possible assignments when feasible.
- **Monte Carlo** (`"monte_carlo"`): draws and filters batched candidates, storing keys for accepted draws.

Users control stringency via `randomization_accept_prob`. Hotelling's T^2 is the default balance metric; a custom vectorized function can be supplied.

`diagnose_rerandomization()`. Maps the diagnostics in Section 2.2 into a single planning tool. Given either an observed imbalance summary or design-stage inputs ($n_T, n_C, d, R^2, \sigma, |\tau|$), the function reports (i) the implied realized RMSE and a conservative upper bound, (ii) the expected ex-ante RMSE under a given acceptance probability q , and (iii) the largest q that meets a user-supplied precision or power target.

`randomization_test()`. Implements design-respecting inference conditional on acceptance: the observed statistic (e.g., difference-in-means) is compared to its distribution across accepted assignments, yielding an exact or approximate p -value for the sharp null [1,2,15] (depending on whether full enumeration or Monte Carlo is used). Optional inversion returns a fiducial interval [16]. Because only accepted assignments enter the reference distribution, p -values are lower-bounded by $1/M_{\text{Accept}}$, where M_{Accept} is the number of accepted draws [15].

3. Illustrative example and workflow

High-dimensional pre-treatment features—e.g., text embeddings, satellite-image descriptors, or network-derived covariates—are increasingly common [3–5]. `fastrerandomize` is designed to enable stringent rerandomization in these settings while preserving design-based guarantees.

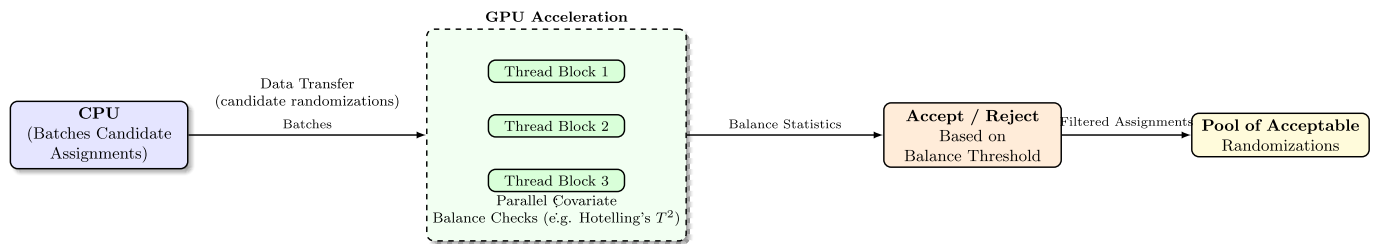
To fix ideas, consider a field experiment that randomizes roughly $n \approx 1000$ villages into treatment and control. For each village, investigators construct a high-dimensional vector X_i from strictly pre-treatment satellite imagery, using a CLIP-based neural network image embedding model to extract the 768 features summarizing settlement structure, roads, vegetation, and other contexts [4,18]. Researchers would like villages to be well-matched on these remotely sensed covariates, so that outcomes are not confounded by local conditions visible from above. Achieving this with a tight rerandomization rule—for example, keeping

only the best 1% of candidate assignments by imbalance in X —can substantially shrink the imbalance component of the MSE (2.2), but requires generating on the order of $1/q$ candidate draws per accepted assignment. With d in the hundreds, such regimes quickly become intractable for existing methods. A typical `fastrerandomize` workflow in this context is:

1. **Prepare covariates.** Assemble the $n \times d$ matrix X (rows are units, columns are features). Column-wise standardization is recommended. Covariates predictive of the outcome should be prioritized for inclusion in X .
2. **Choose stringency.** Set `randomization_accept_prob` to the desired acceptance probability, q , following the diagnostic approach outlined in Section 2.2 in order to model balance-compute-power trade-offs. Holding q fixed, increasing the number of draws enlarges the accepted pool (lowering the minimum attainable randomization-test p -value) without changing expected balance. `diagnose_rerandomization()` can be used to select an acceptance probability or threshold consistent with a target precision/power.
3. **Generate candidates at scale.** Call `generate_randomizations()` (usually with `randomization_type = "monte_carlo"`), using batched draws and, where available, GPU/TPU acceleration to evaluate balance statistics in parallel. Internally, `fastrerandomize` relies on key-only storage to avoid holding large binary assignment arrays in memory.
4. **Infer with design respect.** Use `randomization_test()` to compute p -values (and optionally fiducial intervals) from the distribution of the test statistic over the *accepted* assignments, preserving design-based validity under either exact or Monte Carlo generation.

4. Impact

Enabling stringent balance at scale. Rerandomization improves the precision of treatment effect estimators by conditioning on good covariate balance [1,2]. In practice, stringent thresholds can sharply reduce acceptance probabilities, requiring large candidate pools. `fastrerandomize` makes this regime tractable through batched, compiled kernels and optional GPU/TPU acceleration (Fig. 4). In comparative simulations, we observe speedups over baseline implementations exceeding 42× in



Massively parallel threads reduce runtime for large-scale inference.

Fig. 4. GPU acceleration: CPU forms batched candidates; balance checks are parallelized across many compute units on the GPU.

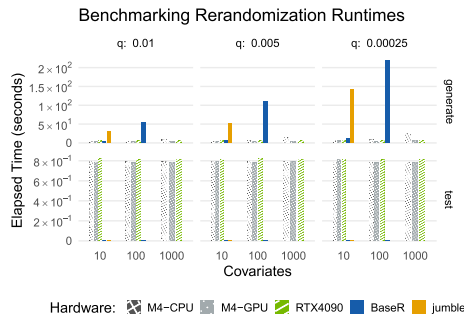


Fig. 5. Baselines and *fastrerandomize*'s CPU/GPU runtimes for $n=100$. Bars show elapsed time (seconds) for (top) generating large pools of randomizations and (bottom) randomization-based inference, across covariate dimensions and rerandomization acceptance thresholds ($q \in \{0.01, 0.005, 0.00025\}$). Base R and *jumble* implementations are shown only for scenarios in which they complete within a reasonable time frame; in larger-scale settings or when $n \leq d$, they become computationally prohibitive or do not produce results with default settings.

high- d and high- n regimes and when requiring more stringent acceptance thresholds.

To demonstrate, we report controlled benchmarks that vary sample size (n), covariate dimension (d), and Monte Carlo draw size (Figs. 5–6). For $n=100$ and $d=100$ with $q=0.005$, the GPU backend (both Apple M4 [METAL] and NVIDIA RTX 4090 [CUDA]) completes pool generation in about 5 s, versus 112 s for a baseline R approach ($\approx 24\times$ speedup). We also benchmarked the recent *jumble* package [13], currently one of the only other R tools that perform rerandomization with a user-specified acceptance rate. *jumble*'s performance lags behind the pure-R baseline, confirming that our gains stem from batched, XLA-based compilation and hardware acceleration rather than minor R implementation differences. Relative to the XLA-optimized *fastrerandomize* CPU backend, GPU execution reduces runtime by up to 96% at $d=1000$. At $n=1000$ and $d=1000$, pool generation falls from 91 s (CPU) to 7 s (GPU) at $q=0.005$. Fig. 7 shows how the GPU advantage increases with problem size in n .

Reducing memory pressure. Key-only storage decouples *evaluation* from *retention*. Assignments are generated and evaluated in place; only the compact keys for accepted draws persist. If a key consists of two integers ($L=2$) and $n=10^3$, retaining keys instead of full 0/1 vectors reduces memory usage by roughly a factor of $n/L \approx 500$ (Fig. 2). This enables the consideration of stringent randomization thresholds without exhausting memory.

Design-based inference. By returning a structured object for accepted draws, *fastrerandomize* integrates well with randomization tests conditioned on the acceptance structure. p -values are computed by comparing the observed statistic to its distribution across *accepted* assignments; inversion yields intervals [15,16].

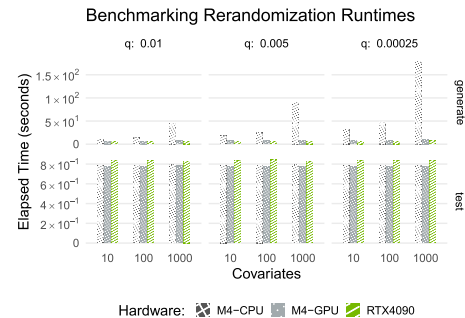


Fig. 6. *fastrerandomize* with CPU vs. GPU for $n=1000$. Bars show elapsed time (seconds) for (top) pool generation and (bottom) inference, across covariate dimensions and acceptance probability for rerandomization.

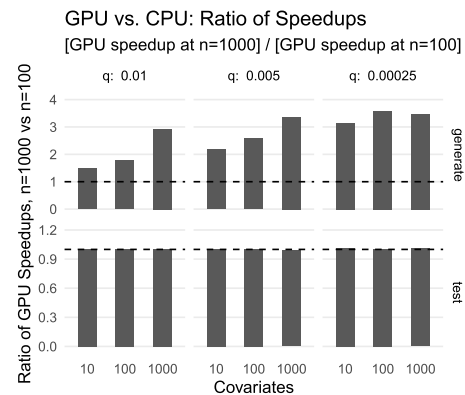


Fig. 7. GPU advantage grows with problem size. Bars plot the ratio of GPU (M4) speedups at $n=1000$ relative to $n=100$ (values > 1 indicate larger gains at $n=1000$), by draw count and step (pool generation vs. inference).

Compatibility and ecosystem. *fastrerandomize* complements established experimental design and inference packages [6–10] by focusing on the scalable construction of *balanced* assignment pools and on accelerated, design-respecting inference. For rerandomization-specific tasks such as power analysis and threshold tuning [11–14], *fastrerandomize* can serve as a downstream engine for candidate generation and testing at scale. The package also includes *fast_distance()* for accelerated pairwise distance calculations in applications beyond randomization, such as phylogenetics [19] and genomics [20].

5. Conclusions

fastrerandomize brings rerandomization into the realm of large-scale, high-dimensional experimentation by combining batched computation, hardware acceleration, and memory-efficient storage with design-based inference. Practitioners can now adopt stringent balance

thresholds—even with thousands of units and covariates—without prohibitive runtime or memory costs. Looking ahead, promising avenues include multi-device parallelism for candidate generation and richer balance criteria tailored to structured text, image, or network covariates [3–5].

CRedit authorship contribution statement

Rebecca Goldstein: Writing – review & editing, Writing – original draft, Methodology, Investigation. **Connor T. Jerzak:** Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Aniket Kamat:** Resources, Formal analysis, Data curation. **Fucheng Warren Zhu:** Writing – original draft, Software, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors are listed in alphabetical order. We thank Kaz Sakamoto for his feedback. The authors report no related grants or competing interests. See [fastrerandomize.github.io](https://github.com/fastrerandomize) for more. For replication data, see huggingface.co/datasets/cjerzak/rerandomization-benchmarks.

Appendix A. Derivations: from observed distance to target MSE

We herein justify the formal expressions found in the main text. Throughout, let $n_T + n_C = n$, and assume Bernoulli(π) or complete randomization with fixed n_T ; replace n_T, n_C by $n\pi, n(1-\pi)$ when a design-stage approximation is desired (i.e., treating $n_T \approx n\pi$ and $n_C \approx n(1-\pi)$). We do not claim novelty in these analyses, but include them here for clarity of exposition; for more information, see the foundational work of [1,2].

A. Conditional (realized) bias, variance, and MSE

Write $\bar{Z}_T = n_T^{-1} \sum_{i:T_i=1} Z_i$ and $\bar{Z}_C$ analogously. Under the linear model,

$$Y_i(t) = \beta^\top X_i + \tau t + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

the difference-in-means is

$$\hat{\tau}_{\text{DiM}} = \bar{Y}_T - \bar{Y}_C = \tau + \beta^\top \Delta_X + (\bar{\varepsilon}_T - \bar{\varepsilon}_C), \quad \Delta_X \equiv \bar{X}_T - \bar{X}_C.$$

Conditioning on (X, T) , the noise term, assumed to be exogenous, has mean 0 and variance $\sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C} \right)$. Consequently, we can write the following conditional error/offset term due to imbalance and a variance term due to sampling:

$$\text{Bias}(\hat{\tau}_{\text{DiM}} | X, T) = \beta^\top \Delta_X,$$

$$\text{Var}(\hat{\tau}_{\text{DiM}} | X, T) = \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C} \right),$$

$$\begin{aligned} \text{MSE}(\hat{\tau}_{\text{DiM}} | X, T) &= \text{Bias}(\hat{\tau}_{\text{DiM}} | X, T)^2 + \text{Var}(\hat{\tau}_{\text{DiM}} | X, T) \\ &= (\beta^\top \Delta_X)^2 + \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C} \right). \end{aligned}$$

Relating $(\beta^\top \Delta_X)^2$ to M . Now, how can we bound or approximate $(\beta^\top \Delta_X)^2$ using only the *realized* multivariate distance M ? Assume columns of X are standardized and pairwise independent (in practice, whiten X so that $\Sigma_X = I_d$). Then $M = \Delta_X^\top \Delta_X = \sum_j \text{SMD}_j^2$, recalling that SMD_j indicates the standardized mean difference (SMD) for column/covariate j .

Two consequences follow.

1. (*Bound*) Cauchy–Schwarz: $(\beta^\top \Delta_X)^2 \leq \|\beta\|_2^2 \|\Delta_X\|_2^2 = \sigma_{\text{Prog}}^2 M$, where $\|\Delta_X\|_2^2 = M$ by definition and since after standardization/whitening, it follows that $\text{Cov}(X_i) = I_d$, and therefore

$$\|\beta\|_2^2 = \beta^\top \beta = \beta^\top \text{Cov}(X_i) \beta = \text{Var}(\beta^\top X_i) \equiv \sigma_{\text{Prog}}^2.$$

The Cauchy-Schwarz bound represents the worst-case prognostic direction (β perfectly aligned with the realized imbalance vector Δ_X) and is therefore highly conservative in high dimensions.

The typical-orientation expression described next is the expected squared bias term when the prognostic direction β is fixed but the imbalance direction is effectively random, which is the quantity of primary relevance in high-dimensional settings.

2. (*Typical orientation*) Assuming Δ_X is approximately multivariate Normal with covariance $(1/n_T + 1/n_C)I_d$ (e.g., by a CLT argument), which makes it spherically symmetric, we see that, conditional on $\|\Delta_X\|_2^2 = M$, the direction $\Delta_X / \|\Delta_X\|_2$ is uniform on the unit sphere. Let $u \equiv \beta / \|\beta\|_2$. Then

$$\begin{aligned} (\beta^\top \Delta_X)^2 &= \left(\|\beta\|_2 (u^\top \Delta_X) \right)^2 \\ &= \|\beta\|_2^2 (u^\top \Delta_X)^2 \\ &= \|\beta\|_2^2 \|\Delta_X\|_2^2 \left(u^\top \frac{\Delta_X}{\|\Delta_X\|_2} \right)^2, \end{aligned}$$

so

$$\begin{aligned} \mathbb{E}[(\beta^\top \Delta_X)^2 | \|\Delta_X\|_2^2 = M] &= \|\beta\|_2^2 M \cdot \mathbb{E}\left[\left(u^\top \frac{\Delta_X}{\|\Delta_X\|_2}\right)^2 | \|\Delta_X\|_2^2 = M\right] \\ &= \|\beta\|_2^2 M \cdot \frac{1}{d} = \frac{\sigma_{\text{Prog}}^2}{d} M, \end{aligned}$$

where the first equality holds from:

The decomposition $\beta = \|\beta\|_2 u$ and $\Delta_X = \|\Delta_X\|_2 (\Delta_X / \|\Delta_X\|_2)$, so $(\beta^\top \Delta_X)^2 = \|\beta\|_2^2 \|\Delta_X\|_2^2 (u^\top (\Delta_X / \|\Delta_X\|_2))^2$; conditioning on $\|\Delta_X\|_2^2 = M$ fixes the factor $\|\Delta_X\|_2^2$ at M .

The second equality follows from the rotational symmetry of the sphere:

If V is uniform on the unit sphere and w is any fixed unit vector, then $\mathbb{E}[(w^\top V)^2] = 1/d$ [21]; the typical squared cosine between w and V is $1/d$; intuitively, a random direction in $\mathbb{R}^d$ spends about a $1/d$ fraction of its squared length in any fixed direction, so the average squared alignment with u is $1/d$.¹ Substitution into the conditional MSE yields the realized RMSE approximation.

Remark (general Σ_X). If $\Sigma_X \succ 0$, let $\tilde{X} = X \Sigma_X^{-1/2}$, $\tilde{\beta} = \Sigma_X^{1/2} \beta$, and $\tilde{\Delta} = \tilde{X}_T - \tilde{X}_C = \Sigma_X^{-1/2} \Delta_X$. Then $M = \Delta_X^\top \Sigma_X^{-1} \Delta_X = \|\tilde{\Delta}\|_2^2$ and $\sigma_{\text{Prog}}^2 = \|\tilde{\beta}\|_2^2$, allowing the application of results from the standardized/whitened case.

B. Ex-ante MSE under complete randomization

Under complete randomization with standardized covariates X (so that $\mathbb{E}[X_i] = 0$ and $\text{Cov}(X_i) = I_d$), the treated and control sample means are

$$\bar{X}_T \equiv \frac{1}{n_T} \sum_{i:T_i=1} X_i, \quad \bar{X}_C \equiv \frac{1}{n_C} \sum_{i:T_i=0} X_i,$$

and the covariate mean difference is $\Delta_X \equiv \bar{X}_T - \bar{X}_C$. By symmetry of the randomization,

$$\mathbb{E}[\bar{X}_T] = \mathbb{E}[\bar{X}_C] = 0 \quad \rightarrow \quad \mathbb{E}[\Delta_X] = 0.$$

¹ More specifically, if V is uniform on S^{d-1} then by symmetry $\mathbb{E}[VV^\top] = c I_d$ for some c , and taking traces gives $\mathbb{E}[\|V\|_2^2] (= 1) = \text{tr}(\mathbb{E}[VV^\top]) = cd$; solving for c yields $\mathbb{E}[VV^\top] = I_d/d$ and hence $\mathbb{E}[(w^\top V)^2] = w^\top \mathbb{E}[VV^\top] w = 1/d$ for any unit vector, w .

Treating units as independent draws from a super-population with $\text{Cov}(X_i) = I_d$, the variance of each sample mean is

$$\text{Var}(\bar{X}_T) = \frac{1}{n_T} I_d, \quad \text{Var}(\bar{X}_C) = \frac{1}{n_C} I_d,$$

and the treated and control groups are independent under the i.i.d. super-population model. Hence,

$$\text{Var}(\Delta_X) = \text{Var}(\bar{X}_T - \bar{X}_C) = \text{Var}(\bar{X}_T) + \text{Var}(\bar{X}_C) = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) I_d.$$

Now consider the scalar random variable $\beta^\top \Delta_X$. Using $\mathbb{E}[\Delta_X] = 0$,

$$\mathbb{E}[\beta^\top \Delta_X] = \beta^\top \mathbb{E}[\Delta_X] = 0,$$

so $\mathbb{E}[(\beta^\top \Delta_X)^2] = \text{Var}(\beta^\top \Delta_X)$. By the bilinearity of variance for linear forms,

$$\text{Var}(\beta^\top \Delta_X) = \beta^\top \text{Var}(\Delta_X) \beta = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \beta^\top I_d \beta = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \|\beta\|_2^2.$$

Recall that, under standardized X , the prognostic variance is

$$\sigma_{\text{Prog}}^2 \equiv \text{Var}(\beta^\top X_i) = \beta^\top \text{Cov}(X_i) \beta = \beta^\top I_d \beta = \|\beta\|_2^2,$$

so

$$\mathbb{E}[(\beta^\top \Delta_X)^2] = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \sigma_{\text{Prog}}^2.$$

From the conditional decomposition (above):

$$\text{MSE}(\hat{\tau}_{\text{DiM}} | X, T) = (\beta^\top \Delta_X)^2 + \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C}\right),$$

taking expectations over the randomization yields

$$\mathbb{E}[\text{MSE}(\hat{\tau}_{\text{DiM}})] = \mathbb{E}[(\beta^\top \Delta_X)^2] + \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C}\right) = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) (\sigma^2 + \sigma_{\text{Prog}}^2).$$

We might, however, want to consider how the MSE varies with the rerandomization acceptance rule, which we examine next.

C. Effect of a mahalanobis acceptance rule

Strategy. Our goal is to understand how a Mahalanobis acceptance rule,

$$\text{accept} \iff M = \Delta_X^\top \Delta_X \leq a,$$

alters the distribution of the treated–control mean difference Δ_X and therefore the ex-ante MSE of $\hat{\tau}_{\text{DiM}}$. Under the Gaussian (CLT) approximation with whitened covariates, Δ_X is approximately isotropic, so conditioning on $M \leq a$ is geometrically just truncating an isotropic Gaussian to the Euclidean ball of radius $\sqrt{a}$. The key observation is that the event $\{M \leq a\}$ depends only on the *radius* of Δ_X and not its *direction*; therefore, the conditional direction remains uniform on the sphere, and the conditional covariance must shrink by a scalar factor found below, yielding the shrunken MSE under rerandomization.

Details. Because under the CLT, $\Delta_X \sim \mathcal{N}\left(0, \left(\frac{1}{n_T} + \frac{1}{n_C}\right) I_d\right)$ is approximately spherically symmetric under the assumptions outlined above, we can write it in spherical coordinates as

$$\Delta_X = \rho U,$$

where

- $\rho \equiv \|\Delta_X\|_2 \geq 0$ is the random radius;
- $U \equiv \Delta_X / \|\Delta_X\|_2$ is the random direction, which is uniform on the unit sphere S^{d-1} .

Because the squared norm of a d –dimensional Gaussian vector has a χ_d^2 distribution,

$$\rho^2 = \|\Delta_X\|_2^2 \sim \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \chi_d^2, \quad (\text{A.1})$$

with ρ independent of the random direction, U . The acceptance event $\{M \leq a\}$ is equivalent to $\{\chi_d^2 \leq c\}$ with $c \equiv a / (\frac{1}{n_T} + \frac{1}{n_C})$. Equipped with this definition of acceptance, we can now evaluate the conditional expectation of the imbalance outer product, which we need because it allows us to compute the expected squared prognostic bias, $(\beta^\top \Delta_X)^2$, via the quadratic form $\beta^\top \mathbb{E}[\Delta_X \Delta_X^\top] \beta$. Using $\mathbb{E}[UU^\top] = I_d/d$ and the independence of ρ and U ,

$$\begin{aligned} \mathbb{E}[\Delta_X \Delta_X^\top | M \leq a] &= \mathbb{E}[\rho^2 U U^\top | \rho^2 \leq a] = \mathbb{E}[\rho^2 | \rho^2 \leq a] \cdot \mathbb{E}[U U^\top] \\ &= \mathbb{E}[\rho^2 | \rho^2 \leq a] \cdot I_d/d. \end{aligned}$$

Rerandomization shrinks the complete-randomization covariance $(\frac{1}{n_T} + \frac{1}{n_C}) I_d$ by a *single scalar* determined by the truncated radial moment $\mathbb{E}[\rho^2 | \rho^2 \leq a]$ (equivalently, by the truncated χ^2 moment). In particular, all directions shrink equally because the acceptance event depends only on the radius and does not privilege any coordinate direction.

To evaluate the scalar $\mathbb{E}[\rho^2 | \rho^2 \leq a]$, use the χ^2 representation in Eq. (A.1). Let $Z \sim \chi_d^2$ so that

$$\rho^2 = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) Z.$$

Because $\{M \leq a\} \iff \{\rho^2 \leq a\}$ and ρ^2 is a rescaling of Z , the acceptance event is equivalently $\{Z \leq c\}$ with

$$c \equiv \frac{a}{\frac{1}{n_T} + \frac{1}{n_C}}.$$

Therefore,

$$\mathbb{E}[\rho^2 | M \leq a] = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) \mathbb{E}[\chi_d^2 | \chi_d^2 \leq c].$$

It is convenient to summarize the effect of truncation by the dimension-normalized shrinkage factor

$$v_a(d) \equiv \frac{\mathbb{E}[\chi_d^2 | \chi_d^2 \leq c]}{d} \in (0, 1), \quad \text{so that} \quad \mathbb{E}[\chi_d^2 | \chi_d^2 \leq c] = d v_a(d).$$

A standard identity for truncated χ^2 moments (see footnote) gives the explicit form²

$$v_a(d) = \frac{\Pr(\chi_{d+2}^2 \leq c)}{\Pr(\chi_d^2 \leq c)}.$$

Substituting this into the previously derived expression $\mathbb{E}[\Delta_X \Delta_X^\top | M \leq a] = \mathbb{E}[\rho^2 | \rho^2 \leq a] \cdot I_d/d$ yields the isotropically shrunken second moment:

$$\mathbb{E}[\Delta_X \Delta_X^\top | M \leq a] = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) v_a(d) I_d.$$

Moreover, $\mathbb{E}[\Delta_X | M \leq a] = 0$ by symmetry (the event $\{M \leq a\}$ depends only on $\|\Delta_X\|_2$), so this is also the conditional covariance $\text{Var}(\Delta_X | M \leq a)$.

² Let $X \sim \chi_d^2$ with density $f_d(x) = \frac{1}{2^{d/2} \Gamma(d/2)} x^{d/2-1} e^{-x/2}$ for $x > 0$. Then

$$\mathbb{E}[X \mathbf{1}\{X \leq c\}] = \int_0^c x f_d(x) dx = \frac{1}{2^{d/2} \Gamma(d/2)} \int_0^c x^{d/2} e^{-x/2} dx.$$

Using $\Gamma(d/2 + 1) = (d/2) \Gamma(d/2)$, the right-hand side can be rewritten as

$$d \int_0^c f_{d+2}(x) dx = d \Pr(\chi_{d+2}^2 \leq c).$$

Dividing by $\Pr(\chi_d^2 \leq c)$ yields $\mathbb{E}[X | X \leq c] = d \Pr(\chi_{d+2}^2 \leq c) / \Pr(\chi_d^2 \leq c)$ and hence $v_a(d) = \Pr(\chi_{d+2}^2 \leq c) / \Pr(\chi_d^2 \leq c)$, as in [1].

Consequently, the expected squared prognostic bias among accepted assignments is

$$\mathbb{E}[(\beta^\top \Delta_X)^2 | M \leq a] = \beta^\top \mathbb{E}[\Delta_X \Delta_X^\top | M \leq a] \beta = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) v_a(d) \beta^\top \beta = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) v_a(d) \sigma_{\text{Prog}}^2,$$

where $\beta^\top \beta = \sigma_{\text{Prog}}^2$ under the standardized/whitened normalization. Adding the residual-noise variance term from Section A yields the ex-ante MSE conditional on acceptance:

$$\mathbb{E}[\text{MSE}(\hat{\tau}_{\text{DiM}}) | M \leq a] = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) (\sigma^2 + v_a(d) \sigma_{\text{Prog}}^2).$$

Finally, the expected imbalance distance among accepted assignments follows immediately by taking a trace (since $M = \Delta_X^\top \Delta_X = \text{tr}(\Delta_X \Delta_X^\top)$ and $\mathbb{E}[\text{tr}(A)] = \text{tr}(\mathbb{E}[A])$):

$$\mathbb{E}[M | M \leq a] = \mathbb{E}[\Delta_X^\top \Delta_X | M \leq a] = \text{tr} \mathbb{E}[\Delta_X \Delta_X^\top | M \leq a] = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) d v_a(d).$$

D. Design-stage inversion for choosing a threshold

For a target two-sided test with size α and power, $1 - \tilde{B}$, a Normal approximation³ implies $|\tau|/\text{RMSE} \gtrsim z_{1-\alpha/2} + z_{1-\tilde{B}}$, meaning that the signal-to-noise ratio $|\tau|/\text{RMSE}$ must be at least $z_{1-\alpha/2} + z_{1-\tilde{B}}$ to achieve size α and power $1 - \tilde{B}$ [22].⁴ Using the conditional-on-acceptance MSE,

$$\text{RMSE}^2(a) = \left(\frac{1}{n_T} + \frac{1}{n_C}\right) (\sigma^2 + v_a(d) \sigma_{\text{Prog}}^2) \leq \frac{\tau^2}{(z_{1-\alpha/2} + z_{1-\tilde{B}})^2},$$

where the MSE we are willing to tolerate after rerandomization cannot be larger than the square of the effect size divided by the usual Normal critical value sum for the desired power. Finally, one solves for the largest a (or, equivalently, acceptance probability q) such that $v_a(d)$ meets the inequality. Because $v_a(d)$ is strictly decreasing in stringency, a scalar line search suffices. In `fastrerandomize`, this calculation is wrapped by `diagnose_rerandomization()` to return a recommended `randomization_accept_prob` given user-specified contextual factors.

Appendix B. Sample code snippets

Installation and backend setup

```
# install.packages("devtools")
devtools::install_github("cjerzak/fastrerandomize-software/fastrerandomize")
library(fastrerandomize)

# Create/update a conda backend for JAX - Done once upon installation
# build_backend(conda_env = "fastrerandomize")
```

Generate assignments via Monte Carlo

```
set.seed(456)
n_units <- 200
n_treated <- 100
X <- matrix(rnorm(n_units * 50), nrow = n_units) # 50 covariates

rand_mc <- generate_randomizations(
  n_units = n_units,
  n_treated = n_treated,
  X = X,
  randomization_type = "monte_carlo",
  randomization_accept_prob = 0.01, # keep top 1%
  max_draws = 2e5,
  batch_size = 1e4,
  approximate_inv = TRUE
)
```

³ Assuming $\hat{\tau} \sim \mathcal{N}(\tau, \text{RMSE}_{\hat{\tau}}^2)$, the standardized statistic $Z = \frac{\hat{\tau} - \tau}{\text{RMSE}_{\hat{\tau}}} \sim \mathcal{N}(0, 1)$. Under $H_0 : \tau = 0$, $Z = \frac{\hat{\tau}}{\text{RMSE}_{\hat{\tau}}} \sim \mathcal{N}(0, 1)$. Under $H_1 : \tau \neq 0$, $\frac{\hat{\tau}}{\text{RMSE}_{\hat{\tau}}} = Z + \frac{\tau}{\text{RMSE}_{\hat{\tau}}} \sim \mathcal{N}\left(\frac{\tau}{\text{RMSE}_{\hat{\tau}}}, 1\right)$.

⁴ For a two-sided z-test that rejects when $|\hat{\tau}|/\text{RMSE} > z_{1-\alpha/2}$, the Normal approximation gives

$$\Pr(\text{reject} | \tau) = \Pr\left(|Z + \frac{\tau}{\text{RMSE}}| > z_{1-\alpha/2}\right), \quad Z \sim \mathcal{N}(0, 1).$$

A conservative sufficient condition for achieving power $1 - \tilde{B}$ is

$$\frac{|\tau|}{\text{RMSE}} \geq z_{1-\alpha/2} + z_{1-\tilde{B}}.$$

Design-respecting randomization test

```
obsW <- rand_mc$randomizations[1, ]
beta <- rnorm(ncol(X))
tau <- 1
obsY <- as.numeric(X %*% beta + tau * obsW + rnorm(n_units, 0, 0.5))

test_out <- randomization_test(
  obsW          = obsW,
  obsY          = obsY,
  candidate_randomizations = rand_mc$randomizations,
  findFI        = TRUE
)
```

References

- [1] Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. *Ann Stat* 2012;40(2):1263–82. <http://projecteuclid.org.ezp-prod1.hul.harvard.edu/euclid.aos/1342625468>.
- [2] Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment–control experiments. *Proc Natl Acad Sci* 2018;115(37):9157–62.
- [3] Keith K, Jensen D, O'Connor B. Text and causal inference: a review of using text to remove confounding from causal estimates. In: *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*; 2020. p. 5332–44.
- [4] Jerzak CT, Johansson F, Daoud A. Image-based treatment effect heterogeneity. In: *Proceedings of the second conference on causal learning and reasoning (CLEaR), Proceedings of machine learning Research (PMLR)*, vol. 213. 2023. p. 531–52.
- [5] Ogburn EL, Sofrygin O, Diaz I, Van der Laan MJ. Causal inference for social network data. *J Am Stat Assoc* 2024;119(545):597–611.
- [6] Bowers J, Fredrickson M, Hansen B. Rltools: randomization inference tools, R package version 0.3-3; 2023. <https://CRAN.R-project.org/package=Rltools>.
- [7] Coppock A. ri2: randomization inference for randomized experiments, R package version 0.4.1; 2025. <https://doi.org/10.32614/CRAN.package.ri2>. <https://CRAN.R-project.org/package=ri2>.
- [8] Coppock A. randomizr: easy-to-use tools for common forms of random assignment and sampling, R package version 1.0.0; 2024. <https://github.com/DeclareDesign/randomizr>.
- [9] Olivares-Gonzalez M, Sarmiento-Barbieri I. Randomization tests, R package version 0.1.4; 2017. <https://github.com/ignaciomsarmiento/RATest>.
- [10] Huang K, Jiang Z, Imai K. RCT2: designing and analyzing two-stage randomized experiments, R package version 0.0.1; 2022. <https://doi.org/10.32614/CRAN.package.RCT2>. <https://CRAN.R-project.org/package=RCT2>.
- [11] Branson Z, Li X, Ding P. Power and sample size calculations for rerandomization. *Biometrika* 2024;111(1):355–63.
- [12] Kapelner A, Krieger AM, Sklar M, Azriel D. Optimal rerandomization designs via a criterion that provides insurance against failed experiments. *J Stat Plan Inference* 2022;219:63–84.
- [13] McConeghy K. Jumble: an R package to perform stratified and Re-randomization procedures and assess covariate balance, R package for clinical trial randomization; 2024. <https://github.com/kmcconeghy/jumble>.
- [14] Banerjee A, Chassang S, Montero S, Snowberg E. A theory of experimenters. Working Paper 23867. National Bureau of Economic Research. Sep 2017. <https://doi.org/10.3386/w23867>. <http://www.nber.org/papers/w23867>.
- [15] Lehmann EL, Romano JP. Testing statistical hypotheses. 3rd ed. New York: Springer Texts in Statistics, Springer; 2005.
- [16] Leemis L. Mathematical statistics. Ascended Ideas; 2020.
- [17] Lu X, Ding P. Rerandomization for covariate balance mitigates p-hacking in regression adjustment. [arXiv preprint] [arXiv:2505.01137](https://arxiv.org/abs/2505.01137). 2025.
- [18] Xiao A, Xuan W, Wang J, Huang J, Tao D, Lu S, Yokoya N. Foundation models for remote sensing and earth observation: a survey. *IEEE Geosci Remote Sens Mag* 2025;13(4):297–324. <https://doi.org/10.1109/MGRS.2025.3576766>.
- [19] Elias I, Lagergren J. Fast computation of distance estimators. *BMC Bioinformatics* 2007;8(1):89.
- [20] Akhtyamov P, Nabi A, Gafurov V, Sizykh A, Favorov A, Medvedeva Y, Stupnikov A. GPU-accelerated kendall distance computation for large or sparse data. *GigaScience* 2024;13:giae088.
- [21] Vershynin R. High-dimensional probability: an introduction with applications in data science. Second ed. Cambridge University Press; 2025.
- [22] Rosner BA, et al. Fundamentals of biostatistics, vol. 6. Belmont, CA: Thomson-Brooks/Cole; 2006.