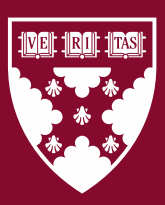
Working Paper 24-060

Anytime-Valid Inference in Linear Models and Regression-Adjusted Inference

Michael Lindon
Dae Woong Ham
Martin Tingley
Iavor Bojinov



Harvard Business School

Anytime-Valid Inference in Linear Models and Regression-Adjusted Inference

Michael Lindon
Netflix

Dae Woong Ham
Harvard University

Martin Tingley
Netflix

Iavor Bojinov
Harvard Business School

Working Paper 24-060

Copyright © 2024 by Michael Lindon, Dae Woong Ham, Martin Tingley, and Iavor Bojinov.

Working papers are in draft form. This working paper is distributed for purposes of comment and discussion only. It may not be reproduced without permission of the copyright holder. Copies of working papers are available from the author.

Funding for this research was provided in part by Harvard Business School.

Anytime-Valid Inference in Linear Models and Regression-Adjusted Causal Inference

Michael Lindon*, Dae Woong Ham†, Martin Tingley*, and Iavor Bojinov◇

February 8, 2024

arXiv:2210.08589v4 [stat.ME] 7 Feb 2024

Abstract

Linear regression adjustment is commonly used to analyse randomised controlled experiments due to its efficiency and robustness against model misspecification. Current testing and interval estimation procedures leverage the asymptotic distribution of such estimators to provide Type-I error and coverage guarantees that hold only at a single sample size. Here, we develop the theory for the anytime-valid analogues of such procedures, enabling linear regression adjustment in the sequential analysis of randomised experiments. We first provide sequential F-tests and confidence sequences for the parametric linear model, which provide time-uniform Type-I error and coverage guarantees that hold for all sample sizes. We then relax all linear model parametric assumptions in randomised designs and provide nonparametric model-free sequential tests and confidence sequences for treatment effects. This formally allows experiments to be continuously monitored for significance, stopped early, and safeguards against statistical malpractices in data collection. A particular feature of our results is their simplicity. Our test statistics and confidence sequences all emit closed-form expressions, which are functions of statistics directly available from a standard linear regression table. We illustrate our methodology with the sequential analysis of software A/B experiments at Netflix, performing regression adjustment with pre-treatment outcomes.

Keywords: e-processes, safe testing, martingales, Bayes factors, sequential testing, confidence sequences, anytime-valid

*Netflix, 121 Albright Way, Los Gatos, CA 90302. michael.s.lindon@gmail.com, mtingley@netflix.com

†Department of Statistics, Harvard, MA. daewoongham@g.harvard.edu

◇Harvard School of Business, Harvard, MA. ibojinov@hbs.edu

1 Introduction

Linear models are fundamental tools across many scientific disciplines. In causal inference, linear regression adjustment is widely utilized to analyse randomised trials, leveraging pre-treatment covariates to obtain treatment effect estimators of reduced variance (Lin 2013). Due to their robustness to model misspecification and efficiency, their use is widely recommended across many official research guidelines (Ye et al. 2022). However, tests and interval estimation are based on the asymptotic distribution of such estimators, which provide Type-I error and coverage guarantees at only a single fixed sample size (fixed-n). This can be a limitation at best and problematic for published research at worst.

A problem arises in connection with the replication crisis (Wasserstein & Lazar 2016, Benjamin et al. 2018) that pervades published research (Sagarin et al. 2014). It is estimated that only one-third of statistically significant results published in top psychology journals can be reproduced in follow-up experiments (Collaboration 2015). One recognized explanation of this phenomenon is an omission at the time of submission of how the data was collected (Ioannidis 2005) - whether a sample size was determined ahead of time or whether a data-dependent stopping rule was used, such as collecting data until a statistically significant result is obtained. In an anonymous survey of 2000 published researchers in psychology, approximately 60% of respondents admitted to collecting more data after seeing whether the results were significant, with approximately 20% admitting to stopping data collection after achieving a significant result (John et al. 2012).

A related risk arises in applications where observations arrive sequentially, instead of simultaneously, providing multiple opportunities to test hypotheses instead of a single opportunity. At large internet companies performing A/B tests, for example, outcomes are realized sequentially at extremely high rates. In signup or payment experiments, customers

produce Bernoulli outcomes whenever a registration or purchase occurs, which can happen many times per second. In these applications, it is critical to be able to detect large treatment effects *early*, while also being able to detect small treatment effects *eventually*. Herein lies the dichotomy for fixed-n hypothesis tests. A fixed-n test performed at a small n may detect large effects early, but is likely not powered to detect small effects. Conversely, a fixed-n test performed at a large n may be powered to detect small effects but risks large, potentially harmful, treatment effects remaining undetected for too long. Technology companies have a wealth of pre-treatment data that can be successfully leveraged for regression adjustment, almost always via a linear regression adjustment for its scalability (Deng et al. 2013). However, they currently lack the statistical methodology to perform this analysis sequentially over streaming data. Naively applying fixed-n tests repeatedly, without correction, invalidates the Type-I error guarantee and is guaranteed to reject the null eventually. Consequently, researchers intent on rejecting the null hypothesis can simply keep collecting new observations until this is so (Armitage et al. 1969, Armitage 1993).

Our contributions target this methodological gap by providing anytime-valid counterparts to fixed-n procedures. We begin parametrically by developing an anytime-valid theory for linear models, generalising Type-I error and coverage guarantees to hold at all sample sizes instead of only single sample sizes. Sequential F-tests provide these guarantees for subsets of regression coefficients, yielding sequential t-tests for single coefficients as a special case. These results generalise existing sequential t-tests in which there are no additional covariates (Hendriksen et al. 2021). The sequential F-tests are then inverted to provide confidence *sequences* for subsets of coefficients. A key feature of these results is their simplicity. The sequential tests and confidence sequences are closed-form expressions of the same statistics used in the fixed-n analysis, meaning they are no harder to implement

procedurally.

We then provide nonparametric asymptotic results by relaxing the linear model assumptions in randomised experiments. These provide asymptotic sequential tests and confidence sequences for treatment effects even when the model is misspecified. In particular, we relax the following four assumptions: linearity, Gaussianity of residuals, homoskedasticity of the residuals, and no omitted variable bias. We argue that these results are a direct anytime-valid analogy of the asymptotic fixed-n procedure. The confidence sequence is centered at the same estimate, and the width remains a multiple of the standard error. The only difference is that this multiple is now a function of the sample size instead of a constant quantile.

These results simplify migrating common inferential procedures in randomised trials from fixed-n guarantees to time-uniform anytime-valid guarantees. We illustrate this with an application from real-time software A/B testing at Netflix. Treatment users (hardware devices) are upgraded to a new software version, with control users running an older existing version. Continuous post-treatment performance outcomes are measured, and regression adjustment is performed with pre-treatment outcomes. The distribution of the stopping time without regression adjustment stochastically dominates the distribution with regression adjustment. On average, the sequential test with regression adjustment required only half as many samples to finish. Using bootstrap resampling from the empirical distribution of the control data, we demonstrate that the Type-I error remains below the nominally chosen α.

The paper is structured as follows. We begin in Section 2 with a review of the literature on regression adjustment, anytime-valid inference, and safe testing. Our main results are front-loaded toward the beginning of the paper for readers unconcerned with the proofs

and theory. Section 3 provides nonasymptotic anytime-valid inference for linear models. Section 3.2 introduces sequential F-tests and confidence sequences for collections of coefficients. Section 3.3 devotes special attention to the special case of sequential t-tests. Section 3.4.1 provides a complementary group-theoretic construction of our test statistic in terms of maximal invariant test statistics. Section 3.4 describes how to choose the tuning parameter of our procedure and provides the optimal choice under both frequentist and Bayesian alternatives. Section 4 describes the expected stopping time of our anytime-valid procedure in comparison to fixed-n designs, accompanied by simulations. Section 5 studies the asymptotic behaviour of our test statistic. This section is necessary for us to compare nonasymptotic parametric confidence sequences for the linear model and asymptotic nonparametric confidence sequences for treatment effects using regression adjustment in randomised experiments. Section 6 presents the asymptotic nonparametric results for regression-adjusted anytime-valid inference of treatment effects in randomised experiments, followed by a comparison of the parametric results of the preceding section. Section 7 provides a simulation under a nonlinear model to demonstrate that our asymptotic results are well calibrated at the nominal α-level. Section 8 applies these methods to the sequential analysis of software A/B testing. The paper concludes with a discussion in section 9. Sections 3 and 6 require technical proofs that leverage group theory and strong approximations. We have deferred almost all of the details to the appendix to make the exposition of our results clearer. For the reader interested in how these test martingales and approximations are constructed, see appendix sections A.4 and A.11.

2 Related Literature

In causal inference, particularly in randomised experiments, linear regression has become a staple methodology for performing inference on average treatment effects. The primary rationale is that the effect of the treatment can be better isolated by regressing on pre-treatment covariates that are correlated with the response, leading to estimators of reduced variance. Freedman (2008) showed that the regression-adjusted difference in means estimator obtained from a linear model with pre-treatment covariates and a main effect, in the case of equal assignment probabilities to treatment and control groups, improves or does not hurt asymptotic precision relative to a simple difference in means estimator. Moreover, the standard error from the linear model, in this case, is consistent or asymptotically conservative. Relaxing the requirement of equal assignment probabilities, Lin (2013) demonstrates that asymptotic efficiency is no worse when a complete set of treatment-covariate interactions is included and that the Huber-White (White 1980) sandwich standard error is consistent or asymptotically conservative. Reluga et al. (2022) provides the variance ordering of difference-in-means and regression-adjusted difference-in-means estimators with and without interaction terms in cases where the means of covariates are known and unknown. Critically, these results do not assume that the linear model is correctly specified. The randomization in the assignment mechanism protects against omitted variable bias, non-linear functional dependence, and non-Gaussian residuals (Imbens & Rubin 2015). This robustness to model misspecification in combination with efficiency gains has led to these estimators becoming a gold standard in randomised trials (Ye et al. 2022), appearing in many official research guidelines (U.S. Food and Drug Administration 2023, European Medicines Agency 2015, WWC 2022, Higgins et al. 2019).

Several alternative approaches have been proposed to address the shortcomings of clas-

sical fixed-n hypothesis tests. Grünwald et al. (2021) proposed "safe" hypothesis tests constructed by e-process which preserve Type-I error guarantees under optional stopping (Turner et al. 2021, Pérez-Ortiz et al. 2022, Turner & Grünwald 2023, Hao et al. 2023). Simultaneously, a body of literature on "anytime-valid" inference has evolved, using novel martingale techniques (Ramdas et al. 2020) to provide time-uniform (for all n) Type-I error and confidence guarantees (Howard et al. 2020, 2021, Waudby-Smith & Ramdas 2020, Waudby-Smith et al. 2021, Howard & Ramdas 2022, Lindon & Malek 2022, ter Schure et al. 2020, Wang & Ramdas 2023, Bibaut et al. 2022). Both kinds of literature are closely related (Ramdas et al. 2023), allowing experiments to be continuously monitored and hypotheses to be tested sequentially. Our construction is based on mixture-martingales obtained using the right-Haar mixture (Wijsman 1990, Eaton 1989). This exploits group invariance properties of the linear model to provide frequentist Type-I error and coverage guarantees that hold not just for all n but also for all possible values of the nuisance parameters (de Heide & Grünwald 2021, Pérez-Ortiz et al. 2022).

Sequential analysis is a natural solution for technology companies performing A/B tests over streams of outcomes, given the observation that the malpractice of "peeking" (repeatedly applying fixed-n hypothesis tests to stop experiments early) is rife within the industry (Johari et al. 2021, 2017). Recently, these methods have seen growing adoption from the technology sector. Lindon et al. (2022) use anytime-valid methods to continuously monitor software performance in A/B tests. Ham et al. (2022) provide a design-based approach to anytime-valid causal inference, with applications toward risk mitigation in the experimentation of digital products such as signup-funnel experiments. Lindon & Malek (2022) provide anytime-valid tests for time-inhomogeneous Poisson counting processes, applied to continuously monitor error rates produced by bugs in the software rollout process. See also

the integration of anytime-valid testing procedures in experimentation software products (Waudby-Smith et al. 2022, Services 2023, Maharaj et al. 2023, Pekelis et al. 2015, Beasley 2023, Eppo 2023).

Our results are relevant to digital experimentation in the technology sector for the following reasons. Firstly, technology companies have a wealth of pre-treatment data, which is often leveraged for covariate adjustment. Secondly, linear regression adjustment remains dominant across the industry because it strikes a good balance between variance reduction and computational burden (Deng et al. 2013). Current methods are, however, limited to fixed-n analyses. Our results enable such approaches to be seamlessly migrated to sequential analyses. Moreover, linear regression lends itself to sequential analysis because the model can be updated iteratively and online as new data arrives, which is crucial for streaming data.

3 Anytime-Valid Inference for Linear Models

3.1 Linear Models and Notation

Consider a sequence of i.i.d. observations

$$y_i|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2 \sim N(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta}, \sigma^2), \tag{1}$$

with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\delta} \in \mathbb{R}^d$. The parameterization in (1) splits the parameters into those of interest ($\boldsymbol{\delta}$), and those that are nuisance ($\boldsymbol{\beta}, \sigma^2$), which includes a variety of interesting models. For example, $\boldsymbol{\delta}$ could denote main effects terms, while pre-treatment covariates and their centered interactions with treatments can be absorbed by $\boldsymbol{\beta}$. Alternatively, $\boldsymbol{\delta}$ could denote the covariate-treatment interaction terms, absorbing pre-treatment covariates and main effects into $\boldsymbol{\beta}$, enabling tests for treatment effect heterogeneity. Let the value of

$\boldsymbol{\delta}$ specified by the null hypothesis be denoted by $\boldsymbol{\delta}_0$. Throughout, we take $\boldsymbol{\delta}_0 = \mathbf{0}_d$ as this is the most common null encountered in practice and simplifies our expressions. We can do this without losing generality, as the test of $\boldsymbol{\delta}_0 \neq \mathbf{0}_d$ can always be recovered by replacing y_i with $y_i - \boldsymbol{z}_i'\boldsymbol{\delta}_0$.

We will express equation (1) in matrix notation as $\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{Z}_n\boldsymbol{\delta} + \boldsymbol{\varepsilon}_n$ where $\boldsymbol{Y}_n = (y_1, y_2, \ldots, y_n)$, $\boldsymbol{X}_n$ is the matrix with rows $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{Z}_n$ is the matrix with rows $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)$. Let $\boldsymbol{W}_n = [\boldsymbol{X}_n, \boldsymbol{Z}_n]$ and $\boldsymbol{\gamma}' = [\boldsymbol{\beta}', \boldsymbol{\delta}']$ so that $\boldsymbol{Y}_n|\boldsymbol{\gamma}, \sigma^2 \sim N(\boldsymbol{W}_n\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}_n)$. Let $\mathcal{C}(\boldsymbol{A}) = \{\sum_i c_i\boldsymbol{A}_i : c_i \in \mathbb{R}, \boldsymbol{A}_i = \boldsymbol{A}\boldsymbol{e}_i\}$ denote the column space of a given matrix $\boldsymbol{A}$, $\mathcal{C}(\boldsymbol{A})^\perp$ the orthogonal complement of $\mathcal{C}(\boldsymbol{A})$, $P_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ denote the orthogonal projection operator onto $\mathcal{C}(\boldsymbol{A})$, $r(\boldsymbol{A})$ the rank of $\boldsymbol{A}$ and $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 = \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}$. Let $s^2(\boldsymbol{Y}) = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}/\nu_n$ where $\nu_n = n - p - d$ denote the usual unbiased estimator of σ^2. Let $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n = \boldsymbol{Z}_n'(\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n = \boldsymbol{Z}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n$. Note that $(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}$ is the submatrix of $(\boldsymbol{W}_n\boldsymbol{W}_n)^{-1}$ corresponding to $\boldsymbol{\delta}$ (see appendix section A.1). Let $F(d_1, d_2, \lambda)$ denote the non-central F-distribution with degrees of freedom d_1 and d_2 with non-centrality parameter λ. Similarly, let $\chi^2(d, \lambda)$ denote a noncentral chi-squared distribution with d degrees of freedom and noncentrality parameter λ. Let $\hat{\theta}_n(\boldsymbol{Y}_n)$ denote the maximum likelihood estimator of a parameter θ after n observations. We use a zero subscript to denote null hypothesis parameter values. Let $\boldsymbol{\Theta}$ denote the parameter space of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2)$ and $\boldsymbol{\Theta}_0$ denote the subset restricted to the null hypothesis.

3.2 Sequential F-Tests

We seek to test the null hypothesis after every new observation while maintaining a Type-I error guarantee of $\boldsymbol{\alpha}$. Mathematically, we seek a sequential p-value $p_n(\boldsymbol{Y}_n)$ such that $\mathbb{P}_{\boldsymbol{\theta}}\left[\inf_{n\in\mathbb{N}} p_n(\boldsymbol{Y}_n) \leq \alpha\right] \leq \alpha$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Equivalently, we seek to cover $\boldsymbol{\delta}$ by a sequence of

sets $C_n(\boldsymbol{Y}_n; \alpha)$, such that the probability $\boldsymbol{\delta}$ is covered by *all* sets is at least $1-\alpha$, that is, $\mathbb{P}_{\boldsymbol{\theta}}\left[\boldsymbol{\delta} \in \bigcap_{n=1}^{\infty} C_n(\boldsymbol{Y}_n; \alpha)\right] \geq 1-\alpha$.

Sequential tests for composite alternatives are often constructed by taking mixtures over alternatives (Wald 1945, Robbins & Siegmund 1970, de la Peña et al. 2004). When the null hypothesis is also composite, special care is required to ensure that Type-I error probabilities hold for all possible values of the nuisance parameters. In models possessing a group structure, invariance arguments can be leveraged to simplify a composite null to a simple null (Lehmann & Romano 2005, Eaton 1989, Wijsman 1990). Consider an i.i.d. sequence of observations $y_i \sim N(\boldsymbol{w}_i\boldsymbol{\gamma}, \sigma^2)$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$ and $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\delta})$. We define our test martingale as follows.

Definition 3.1. *Group Invariant Mixture-SPRT/Bayes-Factor*

$$B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi}) = \frac{\int \Pi_{i=1}^n \frac{1}{\sigma}\psi(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{\delta})\frac{1}{\sigma}dF^{\boldsymbol{\Phi}}(\boldsymbol{\delta})d\boldsymbol{\beta}d\sigma}{\int \Pi_{i=1}^n \frac{1}{\sigma}\psi(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\frac{1}{\sigma}d\boldsymbol{\beta}d\sigma} \tag{2}$$

when $\boldsymbol{W}_n'\boldsymbol{W}_n$ is full rank, otherwise $B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi}^{-1}) = 1$, where ψ is a standard Gaussian density and $F^{\boldsymbol{\Phi}}$ is a $N(\boldsymbol{0}_d, \sigma^2\boldsymbol{\Phi}^{-1})$ probability measure.

The statistic $B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi}^{-1})$ can be interpreted as an invariant mixture sequential probability ratio test statistic or a Bayes factor resulting from a $N(\boldsymbol{0}, \sigma^2\boldsymbol{\Phi}^{-1})$ prior on $\boldsymbol{\beta}$ and the right-Haar prior on the nuisance parameters ($p(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$). The invariant right-Haar measure leverages the group invariance properties of the linear model and is motivated by the requirement that the time-uniform Type-I error guarantee holds for all possible values of the nuisance parameters (Hendriksen et al. 2021, Pérez-Ortiz et al. 2022). We discuss this in detail in section A.4. The choice of $\boldsymbol{\Phi}$ will be discussed in sections 3.4 and 5. Until then we will simply write $B_n(\boldsymbol{Y}_n)$ instead of $B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi})$ for brevity. We do not, however, make Bayesian assumptions in this work. Instead, we examine the behaviour of $B_n(\boldsymbol{Y}_n)$

from a frequentist perspective, that is, assuming there exists some fixed "true" unknown values of the parameters. To be explicit, we assume that the distribution of $\boldsymbol{Y}_n$ under the null hypothesis is $N(\boldsymbol{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ for some value of $\boldsymbol{\beta}$ and σ. The statistic $B_n(\boldsymbol{Y}_n)$ has the following closed-form expression, the proof of which is given in Appendix A.2.

Proposition 3.2. *Under our setup,*

$$B_n(\boldsymbol{Y}_n) = \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}} \frac{\left(1 + \frac{\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)'(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n - \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)}{s_n^2(\boldsymbol{Y}_n)}\right)^{-\frac{\nu_n+d}{2}}}{\left(1 + \frac{\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)'\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)}{s_n^2(\boldsymbol{Y}_n)}\right)^{-\frac{\nu_n+d}{2}}} \tag{3}$$

where $\nu_n = n - p - d$ is the degrees of freedom, $\hat{\delta}_n(\boldsymbol{Y}_n)$ is the ordinary least squares (OLS) estimator of δ, $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n = \boldsymbol{Z}_n'(\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n = \boldsymbol{Z}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n$ and $s_n^2(\boldsymbol{Y}_n) = \boldsymbol{Y}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_n})\boldsymbol{Y}_n/\nu_n$.

The expression in equation (3) for $B_n(\boldsymbol{Y}_n)$ may appear complex at first; however, it requires only the same three statistics as a classical f-test[1]. Recall in the fixed-n case

$$\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta} \sim N(0, \sigma^2(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}),$$

$$\boldsymbol{L}^{-1}(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta})/\sigma \sim N(\boldsymbol{0}, \boldsymbol{I}_d),$$

$$\boldsymbol{L}^{-1}(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta})/s_n(\boldsymbol{Y}_b) \sim t_{\nu_n}(\boldsymbol{0}, \boldsymbol{I}_d),$$

$$(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta})'\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta})/ds_n^2(\boldsymbol{Y}_n) \sim F(d, \nu_n),$$

where $\boldsymbol{L}\boldsymbol{L}' = \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n$. As this illustrates, our new statistic is primarily a function of the fixed-n statistics n, $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n$, $\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)$ and $s_n^2(\boldsymbol{Y}_n)$, making it simple to compute. The only new term is $\boldsymbol{\Phi}$, which is a tuning parameter discussed in Section 3.4.

The statistic $B_n(\boldsymbol{Y}_n)$ is a nonnegative supermartingale under the null hypothesis, which in combination with Ville's inequality (Ville 1939) provides the following time-uniform bound.

[1]For a review of classical f-tests, see Appendix A

Theorem 3.3. *Consider an i.i.d. sequence of observations* $y_i \sim N(\boldsymbol{w}_i\boldsymbol{\gamma}, \sigma^2)$ *where* $\boldsymbol{w}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$ *and* $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \delta)$. *For* $B_n(\boldsymbol{Y}_n)$ *in definition 3.1 with any symmetric positive definite matrix* $\boldsymbol{\Phi}$ *then*

$$\mathbb{P}_\theta[\sup_n B_n(\boldsymbol{Y}_n) \geq \alpha^{-1}] \leq \alpha, \tag{4}$$

for all $\theta \in \Theta_0$.

We defer the technical details to Appendix A.4. A sequential p-value can be obtained by simply taking the reciprocal of $B_n(\boldsymbol{Y}_n)$.

Corollary 3.4. $p_n(\boldsymbol{Y}_n) = 1/B_n(\boldsymbol{Y}_n)$ *defines a sequential p-value satisfying*

$$\mathbb{P}_\theta[\inf_n p_n(\boldsymbol{Y}_n) \leq \alpha] \leq \alpha,$$

for all $\theta \in \boldsymbol{\Theta}_0$.

Testing a specific null $\boldsymbol{\delta}_0 \neq \mathbf{0}_d$ can be obtained by replacing $\boldsymbol{Y}_n$ with $\boldsymbol{Y}_n - \boldsymbol{Z}_n\boldsymbol{\delta}_0$. Define $B_n(\boldsymbol{Y}_n; \boldsymbol{\delta}_0) = B_n(\boldsymbol{Y}_n - \boldsymbol{Z}_n\boldsymbol{\delta}_0)$ and $p_n(\boldsymbol{Y}_n; \boldsymbol{\delta}_0) = p_n(\boldsymbol{Y}_n - \boldsymbol{Z}_n\boldsymbol{\delta}_0)$, noting $\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n - \boldsymbol{Z}_n\boldsymbol{\delta}_0) = \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0$ and $s_n^2(\boldsymbol{Y}_n - \boldsymbol{Z}_n\boldsymbol{\delta}_0) = s_n^2(\boldsymbol{Y}_n)$. The duality between p-values and confidence sets can be exploited to construct a confidence sequence for $\boldsymbol{\delta}$ by finding all the null hypotheses that would not have been rejected.

Corollary 3.5. *Let* $C_n(\boldsymbol{Y}_n; \alpha) = \{\boldsymbol{\delta} \in \mathbb{R}^d : B_n(\boldsymbol{Y}_n; \boldsymbol{\delta}) \leq \alpha^{-1}\} = \{\boldsymbol{\delta} \in \mathbb{R}^d : p_n(\boldsymbol{Y}_n; \boldsymbol{\delta}) > \alpha\}$.

$$C_n(\boldsymbol{Y}_n; \alpha) = \left\{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{A}_n} \leq \nu_n s_n^2(\boldsymbol{Y}_n) \left(1 - \left(\frac{\alpha^2 \det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}\right)^{\frac{1}{\nu_n + d}}\right)\right\} \tag{5}$$

where

$$\boldsymbol{A}_n = \left(\left(\frac{\alpha^2 \det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}\right)^{\frac{1}{\nu_n + d}} - 1\right) \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n.$$

Then $C_n(\boldsymbol{Y}_n; \alpha)$ *defines a confidence sequence satisfying* $\mathbb{P}_\theta[\boldsymbol{\delta} \in \bigcap_n C_n(\boldsymbol{Y}_n; \alpha)] \geq 1 - \alpha$ *for all* $\theta \in \boldsymbol{\Theta}$.

The proof is in Appendix A.5. Once again, we only need the three statistics ($\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)$, $s_n^2(\boldsymbol{Y}_n)$ and $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n$) from the data to compute these confidence sets, which are all directly available from standard linear regression outputs.

Remark 3.6. *Confidence intervals for $g(\boldsymbol{\delta})$, g convex, can be obtained through convex programming by maximizing and minimizing $g(\boldsymbol{\delta})$ subject to the quadratic constraint $\boldsymbol{\delta} \in C_n(\boldsymbol{Y}_n; \alpha)$. A hypothesis H_i, associated with a subset of the parameter space $\boldsymbol{\Theta}_i$, can be rejected at the α level as soon as $C(\boldsymbol{Y}_n; \alpha) \cap \Theta_i = \emptyset$.*

3.3 Sequential t-Tests

When $d = 1$ our results can be considerably simplified. For instance, 3.2 can be rewritten as follows.

Proposition 3.7.

$$B_n(\boldsymbol{Y}_n) = \sqrt{\frac{\phi}{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}} \frac{\left(1 + \frac{\phi}{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2} \frac{t_n(\boldsymbol{Y}_n)^2}{\nu_n}\right)^{-\frac{\nu_n+1}{2}}}{\left(1 + \frac{t_n(\boldsymbol{Y}_n)^2}{\nu_n}\right)^{-\frac{\nu_n+1}{2}}} \tag{6}$$

where $t_n(\boldsymbol{Y}_n) = \hat{\delta}_n(\boldsymbol{Y}_n)/se(\hat{\delta}(\boldsymbol{Y}_n))$ is the classical t statistic, $\nu_n = n - p - 1$ is the degrees of freedom, $\hat{\delta}_n(\boldsymbol{Y}_n)$ is the OLS estimator of δ, $se(\hat{\delta}(\boldsymbol{Y}_n)) = \sqrt{s_n^2(\boldsymbol{Y}_n)/\|\tilde{\boldsymbol{Z}}_n\|_2^2}$ is the standard error, $\|\tilde{\boldsymbol{Z}}_n\|_2^2 = \boldsymbol{Z}_n'(\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n$ and $s_n^2(\boldsymbol{Y}_n) = \boldsymbol{Y}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_n})\boldsymbol{Y}_n/\nu_n$ the usual unbiased estimator of σ^2.

We use ϕ instead of $\boldsymbol{\Phi}$ to stress that this parameter is now a scalar and no longer a positive definite symmetric matrix. The term $\|\tilde{\boldsymbol{Z}}_n\|_2^2$ is simply the reciprocal of the lower right element of the $(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}$ matrix. The elliptical confidence sets of corollary 3.5 can now be further simplified to intervals.

Corollary 3.8. *Let* $C_n(\boldsymbol{Y}_n;\alpha) = (\hat{\delta}_n(\boldsymbol{Y}_n) - r_n(\boldsymbol{Y}_n), \hat{\delta}_n(\boldsymbol{Y}_n) + r_n(\boldsymbol{Y}_n))$*, where*

$$r_n(\boldsymbol{Y}_n) = \frac{s_n(\boldsymbol{Y}_n)}{\|\tilde{\boldsymbol{Z}}_n\|_2}\sqrt{a_n(\alpha, \|\tilde{\boldsymbol{Z}}_n\|_2^2, \phi)}$$

$$a_n(\alpha, \|\tilde{\boldsymbol{Z}}_n\|_2^2, \phi) = \nu_n \left(\frac{1 - \left(\frac{\phi}{\phi+\|\tilde{\boldsymbol{Z}}_n\|_2^2}\alpha^2\right)^{\frac{1}{\nu_n+1}}}{0 \vee \left(\left(\frac{\phi}{\phi+\|\tilde{\boldsymbol{Z}}_n\|_2^2}\alpha^2\right)^{\frac{1}{\nu_n+1}} - \frac{\phi}{\phi+\|\tilde{\boldsymbol{Z}}_n\|_2^2}\right)} \right). \tag{7}$$

$C_n(\boldsymbol{Y}_n;\alpha)$ *then defines a confidence sequence satisfying* $\mathbb{P}_\theta[\delta \in \bigcap_n C_n(\boldsymbol{Y}_n;\alpha)] \geq 1-\alpha$*, for all* $\theta \in \boldsymbol{\Theta}$*.*

The proof is in the appendix section A.5. Again, these expressions are no more difficult to evaluate than their fixed-n counterparts. In section B we provide R code to demonstrate how the fixed-n linear model analysis can be converted to anytime-valid in a few lines.

3.4 Choosing $\boldsymbol{\Phi}$

To prove that the time uniform bound in theorem 3.3 holds for all possible values of the nuisance parameters, and also for guidance on how to choose the parameter $\boldsymbol{\Phi}$, it is convenient to work with the following alternative representation of the statistic $B_n(\boldsymbol{Y}_n)$.

3.4.1 Maximal Invariant Representation

The projection matrix $\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n}$ has d eigenvalues equal to one and p eigenvalues equal to zero. The eigendecomposition admits a matrix square root of the form $\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n} = \boldsymbol{V}_n\boldsymbol{V}_n'$, where $\boldsymbol{V}_n'$ is of dimension $d \times n$, allowing us to define $\tilde{\boldsymbol{Z}}_n = \boldsymbol{V}_n'\boldsymbol{Z}_n$, $\boldsymbol{t}_n(\boldsymbol{Y}_n) = \boldsymbol{V}_n'\boldsymbol{Y}_n/s_n(\boldsymbol{Y}_n) = \tilde{\boldsymbol{Z}}_n\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)/s_n(\boldsymbol{Y}_n)$. Note the connection to the classical F statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)/d = f(\boldsymbol{Y}_n)$. The statistic $B_n(\boldsymbol{Y}_n)$ in proposition 3.2 can now be expressed as

$$B_n(\boldsymbol{Y}_n) = \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}} \frac{\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n}\right)^{-\frac{\nu_n+d}{2}}}{\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n}\right)^{-\frac{\nu_n+d}{2}}},$$

where that the data enter only through $t_n(\boldsymbol{Y}_n)$. The statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ is a maximal invariant statistic under the group transformations that leave the linear model invariant. It is helpful to reparameterize the linear model in terms of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\xi})$, where $\boldsymbol{\xi} = \boldsymbol{\delta}/\sigma$ are standardized coefficients. In this parameterization $\boldsymbol{Y}_n|\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma \sim N(\boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{Z}_n\sigma\boldsymbol{\xi}, \sigma^2\boldsymbol{I}_n)$.

Consider a Bayesian alternative in which $\boldsymbol{\xi}$ is random with $\boldsymbol{\xi}|H_1 \sim N(\boldsymbol{0}_d, \boldsymbol{\Phi})$. The distribution of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ under the frequentist alternative, Bayesian alternative, and null distributions are as follows.

Proposition 3.9.

$$\begin{aligned} \boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{\xi}, H_1 &\sim t^{nc}_{\nu_n}(\tilde{\boldsymbol{Z}}_n\boldsymbol{\xi}, \boldsymbol{I}_d) \\ \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1 &\sim t_{\nu_n}(\boldsymbol{0}_d, \boldsymbol{I}_d + \tilde{\boldsymbol{Z}}'_n\boldsymbol{\Phi}^{-1}\tilde{\boldsymbol{Z}}_n) \\ \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0 &\sim t_{\nu_n}(\boldsymbol{0}_d, \boldsymbol{I}_d). \end{aligned} \tag{8}$$

The first two distributions are the frequentist and Bayesian alternatives. The last distribution $\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0$ is the distribution under the null, about which both Bayesians and frequentists agree. In all cases, the distribution of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$, and consequently $B_n(\boldsymbol{Y}_n)$, is free of the nuisance parameters $(\boldsymbol{\beta}, \sigma)$. This is a specific example of a general Thoerem from Dass & Berger (2003, Theorem 1), which states that when using the right-Haar prior the distribution of the Bayes factor does not depend on the nuisance parameters under both H_1 and H_0. The following proposition states that the Bayes factor in equation (3) is not merely a function of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$, it is in fact equal to the Bayes factor obtained by conditioning on $\boldsymbol{t}_n(\boldsymbol{Y}_n)$, the proof is in the Appendix.

Proposition 3.10.

$$B_n(\boldsymbol{Y}_n) = \frac{p(\boldsymbol{Y}_n|H_1)}{p(\boldsymbol{Y}_n|H_0)} = \frac{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} = \frac{\int p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{\xi}, H_1)p(\boldsymbol{\xi}|H_1)d\boldsymbol{\xi}}{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} \tag{9}$$

where $p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{\xi}, H_1)$ *denotes a* $t^{nc}_{\nu_n}(\tilde{\boldsymbol{Z}}_n\boldsymbol{\xi}, \boldsymbol{I}_d)$ *density,* $p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)$ *denotes a* $t_{\nu_n}(\boldsymbol{0}_d, \boldsymbol{I}_d +$

$\tilde{\boldsymbol{Z}}_n'\boldsymbol{\Phi}^{-1}\tilde{\boldsymbol{Z}}_n)$ *density,* $p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)$ *denotes a* $t_{\nu_n}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ *density, and* $p(\boldsymbol{\xi}|H_1)$ *denotes a* $N(\boldsymbol{0}_d, \boldsymbol{\Phi}^{-1})$ *density.*

The Proposition states that the Bayes factor obtained from observing $\boldsymbol{Y}_n$ equals the Bayes factor obtained from observing $\boldsymbol{t}_n(\boldsymbol{Y}_n)$. This is a specific example of Berger et al. (1998, Theorem 2.1), which leverages the Wijsman representation theorem (Wijsman 1990) to equate the Bayes factor based on the raw observations to the Bayes factor based on the maximal invariant statistic when using the right-Haar prior. As the distribution of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ is free of the nuisance parameters under the null hypothesis, we can conclude that $\mathbb{E}_\theta[B_n(\boldsymbol{Y}_n)] = 1$ for all $\theta \in \Theta_0$, which combined with Markov's inequality gives $\mathbb{P}_\theta[B_n(\boldsymbol{Y}_n) \geq \alpha^{-1}] \leq \alpha$ for all $\theta \in \Theta_0$. This already provides a fixed-n Type-I error guarantee, but also provides a stronger time-uniform guarantee that we prove in Appendix A.4.2.

3.4.2 Selecting $\boldsymbol{\Phi}$ to maximize $\mathbb{E}[\log B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi})]$ under the alternative

As shown in definition 3.1, it is clear that the practitioner must pre-specify a value of $\boldsymbol{\Phi}$ to compute $B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi})$. We seek a choice of $\boldsymbol{\Phi}$ that provides the largest value of $\mathbb{E}[\log B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi})]$ under the alternative. We begin with the one-dimensional case with $\boldsymbol{\Phi} = \phi$ and use the shorthand notation $t_n = t_n(\boldsymbol{Y}_n)$.

Proposition 3.11. *For large* n, $\mathbb{E}[\log B_n(\boldsymbol{Y}_n)]$ *is approximately maximized with the choice*

$$\phi_{freq}^{opt} = \xi^{-2}.$$

To see this note that

$$\mathbb{E}[\log B_n(\boldsymbol{Y}_n; \phi)] = \int p(t_n|\xi) \log p(t_n|H_1) dt_n - \int p(t_n|\xi) \log p(t_n|H_0) dt_n,$$

where $p(t_n|\xi)$ is a noncentral $t_{\nu_n}^{nc}(\|\tilde{\boldsymbol{Z}}_n\|_2 \xi)$ density, $p(t_n|H_1)$ is a location scale $t_{\nu_n}(0, 1 + \|\tilde{\boldsymbol{Z}}_n\|_2/\phi)$ and $p(t_n|H_0)$ is a standard t_{ν_n}. The expected value of $\log B_n(\boldsymbol{Y}_n; \phi)$ is, therefore,

the difference between two cross entropies; namely, the cross entropy of the true and null minus the cross entropy of the true and alternative distributions. We are unaware of a closed-form expression for the cross entropy between a noncentral and a location-scale t. Instead, when ν_n is large, we can approximate these distributions as $N(\|\tilde{\boldsymbol{Z}}_n\|_2\xi, 1)$, $N(0, 1+\|\tilde{\boldsymbol{Z}}_n\|_2/\phi)$, and $N(0,1)$, respectively. The expected value can then be approximated by

$$\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi)] \approx \frac{1}{2}\log\left(\frac{\phi}{\phi+\|\tilde{\boldsymbol{Z}}_n\|_2^2}\right) + \left(1-\frac{\phi}{\phi+\|\tilde{\boldsymbol{Z}}_n\|_2^2}\right)(1+\|\tilde{\boldsymbol{Z}}_n\|_2^2)\xi,$$

which is maximized by the choice $\phi = \phi_{freq}^{opt} = \xi^{-2}$. Through different arguments, this parameter choice was also suggested by Bibaut et al. (2022) for their nonparametric mixture sequential probability ratio test.

In practice, ξ is unknown, but practitioners often have a sense of its magnitude. In a frequentist fixed-n power analysis, for example, practitioners must specify a minimum detectable effect (MDE). We suggest equating ϕ to the reciprocal of the squared MDE, effectively equating the standard deviation of the Gaussian mixture on ξ to the MDE.

3.4.3 Selecting Φ to maximize $\mathbb{E}^{\Pi}[\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi)]]$ under the Bayesian alternative

Suppose $\xi \sim \Pi$ where Π is a $N(0,\zeta^{-1})$ distribution. In this case, we seek to maximize $\mathbb{E}^{\Pi}[\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi)]]$, where the expectation is taken with respect to the observations and the random δ jointly.

Proposition 3.12. *For all $n \in \mathbb{N}$,*

$$\mathbb{E}^{\Pi}[\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi_{Bayes}^{opt})]] \geq \mathbb{E}^{\Pi}[\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi)]]$$

for all $\phi > 0$, where

$$\phi_{Bayes}^{opt} = \zeta.$$

This is a stronger proposition than 3.11 as it holds for all n, instead of only holding approximately for large n. To see this note that,

$$\mathbb{E}^{\Pi}[\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi)]] = \int p(t_n)\log p(t_n|H_1)dt_n - \int p(t_n)\log p(t_n|H_0)dt_n, \tag{10}$$

where $p(t_n)$ is the density of a location-scale $t_{\nu_n}(0, 1+\|\tilde{\boldsymbol{Z}}_n\|_2^2\zeta^{-1})$ distribution and $p(t_n|H_1)$ is the density of a location-scale $t_{\nu_n}(0, 1+\|\tilde{\boldsymbol{Z}}_n\|_2^2\phi^{-1})$. Critically, the two distributions are of the same form, differing only by the parameters ϕ and ζ. We are unaware of a closed form expression for equation (10). Fortunately, it is not necessary to determine the optimal value of ϕ. Instead of employing approximations, as before, we can leverage the result that for two distributions p and q, the (negative) cross-entropy is bounded from above by the (negative) entropy,

$$\begin{aligned}\int p(x)\log p(x)dx &= \int p(x)\log q(x)dx - \int p(x)\log\frac{p(x)}{q(x)}dx \\ &\geq \int p(x)\log q(x)dx,\end{aligned}$$

where the last inequality holds because the Kullback-Leibler divergence is non-negative. Hence $\mathbb{E}^{\Pi}[\mathbb{E}[\log B_n(\boldsymbol{Y}_n;\phi)]]$ is maximized by equating ϕ to ζ.

Although the Bayesian and frequentist schools of thought have different approaches, they both require some information about the magnitude of the hypothesized treatment effect. The frequentist side specifies a minimum detectable effect, which requires an understanding of the magnitude of practically meaningful effects. On the other hand, the Bayesian approach achieves an average optimality by using the prior to weigh the performance of the test towards plausible values of the treatment effect. In many practical cases, both $\zeta^{-\frac{1}{2}}$ and the MDE on ξ capture the same scale.

4 Sequential Power Analysis: Comparing Fixed n with Anytime Valid Inference

We now compare our anytime-valid approach to a fixed-n procedure that leverages a minimum detectable effect. Our goal is to test whether a single coefficient is non-zero. Following standard power analysis calculations, we divide $\mathbb{R}$ into a region of relevance $\{\xi \in \mathbb{R} : |\xi| > \underline{\xi}\}$, where $\underline{\xi}$ denotes the minimum standardized effect size. As $d = 1$, $t_n(\boldsymbol{Y}_n)^2 = f_n(\boldsymbol{Y}_n)$, where f_n is the standard f statistic. The minimum sample size to provide power greater than c for all ξ with $|\xi| > \underline{\xi}$ in an α-level test, is the smallest n satisfying,

$$\mathbb{P}[f_n(\boldsymbol{Y}_n) > f_{\nu_n,1-\alpha}] = c,$$

where $f_n(\boldsymbol{Y}_n) \sim f(1, \nu_n, \|\tilde{\boldsymbol{Z}}_n\|_2^2 \underline{\xi}^2)$ and $f_{\nu_n,1-\alpha}$ is the $1-\alpha$ quantile of the null distribution $f(1, \nu_n, 0)$. In fixed designs, $\tilde{\boldsymbol{Z}}_n$ is known ahead of time, whereas in the case of stochastic regressors, discussed later, it is not.

For the anytime valid procedure we care about the probability of rejecting the null hypothesis *by* time n; that is, $\mathbb{P}\left[\inf_{i \le n} p_i(\boldsymbol{Y}_i) \le \alpha\right]$ under the alternative $f(1, \nu_n, \|\tilde{\boldsymbol{Z}}_n\|_2^2 \xi^2)$. This is difficult to compute directly, but a naive lower bound can be obtained from $\mathbb{P}\left[\inf_{i \le n} p_i(\boldsymbol{Y}_i) \le \alpha\right] > \mathbb{P}\left[p_n(\boldsymbol{Y}_n) \le \alpha\right]$. This ignores the possibility of rejecting at the previous $n-1$ times and focuses only on the probability of rejecting at time n. From the discussion of Section 3.4, we choose $\phi = \underline{\xi}^{-2}$. The probability of rejecting the null *at* time n is then

$$\mathbb{P}[f_n(\boldsymbol{Y}_n) > a_n(\alpha, \|\tilde{\boldsymbol{Z}}_n\|_2^2, \underline{\xi}^{-2})] \tag{11}$$

where $f_n(\boldsymbol{Y}_n) \sim f(1, \nu_n, \|\tilde{\boldsymbol{Z}}_n\|_2^2 \xi)$ and $a_n(\alpha, \|\tilde{\boldsymbol{Z}}_n\|_2^2, \underline{\xi}^{-2})$ is from equation (7). This can easily be evaluated by using the quantile function of a noncentral F distribution. We can di-

rectly observe that $\mathbb{P}[f_n(\boldsymbol{Y}_n) > f_{\nu_n,1-\alpha}] > \mathbb{P}[f_n(\boldsymbol{Y}_n) > a_n(\alpha, \|\tilde{\boldsymbol{Z}}_n\|_2^2, \underline{\xi}^{-2})] = \mathbb{P}\left[p_n(\boldsymbol{Y}_n) \leq \alpha\right]$. However, we cannot analytically compute the probability of the fully sequential procedure $\mathbb{P}\left[\inf_{i\leq n} p_i(\boldsymbol{Y}_i) \leq \alpha\right]$; instead, we examine it via a simulation.

Consider a deterministic design with $\boldsymbol{w}_i = [1, 1]$ if i is even else $\boldsymbol{w}_i = [1, 0]$. Let $\beta = 1$, $\delta = 0.2$, and the residual variance $\sigma^2 = 1.5$, then the standardized effect $\xi = \delta/\sigma \approx 0.16$. Suppose first that the experimenter has some oracle information about ξ and sets the minimum detectable effect $\underline{\xi} = \xi$ — this is the setting most favorable to a fixed-n test. When n is even $\|\tilde{\boldsymbol{Z}}_n\|_2^2 = \frac{n}{4}$, otherwise $\|\tilde{\boldsymbol{Z}}_n\|_2^2 = \frac{n^2-1}{4n}$. We take $\alpha = 0.01$ and seek a power of 0.95. This yields a minimum sample size requirement for the fixed-n test of 2676.

To simulate the approximate stopping time, we generate 10000 realizations under the above-described data-generating process and use the following stopping rule $\tau_s = \inf\{n \in \mathbb{N} : p_n(\boldsymbol{Y}_n) < \alpha\}$. The distribution of the stopping time is equal to the probability of rejecting the null by time n, $\mathbb{P}[\tau_s \leq n] = \mathbb{P}\left[\inf_{i\leq n} p_i(\boldsymbol{Y}_i) \leq \alpha\right]$. We use the empirical distribution of τ_s to estimate $\mathbb{P}\left[\inf_{i\leq n} p_i(\boldsymbol{Y}_i) \leq \alpha\right]$. This is shown in figure 1 in addition to the lower bound provided in (11).

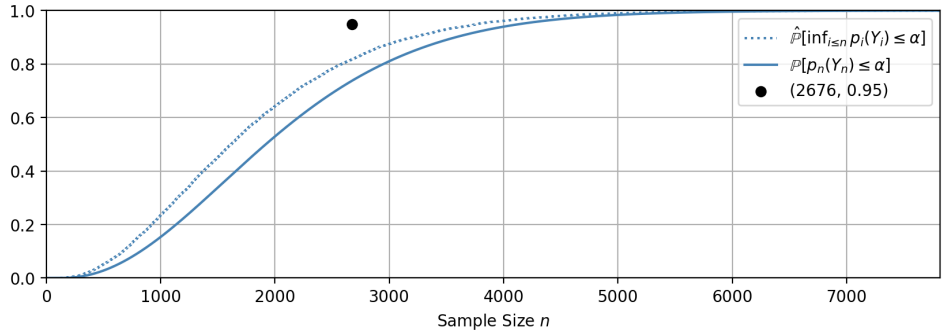


Figure 1: Empirical distribution of the stopping time τ_s based on 10000 draws (dotted), the lower bound from equation 11, and the fixed-n sample size that achieves a power greater than 0.95 for all ξ with $\xi > \underline{\xi}$.

Figure 1 provides some interesting insights. First, consider using the sequential test in a fixed-n modality; that is, rejecting the null at time $n = 2676$ if $p_n(\boldsymbol{Y}_n) \leq \alpha$ and ignoring all other n. From (11), this has a power of 0.73 whereas the fixed-n test has a power of 0.95. The fixed-n procedure always takes $n = 2676$ observations and has a Type-II error of 0.05. On the other hand, the fully sequential test requires *on average* only $n = 1806$ observations and has a Type-II error of *zero* (with a median stopping time of 1616). The Type-II error is zero because the test is guaranteed to reject the null eventually (sometimes referred to as *asymptotically power 1)*. The sequential test, therefore, takes *fewer samples on average* and has *preferable Type-II error performance*.

Now consider a more realistic example in which the minimum detectable effect is smaller than the true effect. Keeping the minimum detectable effect at $\underline{\xi} = 0.16$ and $\phi = \underline{\xi}^{-2}$ we repeat the same simulation as before, but instead set ξ to 0.32.

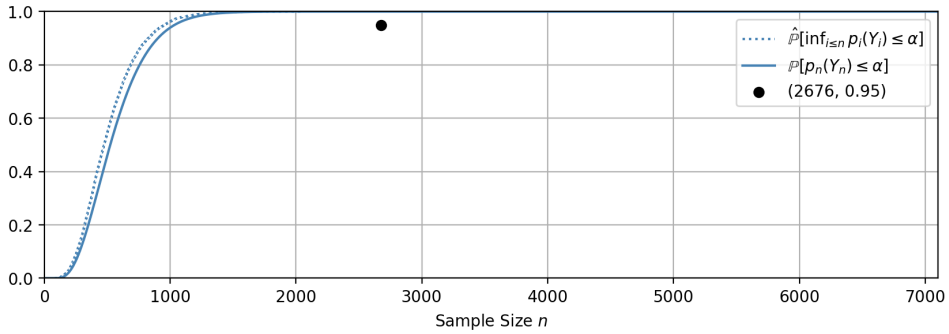


Figure 2: Empirical distribution of the stopping time τ_s based on 10000 draws (dotted), the lower bound from equation 11, and the fixed-n sample size that achieves a power greater than 0.95 for all ξ with $\xi > \underline{\xi}$.

Figure 2 shows the substantial speedup over fixed-n tests when the minimum detectable effect is underestimated. This forces the fixed-n test to collect more samples than is required. The average stopping time is now 515, with a median stopping time of 475.

5 Asymptotic Approximations

The results of the previous sections hold for all sample sizes as long as the linear model assumption is true. We now explore the asymptotic behavior of $B_n(\boldsymbol{Y}_n)$ and $C_n(\boldsymbol{Y}_n; \alpha)$, which will be particularly useful in the subsequent section where we consider randomized experiments. The proofs can be found in the Appendix.

Definition 5.1.

$$B_n^{\sigma}(\boldsymbol{Y}_n; \boldsymbol{\Phi}) = \frac{\int \Pi_{i=1}^n \frac{1}{\sigma} \psi(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{\delta}) dF^{\boldsymbol{\Phi}}(\boldsymbol{\delta}) d\boldsymbol{\beta}}{\int \Pi_{i=1}^n \frac{1}{\sigma} \psi(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) d\boldsymbol{\beta}} \tag{12}$$

when $\boldsymbol{W}_n'\boldsymbol{W}_n$ is full rank, otherwise $B_n^{\sigma}(\boldsymbol{Y}_n; \boldsymbol{\Phi}) = 1$, where ψ is a standard Gaussian density and $F^{\boldsymbol{\Phi}}$ is a $N(\boldsymbol{0}_d, \sigma^2\boldsymbol{\Phi}^{-1})$ probability measure. Let

$$\tilde{B}_n(\boldsymbol{Y}_n; \boldsymbol{\Phi}) = B_n^{s_n(\boldsymbol{Y}_n)}(\boldsymbol{Y}_n; \boldsymbol{\Phi}) \tag{13}$$

The statistic $B_n^{\sigma}(\boldsymbol{Y}_n; \boldsymbol{\Phi})$ is the mixture martingale or Bayes factor for the case when σ is known. The statistic $\tilde{B}_n(\boldsymbol{Y}_n)$ is obtained by using a plugin estimator $s_n(\boldsymbol{Y}_n)$ in place of σ. The difference between $B_n(\boldsymbol{Y}_n)$ and $\tilde{B}_n(\boldsymbol{Y}_n)$ is that the former properly marginalizes over the unknown scale parameter σ with respect to the right-Haar measure, whereas the latter assumes it is known and then uses a plugin estimator.

Theorem 5.2.

$$\tilde{B}_n(\boldsymbol{Y}_n; \boldsymbol{\Phi}) = \left(\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)} \right)^{\frac{1}{2}} e^{\frac{1}{2}\boldsymbol{t}_n(\boldsymbol{Y}_n)'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n'\boldsymbol{t}_n(\boldsymbol{Y}_n)}. \tag{14}$$

Suppose $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n = \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i' \to \boldsymbol{\Omega}_{\boldsymbol{w}}$ almost surely. Under the null hypothesis

$$\frac{\tilde{B}_n(\boldsymbol{Y}_n; \boldsymbol{\Phi})}{B_n(\boldsymbol{Y}_n; \boldsymbol{\Phi})} \xrightarrow{a.s.} 1.$$

Using $\tilde{B}_n(\boldsymbol{Y}_n)$ does not lead to a fully anytime-valid procedure, as it neglects uncertainty in estimating σ, but it converges to $B_n(\boldsymbol{Y}_n)$ almost surely under the null hypothesis[2]. As

[2]This is not the case under the alternative hypothesis

$\tilde{B}_n(\boldsymbol{Y}_n)$ is asymptotically equivalent to $B_n(\boldsymbol{Y}_n)$, we can define an *asymptotic confidence sequence* (Waudby-Smith et al. 2021) $\tilde{C}_n(\boldsymbol{Y}_n;\alpha) := \{\boldsymbol{\delta} \in \mathbb{R}^d : \tilde{B}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) \leq \alpha^{-1}\}$ which is asymptotically equivalent to $C_n(\boldsymbol{Y}_n;\alpha)$ in the sense $C_n(\boldsymbol{Y}_n;\alpha) \triangle \tilde{C}_n(\boldsymbol{Y}_n;\alpha) \overset{a.s.}{\to} \emptyset$, where $\triangle$ denotes the symmetric set difference.

Corollary 5.3. *Define*

$$\begin{aligned}\tilde{C}_n(\boldsymbol{Y}_n;\alpha) &= \{\boldsymbol{\delta} \in \mathbb{R}^d : \tilde{B}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) \leq \alpha^{-1}\} \\ &= \left\{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{D}_n} \leq s_n(\boldsymbol{Y}_n)^2 \log\left(\frac{\det(\boldsymbol{\Phi} + \boldsymbol{Z}_n'\boldsymbol{Z}_n)}{\alpha^2 \det(\boldsymbol{\Phi})}\right)\right\}\end{aligned} \tag{15}$$

where $\boldsymbol{D}_n = \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \boldsymbol{\Phi})^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n$. *Suppose* $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n = \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i' \to \boldsymbol{\Omega}_{\boldsymbol{w}}$ *almost surely, then* $\tilde{C}_n(\boldsymbol{Y}_n;\alpha)$ *is an asymptotic confidence sequence for* $C_n(\boldsymbol{Y}_n;\alpha)$.

In theorem 5.2 we worked in a stochastic regressor context assuming $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n = \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i' \to \boldsymbol{\Omega}_{\boldsymbol{w}}$. Consequently $\frac{1}{n}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n \overset{a.s.}{\to} \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}} = \boldsymbol{\Omega}_{\boldsymbol{z}} - \boldsymbol{\Omega}_{\boldsymbol{zx}}\boldsymbol{\Omega}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{xz}}$. Equation (14) can then be simplified even further by the choice of $\boldsymbol{\Phi} = \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$. This choice effectively replaces $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \boldsymbol{\Phi}$ in equation (14) by $(\lambda + n)\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$.

Theorem 5.4. *Suppose* $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n = \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i' \to \boldsymbol{\Omega}_{\boldsymbol{w}}$ *almost surely. Let* $\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}} = \boldsymbol{\Omega}_{\boldsymbol{z}} - \boldsymbol{\Omega}_{\boldsymbol{zx}}\boldsymbol{\Omega}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{xz}}$, *choose* $\boldsymbol{\Phi} = \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$, *and define*

$$B_n^{\infty}(\boldsymbol{Y}_n) = \left(\frac{\lambda}{\lambda + n}\right)^{\frac{d}{2}} e^{\frac{1}{2}\frac{n^2}{n+\lambda}\frac{\boldsymbol{\delta}_n^{ols}(\boldsymbol{Y}_n)'\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\boldsymbol{\delta}_n^{ols}(\boldsymbol{Y}_n)}{s_n^2(\boldsymbol{Y}_n)}}. \tag{16}$$

Then, under the null hypothesis,

$$\frac{B_n^{\infty}(\boldsymbol{Y}_n)}{B_n(\boldsymbol{Y}_n;\lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})} \overset{a.s.}{\to} 1.$$

The proof is given in Appendix A.8. In this new parameterization $\boldsymbol{\Phi} = \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$, then $\lambda_{freq}^{opt} = \xi^{-2}\Omega_{\tilde{\boldsymbol{z}}}^{-1}$ and $\lambda_{Bayes}^{opt} = \zeta\Omega_{\tilde{\boldsymbol{z}}}^{-1}$. As noted before this asymptotic test statistic can be inverted to obtain an asymptotic confidence sequence $C_n^{\infty}(\boldsymbol{Y}_n;\alpha) := \{\boldsymbol{\delta} \in \mathbb{R}^d : B_n^{\infty}(\boldsymbol{Y}_n;\boldsymbol{\delta}) \leq \alpha^{-1}\}$. In one dimension, this yields the following corollary.

Corollary 5.5.

$$\mathbb{P}\left[\forall n \in \mathbb{N} : |\hat{\delta}_n^{ols} - \delta| \leq \sqrt{\frac{s_n^2(\boldsymbol{Y}_n)}{n\Omega_{\tilde{\boldsymbol{z}}}}}\sqrt{\frac{n+\lambda}{n}\log\left(\frac{\lambda+n}{\lambda\alpha^2}\right)} + r_n\right] \geq 1-\alpha$$

where $r_n = o_{a.s.}(n^{-\frac{1}{2}})$

The proof is given in Appendix A.9. In randomised experiments with treatment assignment ρ, $\Omega_{\tilde{\boldsymbol{z}}}$ is known in advance with $\Omega_{\tilde{\boldsymbol{z}}} = \rho(1-\rho)$. If, on the other hand, $\Omega_{\tilde{\boldsymbol{z}}}$ is not known in advance, it can be estimated via $\frac{1}{n}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n$, in which case we have a last asymptotic approximation.

Theorem 5.6. *Suppose* $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n = \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i' \to \boldsymbol{\Omega}_{\boldsymbol{w}}$ *almost surely and define*

$$B_n^g(\boldsymbol{Y}_n) = \left(\frac{\lambda}{\lambda+n}\right)^{\frac{d}{2}} e^{\frac{d}{2}\frac{n}{n+\lambda}f(\boldsymbol{Y}_n)}. \tag{17}$$

Then, under the null hypothesis,

$$\frac{B_n^g(\boldsymbol{Y}_n)}{B_n(\boldsymbol{Y}_n; \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})} \xrightarrow{a.s.} 1$$

Equation 17 is arguably the simplest expression for a sequential F-test, albeit asymptotic. It is appealing because it is considerably simpler than equation (3) (at the cost of being asymptotic) and does not require knowledge of the limiting $\Omega_{\tilde{\boldsymbol{z}}}$ — only the classical F statistic. $B_n^g(\boldsymbol{Y}_n)$ is closely related to the Bayes factor resulting from Zellner's g-prior (Zellner 1986). The proof is given in appendix A.10.

6 Nonparametric Asymptotic Anytime-Valid Inference for Treatment Effects in Randomised Experiments

The results presented earlier are valid within the framework of the linear model. When used to perform inferences on average treatment effect in an experiment; however, several

of these assumptions come into question. Is the functional dependence of the response on the covariates truly linear? Have all relevant covariates been included in the regression? Are the residuals Gaussian, and are they homoskedastic? Randomized assignment protects against many of these issues, albeit asymptotically, making the linear model remarkably robust to model misspecification (Lin 2013).

Adopting a potential outcomes framework, let each experimental unit i have two potential outcomes $y_i(1)$ and $y_i(0)$, corresponding to the outcome that would be observed if the unit was assigned to treatment or control, respectively (Imbens & Rubin 2015). Each unit is independently assigned to treatment with probability ρ, which we express with a treatment indicator $T_i \sim \text{Bernoulli}(\rho)$, and a potential outcome is observed $y_i^{obs} = T_i y_i(1) + (1 - T_i) y_i(0)$.

Define the superpopulation ATE as $\tau_{sp} = \mathbb{E}_{sp}[y_i(1) - y_i(0)]$. Consider an interacted linear model $y_i = \alpha + (\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\beta} + T_i\tau + T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\eta} + \varepsilon_i$ where $\boldsymbol{\mu_m} = \mathbb{E}_{sp}[\boldsymbol{m}_i]$. In practice, centring the $(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\beta}$ term is not necessary as it does not affect inference for $\boldsymbol{\delta}$ or $\boldsymbol{\gamma}$. It merely simplifies the exposition of the theory. Centering of the interaction terms $T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\eta}$ is, however, necessary (Guo & Basse 2023). Including the interaction terms is a matter of preference. It can improve asymptotic efficiency and provide tests of treatment effect heterogeneity but requires knowledge of $\boldsymbol{\mu_m}$. We include this term for completeness, though our analysis also holds for the case where interaction terms are omitted.

Define the population least squares parameter values as

$$(\alpha^\star, \boldsymbol{\beta}^\star, \tau^\star, \boldsymbol{\eta}^\star) = \operatorname{argmin} \mathbb{E}[(y_i - \alpha - \boldsymbol{m}_i'\boldsymbol{\beta} - T_i\tau - T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\eta})^2]. \tag{18}$$

It is easily verified that $\tau^\star = \tau_{sp}$. Under mild conditions

$$\sqrt{n}(\hat{\tau}_n^{ols} - \tau_{sp}) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\rho(1-\rho)}\right), \tag{19}$$

where

$$\sigma^2 = \frac{\mathbb{E}_{sp}[(T_i - \rho)^2(y_i - \alpha^\star - (T_i - \rho)\tau_{sp} - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\beta}^\star)^2]}{\rho(1-\rho)},$$

providing asymptotic *fixed-n* hypothesis tests and confidence intervals for τ_{sp}. Given a consistent estimator $\hat{\sigma}_n$, an asymptotic confidence interval can be constructed with radius

$$r_n(\boldsymbol{Y}_n) = \frac{\hat{\sigma}_n}{\sqrt{n\rho(1-\rho)}}\sqrt{\chi^2_{1-\alpha}},$$

where $\chi^2_{1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared random variable with 1 degree of freedom. We now provide the analogous asymptotic confidence sequence.

Theorem 6.1. *Assume* $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n \overset{a.s.}{\to} \boldsymbol{\Omega}_{\boldsymbol{w}}$ *and assume* $\mathbb{P}[T_i = 1|\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{W}] = \rho$. *Let* $\hat{\tau}_n(\boldsymbol{Y}_n)$ *be the ordinary least squares estimate of the main effect,* $\hat{\sigma}_n(\boldsymbol{Y}_n)$ *a strongly consistent estimator of* $\sigma(\boldsymbol{Y}_n)$ *and let*

$$r_n(\boldsymbol{Y}_n) = \frac{\hat{\sigma}_n(\boldsymbol{Y}_n)}{\sqrt{n\rho(1-\rho)}}\sqrt{\frac{\lambda+n}{n}\log\left(\frac{\lambda+n}{\lambda\alpha^2}\right)}. \tag{20}$$

Then $C_n(\boldsymbol{Y}_n;\alpha) := (\hat{\tau}_n(\boldsymbol{Y}_n) - r_n(\boldsymbol{Y}_n), \hat{\tau}_n(\boldsymbol{Y}_n) + r_n(\boldsymbol{Y}_n))$ *is an asymptotic confidence sequence for the superpopulation average treatment effect* $\tau_{sp} = \mathbb{E}_{sp}[y_i(1) - y_i(0)]$ *with rate* $o_{a.s.}(\sqrt{\log n/n})$.

Note that the only difference between the fixed-n confidence interval in equation (6) and the time-uniform confidence sequence in equation (20) is the term multiplying the standard error. Simply replacing $\sqrt{\chi^2_{1-\alpha}}$ by $\sqrt{((\lambda+n)/n)\log((\lambda+n)/(\lambda\alpha^2))}$ generalizes the classical confidence interval to an asymptotic confidence sequence; the relative width is shown in Figure 3.

In a fully sequential setting, $n-1$ additional confidence sets at time n cover the true parameter, not just the current n. For the same reason $\mathbb{P}\left[\inf_{i\leq n} p_i(\boldsymbol{Y}_i) \leq \alpha\right] > \mathbb{P}\left[p_n(\boldsymbol{Y}_n) \leq \alpha\right]$, we also have $\bigcap_{i=1}^n C_i(\boldsymbol{Y}_i;\alpha) \subset C_n(\boldsymbol{Y}_n;\alpha)$. The relative width achieves a minimum at differ-

ent n for different λ. In this sense, the role of λ is similar to that of the "α-spending function" in group sequential tests (Jennison & Turnbull 1999). A lower value of λ spends more α at the beginning, resulting in an initially tighter confidence sequence relative to a higher value. Following the discussion in Section 3.4. we recommend using $\lambda = \frac{\hat{\sigma}^2}{\underline{\tau}^2 \rho(1-\rho)}$ where $\underline{\tau}$ is the absolute minimum detectable effect, $\hat{\sigma}^2$ is a pre-experimental estimate of the variance σ^2, and ρ is the treatment assignment probability (expressed otherwise, $\lambda = \frac{1}{\underline{\xi}^2 \rho(1-\rho)}$ where ξ is the absolute minimum detectable effect on the *standardized* effect size).

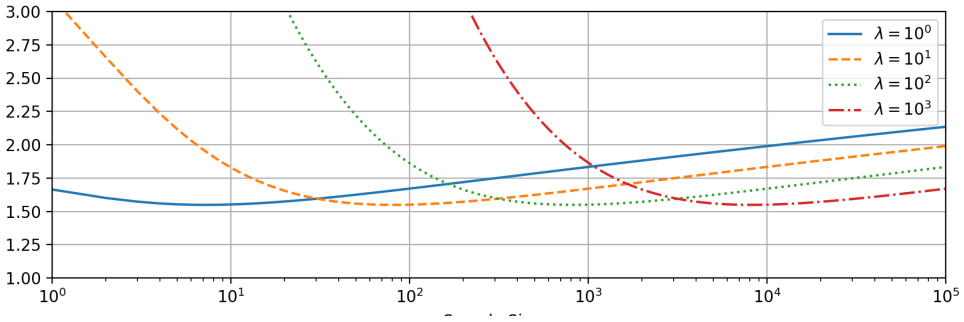


Figure 3: Relative width of anytime-valid confidence sequences to fixed-n confidence intervals $\sqrt{((\lambda+n)/n)\log((\lambda+n)/(\lambda\alpha^2))}/\sqrt{\chi^2_{1-\alpha}}$ for various choices of λ

6.1 Comparison with Parametric Linear Model Confidence Sequences

The nonparametric asymptotic confidence sequence in Equation (20) is almost identical to the parametric confidence sequence in Equation (5.5). The only difference between these confidence sequences is that $s_n(\boldsymbol{Y}_n)$ is replaced by $\hat{\sigma}_n(\boldsymbol{Y}_n)$ (note that $\Omega_{\tilde{\boldsymbol{z}}} = \rho(1-\rho)$). Under what circumstances, then, does the confidence sequence for the τ, derived under the assumptions of the linear model, also provide an asymptotic confidence sequence for the superpopulation average treatment effect τ_{sp}? One key assumption of the linear model

is the homogeneity of the residual variance. When this assumption holds, the $(T_i - \rho)^2$ term in equation (19) is independent of the residual term, and $s_n^2(\boldsymbol{Y}_n) \xrightarrow{a.s.} \sigma^2$. Under these circumstances, the linear model confidence sequence for τ is also an asymptotic confidence sequence for τ_{sp}. Even when the homogeneous residual assumption does not hold, Freedman (2008) showed that the linear model standard error, in the case where interaction terms are omitted, is consistent or conservative, provided that $\rho = 1/2$, which would also yield a valid asymptotic confidence sequence for the ATE. In other cases, however, robust estimators such as the Huber-White estimator must be used.

7 Simulation

There is a reasonable concern that asymptotic sequential tests and confidence sequences may not truly provide time-uniform Type-I and coverage guarantees, especially for small n, as they leverage asymptotic arguments. Bibaut et al. (2022) study this issue and find that mixture sequential probability ratio tests are well calibrated in the asymptotic regimes $\alpha \searrow 0$ and $\lambda \nearrow \infty$. In both cases, the stopping time is large, allowing enough time for the asymptotic convergence to occur. We do not attempt to recreate the detailed arguments here but instead provide intuition and demonstrate via simulation. When α is small, the rejection threshold α^{-1} is large, resulting in a later stopping time, allowing time for the asymptotics. Given the discussion in section 6 relating λ to α-spending, a larger λ spends less α at the beginning of the sequential test yielding initially larger confidence sequences, resulting in later rejection times and again leaving time for the asymptotics. The worst case behaviour is when α is large and λ is small. We use a standard $\alpha = 0.05$ and apply little regularization in the following.

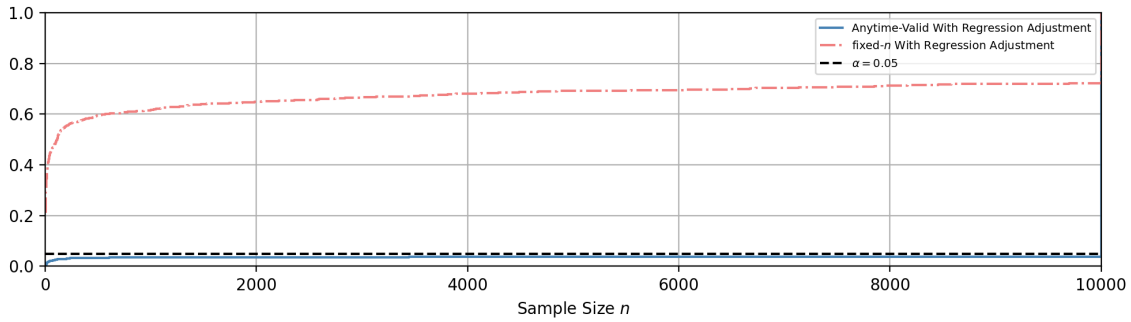


Figure 4: Empirical distribution of the stopping time $\tau_s = \inf\{n \in \mathbb{N} : p_n(\boldsymbol{Y}_n; 0) < 0.05\}$ based on 1000 simulations under the null hypothesis.

Suppose the true data-generating process is

$$y_i = 1 - 2x_{i1}^2 - 2\sin x_{i2} + 3|x_{i3}| + z_i\delta + \epsilon_i, \tag{21}$$

which we have modified from (Waudby-Smith et al. 2021), where $\epsilon_i \sim t_5(0, 1.5)$ and $\boldsymbol{x}_i \sim N(\boldsymbol{0}_3, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$, $z_i + \rho \sim$ Bernoulli(ρ) with $\rho = \frac{1}{2}$ and $\delta = \delta_0 = 0$. In this example, the average treatment effect is zero. We create 1000 realizations, and for each realization record the smallest n such that the sequential p-value from corollary 3.4 (with $\lambda = 1$) and the fixed-n p-value is less than α. This procedure simulates a draw from the stopping time distribution for the anytime-valid procedure and the fixed-n procedure with continuous monitoring. These draws are used to estimate the distribution of the stopping time over the interval $n = 1$ to $n = 10000$ shown in figure 4.

Figure 4 shows that the Type-I error rapidly increases when naively applying the fixed-n test after each new datapoint, as is expected. The Type-I error for the anytime-valid procedure also increases quickly at the beginning of the experiment but plateaus below the nominal $\alpha = 0.05$ level. As a small value of λ spends most α in the beginning, this rapid rise towards α is expected. A larger value of λ would result in a slower rise towards α. Among the 1000 realizations, only 38 incorrectly rejected the null using the anytime-valid

procedure, in contrast to 722. At first, this may seem undesirably less than the nominal $\alpha = 0.05$, but we must point out that these simulations were sequential tests truncated prematurely at $n = 10^4$. In contrast, the test is designed to spend α over $\mathbb{N}$.

8 Application

Sequential tests have been successfully employed in large technology companies, such as Netflix, to mitigate risk in the software deployment process (Lindon et al. 2022, Ham et al. 2022). For instance, when an Internet streaming company releases a new version of its application, it often uses canary testing, a strategy in which software is initially released only to a tiny subset of randomly selected users to test the performance. Suppose that the measured performance is worse as indicated by, for instance, a higher rate of errors is produced. In that case, the release is halted, and the software is reverted to the previous version. This vital quality control prevents software bugs and performance issues from reaching the entire user population and dramatically reduces the cost of innovation. Detecting differences as early as possible necessitates the use of anytime-valid methods.

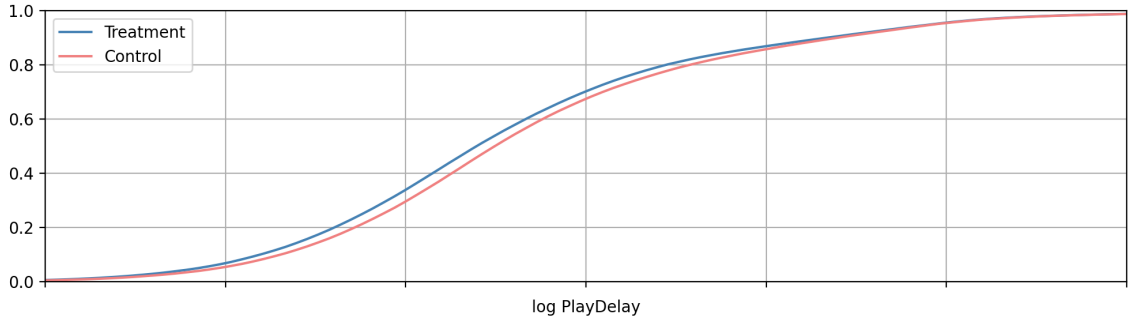


Figure 5: Empirical distributions of *Log-PlayDelay*

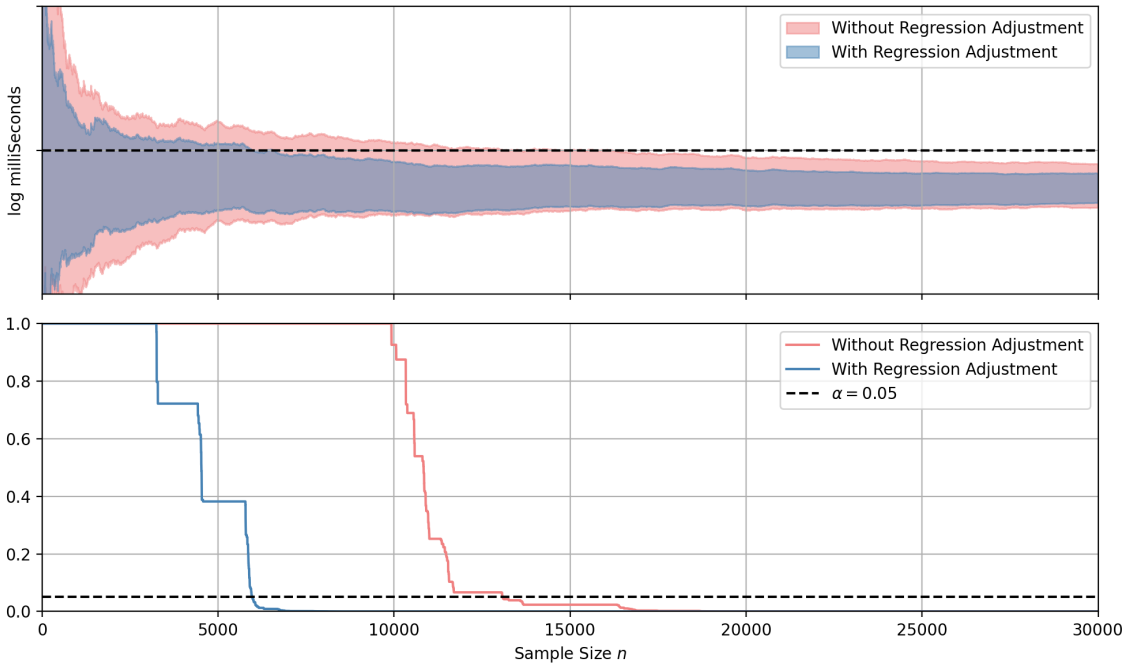


Figure 6: 0.95 Confidence sequences from Equation (7) with sequential p-values from Equation (6)

One of the most critical performance metrics for a streaming company to measure is *Log-PlayDelay* - the amount of time it takes for a stream to start playback. This load-time is a function of the device hardware and the speed of the internet connection. The experimental unit in such canary tests is the device. With the exception of cellular devices, the internet connection and hardware of a device generally do not change with time. Consequently, device measurements of *Log-PlayDelay* before treatment assignment are highly correlated with measurements after treatment assignment, providing an opportunity for variance reduction by regressing post-treatment outcomes on pre-treatment outcomes.

The following dataset is taken from a streaming canary A/B experiment. The release candidate contained newly optimized code which reduced *Log-PlayDelay*. The empirical distributions of *Log-PlayDelay* are shown in Figure 5. The correlation between post and pre-treatment outcomes is approximately 0.7. As the propensity score is 1/2, we use the

linear model confidence sequences from Corollary 3.8 which, by the discussion in Section 6, form an asymptotic confidence sequence for the average treatment effect.

The confidence sequences with and without regressing on pre-treatment outcomes are shown in Figure 6. The reduction in uncertainty obtained by regressing on pre-treatment outcomes is clearly reflected in the time required to conclude that the treatment effect is negative. This is also visible through the sequential p-values in Figure 6. Regression adjustment clearly reduces the time taken by canary tests, increasing the velocity at which software can be rolled out to users.

8.1 Simulation based on the empirical data

The distribution of stopping times, defined as the first n for which $p_n(\boldsymbol{Y}_n) < \alpha$, can be explored via simulations. The purpose of the first simulation is to demonstrate that the Type-I error remains well calibrated below the nominal $\alpha = 0.05$ level despite using an asymptotic confidence sequence. We first simulate 1000 A/A tests in which the null hypothesis is true by sampling outcomes for both treatment and control units from the empirical distribution of the control data shown in figure 5. The empirical distribution of the regression adjusted and unadjusted stopping times for both the fixed-n and anytime-valid tests of section 6 in Figure 7. This figure demonstrates that the Type-I error increases rapidly when continuously monitoring the experiment with fixed-n tests, whereas the anytime-valid Type-I error remains below the nominal α. In 1000 simulations at the 0.05 level, there are 26 and 24 incorrect rejections of the null for regression adjusted and unadjusted tests within the *first* 25k observations. Although this appears lower than what we might expect for a $\alpha = 0.05$ level test, note that these simulations were prematurely truncated after $25k$ outcomes, which is only a subset of $\mathbb{N}$ over which Type-I error is intended to be controlled.

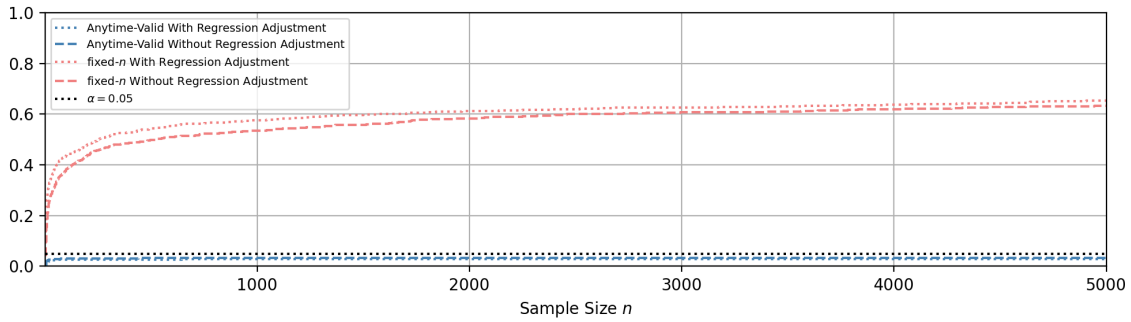


Figure 7: Empirical distribution of the stopping time $\tau_s = \inf\{n \in \mathbb{N} : p_n(\boldsymbol{Y}_n; 0) < 0.05\}$ based on 1000 simulations under the null hypothesis.

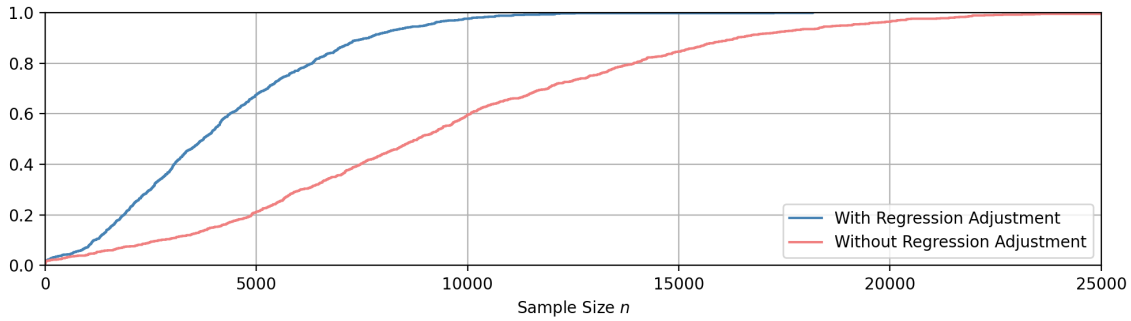


Figure 8: Empirical distribution of the stopping time $\tau_s = \inf\{n \in \mathbb{N} : p_n(\boldsymbol{Y}_n; 0) < 0.05\}$ based on 1000 simulations under the alternative hypothesis.

To explore the stopping time distributions when the alternative is true, we simulate data for treatment and control from their respective empirical distributions in figure 5. Figure 8 shows that the distribution of stopping time without regression adjustment stochastically dominates the distribution of stopping times with regression adjustment. The mean stopping time with regression adjustment was approximately 4120, compared to 9350 without. A comparison of our procedure with the empirical Bernstein confidence sequences of Howard et al. (2021) has been included in Appendix C.

9 Conclusion

Linear Regression adjustment with interaction terms is considered a gold standard for analyzing randomised controlled trials. Our results generalise the guarantees of existing fixed-n procedures to provide *time-uniform* guarantees, i.e., Type-I error and coverage guarantees that hold for all sample sizes. This enables trials to be stopped using data-dependent stopping rules, and can help safeguard publications from spurious significant results. For example, suppose a hypothesis is rejected at the α not only by the fixed-n test but also by the anytime-valid test. In that case, the reader can be more confident that this is not an artefact from dubious data collection. Moreover, the anytime-valid procedure is no more difficult than the fixed-n procedure, requiring only the same statistics. Published works that report these statistics can easily be reevaluated through the anytime-valid procedure, and software libraries can easily be converted.

In the context of large-scale A/B tests in digital settings, linear regression remains the primary tool because of its scalability. Our results now enable practitioners to perform this analysis sequentially. This allows A/B tests to be performed autonomously and stopped early, safeguarding users from harmful treatment effects and productionizing beneficial treatment effects sooner. This leads to a lower risk and more agile culture of experimentation.

References

Armitage, P. (1993), *Interim Analyses in Clinical Trials*, Multiple Comparisons, Selection and Applications in Biometry, CRC Press.

Armitage, P., McPherson, C. K. & Rowe, B. C. (1969), 'Repeated significance tests on accu-

mulating data', *Journal of the Royal Statistical Society. Series A (General)* **132**(2), 235–244.

URL: *http://www.jstor.org/stable/2343787*

Beasley, D. (2023), 'Safe testing for large-scale experimentation platforms', *arXiv preprint arXiv:2310.20450* .

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. & Johnson, V. E. (2018), 'Redefine statistical significance', *Nature Human Behaviour* **2**(1), 6–10.

URL: *https://doi.org/10.1038/s41562-017-0189-z*

Berger, J. O., Pericchi, L. R. & Varshavsky, J. A. (1998), 'Bayes factors and marginal distributions in invariant situations', *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **60**(3), 307–321.

URL: *http://www.jstor.org/stable/25051210*

Bibaut, A., Kallus, N. & Lindon, M. (2022), 'Near-optimal non-parametric sequential tests and confidence sequences with possibly dependent observations'.

Collaboration, O. S. (2015), 'Estimating the reproducibility of psychological science', *Science* **349**(6251), aac4716.
URL: *https://www.science.org/doi/abs/10.1126/science.aac4716*

Dass, S. C. & Berger, J. O. (2003), 'Unified conditional frequentist and bayesian testing of composite hypotheses', *Scandinavian Journal of Statistics* **30**(1), 193–210.
URL: *http://www.jstor.org/stable/4616757*

de Heide, R. & Grünwald, P. D. (2021), 'Why optional stopping can be a problem for bayesians', *Psychonomic Bulletin & Review* **28**(3), 795–812.
URL: *https://doi.org/10.3758/s13423-020-01803-x*

de la Peña, V. H., Klass, M. J. & Lai, T. L. (2004), 'Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws', *The Annals of Probability* **32**(3), 1902 – 1933.
URL: *https://doi.org/10.1214/009117904000000397*

Deng, A., Xu, Y., Kohavi, R. & Walker, T. (2013), Improving the sensitivity of online controlled experiments by utilizing pre-experiment data, *in* 'Proceedings of the Sixth ACM International Conference on Web Search and Data Mining', WSDM '13, Association for Computing Machinery, New York, NY, USA, p. 123–132.
URL: *https://doi.org/10.1145/2433396.2433413*

Eaton, M. L. (1989), 'Group invariance applications in statistics', *Regional Conference Series in Probability and Statistics* **1**, i–133.
URL: *http://www.jstor.org/stable/4153172*

Eppo (2023), 'Statistical nitty-gritty', https://docs.geteppo.com/statistics/confidence-intervals/statistical-nitty-gritty. Accessed: 2023-11-30.

European Medicines Agency (2015), 'Guideline on adjustment for baseline covariates in clinical trials', https://www.ema.europa.eu/en/adjustment-baseline-covariates-clinical-trials-scientific-guideline. EMA/CHMP/295050/2013.

Freedman, D. A. (2008), 'On regression adjustments to experimental data', *Advances in Applied Mathematics* **40**(2), 180–193.

Grünwald, P., de Heide, R. & Koolen, W. (2021), 'Safe testing'.

Guo, K. & Basse, G. (2023), 'The generalized oaxaca-blinder estimator', *Journal of the American Statistical Association* **118**(541), 524–536.
URL: *https://doi.org/10.1080/01621459.2021.1941053*

Ham, D. W., Bojinov, I., Lindon, M. & Tingley, M. (2022), 'Design-based confidence sequences for anytime-valid causal inference'.

Hao, Y., Grünwald, P., Lardy, T., Long, L. & Adams, R. (2023), 'E-values for k-sample tests with exponential families'.

Hendriksen, A., de Heide, R. & Grünwald, P. (2021), 'Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations', *Bayesian Analysis* **16**(3), 961 – 989.
URL: *https://doi.org/10.1214/20-BA1234*

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J. & Welch, V. A. (2019), *Cochrane handbook for systematic reviews of interventions*, John Wiley & Sons.

Howard, S. R. & Ramdas, A. (2022), 'Sequential estimation of quantiles with applications to A/B testing and best-arm identification', *Bernoulli* **28**(3), 1704 – 1728.
URL: *https://doi.org/10.3150/21-BEJ1388*

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2020), 'Time-uniform Chernoff bounds via nonnegative supermartingales', *Probability Surveys* **17**(none), 257 – 317.
URL: *https://doi.org/10.1214/18-PS321*

Howard, S. R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2021), 'Time-uniform, nonparametric, nonasymptotic confidence sequences', *The Annals of Statistics* **49**(2), 1055 – 1080.
URL: *https://doi.org/10.1214/20-AOS1991*

Imbens, G. W. & Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLoS medicine* **2**(8), e124.

Jennison, C. & Turnbull, B. W. (1999), *Group sequential methods with applications to clinical trials*, CRC Press.

Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2017), Peeking at a/b tests: Why it matters, and what to do about it, *in* 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '17, Association for Computing Machinery, New York, NY, USA, p. 1517–1525.
URL: *https://doi.org/10.1145/3097983.3097992*

Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2021), 'Always valid inference: Continuous

monitoring of a/b tests', *Operations Research* .

URL: *https://doi.org/10.1287/opre.2021.2135*

John, L. K., Loewenstein, G. & Prelec, D. (2012), 'Measuring the prevalence of questionable research practices with incentives for truth telling', *Psychological science* **23**(5), 524–532.

Lai, T. L. (1981), 'Asymptotic Optimality of Invariant Sequential Probability Ratio Tests', *The Annals of Statistics* **9**(2), 318 – 333.

URL: *https://doi.org/10.1214/aos/1176345398*

Lehmann, E. L. & Romano, J. P. (2005), *Testing statistical hypotheses*, Springer Texts in Statistics, third edn, Springer, New York.

Lin, W. (2013), 'Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique', *The Annals of Applied Statistics* **7**(1), 295 – 318.

URL: *https://doi.org/10.1214/12-AOAS583*

Lindon, M. & Malek, A. (2022), Anytime-valid inference for multinomial count data, *in* A. H. Oh, A. Agarwal, D. Belgrave & K. Cho, eds, 'Advances in Neural Information Processing Systems'.

URL: *https://openreview.net/forum?id=a4zg0jiuVi*

Lindon, M., Sanden, C. & Shirikian, V. (2022), Rapid regression detection in software deployments through sequential testing, *in* 'Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining', KDD '22, Association for Computing Machinery, New York, NY, USA, p. 3336–3346.

URL: *https://doi.org/10.1145/3534678.3539099*

Maharaj, A., Sinha, R., Arbour, D., Waudby-Smith, I., Liu, S. Z., Sinha, M., Addanki, R.,

Ramdas, A., Garg, M. & Swaminathan, V. (2023), Anytime-valid confidence sequences in an enterprise a/b testing platform, *in* 'Companion Proceedings of the ACM Web Conference 2023', ACM.

Pekelis, L., Walsh, D. & Johari, R. (2015), 'The new stats engine', *Internet. Retrieved December* **6**, 2015.

Pérez-Ortiz, M. F., Lardy, T., de Heide, R. & Grünwald, P. (2022), 'E-statistics, group invariance and anytime valid testing'.
URL: *https://arxiv.org/abs/2208.07610*

Ramdas, A., Grünwald, P., Vovk, V. & Shafer, G. (2023), 'Game-theoretic statistics and safe anytime-valid inference'.

Ramdas, A., Ruf, J., Larsson, M. & Koolen, W. (2020), 'Admissible anytime-valid sequential inference must rely on nonnegative martingales'.
URL: *https://arxiv.org/abs/2009.03167*

Reluga, K., Ye, T. & Zhao, Q. (2022), 'A unified analysis of regression adjustment in randomized experiments', *arXiv preprint arXiv:2210.04360* .

Robbins, H. & Siegmund, D. (1970), 'Boundary Crossing Probabilities for the Wiener Process and Sample Sums', *The Annals of Mathematical Statistics* **41**(5), 1410 – 1429.
URL: *https://doi.org/10.1214/aoms/1177696787*

Sagarin, B. J., Ambler, J. K. & Lee, E. M. (2014), 'An ethical approach to peeking at data', *Perspectives on Psychological Science* **9**(3), 293–304.

Services, A. W. (2023), 'Calculating expected results for cloudwatch anomaly detection'. Accessed on May 10, 2023.

URL: *https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch-Evidently-calculate-results.html*

Strassen, V. (1964), 'An invariance principle for the law of the iterated logarithm', *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **3**(3), 211–226.

ter Schure, J., Pérez-Ortiz, M. F., Ly, A. & Grunwald, P. (2020), 'The safe logrank test: Error control under continuous monitoring with unlimited horizon', *arXiv preprint arXiv:2011.06931* .

Turner, R. J. & Grünwald, P. D. (2023), 'Safe sequential testing and effect estimation in stratified count data'.

Turner, R., Ly, A. & Grünwald, P. (2021), 'Generic e-variables for exact sequential k-sample tests that allow for optional stopping', *arXiv preprint arXiv:2106.02693* .

U.S. Food and Drug Administration (2023), 'Adjusting for covariates in randomized clinical trials for drugs and biological products', `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products`. Docket Number: FDA-2019-D-0934.

Ville, J. (1939), Étude critique de la notion de collectif, PhD thesis, L'École Polytechnique.

URL: *http://eudml.org/doc/192893*

Wald, A. (1945), 'Sequential tests of statistical hypotheses', *Ann. Math. Statist.* **16**(2), 117–186.

URL: *https://doi.org/10.1214/aoms/1177731118*

Wang, H. & Ramdas, A. (2023), 'Catoni-style confidence sequences for heavy-tailed mean estimation', *Stochastic Processes and their Applications* **163**, 168–202.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The asa statement on p-values: Context, process, and purpose', *The American Statistician* **70**(2), 129–133.

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H. & Ramdas, A. (2021), 'Time-uniform central limit theory, asymptotic confidence sequences, and anytime-valid causal inference'.

URL: *https://arxiv.org/abs/2103.06476*

Waudby-Smith, I. & Ramdas, A. (2020), Confidence sequences for sampling without replacement, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin, eds, 'Advances in Neural Information Processing Systems', Vol. 33, Curran Associates, Inc., pp. 20204–20214.

URL: *https://proceedings.neurips.cc/paper/2020/file/e96c7de8f6390b1e6c71556e4e0a4959-Paper.pdf*

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N. & Mineiro, P. (2022), 'Anytime-valid off-policy inference for contextual bandits'.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica: Journal of the Econometric Society* pp. 817–838.

Wijsman, R. (1990), *Invariant Measures on Groups and Their Use in Statistics*, IMS Lecture Notes, Institute of Mathematical Statistics.

URL: *https://books.google.com/books?id=GSk4ueHzo30C*

WWC (2022), 'Procedures and standards handbook, version 5.0'. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf.

Ye, T., Shao, J., Yi, Y. & Zhao, Q. (2022), 'Toward better practice of covariate adjustment in analyzing randomized clinical trials', *Journal of the American Statistical Association* pp. 1–13.

Zellner, A. (1986), 'On assessing prior distributions and bayesian regression analysis with g-prior distributions', *Bayesian inference and decision techniques* .

SUPPLEMENTARY MATERIAL

A Review: The classical fixed-n F-Test

As this section concerns fixed-n case, we drop the n superscript from $\boldsymbol{Y}_n$ to simplify the exposition. An α-level test of $H_0 : \boldsymbol{\delta} = 0$, without loss of generality, can be obtained by examining the likelihood ratio test statistic

$$\Lambda(\boldsymbol{Y}) := \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1} p(\boldsymbol{Y}|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} p(\boldsymbol{Y}|\boldsymbol{\theta})} \tag{22}$$

and rejecting the null hypothesis when $\Lambda(\boldsymbol{Y}) > c_\alpha$ for some constant $c_\alpha > 0$ suitably chosen to provide a Type-I error probability of at most α. The following lemma recalls the classical likelihood ratio test construction of the F-test.

Theorem A.1.

$$\Lambda(\boldsymbol{Y}) = 1 + \frac{d}{n-p-d} f(\boldsymbol{Y}) \tag{23}$$

where the f-statistic is defined as

$$f(\boldsymbol{Y}) = \frac{\frac{\boldsymbol{Y}'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}}{d}}{\frac{\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}}{n-p-d}} = \frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{d s^2(\boldsymbol{Y})} = \frac{\boldsymbol{t}(\boldsymbol{Y})'\boldsymbol{t}(\boldsymbol{Y})}{d} \tag{24}$$

Then $\Lambda(\boldsymbol{Y}) > c_\alpha \iff f(\boldsymbol{Y}) > f_\alpha$ for some $f_\alpha > 0$. The distributions of the f-statistic under H_1 and H_0 are

$$\begin{aligned} f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, H_1 &\sim F(d, n-p-d, \|\tilde{\boldsymbol{Z}}\boldsymbol{\delta}\|_2^2/\sigma^2) \\ f(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0 &\sim F(d, n-p-d, 0) \end{aligned} \tag{25}$$

Rejecting when $f(Y) > f_\alpha$, with f_α denoting the $1-\alpha$ quantile $F(d, n-p-d, 0)$ yields a fixed-n test with Type-I error probability α.

Note that the f statistic can be written in terms of the maximal invariant statistic $\boldsymbol{t}(\boldsymbol{Y})$. In the case of $d = 1$, when there is only a single main effect, then f can be identified

as the square of the usual t-statistic ($t \sim t_{n-p-1} \Rightarrow t^2 \sim F(1, n-p-1)$). A p-value can be calculated by computing $\mathbb{P}[f \geq f(Y)]$ under the null $F(d, n-p-d, 0)$ distribution.

Proof. Starting with the denominator in (22), consider first expressing the quadratic form in the Guassian likelihood as a component in $\mathcal{C}(\boldsymbol{X})$ and a component in $\mathcal{C}(\boldsymbol{X})^{\perp}$.

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 &= \|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\|_2^2 \\ &= \|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\|_2^2 \end{aligned} \tag{26}$$

This is minimized by setting $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}'\boldsymbol{Y}$, which sets the first term to zero. The likelihood is then maximized by setting $\hat{\sigma^2} = \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\|_2^2/n$. It follows that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} p(\boldsymbol{Y}|\boldsymbol{\theta}) = \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\|_2^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}} \tag{27}$$

Now consider the numerator in (22), expressing the quadratic form in the Gaussian likelihood as a component in $\mathcal{C}(\boldsymbol{W})$ a component in $\mathcal{C}(\boldsymbol{W})^{\perp}$.

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\delta}\|_2^2 = \|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}\|_2^2 &= \|\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 \\ &= \|\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2 \end{aligned} \tag{28}$$

where $\boldsymbol{W} = [\boldsymbol{X}, \boldsymbol{Z}]$ and $\boldsymbol{\gamma}' = (\boldsymbol{\beta}', \boldsymbol{\delta}')$. Applying the same reasoning as before, this is minimized by setting $\hat{\boldsymbol{\gamma}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'Y$, which sets the first term to zero. The likelihood is then maximized by setting $\hat{\sigma}^2 = \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2/n$. It follows that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1} p(\boldsymbol{Y}|\boldsymbol{\theta}) = \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}} \tag{29}$$

and therefore

$$\Lambda(\boldsymbol{Y}) = \left(\frac{\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\|_2^2}{\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2}\right)^{\frac{n}{2}}. \tag{30}$$

However, the vector in the numerator be expressed as a component in $\mathcal{C}(\boldsymbol{W})$ and a com-

ponent in $\mathcal{C}(\boldsymbol{W})^{\perp}$.

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 &= \|\boldsymbol{P_W}(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 \\ &= \|(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2 \\ &= \|(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2\end{aligned} \tag{31}$$

and so the likelihood ratio can be written in terms of the f-statistic as

$$\Lambda(\boldsymbol{Y}) = 1 + \frac{d}{n-p-d} f(\boldsymbol{Y}). \tag{32}$$

To show $f(\boldsymbol{Y})$ can be expressed in terms of $\hat{\boldsymbol{\delta}}(\boldsymbol{Y})$ as in equation (24), note simply that $(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y} = (\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{P_W}\boldsymbol{Y} = (\boldsymbol{I}-\boldsymbol{P_X})(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{Y}) + \boldsymbol{Z}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})) = (\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Z}\hat{\boldsymbol{\delta}}(\boldsymbol{Y}) = \tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})$. □

A test of the null hypothesis $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ can easily be obtained from a hypothesis test of $\boldsymbol{\delta} = \boldsymbol{m}0$ by replacing $\boldsymbol{Y}$ with $\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\delta}_0$. In this case, the f- statistic becomes

$$f(\boldsymbol{Y}) = \frac{\frac{(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)'(\boldsymbol{P_W}-\boldsymbol{P_X})(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)}{d}}{\frac{(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)'(\boldsymbol{I}-\boldsymbol{P_W})(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)}{n-p-d}} = \frac{(\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta}_0)}{ds^2(\boldsymbol{Y})} \tag{33}$$

By finding the set of null-values that would not be rejected by this test one obtains a confidence set for the vector $\boldsymbol{\delta}$.

Corollary A.2. *A $1-\alpha$ confidence set for $\boldsymbol{\delta}$ is provided by*

$$\mathcal{C}_\alpha(\boldsymbol{Y}) := \{\boldsymbol{\delta} : (\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta}) \leq ds^2(\boldsymbol{Y}) f_\alpha\}, \tag{34}$$

A.1 Relationship between $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n$ and $\boldsymbol{W}_n\boldsymbol{W}_n$

$$\begin{aligned}(\boldsymbol{W}'\boldsymbol{W})^{-1} &= \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} * & * \\ * & (\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z})^{-1} \end{pmatrix}.\end{aligned}$$

The last line follows from the inverse of block matrices. Note $(\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}) = \boldsymbol{Z}(\boldsymbol{I} - \boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Z} = \boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Z} = \boldsymbol{Z}'(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Z} = \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}$, and so

$$(\boldsymbol{W}'\boldsymbol{W})^{-1} = \begin{pmatrix} * & * \\ * & (\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1} \end{pmatrix}$$

A.2 Proof of Proposition 3.2

Proof. It's best to break this down into a number of steps. First decompose the quadratic form in the likelihood into two components, one in $\mathcal{C}(\boldsymbol{W})$ and the other in $\mathcal{C}(\boldsymbol{W})^{\perp}$. Then, further subdivide the component in $\mathcal{C}(\boldsymbol{W})$ into two subcomponents, one in $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{C}(\boldsymbol{X})^{\perp}$. This helps to isolate terms in $\boldsymbol{\beta}, \boldsymbol{\delta}$ and σ^2 to make computing the marginals easier.

$$\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}\|_2^2 &= \|\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 \\
&= \|\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2 \\
&= \|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(I - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2 \\
&= \|X\hat{\boldsymbol{\beta}} + \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Z}\hat{\boldsymbol{\delta}} - X\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Z}\boldsymbol{\delta})\|_2^2 + \|(I - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Z}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2 \\
&= \|\boldsymbol{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}(\boldsymbol{Y}, \boldsymbol{\delta}))\|_2^2 + \|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2
\end{aligned} \tag{35}$$

where $\tilde{\boldsymbol{\beta}}(\boldsymbol{Y}, \boldsymbol{\delta})) = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}))$.

Step 1) Compute $p(\boldsymbol{Y}|H_1)$

Let's proceed first by computing the marginal under H_1

$$p(\boldsymbol{Y}|H_1) = \int\int\int p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, H_1)p(\boldsymbol{\delta}|\sigma^2, H_1)p(\boldsymbol{\beta}, \sigma^2|H_1)d\boldsymbol{\beta}d\boldsymbol{\delta}d\sigma^2. \tag{36}$$

We can handle these three marginalizations in three consecutive steps.

Step 1)i) Compute $p(\boldsymbol{Y}|\boldsymbol{\delta}, \sigma^2, H_1)$

Handling the marginalization for $\boldsymbol{\beta}$ first gives

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\delta}, \sigma^2, H_1) &= \int p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, H_1) p(\boldsymbol{\beta}|H_1) d\beta \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left(\|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2\right)} \int e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}(\boldsymbol{Y},\boldsymbol{\delta}))\|_2} d\boldsymbol{\beta} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})\|_2^2} e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2}, \end{aligned} \tag{37}$$

where the last line follows from recognizing the integrand as the kernel of a multivariate Gaussian in $\boldsymbol{\beta}$ with precision matrix $\boldsymbol{X}'\boldsymbol{X}/\sigma^2$.

Step 1)ii) Compute $p(\boldsymbol{Y}|\sigma^2, H_1)$

We now move onto performing the marginalization with respect to $\boldsymbol{\delta} \sim N(\boldsymbol{\delta}_0, \sigma^2\boldsymbol{\Phi}^{-1})$. Before doing this we complete the square in the following sense

$$\|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2^2 + \boldsymbol{\delta}'\boldsymbol{\Phi}\boldsymbol{\delta} = (\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}})'(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}) + \hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}} \tag{38}$$

where $\tilde{\boldsymbol{\delta}} = (\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\delta}$ is the posterior mean and $(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})/\sigma^2$ the posterior precision. Performing the marginalization then yields

$$\begin{aligned} p(\boldsymbol{Y}|\sigma^2, H_1) &= \int p(\boldsymbol{Y}|\boldsymbol{\delta}, \sigma^2, H_1) p(\boldsymbol{\delta}|\sigma^2, H_1) d\boldsymbol{\delta} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2} e^{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}} \\ &\quad \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d}{2}} \det(\boldsymbol{\Phi})^{\frac{1}{2}} \int e^{-\frac{1}{2\sigma^2}(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})'(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})} d\boldsymbol{\delta} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \left(\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}\right)^{\frac{1}{2}} \\ &\quad e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2} e^{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}} \end{aligned} \tag{39}$$

where the last step is achieved by recognizing the kernel of a d-dimensional multivariate Gaussian density in $\boldsymbol{\delta}$ with precision $\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}$.

Step 1)iii) Compute $p(\boldsymbol{Y}|H_1)$

Now we perform the final marginalization over σ^2

$$\begin{aligned}
p(\boldsymbol{Y}|H_1) &= \int p(\boldsymbol{Y}|\sigma^2, H_1)p(\sigma^2|H_1)d\sigma^2 \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \\
&\int \left(\frac{1}{\sigma^2}\right)^{\frac{n-p}{2}+1} e^{-\frac{1}{2\sigma^2}\left(\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2+\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}\right)} d\sigma^2 \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \\
&\Gamma\left(\frac{n-p}{2}\right)\left(\frac{\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2}{2} + \frac{\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}}{2}\right)^{-\frac{n-p}{2}}
\end{aligned} \tag{40}$$

where the last line follows from recognizing the kernel of an Inverse Gamma. Tidying the expression up yields

$$\begin{aligned}
p(\boldsymbol{Y}|H_1) = &\left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}}\Gamma\left(\frac{n-p}{2}\right) \\
&\left(\frac{s^2(\boldsymbol{Y})}{2}\right)^{-\frac{n-p}{2}} (n-p-d)^{-\frac{n-p}{2}} \left(1+\frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}}
\end{aligned} \tag{41}$$

The derivation for $p(\boldsymbol{Y}|H_0)$ proceeds similarly as $p(\boldsymbol{Y}|H_1)$, except for the marginalization over $\boldsymbol{\delta}$. Performing the marginalization with respect to $\boldsymbol{\beta}$ first yields

$$\begin{aligned}
p(\boldsymbol{Y}|\sigma^2, H_0) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\delta}}'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}} e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}\|_2^2} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{s^2(\boldsymbol{Y})(n-p-d)}{2\sigma^2}\left(1+\frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)}
\end{aligned} \tag{42}$$

Performing the marginalization finally with respect to σ^2 yields

$$\begin{aligned}
p(\boldsymbol{Y}|H_0) = &\left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \Gamma\left(\frac{n-p}{2}\right) \\
&\left(\frac{s^2(\boldsymbol{Y})}{2}\right)^{-\frac{n-p}{2}} (n-p-d)^{-\frac{n-p}{2}} \left(1+\frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}}
\end{aligned} \tag{43}$$

The Bayes factor (or "likelihood ratio mixture") is given by

$$\frac{p(\boldsymbol{Y}|H_1)}{p(\boldsymbol{Y}|H_0)} = \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \frac{\left(1+\frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}}}{\left(1+\frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}}} \tag{44}$$

□

A.3 Proof of Propositions 3.9 and 3.10

Proof. Starting with the model under H_0

$$\begin{aligned}
\boldsymbol{t}(\boldsymbol{Y}) = (\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y} &= \tilde{Z}\hat{\boldsymbol{\delta}}(\boldsymbol{Y}) \\
\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_0 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I) \\
\Rightarrow \boldsymbol{V}_d'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_0 &\sim N_d(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d) \\
\Rightarrow \boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0 &\sim t_{n-p-d}(\boldsymbol{0}, \boldsymbol{I}_d)
\end{aligned}$$

Therefore the density is

$$p(\boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0) = \frac{\Gamma(\frac{n-p}{2})}{\Gamma(\frac{n-p-d}{2})} \frac{1}{(n-p-d)^{\frac{d}{2}}\pi^{\frac{d}{2}}} \left(1 + \frac{\boldsymbol{t}(\boldsymbol{Y})'\boldsymbol{t}(\boldsymbol{Y})}{n-p-d}\right)^{\frac{n-p}{2}} \tag{45}$$

Now considering the model under H_1

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_1 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2(I + \boldsymbol{Z}\Phi^{-1}\boldsymbol{Z})) \\
\Rightarrow \boldsymbol{V}_d'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_1 &\sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_d'(I + \boldsymbol{Z}\Phi^{-1}\boldsymbol{Z}')\boldsymbol{V}_d) \\
\Rightarrow \boldsymbol{V}_d'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_1 &\sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{V}_d'\boldsymbol{V}_d + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}}')) \\
\Rightarrow \boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_1 &\sim t_{n-p-d}(\boldsymbol{0}, (\boldsymbol{I}_d + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}}')),
\end{aligned}$$

where the last line follows from $\boldsymbol{I}_d = \boldsymbol{V}_d'\boldsymbol{V}_d$. From the Sherman-Morrison-Woodbury Identity

$$(\boldsymbol{I}_d + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}})^{-1} = \boldsymbol{I}_d - \tilde{\boldsymbol{Z}}(\Phi + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'.$$

By the matrix determinant lemma

$$\frac{1}{\det(\boldsymbol{I} + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}}')} = \frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}$$

Therefore the density is

$$\begin{aligned} p(\boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0) = & \frac{\Gamma(\frac{n-p}{2})}{\Gamma(\frac{n-p-d}{2})} \frac{1}{(n-p-d)^{\frac{d}{2}}\pi^{\frac{d}{2}}} \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}} \\ & \left(1 + \frac{\boldsymbol{t}(\boldsymbol{Y})(I - \tilde{\boldsymbol{Z}}(\Phi + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}')\boldsymbol{t}(\boldsymbol{Y})}{n-p-d}\right)^{\frac{n-p}{2}} \end{aligned} \tag{46}$$

□

A.4 Group Invariant Bayes Factors for Linear Models

The main challenge in performing anytime-valid inference in linear models is the presence of nuisance parameters - we require time-uniform Type-I error and coverage guarantees for all possible values of the nuisance parameters $\boldsymbol{\beta}$ and σ^2. If a statistical model possesses a group structure, then invariance arguments can be effective in reducing a composite null hypothesis to a simple null hypothesis by constructing tests using a maximal invariant test statistic (Eaton 1989, Lehmann & Romano 2005, Wijsman 1990). To generate some intuition, suppose that instead of observing the sequence of raw observations, the statistician is provided instead with a sequence $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots$ of t-statistics. The statistician could perform a sequential probability ratio test using the sequence of t-statistics instead of the original sequence of observations. The advantage is that the sequence of t-statistics depends neither on the nuisance parameters $\boldsymbol{\beta}$ and σ^2 under the null nor under the alternative, and so a Type-I error guarantee is obtained regardless of the nuisance parameters. This holds in general for sequences of maximal invariant statistics, and sequential probability ratio tests constructed via these arguments are called *invariant SPRTs* (Lai 1981). A further convenience is that the maximal invariant statistic summarizes all information in the sequence, and so the

likelihood ratio of the most recent maximal invariant statistic can be used, instead of the whole sequence. For testing a composite alternative, a mixture of invariant SPRT's can be used. This is also equivalent to computing a Bayes-factor/mixture-martingale on the raw sequence of observations when using a particular prior/mixture, namely, the *right-Haar* prior/mixture on the nuisances parameters. Optional stopping behaviour of Bayes factors under group invariant models are studied in Hendriksen et al. (2021). Similarly, e-values for group invariant models are studied in Pérez-Ortiz et al. (2022).

A.4.1 Group Theory Preliminaries

To ease the notation we assume without loss of generality that $\boldsymbol{\delta}_0 = \mathbf{0}$ and simply write $B_n(\boldsymbol{Y}_n)$. It is helpful to reparameterize the linear model in terms of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\xi})$, where $\boldsymbol{\xi} = \boldsymbol{\delta}/\sigma$ are standardized coefficients. Let the parameters be denoted by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \boldsymbol{\xi}) \in \boldsymbol{\Theta}$, with $\boldsymbol{\Theta} = \mathbb{R}^p \times \mathbb{R}^+ \times \mathbb{R}^d$. The null parameter space $\boldsymbol{\Theta}_0 = \mathbb{R}^p \times \mathbb{R}^+ \times \{\mathbf{0}\}$ and $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. The model is invariant under the following transformations

$$\begin{aligned} g_{n,\boldsymbol{\alpha},c} &: \boldsymbol{Y}_n \mapsto c\boldsymbol{Y}_n + \boldsymbol{X}_n\boldsymbol{\alpha}, \\ \bar{g}_{\boldsymbol{\alpha},c} &: (\boldsymbol{\beta}, \sigma, \boldsymbol{\xi}) \mapsto (c\boldsymbol{\beta} + \boldsymbol{\alpha}, c\sigma.\boldsymbol{\xi}). \end{aligned} \tag{47}$$

In other words, the transformed observation $g_{n,\boldsymbol{\alpha},c}(\boldsymbol{Y})$ belongs to the same family of Gaussian linear models with transformed parameters $\bar{g}_{\boldsymbol{\alpha},c}(\boldsymbol{\theta})$. Let the group of transformations that act on the outcome and parameter space be denoted $G_n = \{g_{n,\boldsymbol{\alpha},c} : \alpha \in \mathbb{R}^p, c \in \mathbb{R}\}$ and $\bar{G} = \{\bar{g}_{\boldsymbol{\alpha},c} : \alpha \in \mathbb{R}^p, c \in \mathbb{R}\}$ respectively, noting that these are common to both the null and alternative hypotheses and leave $\boldsymbol{\xi}$ unchanged.

The *orbit* of $\boldsymbol{Y}_n$ is defined as $\mathcal{O}(\boldsymbol{Y}_n) = \{g(\boldsymbol{Y}_n), g \in G_n\}$. A function ϕ is G_n-invariant if $\lambda(\boldsymbol{Y}_n) = \lambda(g(\boldsymbol{Y}_n))$ for all $g \in G_n$, that is, it is constant on orbits. A *test function* is a function used to reject the null hypothesis when $\lambda(\boldsymbol{Y}) > c$. As an equivalent model is obtained under transformations (47), we should reasonably expect that a test function

is G_n-invariant. For instance, it should not matter if the units of $\boldsymbol{Y}_n$ are changed or if the component of $\boldsymbol{Y}_n$ in $\mathcal{C}(\boldsymbol{X}_n)$ is changed given we are testing a hypothesis about the component in $\mathcal{C}(\boldsymbol{W}_n) \setminus \mathcal{C}(\boldsymbol{X}_n)$. It is helpful to regard all elements of an orbit as carrying the same amount of evidence against the null hypothesis.

Definition A.3. *A maximal invariant function M is a function that is constant on orbits and takes distinct values on each orbit, that is, $M(\boldsymbol{Y}_{n,1}) = M(\boldsymbol{Y}_{n,2})$ implies $\boldsymbol{Y}_{n,1} = g(\boldsymbol{Y}_{n,2})$ for some $g \in G_n$.*

A maximal invariant statistic is simply a maximal invariant function of the data.

Lemma A.4. *A test function $\lambda(\boldsymbol{Y})$ is invariant if and only if it is a function of a maximal invariant statistic.*

Proof. ($\Rightarrow$) Assume $\lambda(\boldsymbol{Y}) = h(M(\boldsymbol{Y}))$ where M is a maximal invariant. For all $g \in G$, $\lambda(g(\boldsymbol{Y})) = h(M(g(\boldsymbol{Y})) = h(M(\boldsymbol{Y})) = \lambda(\boldsymbol{Y})$ and therefore ϕ is an invariant function.
($\Leftarrow$) Assume ϕ is invariant, then ϕ is a constant on orbits. The maximal invariant is also constant on orbits and takes a unique value on each orbit, by definition, hence their exists a surjective function that maps the values taken by the maximal invariant on orbits to the values taken by ϕ on orbits. □

A.4.2 Proof of Time-Uniform Guarantees

Theorem A.5. *The statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ is a maximal invariant statistic under G_n.*

Proof. We drop the n subscript for simplicity. We proceed by proving the contrapositive,

namely, if $\boldsymbol{Y}_1 \neq g(\boldsymbol{Y}_2)$ for all $g \in G$ then $\boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)$. Write

$$\boldsymbol{Y}_1 = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}_1 + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_1$$

$$\boldsymbol{Y}_2 = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}_2 + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_2$$

If $\boldsymbol{Y}_1 \neq g(\boldsymbol{Y}_2)$ for all $g \in G$ then we know

1. $\boldsymbol{Y}_1 \neq c\boldsymbol{Y}_2$ for all $c \in \mathbb{R}$
2. $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_1 \neq (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_2$

The latter must be true because if both vectors only differed by their component in $\mathcal{C}(\boldsymbol{X})$, then one could easily be expressed in terms of the other plus an appopriate term $\boldsymbol{X}\boldsymbol{\alpha}^\star$ for some $\boldsymbol{\alpha}^\star$. It must be the components in $\mathcal{C}(\boldsymbol{X})^\perp$ that are different. Let's take the component of each vector in $\mathcal{C}(\boldsymbol{X})^\perp$ and further decompose it into a component in $\mathcal{C}(\boldsymbol{W})$ and a component in $\mathcal{C}(\boldsymbol{W})^\perp$,

$$(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_i = (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_i + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_i, \tag{48}$$

for $i \in \{1, 2\}$. There are now three cases to consider

Case 1: $(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_1 \neq (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_2$ and $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_1 = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_2$.

Clearly $s^2(\boldsymbol{Y}_1) = s^2(\boldsymbol{Y}_2)$, but $\boldsymbol{Y}_1'(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_1 \neq \boldsymbol{Y}_2'(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_2$, which implies $\boldsymbol{t}(\boldsymbol{Y}_1)'\boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)'\boldsymbol{t}(\boldsymbol{Y}_2)$ which implies $\boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)$.

Case 2: $(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_1 = (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_2$ and $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_1 \neq (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_2$.

Clearly $s^2(\boldsymbol{Y}_1) \neq s^2(\boldsymbol{Y}_2)$, which implies $\boldsymbol{t}(\boldsymbol{Y}_1)'\boldsymbol{t}(\boldsymbol{Y}_1) = s^2(\boldsymbol{Y}_2)\boldsymbol{t}(\boldsymbol{Y}_2)'\boldsymbol{t}(\boldsymbol{Y}_2)/s^2(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)'\boldsymbol{t}(\boldsymbol{Y}_2)$ $\Rightarrow \boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)$

Case 3: $(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_1 \neq (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}_2$ and $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_1 \neq (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_2$.

Clearly $s^2(\boldsymbol{Y}_1) \neq s^2(\boldsymbol{Y}_2)$. Proof by contradiction. If $\boldsymbol{t}(\boldsymbol{Y}_1) = \boldsymbol{t}(\boldsymbol{Y}_2)$ then

$$\boldsymbol{Y}_1'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_1 = \frac{s^2(\boldsymbol{Y}_1)}{s^2(\boldsymbol{Y}_2)}\boldsymbol{Y}_2(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_2$$

which would imply $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_1 = (s(\boldsymbol{Y}_1)/s(\boldsymbol{Y}_2))(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}_2$, but this is a contradiction because $\boldsymbol{Y}_1 \neq c\boldsymbol{Y}_2$ for any c. □

Theorem A.5 provides the corollary that $B_n(\boldsymbol{Y}_n)$ is an invariant test function, as it is simply a function of a maximal invariant test statistic. Suppose instead of observing a sequence $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ of raw observations we observe a sequence $\boldsymbol{t}_1(\boldsymbol{Y}_1),\ \boldsymbol{t}_2(\boldsymbol{Y}_2), \ldots$ of multivariate t statistics instead. This involves no loss of information, at least from the Bayesian perspective, by proposition 3.10. Consider instead of $B_n(\boldsymbol{Y}_n)$, the Bayes factor based on the last t statistic, the Bayes factor based on the full sequence of t-statistics

$$A_n(\boldsymbol{Y}_n) := \frac{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)}, \tag{49}$$

which can be factored as

$$A_n(\boldsymbol{Y}_n) = \frac{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{t}_{n-1}(\boldsymbol{Y}_{n-1}), \ldots, \boldsymbol{t}_1(\boldsymbol{Y}_1), H_1)}{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{t}_{n-1}(\boldsymbol{Y}_{n-1}), \ldots, \boldsymbol{t}_1(\boldsymbol{Y}_1), H_0)}A_{n-1}(\boldsymbol{Y}_{n-1}). \tag{50}$$

Under the null hypothesis, $\mathbb{E}_{H_0}[A_n|\sigma(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_{n-1}(\boldsymbol{Y}_{n-1}))] = A_{n-1}$, and so the process A_n is a nonnegative supermartingale. By Ville's inequality (Ville 1939) for nonegative supermartingales

$$\mathbb{P}_{H_0}[\exists n \in \mathbb{N} : A_n(\boldsymbol{Y}_n) \geq \alpha^{-1}] \leq \alpha, \tag{51}$$

and so a time-uniform Type-I error guarantee could be obtained using the statistic $A_n(\boldsymbol{Y}_n)$. The following theorem, however, shows that the Bayes factor obtained from observing the full sequence of t statistics is equal to the Bayes factor obtained from observing only the last t-statistic, which in turn is equal to the Bayes factor obtained from observing the full sequence of raw observations by proposition 3.10. First, however, we need a lemma.

Lemma A.6. *The statistics $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ $(i \leq n)$ can be written as a functions of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$.*

In other words, knowledge of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ implies knowledge of $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ for all $i < n$ also.

Proof. To see this note that $\boldsymbol{Y}_i = \boldsymbol{P}_{in}\boldsymbol{Y}_n$ where $\boldsymbol{P}_{in}$ is the projection from $\mathbb{R}^n$ to $\mathbb{R}^i$ obtained by retaining only the first i elements of the vector $\boldsymbol{Y}_n$. Then $\boldsymbol{t}_i(\boldsymbol{Y}_i) = \boldsymbol{t}_i(\boldsymbol{P}_{in}\boldsymbol{Y}_n)$, which we write $\boldsymbol{t}_i(\boldsymbol{Y}_i) = \boldsymbol{u}_{in}(\boldsymbol{Y}_n)$. Each function $\boldsymbol{u}_{in}$ is i) a function of $\boldsymbol{Y}_n$ that is also ii) invariant under transformations $\boldsymbol{Y}_n \to c\boldsymbol{Y}_n + \boldsymbol{X}_n\boldsymbol{\alpha}$. It follows from lemma A.4 that because each $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ is a G_n-invariant function of $\boldsymbol{Y}_n$, it can be written as a function of the maximal invariant $\boldsymbol{t}_n(\boldsymbol{Y}_n)$. □

Theorem A.7.

$$B_n(\boldsymbol{Y}_n) = \frac{p(\boldsymbol{Y}_n|H_1)}{p(\boldsymbol{Y}_n|H_0)} = \frac{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} = \frac{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} = A_n(\boldsymbol{Y}_n) \tag{52}$$

Proof. The second equality follows from proposition 3.10. Lemma A.6 implies that each $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ can be written as functions of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ for $i \leq n$. This implies the third equality $p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_i) = p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_i)$, □

Theorem A.7 leverages the fact that $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ is a maximal invariant test statistic and, from lemma A.4, a statistic is invariant if any only if it is a function of a maximal invariant test statistic. The statistics $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ for $i \leq n$ are G_n-invariant, and must therefore be functions of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$. Equating $B_n(\boldsymbol{Y}_n)$ and $A_n(\boldsymbol{Y}_n)$, together with the time-uniform bound in equation 51 yields the time-uniform bound in theorem 3.3

$$\mathbb{P}_\theta[\exists n \in \mathbb{N} : B_n(\boldsymbol{Y}_n) \geq \alpha^{-1}] \leq \alpha. \tag{53}$$

A.5 Proof of Corollary 3.8

Proof. Let $r_n = \frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}$

$$\sqrt{r_n}\frac{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'(\boldsymbol{I}_n-\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}\right)^{-\frac{\nu_n+d}{2}}}{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}\right)^{-\frac{\nu_n+d}{2}}} < \alpha^{-1}$$

$$\Rightarrow\sqrt{r_n}\alpha < \frac{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'(\boldsymbol{I}_n-\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}\right)^{\frac{\nu_n+d}{2}}}{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}\right)^{\frac{\nu_n+d}{2}}}$$

$$\Rightarrow(r_n\alpha^2)^{\frac{1}{\nu_n+d}} < \frac{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'(\boldsymbol{I}_n-\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}\right)}{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}\right)}$$

Let $k_n = (r_n\alpha^2)^{\frac{1}{\nu_n+d}}$, then we have

$$k_n + k_n\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n} < 1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'(\boldsymbol{I}_n-\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n}$$

$$\Rightarrow\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n} + (k_n-1)\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})}{\nu_n} < 1-k_n$$

$$\Rightarrow\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) + (k_n-1)\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta})'\boldsymbol{t}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) < \nu_n(1-k_n)$$

□

A.6 Proof of Theorem 5.2

One may recognize that $B_n(\boldsymbol{Y}_n)$ resembles the ratio of two t densities whereas $\tilde{B}_n(\boldsymbol{Y}_n)$ resembles the ratio of two Gaussian densities. To sketch the proof, our goal is to show that, under the null hypothesis, $\log B_n(\boldsymbol{Y}_n) = \log\tilde{B}_n(\boldsymbol{Y}_n) + o_{a.s.}(1)$, which implies that $B_n(\boldsymbol{Y}_n) = \tilde{B}_n(\boldsymbol{Y}_n)(1+o_{a.s.}(1))$ i.e. $B_n(\boldsymbol{Y}_n) \overset{a.s.}{\sim} \tilde{B}_n(\boldsymbol{Y}_n)$. Working on the log-scale, the path forward is to exploit the series expansion $\log(1+x) = \sum_{i=1}^{\infty}(-1)^{(i-1)}\frac{1}{i}x^i$ for $|x|<1$ and retain the leading orders, where $x = \frac{1}{\nu_n}\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)$ in the denominator and $x = \frac{1}{\nu_n}\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n-\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n)$ in the numerator. We will need both of these terms to converge to zero almost surely. This is not always true under the alternative, but

is true under the null under certain conditions. While this may occur under more general conditions, we prove it under Gauss-Markov assumptions.

Proof. $\frac{1}{n}\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i'\varepsilon_i \to \mathbf{0}$ almost surely by the strong law because $\mathbb{E}[\boldsymbol{w}_i\varepsilon_i] = \mathbf{0}$. By positive definitness of $\boldsymbol{\Omega}_{\boldsymbol{w}}$ and the continuous mapping theorem, $(\boldsymbol{W}_n'\boldsymbol{W}_n/n)^{-1} \to \boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}$ almost surely. By definition of the OLS estimator $\hat{\boldsymbol{\gamma}}_n^{ols} = \boldsymbol{\gamma} + (\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}nn^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon} \overset{a.s.}{\to} \boldsymbol{\gamma} + \boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\mathbf{0} = \boldsymbol{\gamma}$. For strong consistency of σ^2, $\hat{\sigma}_n^2 = (\boldsymbol{\varepsilon}_n'\boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n'\boldsymbol{W}_n(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n)/\nu_n$. Considering the last term, $\boldsymbol{\varepsilon}_n'\boldsymbol{W}_n(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n/\nu_n = \boldsymbol{\varepsilon}_n'\boldsymbol{W}_n n^{-1}n(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}n^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n(n/\nu_n) \overset{a.s.}{\to} \mathbf{0}$ while $\boldsymbol{\varepsilon}_n'\boldsymbol{\varepsilon}_n/\nu_n \overset{a.s.}{\to} \sigma^2$ by the strong law ($\mathbb{E}[\varepsilon_i] = \sigma^2$). In addition

$$\frac{1}{n}\boldsymbol{W}_n'\boldsymbol{W}_n = \frac{1}{n}\begin{pmatrix} \boldsymbol{X}_n'\boldsymbol{X}_n & \boldsymbol{X}_n'\boldsymbol{Z}_n \\ \boldsymbol{Z}_n'\boldsymbol{X}_n & \boldsymbol{Z}_n'\boldsymbol{Z}_n \end{pmatrix} \overset{a.s.}{\to} \begin{pmatrix} \boldsymbol{\Omega}_{\boldsymbol{X}} & \boldsymbol{\Omega}_{\boldsymbol{XZ}} \\ \boldsymbol{\Omega}_{\boldsymbol{ZX}} & \boldsymbol{\Omega}_{\boldsymbol{Z}} \end{pmatrix} = \boldsymbol{\Omega}_{\boldsymbol{w}},$$

for positive definite $\boldsymbol{\Omega}_{\boldsymbol{w}}$, $((1/n)\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1} \overset{a.s.}{\to} \boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}$ then $\frac{1}{n}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n \overset{a.s.}{\to} \boldsymbol{\Omega}_{\boldsymbol{Z}} - \boldsymbol{\Omega}_{\boldsymbol{ZX}}\boldsymbol{\Omega}_{\boldsymbol{X}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{XZ}}$. Consider first the term in the denominator of $B_n(\boldsymbol{Y}_n)$,

$$\begin{aligned}\frac{1}{n}\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n) &= \frac{1}{n}\frac{(\hat{\boldsymbol{\delta}}_n^{ols} - \boldsymbol{\delta}_0)\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\hat{\boldsymbol{\delta}}_n^{ols} - \boldsymbol{\delta}_0)}{\hat{\sigma}_n^2} \\ &\overset{a.s.}{\to} \frac{(\boldsymbol{\delta} - \boldsymbol{\delta}_0)(\boldsymbol{\Omega}_{\boldsymbol{Z}} - \boldsymbol{\Omega}_{\boldsymbol{ZX}}\boldsymbol{\Omega}_{\boldsymbol{X}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{XZ}})(\boldsymbol{\delta} - \boldsymbol{\delta}_0)}{\sigma^2}\end{aligned}$$

which implies $\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)/\nu_n \overset{a.s.}{\to} \textit{const.}$ as $n \to \infty$. When the null is true, $\boldsymbol{\delta} = \boldsymbol{\delta}_0$, this term converges to zero almost surely.

Now consider the second term in the denominator of $B_n(\boldsymbol{Y}_n)$. For large enough n, the Neumann series representation of the inverse is

$$\begin{aligned}(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1} &= \sum_{i=0}^{\infty}(-1)^i((\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{\Phi})^i(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1} \\ &= (\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1} - (\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{\Phi}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1} + \sum_{i=2}^{\infty}(-1)^i((\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{\Phi})^i(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\end{aligned}$$

which implies

$$\begin{aligned}&\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n - \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n \\ &= \boldsymbol{\Phi} + \sum_{i=2}^{\infty}(-1)^{i+1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n((\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{\Phi})^i\end{aligned}$$

and

$$\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n} = \frac{(\hat{\boldsymbol{\delta}}_n^{ols} - \boldsymbol{\delta}_0)'\boldsymbol{\Phi}(\hat{\boldsymbol{\delta}}_n^{ols} - \boldsymbol{\delta}_0)}{\nu_n\sigma^2} + r_n \qquad (54)$$

where $r_n = \frac{1}{\nu_n}\sum_{i=2}^{\infty}(-1)^{i+1}(\hat{\boldsymbol{\delta}}_n^{ols} - \boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n((\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{\Phi})^i(\hat{\boldsymbol{\delta}}_n^{ols} - \boldsymbol{\delta}_0)$. Both terms in the right hand side of (54) converge to zero almost surely.

We can now apply the series expansion of the logarithm

$$\begin{aligned}
-\frac{\nu_n + d}{2}\log\left(1 + \frac{x}{\nu_n}\right) &= -\frac{\nu_n + d}{2}\left(\sum_{i=1}^{\infty}(-1)^{(i-1)}\frac{1}{i}\left(\frac{x}{\nu_n}\right)^i\right) \\
&= -\frac{1}{2}\frac{\nu_n + d}{\nu_n}x - \frac{\nu_n + d}{2}\left(\sum_{i=2}^{\infty}(-1)^{(i-1)}\frac{1}{i}\left(\frac{x}{\nu_n}\right)^i\right) \\
&= -\frac{1}{2}x - \frac{d}{2\nu_n}x - \frac{\nu_n + d}{2}\left(\sum_{i=2}^{\infty}(-1)^{(i-1)}\frac{1}{i}\left(\frac{x}{\nu_n}\right)^i\right) \\
&= -\frac{1}{2}x - \frac{d}{2\nu_n}x + \frac{x^2}{4\nu_n} + \frac{dx^2}{4\nu_n^2} - \frac{\nu_n + d}{2}\left(\sum_{i=3}^{\infty}(-1)^{(i-1)}\frac{1}{i}\left(\frac{x}{\nu_n}\right)^i\right) \\
&= -\frac{1}{2}x + \frac{1}{2}\left(\frac{x}{2} - d\right)\frac{x}{\nu_n} - r_n
\end{aligned}$$

For the term in the numerator we have

$$\begin{aligned}
&-\frac{\nu_n + d}{2}\log\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n}\right) = \\
&-\frac{1}{2}\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n) + o_{a.s.}(1)
\end{aligned}$$

For the term in the denominator we have

$$\begin{aligned}
&-\frac{\nu_n + d}{2}\log\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n}\right) = \\
&-\frac{1}{2}\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n) + o_{a.s.}(1).
\end{aligned}$$

Combining these expressions yields

$$\begin{aligned}
&-\frac{\nu_n + d}{2}\left(\log\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n}\right) - \log\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)}{\nu_n}\right)\right) \\
&= -\frac{1}{2}\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n) + \frac{1}{2}\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n) + o_{a.s.}(1) \\
&= \frac{1}{2}\boldsymbol{t}_n(\boldsymbol{Y}_n)'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n'\boldsymbol{t}_n(\boldsymbol{Y}_n) + o_{a.s.}(1).
\end{aligned}$$

Therefore

$$\log B_n(\boldsymbol{Y}_n) = \log \tilde{B}_n(\boldsymbol{Y}) + o_{a.s.}(1).$$

This implies that $B_n(\boldsymbol{Y}_n) = \tilde{B}_n(\boldsymbol{Y}) e^{o_{a.s.}(1)} = \tilde{B}_n(\boldsymbol{Y})(1 + o_{a.s.}(1))$, i.e. $B_n(\boldsymbol{Y}_n) \sim \tilde{B}_n(\boldsymbol{Y})$ almost surely. □

A.7 Proof of Corollary 5.3

The asymptotic confidence sequence $\tilde{C}_n(\boldsymbol{Y}_n; \alpha)$ can be obtained by inverting $\tilde{B}_n(\boldsymbol{Y}; \boldsymbol{\delta}_0)$ to find all the nulls $\boldsymbol{\delta}_0$ not rejected by the data at the α level. This follows immediately from theorem 5.2. We provide, for completeness, an alternative proof starting from the definition of $C_n(\boldsymbol{Y}_n; \alpha)$. As $\frac{1}{n}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n \overset{a.s.}{\to} \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$, we can define a remainder $\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\boldsymbol{R}_n = \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n - n\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$ where $\boldsymbol{R}_n = o_{a.s.}(n)$ (we could have equally defined the remainder term as $\boldsymbol{R}_n = \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n - n\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$, yet our earlier definition simplifies the following expressions). Recall the definition of k_n from section A.5

$$k_n = \left(\frac{\alpha^2 \det \boldsymbol{\Phi}}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}\right)^{\frac{1}{\nu_n + d}}.$$

Let

$$e_n = \log k_n = \frac{1}{\nu_n + d} \log\left(\frac{\alpha^2 \det \boldsymbol{\Phi}}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}\right)$$

This can be approximated for large n to leading order as

$$\begin{aligned}
e_n &= \frac{\log\det \alpha^2\boldsymbol{\Phi}}{\nu_n + d} - \frac{\log\left(\det(\boldsymbol{\Phi} + \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}(n\boldsymbol{I}_d + \boldsymbol{R}_n))\right)}{\nu_n + d} \\
&= \frac{\log\det \alpha^2\boldsymbol{\Phi}}{\nu_n + d} - \frac{\log\left(\det(\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}(n\boldsymbol{I}_d + \boldsymbol{R}_n))\right)}{\nu_n + d} + \frac{\log\left(\det(\boldsymbol{I} + \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}^{-1}(n\boldsymbol{I}_d + \boldsymbol{R}_n)^{-1}\boldsymbol{\Phi})\right)}{\nu_n + d} \\
&= \frac{\log\det \alpha^2\boldsymbol{\Phi}}{\nu_n + d} - \frac{\log\det(\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})}{\nu_n + d} - \frac{\log\det(n\boldsymbol{I}_d + \boldsymbol{R}_n)}{\nu_n + d} + \frac{1}{\nu_n + d}\sum_{i=1} \frac{Tr((\boldsymbol{\Omega}_{\boldsymbol{z}}^{-1}(n\boldsymbol{I}_d + \boldsymbol{R}_n)^{-1}\boldsymbol{\Phi})^i)}{i} \\
&= \frac{\log\det \alpha^2\boldsymbol{\Phi}\boldsymbol{\Omega}_{\boldsymbol{z}}^{-1}}{\nu_n + d} - \frac{\log\det n\boldsymbol{I}_d}{\nu + d} - \frac{\log\det(\boldsymbol{I} + \boldsymbol{R}_n/n)}{\nu_n + d} + \frac{1}{\nu_n + d}\sum_{i=1} \frac{Tr((\boldsymbol{\Omega}_{\boldsymbol{z}}^{-1}(\boldsymbol{I}_d + \boldsymbol{R}_n/n)^{-1}\boldsymbol{\Phi})^i)}{ni} \\
&= -\frac{d\log n}{\nu_n + d} + O_{a.s.}(n^{-1})
\end{aligned}$$

Hence $e_n = O_{a.s.}(\log(n)/n)$. Consequently

$$k_n = \left(\frac{\alpha^2 \det \boldsymbol{\Phi}}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)}\right)^{\frac{1}{\nu_n + d}} = e^{e_n} = 1 + \frac{1}{\nu_n + d} \log\left(\frac{\alpha^2 \det \boldsymbol{\Phi}}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)}\right) + O_{a.s.}(e_n^2)$$

and similarly

$$(1 - k_n) = \frac{1}{\nu_n + d} \log\left(\frac{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)}{\alpha^2 \det \boldsymbol{\Phi}}\right) + O_{a.s.}(e_n^2) = O_{a.s.}(\log(n)/n)$$

The term $\nu_n(1 - k_n)$ appears in the right hand side of the inequality defining $C_n(\boldsymbol{Y}_n; \alpha)$ in equation (5). From the above, $\nu_n(1-k_n) = \log\left(\frac{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}{\alpha^2 \det \boldsymbol{\Phi}}\right) + O_{a.s.}\left(\log(n)^2/n\right)$, where the first term can be recognized as the term in the right hand side of the inequality defining $\tilde{C}_n(\boldsymbol{Y}_n; \alpha)$ in equation (15). On the left hand side of the inequality defining $C_n(\boldsymbol{Y}_n; \alpha)$ we have

$$\begin{aligned}
&\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{D}_n} \left(1 + (k_n - 1)\frac{\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n\|^2_{\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n}}{\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{D}_n}}\right) \\
&= \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{D}_n} \left(1 + O_{a.s.}(\log(n)/n)\right),
\end{aligned}$$

where $\boldsymbol{D}_n = \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n (\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)^{-1} \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n$. We can therefore express $C_n(\boldsymbol{Y}_n; \alpha)$ as

$$C_n(\boldsymbol{Y}_n; \alpha) = \left\{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{D}_n} \leq s_n^2(\boldsymbol{Y}_n) \log\left(\frac{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)}{\alpha^2 \det \boldsymbol{\Phi}}\right) + O_{a.s.}\left(\log(n)^2/n\right)\right\},$$

and thus both $C_n(\boldsymbol{Y}_n; \alpha)$ and $\tilde{C}_n(\boldsymbol{Y}_n; \alpha)$ can be expressed for large enough n as quadratic sets defined by an upper bound on $\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)\|^2_{\boldsymbol{D}_n}$. As n goes to infinity, the upper bounds used to define $C_n(\boldsymbol{Y}_n; \alpha)$ and $\tilde{C}_n(\boldsymbol{Y}_n; \alpha)$ converge.

A.8 Proof of Theorem 5.4

As $B_n \sim \tilde{B}_n$ it suffices to show $\tilde{B}_n \sim B_n^\infty$. As $\frac{1}{n}\tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n \overset{a.s.}{\to} \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$, we can define a remainder $\boldsymbol{\Omega}_{\tilde{z}} \boldsymbol{R}_n = \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n - n\boldsymbol{\Omega}_{\tilde{z}}$ where $\boldsymbol{R}_n = o_{a.s.}(n)$ (we could have equally defined the remainder

term as $\boldsymbol{R}_n = \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n - n\boldsymbol{\Omega}_{\tilde{z}}$, yet our earlier definition simplifies the following expressions). Our goal is to replace the expressions involving $\boldsymbol{Z}_n'\boldsymbol{Z}_n$ with expressions in terms of $\boldsymbol{\Omega}_{\tilde{\boldsymbol{Z}}}$ and $\boldsymbol{R}_n$ and retain leading order terms. Starting with the determinants

$$\frac{\det \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{\det(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})} = \frac{\det \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{\det((n+\lambda)\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}} + \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\boldsymbol{R}_n)} = \left(\frac{\lambda}{\lambda+n}\right)^d \frac{1}{\det(\boldsymbol{I}_d + (n+\lambda)^{-1}\boldsymbol{R}_n)}$$

$$\begin{aligned}
\Rightarrow \frac{1}{2}\log\left(\frac{\det \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{\det(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})}\right) &= \frac{d}{2}\log\left(\frac{\lambda}{\lambda+n}\right) - \frac{d}{2}\log\det\left(\boldsymbol{I}_d + \frac{1}{n+\lambda}\boldsymbol{R}_n\right) \\
&= \frac{d}{2}\log\left(\frac{\lambda}{\lambda+n}\right) - \frac{d}{2}\frac{Tr(\boldsymbol{R}_n)}{n+\lambda} - \frac{d}{2}\sum_{i=1}^{\infty}(-1)^{i+1}\frac{r_n^i}{i(n+\lambda)^i} \\
&= \frac{d}{2}\log\left(\frac{\lambda}{\lambda+n}\right) - \frac{d}{2}\frac{Tr(\boldsymbol{R}_n)}{n+\lambda} + O\left(\frac{Tr(\boldsymbol{R}_n^2)}{n^2}\right).
\end{aligned}$$

In the exponent

$$\begin{aligned}
\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}} + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}} &= (n\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}} + \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\boldsymbol{R}_n)(n\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}} + \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\boldsymbol{R}_n + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})^{-1}(n\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}} + \boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\boldsymbol{R}_n) \\
&= \frac{n^2}{n+\lambda}\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}(\boldsymbol{I} + \frac{1}{n}\boldsymbol{R}_n)(\boldsymbol{I} + \frac{1}{n+\lambda}\boldsymbol{R}_n)^{-1}(\boldsymbol{I} + \frac{1}{n}\boldsymbol{R}_n) \\
&= \frac{n^2}{n+\lambda}\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}\left(\boldsymbol{I} + \left(\frac{2}{n} - \frac{1}{n+\lambda}\right)\boldsymbol{R}_n + O\left(\frac{\boldsymbol{R}_n^2}{n^2}\right)\right),
\end{aligned}$$

where the last line follows from the Neumann series $(\boldsymbol{I}+\frac{1}{n+\lambda}\boldsymbol{R}_n)^{-1} = \sum_{i=0}^{\infty}(-1)^i\boldsymbol{R}_n^i/(n+\lambda)^i$.

Therefore

$$\begin{aligned}
\log\tilde{B}_n(\boldsymbol{Y}_n) - \log B_n^{\infty}(\boldsymbol{Y}_n) &\sim -\frac{d}{2}\frac{Tr(\boldsymbol{R}_n)}{\lambda+n} + \frac{1}{2}\frac{n}{n+\lambda}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0)'\frac{n\boldsymbol{\Omega}_z\boldsymbol{R}_n}{s_n^2}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0)\left(\frac{2}{n} - \frac{1}{n+\lambda}\right) \\
&= o_{a.s.}(1),
\end{aligned}$$

which implies $\tilde{B}_n(\boldsymbol{Y}_n) = B_n^{\infty}(\boldsymbol{Y}_n)(1 + o_{a.s.}(1))$

A.9 Proof of Corollary 5.5

$$\mathbb{P}_{\boldsymbol{\delta}}[\forall n \in \mathbb{N} : \log(B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) < \log(\alpha^{-1})] \geq 1 - \alpha$$

$$\mathbb{P}_{\boldsymbol{\delta}}[\forall n \in \mathbb{N} : \log(B_n(\boldsymbol{Y}_n;\boldsymbol{\delta})) < \log(\alpha^{-1})] = \mathbb{P}_{\boldsymbol{\delta}}[\forall n \in \mathbb{N} : \log(B_n^{\infty}(\boldsymbol{Y}_n;\boldsymbol{\delta})) < \log(\alpha^{-1}) + e_n]$$

where $e_n = o_{a.s.}(1)$.

$$\begin{aligned}
&\frac{1}{2}\log\left(\frac{\lambda}{\lambda+n}\right) + \frac{1}{2}\frac{n^2}{n+\lambda}\frac{\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{s_n^2(\boldsymbol{Y}_n)}(\hat{\delta}_n^{ols} - \delta)^2 < \frac{1}{2}\log(\alpha^{-2}) + o_{a.s.}(1) \\
\Rightarrow &\frac{n^2}{n+\lambda}\frac{\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{s_n^2(\boldsymbol{Y}_n)}(\hat{\delta}_n^{ols} - \delta)^2 < \log\left(\frac{\lambda+n}{n\alpha^2}\right) + o_{a.s.}(1) \\
\Rightarrow &\frac{\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{s_n^2(\boldsymbol{Y}_n)}(\hat{\delta}_n^{ols} - \delta)^2 < \frac{n+\lambda}{n^2}\log\left(\frac{\lambda+n}{n\alpha^2}\right) + o_{a.s.}(n^{-1}) \\
\Rightarrow &|\hat{\delta}_n^{ols} - \delta| < \sqrt{\frac{s_n^2(\boldsymbol{Y}_n)}{\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}\left(\frac{n+\lambda}{n^2}\right)\log\left(\frac{\lambda+n}{n\alpha^2}\right)} + o_{a.s.}(n^{-\frac{1}{2}})
\end{aligned}$$

A.10 Proof of Theorem 5.6

From the proof of theorem 5.4, we have already shown that

$$\frac{1}{2}\log\left(\frac{\det\lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}}{\det(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}} + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})}\right) = \frac{d}{2}\log\left(\frac{\lambda}{\lambda+n}\right) + o_{a.s.}(1)$$

Define $\boldsymbol{R}_n = \tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n - n\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}}$, noting $\boldsymbol{R}_n = o_{a.s.}(n)$

$$(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})^{-1} = \frac{n}{n+\lambda}\left(\boldsymbol{I}_n + \frac{\lambda}{\lambda+n}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{R}_n\right)^{-1}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}$$

For large enough n, $\|\frac{\lambda}{\lambda+n}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{R}_n\| < 1$, and we can employ a Neumann series to express

$$\begin{aligned}
\left(\boldsymbol{I}_n + \frac{\lambda}{\lambda+n}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{R}_n\right)^{-1} &= \sum_{k=0}^{\infty}(-1)^k\left(\frac{\lambda}{\lambda+n}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{R}_n\right)^k \\
&= \boldsymbol{I}_n - \frac{\lambda}{\lambda+n}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{R}_n + O_{a.s.}(n^{-2}),
\end{aligned}$$

This implies

$$(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n + \lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})^{-1} = \frac{n}{n+\lambda}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n^{-1} - \frac{n\lambda}{(\lambda+n)^2}(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\boldsymbol{R}_n(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1} + O_{a.s.}(n^{-3})$$

and therefore

$$\begin{aligned}
\|\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0\|^2_{\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n+\lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n} &= \frac{n}{n+\lambda}\|\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0\|^2_{\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n} \\
&\quad - \frac{n\lambda}{(\lambda+n)^2}\|\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0\|^2_{\boldsymbol{R}_n} \\
&\quad + O_{a.s.}(n^{-1})
\end{aligned}$$

The second and third terms are $o_{a.s.}(1)$ Finally

$$d\frac{(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n+\lambda\boldsymbol{\Omega}_{\tilde{\boldsymbol{z}}})^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)}{ds_n^2(\boldsymbol{Y}_n)}=d\frac{n}{n+\lambda}f(\boldsymbol{Y}_n)+o_{a.s.}(1)$$

which implies $\log\tilde{B}_n(\boldsymbol{Y}_n)-\log B_n^g(\boldsymbol{Y}_n)=o_{a.s.}(1)$ and, consequently, $\tilde{B}_n(\boldsymbol{Y}_n)=B_n^g(\boldsymbol{Y}_n)(1+o_{a.s.}(1))$.

A.11 Asymptotic Anytime-Valid Causal Inference

Define the population least squares values as

$$(\alpha^\star,\boldsymbol{\kappa}^\star,\tau^\star,\boldsymbol{\eta}^\star)=\operatorname{argmin}\mathbb{E}[(y_i-\alpha-\boldsymbol{m}_i'\boldsymbol{\kappa}-(T_i-\rho)\tau+T_i(\boldsymbol{m}-\boldsymbol{\mu}_{\boldsymbol{m}})\boldsymbol{\eta})^2] \tag{55}$$

In appendix section A.12 it is shown that

$$\begin{aligned}
\alpha^\star&=\mathbb{E}[y_i]=\rho\mathbb{E}[y_i(1)]+(1-\rho)\mathbb{E}[y_i(0)],\\
\tau^\star&=\mathbb{E}[y_i(1)]-\mathbb{E}[y_i(0)],\\
\boldsymbol{\eta}^\star&=\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\mathbb{E}[(\boldsymbol{m}_i-\boldsymbol{\mu}_{\boldsymbol{m}})\mathbb{E}[y_i(1)-y_i(0)|\boldsymbol{m}_i]],\\
\boldsymbol{\kappa}^\star&=\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\mathbb{E}[(\boldsymbol{m}_i-\boldsymbol{\mu}_{\boldsymbol{m}})y_i],
\end{aligned} \tag{56}$$

where $\boldsymbol{\Omega}_{\boldsymbol{m}}=\mathbb{E}[(\boldsymbol{m}_i-\boldsymbol{\mu}_{\boldsymbol{m}})(\boldsymbol{m}_i-\boldsymbol{\mu}_{\boldsymbol{m}})']$. In particular, if the average treatment effect is zero then $\tau^\star=0$, and if $\mathbb{E}[y_i(1)-y_i(0)|\boldsymbol{m}_i=\boldsymbol{m}]=0$ for all $\boldsymbol{m}$ then $\boldsymbol{\eta}^\star=\boldsymbol{0}$.

Let $\boldsymbol{\gamma}=(\alpha,\boldsymbol{\kappa},\tau,\boldsymbol{\eta})$ and denote the sample and population least squares estimates as $\hat{\boldsymbol{\gamma}}^{ols}$ and $\boldsymbol{\gamma}^\star$. In addition, let $\boldsymbol{w}_i=[1,\boldsymbol{m}_i-\boldsymbol{\mu}_{\boldsymbol{m}},T_i-\rho,T_i(\boldsymbol{m}_i-\boldsymbol{\mu}_{\boldsymbol{m}})]$ denote the i'th row of the combined design matrix $\boldsymbol{W}_n$. Note that the difference is $\hat{\boldsymbol{\gamma}}_n^{ols}-\boldsymbol{\gamma}^\star=\boldsymbol{W}_n'\boldsymbol{W}_n\sum\boldsymbol{w}_i'y_i-\boldsymbol{w}_i\boldsymbol{w}_i'\boldsymbol{\gamma}^\star$. Focussing first on the summand, we show that this is a sum of independent zero mean random variables.

Proposition A.8. *Each term has zero mean*

$$\mathbb{E}_{sp}[\boldsymbol{w}_iy_i-\boldsymbol{w}_i\boldsymbol{w}_i\boldsymbol{\gamma}^\star]=0,$$

and variance $\boldsymbol{\Delta} := \mathbb{V}[\boldsymbol{w}_i(y_i - \boldsymbol{w}_i'\boldsymbol{\gamma}^\star)] = \mathbb{E}[(y_i - \boldsymbol{w}_i\boldsymbol{\gamma}^\star)^2\boldsymbol{w}_i\boldsymbol{w}_i']$.

The significance of proposition A.8 is that strong invariance principles can be employed to approximate the sum $\sum \boldsymbol{w}_i'y_i - \boldsymbol{w}_i\boldsymbol{w}_i'\boldsymbol{\gamma}^\star$ by a sum of $N(\mathbf{0}, \boldsymbol{\Delta})$ random variables. This strategy was first introduced to the literature on anytime-valid inference by Waudby-Smith et al. (2021) who leveraged a strong approximation due to Strassen (1964) to obtain asymptotic confidence sequences. The following provides a multivariate extension below

Lemma A.9. *Let* $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, ... *be a sequence of i.i.d. random variables with mean* $\mathbf{0}$ *and covariance* $\boldsymbol{V}$*, then, after suitably enriching the probability space,*

$$\sum_i^n \boldsymbol{y}_i = \sum_i^n \boldsymbol{g}_i + \boldsymbol{e}_n \tag{57}$$

where $\boldsymbol{g}_i \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{V})$ *and* $\boldsymbol{e}_n = o_{a.s.}\left(\sqrt{n \log\log n}\right)$

The proof is given in appendix section A.14. For our use case, it may be more relevant to write $\bar{y}_n = \bar{g}_n + \boldsymbol{r}_n$ where $\bar{g}_n \sim N(\mathbf{0}, \boldsymbol{V}/n)$ and $\boldsymbol{r}_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$. So far we have been able to approximate the sum $\sum \boldsymbol{w}_i'y_i - \boldsymbol{w}_i\boldsymbol{w}_i'\boldsymbol{\gamma}^\star$, which is half-way toward our end goal of approximating $\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star$.

Lemma A.10. *Assume* $\frac{1}{n}\boldsymbol{W}_n\boldsymbol{W}_n = \frac{1}{n}\sum \boldsymbol{w}_i\boldsymbol{w}_i' \to \boldsymbol{\Omega}_{\boldsymbol{w}}$ *almost surely with* $\boldsymbol{\Omega}_{\boldsymbol{w}}$ *positive definite, then*

$$\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star = \frac{1}{n}\sum_i^n \boldsymbol{g}_i + \boldsymbol{e}_n \tag{58}$$

where $\boldsymbol{e}_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$ *and* $\boldsymbol{g}_i$ *are independent* $N(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-T})$ *random variables with*

$$\boldsymbol{\Omega}_{\boldsymbol{w}} = \mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_i'] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Omega}_m & 0 & \rho\boldsymbol{\Omega}_n \\ 0 & 0 & \rho(1-\rho) & 0 \\ 0 & \rho\boldsymbol{\Omega}_m & 0 & \rho\boldsymbol{\Omega}_m \end{pmatrix} \tag{59}$$

The proof is given in appendix section A.15. This lemma establishes that the difference between $\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star$ can be strongly approximated by a sum of i.i.d. multivariate Gaussian random variables. If a time-uniform bound is available for the latter, then this can be translated over to the former. The aforementioned time-uniform bound is provided in the following lemma

Lemma A.11. *Let $\{\boldsymbol{g}_i\}_{i=1}^\infty$ be an i.i.d. sequence of multivariate $N_d(\mathbf{0}_d, \boldsymbol{\Sigma})$ random variables, then for any prespecified $\lambda > 0$*

$$\mathbb{P}\left[\forall n \in \mathbb{N} : \|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}} \leq \sqrt{\frac{1}{n}}\sqrt{\frac{\lambda+n}{n}\log\left(\left(\frac{\lambda+n}{\lambda}\right)^d \frac{1}{\alpha^2}\right)}\right] \geq 1-\alpha \tag{60}$$

Theorem A.12. *Let $\{\boldsymbol{g}_i\}_{i=1}^\infty$ be an i.i.d. sequence of d-dimensional random vectors with mean $\mathbf{0}_d$ and covariance $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$, and let $\hat{\boldsymbol{\Sigma}}_n$ be a strongly consistent estimator for $\boldsymbol{\Sigma}$. For any prespecified $\lambda > 0$*

$$\mathbb{P}\left[\forall n \in \mathbb{N} : \|\bar{\boldsymbol{g}}_n\|_{\hat{\boldsymbol{\Sigma}}_n^{-1}} \leq \sqrt{\frac{1}{n}}\sqrt{\frac{\lambda+n}{n}\log\left(\left(\frac{\lambda+n}{\lambda}\right)^d \frac{1}{\alpha^2}\right)} + r_n\right] \geq 1-\alpha \tag{61}$$

where $r_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$.

The proofs of lemma A.11 and theorem A.12 are given in appendix sections A.16 and A.17 respectively. Theorem A.12 effectively provides a confidence sequence for the mean vector for i.i.d. Gaussian random variables with unknown covariance matrix, provided a strongly consistent estimator is known. On the other hand, $\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star$ can be strongly approximated by a sum of multivariate Gaussians. Combining these elements yields the following result

Theorem A.13. *Let $\hat{\boldsymbol{\Sigma}}_n$ be a strongly consistent estimator of $\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-T}$, then for any prespecified $\lambda > 0$*

$$\mathbb{P}\left[\forall n \in \mathbb{N} : \|\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star\|_{\hat{\boldsymbol{\Sigma}}_n^{-1}} \leq \sqrt{\frac{1}{n}}\sqrt{\frac{\lambda+n}{n}\log\left(\left(\frac{\lambda+n}{\lambda}\right)^d \frac{1}{\alpha^2}\right)} + r_n\right] \geq 1-\alpha \tag{62}$$

where $r_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$ *and* d *is the dimenion of* $\boldsymbol{\gamma}^\star$.

It is worth noting that the arguments of theorem A.13 can be applied along any subset of $\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star$. For example, $\alpha^\star$ and $\boldsymbol{\kappa}^\star$ are rarely of primary interest. Instead, $\tau^\star$ is often of interest for testing average treatment effects, and $\boldsymbol{\eta}^\star$ is often of interest for testing conditional average treatment effects. A joint confidence sequence, which can be used to test the hypothesis of no treatment effects, can be obtained via

Theorem A.14. *Partition* $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\delta})$ *where* $\boldsymbol{\beta} = (\alpha, \boldsymbol{\kappa})$ *and* $\boldsymbol{\delta} = (\tau, \boldsymbol{\eta})$. *Let* $\boldsymbol{\Sigma}$ *be the submatrix of* $\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-T}$ *corresponding to* $\boldsymbol{\delta}$ *and let* $\hat{\boldsymbol{\Sigma}}_n$ *be a strongly consistent estimator of* $\boldsymbol{\Sigma}$, *then for any prespecified* $\lambda > 0$

$$\mathbb{P}\left[\forall n \in \mathbb{N} : \|\boldsymbol{\delta}_n^{ols} - \boldsymbol{\delta}^\star\|_{\hat{\boldsymbol{\Sigma}}_n^{-1}} \leq \sqrt{\frac{1}{n}}\sqrt{\frac{\lambda+n}{n}\log\left(\left(\frac{\lambda+n}{\lambda}\right)^d \frac{1}{\alpha^2}\right)} + r_n\right] \geq 1-\alpha \tag{63}$$

where $r_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$ *and* d *is the dimenion of* $\boldsymbol{\delta}$.

Alternatively, suppose we are only interested in an asymptotic confidence sequence for the average treatment effect.

Theorem A.15. *Let* σ^2 *be the element of* $\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-T}$ *corresponding to* τ *and let* $\hat{\sigma}_n^2$ *be a strongly consistent estimator of* σ^2, *then for any prespecified* $\lambda > 0$

$$\mathbb{P}\left[\forall n \in \mathbb{N} : |\tau_n^{ols} - \tau^\star| \leq \sqrt{\frac{\hat{\sigma}_n^2}{n}}\sqrt{\frac{\lambda+n}{n}\log\left(\left(\frac{\lambda+n}{\lambda}\right)\frac{1}{\alpha^2}\right)} + r_n\right] \geq 1-\alpha \tag{64}$$

where $r_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$.

A.12 Population Least Squares

The first order conditions satisfy

$$\begin{pmatrix} \mathbb{E}[y_i - \alpha - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\kappa} - (T_i - \rho)\tau - T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\eta}] \\ \mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})(y_i - \alpha - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\kappa} - (T_i - \rho)\tau - T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\eta})] \\ \mathbb{E}[(T_i - \rho)(y_i - \alpha - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\kappa} - (T_i - \rho)\tau - T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\eta})] \\ \mathbb{E}[T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})(y_i - \alpha - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\kappa} - (T_i - \rho)\tau - T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\eta})] \end{pmatrix} = \mathbf{0}$$

These expressions simplify to

$$\begin{pmatrix} \mathbb{E}[y_i] - \alpha \\ \mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i] - \boldsymbol{\Omega}_{\boldsymbol{m}}\boldsymbol{\kappa} - \rho\boldsymbol{\Omega}_{\boldsymbol{m}}\boldsymbol{\eta})] \\ \mathbb{E}[(T_i - \rho)y_i] - \rho(1-\rho)\tau \\ \mathbb{E}[T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i] - \rho\boldsymbol{\Omega}_{\boldsymbol{m}}\boldsymbol{\kappa} - \rho\boldsymbol{\Omega}_{\boldsymbol{m}}'\boldsymbol{\eta}) \end{pmatrix} = \mathbf{0}$$

Clearly $\alpha^\star = \mathbb{E}[y_i] = \rho\mathbb{E}[y_i(1)] + (1-\rho)\mathbb{E}[y_i(0)]$ and $\tau^\star = \mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)]$. Hence if the average treatment effect is zero, then $\tau^\star = 0$. From the last element we have $\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(1)] = \boldsymbol{\kappa} + \boldsymbol{\eta}$, and from the second line we have $\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i] = \boldsymbol{\kappa} + \rho\boldsymbol{\eta}$. Therefore

$$\begin{aligned} \boldsymbol{\eta}^\star &= (1-\rho)^{-1}\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}(\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(1)] - \mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i]) \\ &= \boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}(\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(1)] - \mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(0)]) \\ \boldsymbol{\kappa}^\star &= (1-\rho)^{-1}\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}(\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i] - \rho\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(1)]) \\ &= (1-\rho)^{-1}\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\mathbb{E}[(1 - T_i)(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i] \\ &= \boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i] \end{aligned}$$

An important note on $\boldsymbol{\eta}^\star$ regarding conditional average treatment effects, if $\mathbb{E}[y_i(1)|\boldsymbol{m}_i = \boldsymbol{m}] - \mathbb{E}[y_i(0)|\boldsymbol{m}_i = \boldsymbol{m}] = 0$ for all $\boldsymbol{m}$, then

$$\begin{aligned}\boldsymbol{\eta}^\star &= \boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}(\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(1)] - \mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})y_i(0)]) \\ &= \boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}(\mathbb{E}[(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})\mathbb{E}[y_i(1) - y_i(0)|\boldsymbol{m}_i]]) \\ &= \mathbf{0}\end{aligned}$$

and so if all condtional average treatment effects are zero, then $\boldsymbol{\eta}^\star = \mathbf{0}$.

A.13 Proof of Proposition A.8

Proof.

$$\mathbb{E}_{sp}[\boldsymbol{w}_i(y_i - \boldsymbol{w}_i'\boldsymbol{\gamma}^\star)] =$$
$$\mathbb{E}_{sp}\left[\begin{pmatrix} 1 \\ \boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}} \\ T_i - \rho \\ T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}}) \end{pmatrix}(y_i - \alpha^\star - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\kappa}^\star - (T_i - \rho)\tau^\star - T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\eta}^\star)\right]$$

Let's compute each element in turn, starting with the first

1.
$$\begin{aligned}&\mathbb{E}_{sp}[y_i - \alpha^\star - (\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\kappa}^\star - (T_i - \rho)\delta^\star - T_i(\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}})'\boldsymbol{\eta}^\star)] \\ &= \mathbb{E}_{sp}[y_i] - \alpha^\star - 0 - 0 \\ &= \rho\mathbb{E}_{sp}[y_i(1)] + (1-\rho)\mathbb{E}_{sp}[y_i(0)] - \alpha^\star = 0\end{aligned}$$

   Line 2 follows from $\mathbb{E}_{sp}[T_i - \rho] = 0$, $\mathbb{E}_{sp}[\boldsymbol{m}_i - \boldsymbol{\mu}_{\boldsymbol{m}}] = 0$, and the final line from the definition of $\alpha^\star$.

2.

$$\begin{aligned}
&\mathbb{E}_{sp}[(\boldsymbol{m}_i - \boldsymbol{\mu_m})(y_i - \alpha^\star - (\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\kappa}^\star - (T_i - \rho)\delta^\star - T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\eta}^\star)] \\
&= \mathbb{E}_{sp}[(\boldsymbol{m}_i - \boldsymbol{\mu_m})y_i] - 0 - \boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\boldsymbol{\kappa}^\star - 0 - \rho\boldsymbol{\Omega_m}\boldsymbol{\eta}^\star \\
&= 0
\end{aligned}$$

The first equality colds because $\mathbb{E}_{sp}[(\boldsymbol{m}_i - \boldsymbol{\mu_m})\alpha^\star] = 0$ and $\mathbb{E}_{sp}[(\boldsymbol{m}_i - \boldsymbol{\mu_m})(T_i - \rho)] = 0$ by independence of T_i and $\boldsymbol{m}_i$. The last equality holds from the definition of $\boldsymbol{\kappa}^\star$ and $\boldsymbol{\eta}^\star$.

3.

$$\begin{aligned}
&\mathbb{E}_{sp}[(T_i - \rho)y_i - \alpha^\star - (\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\kappa}^\star - (T_i - \rho)\delta^\star - T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\eta}^\star)] \\
&= \mathbb{E}_{sp}[(T_i - \rho)y_i] - 0 - 0 - \mathbb{E}_{sp}[(T_i - \rho)^2]\delta^\star - 0 \\
&= \rho(1-\rho)\mathbb{E}_{sp}[y_i(1)] + \rho(1-\rho)\mathbb{E}_{sp}[y_i(0)] - \rho(1-\rho)\delta^\star \\
&= 0
\end{aligned}$$

The first line follows because $\mathbb{E}_{sp}[(T_i - \rho)\alpha^\star] = 0$ and $\mathbb{E}_{sp}[(\boldsymbol{m}_i - \boldsymbol{\mu_m})(T_i - \rho)] = 0$ by independence of T_i and $\boldsymbol{m}_i$. The last line follows from the definition of $\delta^\star$.

4.

$$\begin{aligned}
&\mathbb{E}_{sp}[T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})(y_i - \alpha^\star - (\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\kappa}^\star - (T_i - \rho)\delta^\star - T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})'\boldsymbol{\eta}^\star)] \\
&= \mathbb{E}_{sp}[T_i(\boldsymbol{m}_i - \boldsymbol{\mu_m})y_i(1)] - 0 - \rho\boldsymbol{\Omega}_{\boldsymbol{m}}^{-1}\boldsymbol{\kappa}^\star - 0 - \rho\boldsymbol{\Omega_m}\boldsymbol{\eta}^\star \\
&= 0
\end{aligned}$$

where the last line follows from the definitions of $\boldsymbol{\eta}^\star$ and $\boldsymbol{\kappa}^\star$.

Having established that this vector is zero mean, $\mathbb{V}[\boldsymbol{w}_i(y_i - \boldsymbol{w}_i'\boldsymbol{\gamma}^\star)] = \mathbb{E}[(y_i - \boldsymbol{w}_i'\boldsymbol{\gamma}^\star)^2\boldsymbol{w}_i\boldsymbol{w}_i'] = \mathbb{E}[\mathbb{E}[(y_i - \boldsymbol{w}_i'\boldsymbol{\gamma}^\star)^2|\boldsymbol{w}_i]\boldsymbol{w}_i\boldsymbol{w}_i']$ □

A.14 Proof of lemma A.9

Proof. Let $\boldsymbol{V} = \boldsymbol{L}\boldsymbol{L}'$ and apply a whitening transformation by premultiplying each $\boldsymbol{y}_i$ by $\boldsymbol{L}^{-1}$ to obtain $\boldsymbol{x}_i := \boldsymbol{L}^{-1}\boldsymbol{y}_i$. The elements of $\boldsymbol{x}_i$ are independent with mean 0 and variance 1. Applying Strassen's approximation seperately along each dimension

$$\sum_i^n \boldsymbol{x}_i = \sum_i^n \boldsymbol{z}_i + \boldsymbol{e}_n$$

where $\boldsymbol{z}_i$ are independent $N(\boldsymbol{0}, \boldsymbol{I})$ random variables and $\boldsymbol{e}_n = o_{a.s.}(\sqrt{n \log\log n})$. Premultiplying by $\boldsymbol{L}$ recovers the original $\boldsymbol{y}_i$

$$\sum_i^n \boldsymbol{y}_i = \sum_i^n \boldsymbol{L}\boldsymbol{x}_i = \sum_i^n \boldsymbol{L}\boldsymbol{z}_i + \boldsymbol{L}\boldsymbol{e}_n$$

where $\boldsymbol{L}\boldsymbol{x}_i$ are independent $N(\boldsymbol{0}, \boldsymbol{V})$ random varibles and $\boldsymbol{L}\boldsymbol{e}_n = o_{a.s.}(\sqrt{n \log\log n})$. □

A.15 Proof of lemma A.10

Proof. By the strong law $\frac{1}{n}\boldsymbol{W}_n'\boldsymbol{W}_n \to \mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_i'] = \boldsymbol{\Omega}_{\boldsymbol{w}}$ almost surely. We can therefore consider the remainder term $\boldsymbol{\Omega}_{\boldsymbol{w}}\boldsymbol{R}_n := \boldsymbol{W}_n'\boldsymbol{W}_n - n\boldsymbol{\Omega}_{\boldsymbol{w}}$. Appealing elementwise to the law of iterated logarithm $\boldsymbol{\Omega}_{\boldsymbol{w}}\boldsymbol{R}_n = O_{a.s.}(\sqrt{n \log\log n})$. In particular this implies that $\|\boldsymbol{T}_n - \rho\boldsymbol{1}_n\|_2^2 = n\rho(1-\rho) + O_{a.s.}(\sqrt{n \log\log n})$.

$$\begin{aligned}
(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1} &= (n\boldsymbol{\Omega}_{\boldsymbol{w}} + \boldsymbol{\Omega}_{\boldsymbol{w}}\boldsymbol{R}_n)^{-1} \\
&= \left(\boldsymbol{I} + \frac{1}{n}\boldsymbol{R}_n\right)^{-1} \frac{1}{n}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1} \\
&= \sum_{i=0}^{\infty} (-1)^i \frac{\boldsymbol{R}_n^i \boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}}{n^{i+1}} \\
&= \frac{1}{n}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1} + \boldsymbol{r}_n
\end{aligned}$$

where $\boldsymbol{r}_n = O_{a.s.}\left(\frac{\sqrt{\log\log n}}{n^{\frac{3}{2}}}\right)$. Combining this result with lemma A.9 we have

$$\begin{aligned}\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star &= (\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\sum_i^n \boldsymbol{w}_i y_i - \boldsymbol{w}_i\boldsymbol{w}_i'\boldsymbol{\gamma}^\star \\ &= (\frac{1}{n}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1} + \boldsymbol{r}_n)(\sum_i^n \boldsymbol{g}_i + \boldsymbol{e}_n) \\ &= \frac{1}{n}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\sum_i^n \boldsymbol{g}_i + \frac{1}{n}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\boldsymbol{e}_n + \boldsymbol{r}_n\sum_i^n \boldsymbol{g}_i + \boldsymbol{r}_n\boldsymbol{e}_n\end{aligned}$$

where $\boldsymbol{e}_n = o_{a.s.}(\sqrt{n\log\log n})$ and $\boldsymbol{g}_i$ are i.i.d. $N(0, \boldsymbol{\Delta})$. Collecting terms to leading order yields

$$\boldsymbol{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star = \frac{1}{n}\sum_{i=1}^n \tilde{g}_i + \tilde{\boldsymbol{r}}_n \tag{65}$$

where $\tilde{g}_i$ are i.i.d. $N(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{w}}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}_{\boldsymbol{w}}^{-1})$ and $\tilde{\boldsymbol{r}}_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$.

$$\boldsymbol{\Omega}_{\boldsymbol{w}} = \mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_i'] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Omega}_m & 0 & \rho\boldsymbol{\Omega}_n \\ 0 & 0 & \rho(1-\rho) & 0 \\ 0 & \rho\boldsymbol{\Omega}_m & 0 & \rho\boldsymbol{\Omega}_m \end{pmatrix} \tag{66}$$

□

A.16 Proof of Lemma A.11

Proof. Let

$$M_n(\boldsymbol{\mu}) := e^{\sum_{i=1}^n (\boldsymbol{g}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})} = e^{(\boldsymbol{g}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})} M_{n-1}(\boldsymbol{\mu}) \tag{67}$$

Taking the conditional expectation gives

$$\mathbb{E}[M_n(\boldsymbol{\mu})|\mathcal{F}_{n-1}] = \mathbb{E}\left[e^{\boldsymbol{g}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}|\mathcal{F}_{n-1}\right] e^{-\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} M_{n-1}(\boldsymbol{\mu}) \leq M_{n-1}(\boldsymbol{\mu}) \tag{68}$$

where the last line following from the moment generating function. This establishes that M_n is a nonnegative supermartingale. Mixtures of martingales remain martingales, and so

$$M_n := \int M_n(\boldsymbol{\mu}) dF^{\lambda}(\boldsymbol{\mu}) = \left(\frac{\lambda}{\lambda+n}\right)^{\frac{d}{2}} e^{\frac{1}{2}\frac{n^2}{\lambda+n}\left(\bar{\boldsymbol{g}}_n \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{g}}_n\right)}, \tag{69}$$

where $\bar{\boldsymbol{g}}_n = (1/n)\sum_i \boldsymbol{g}_i$ and F^{λ} is a Gaussian measure with mean 0 and variance $\lambda^{-1}\boldsymbol{\Sigma}$, is also a nonnegative supermartingale. From Ville's inequality for nonnegative supermartingales (Ville 1939)

$$\mathbb{P}[\forall n \in \mathbb{N} : \log M_n \leq -\log \alpha] > 1 - \alpha.$$

Rearranging for $\sqrt{\bar{\boldsymbol{g}}_n \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{g}}_n} = \|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}}$ yields the desired result. □

A.17 Proof of Theorem A.12

Proof. As $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix, it's inverse exists. By the continuous mapping theorem, $\hat{\boldsymbol{\Sigma}}_n^{-1} \to \boldsymbol{\Sigma}^{-1}$ almost surely, and we can write $\hat{\boldsymbol{\Sigma}}_n^{-1} \to \boldsymbol{\Sigma}^{-1} + e_n \boldsymbol{\Sigma}^{-1}$ where $e_n = o_{a.s.}(1)$. By the triangle inequality $\|\bar{\boldsymbol{g}}_n\|_{\hat{\boldsymbol{\Sigma}}_n^{-1}} \leq \|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}} + e_n \|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}}$. Note also that $\|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}} = \sqrt{\bar{\boldsymbol{g}}_n^T \boldsymbol{L}^{-T} \boldsymbol{L}^{-1} \bar{\boldsymbol{g}}_n}$, with $\boldsymbol{L}^{-1}\bar{\boldsymbol{g}}_n = \frac{1}{n}\sum_i \boldsymbol{L}^{-1}\boldsymbol{g}_i$. Pre-multiplying by $\boldsymbol{L}^{-1}$ whitens the vectors, i.e. $\boldsymbol{L}^{-1}\boldsymbol{g}_i \sim N_d(\boldsymbol{0}_d, \boldsymbol{I}_d)$, and so each element of $\sum_i \boldsymbol{L}^{-1}\boldsymbol{g}_i$ is a sum of zero mean unit variance random variables. By the law of iterated logarithm each element is $o_{a.s.}(\sqrt{n \log\log n})$, which implies

$$\|\bar{\boldsymbol{g}}_n\|_{\hat{\boldsymbol{\Sigma}}_n^{-1}} \leq \|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}} + r_n$$

where $r_n = o_{a.s.}\left(\sqrt{\frac{\log\log n}{n}}\right)$. Applying the time-uniform bound to $\|\bar{\boldsymbol{g}}_n\|_{\boldsymbol{\Sigma}^{-1}}$ from lemma A.11 then yields the result.

□

Proof. We adopt a similar proof strategy to Waudby-Smith et al. (2021) who leverage the following strong approximation due to Strassen (1964). Let $x_1, x_2, \ldots$ be a sequence of

random variables with mean 0 and variance 1, then

$$\sum_i^n x_i = \sum_i^n z_i + e_n$$

where z_i are independent standard Gaussians and $e_n = o_{a.s.}(\sqrt{n \log\log n})$. To generalise to the multivariate setting, whiten the $\boldsymbol{h}_i$ by premultiplying by $\boldsymbol{L}^{-1}$, defining $\boldsymbol{x}_i = \boldsymbol{L}^{-1}\boldsymbol{h}_i$. The elements of $\boldsymbol{h}_i$ are independent and we can apply Strassen's result seperately along each dimension

$$\sum_i^n \boldsymbol{x}_i = \sum_i^n \boldsymbol{z}_i + \boldsymbol{e}_n$$

Pre-multiplying by $\boldsymbol{L}$ recovers the original $\boldsymbol{h}_i$ and rescaling by $1/n$ yields

$$\begin{aligned}\bar{\boldsymbol{h}}_n &= \frac{1}{n}\sum_i^n \boldsymbol{h}_i = \frac{1}{n}\sum_i^n \boldsymbol{L}\boldsymbol{x}_i = \frac{1}{n}\sum_i^n \boldsymbol{L}\boldsymbol{z}_i + \frac{1}{n}\boldsymbol{L}\boldsymbol{e}_n \\ &= \bar{\boldsymbol{g}}_n + \boldsymbol{r}_n\end{aligned}$$

where $\tilde{\boldsymbol{z}}_i$ are independent $N(\boldsymbol{0}_d, \boldsymbol{\Sigma})$ random variables and $\boldsymbol{r}_n = o_{a.s.}(\sqrt{\frac{\log\log n}{n}})$. This has enabled us to approximate $\sum_i \boldsymbol{h}_i$ by a sum $\sum_i \boldsymbol{g}_i$ where $\boldsymbol{g}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma})$. By the triangle inequality

$$\|\bar{\boldsymbol{h}}_n\|_{\hat{\boldsymbol{\Sigma}}_n} \leq \|\bar{\boldsymbol{g}}_n\|_{\hat{\boldsymbol{\Sigma}}_n} + r_n,$$

where $r_n = o_{a.s.}(\sqrt{\frac{\log\log n}{n}})$. Applying the time-uniform bound to $\|\bar{\boldsymbol{g}}_n\|_{\hat{\boldsymbol{\Sigma}}_n}$ from theorem A.12 then yields the result. □

B Converting Fixed-n to Anytime-Valid (with R Code)

Algorithm 1 Anytime Valid Linear Model Summary in R

```
mod <- summary(lm(outcome ~ . + trt*., data=df))

stderrs <- mod$coefficients[, 'Std. Error']

t2 <- mod$coefficients[, 't value']^2

ols <- mod$coefficients[, 'Estimate']

nu <- nrow(df) - length(ols)

z2 <- (mod$sigma / stderrs)^2

r <- phi / (phi + z2)

spvals <- min(1, sqrt(r) *

    ((1 + r * (t2 / nu))^(-(nu + 1)/2)) /

    ((1 + (t2 / nu))^(-(nu + 1)/2)))

alpha <- 0.05

radii <- stderrs * sqrt(nu * ((1 - (r * alpha^2)^(1 / (nu + 1))) /

    max(0, ((r * alpha^2)^(1 / (nu + 1))) - r)))

lower_cis <- ols - radii

upper_cis <- ols + radii
```

C Comparison with Empirical Bernstein Confidence Sequences

It is interesting to consider how our asymptotic confidence sequence for the average treatment effect compares with a nonasymptotic confidence sequence. Such a nonasymptotic confidence sequence is provided by *Empirical Bernstein* confidence sequence in Theorem 4 of Howard et al. (2021). It requires, however, that outcomes are bounded. Fortunately, the outcomes of *PlayDelay* in section 8 are in fact bounded. While it may seem that the support of *PlayDelay* is $[0, \infty)$, in practice the application does not wait indefinitely for playback to start. If a playback does not occur within the time interval $[0, b]$, then that outcome is recorded as b, effectively trimming or projecting outcomes onto the interval $[0, b]$. To perform regression adjustment, we consider an AIPW style transformation of outcomes

$$G_i = \hat{\mu}^1_{i-1}(\boldsymbol{X}_i) - \hat{\mu}^0_{i-1}(\boldsymbol{X}_i) + \frac{T_i - \rho}{\rho(1-\rho)}(Y_i - \hat{\mu}^{T_i}_{i-1}(\boldsymbol{X}_i))$$

where $\hat{\mu}^1_{i-1}$ and $\hat{\mu}^0_{i-1}$ are $\mathcal{F}_{i-1}$ measurable linear predictors of treatment and control outcomes respectively, projected onto the interval $[0, b]$. If the average treatment effect is zero, then $\mathbb{E}[G_i|\mathcal{F}_{i-1}] = 0$. With $\rho = 1/2$, the transformed outcomes are bounded below by $-2b$ and above by $2b$. For the empirical Bernstein confidence sequence we use a scale of $4b$ and use the Gamma-Exponential conjugate mixture (Howard et al. 2021, Proposition 9) with tuning parameter 100 to provide an equitable comparison to the Gaussian mixture we employ with precision $\phi = 100$, as is suggested in Howard et al. (2021, Section 3.5). Figure 9 is identical to figure 8, except that the empirical distribution of the stopping time for the empirical Bernstein confidence sequence has been added. The shift toward larger stopping

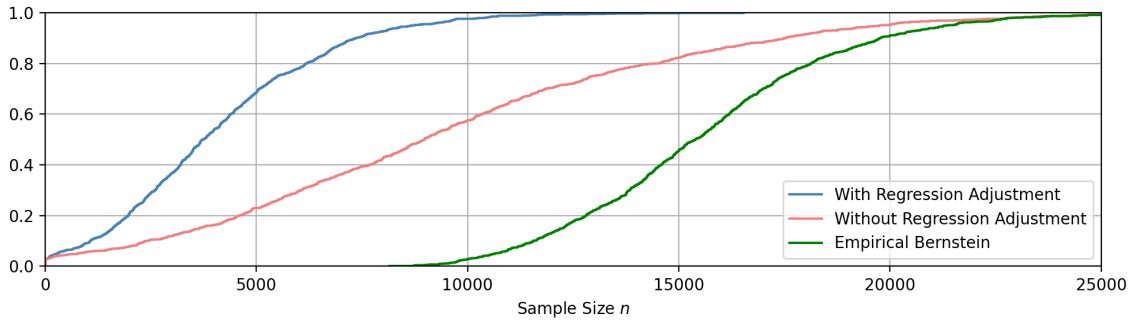


Figure 9: Empirical distribution of the stopping time $\tau_s = \inf\{n \in \mathbb{N} : p_n(\boldsymbol{Y}_n; 0) < 0.05\}$ based on 1000 simulations under the alternative hypothesis.

times exhibited by the empirical distribution of the empirical Bernstein procedure suggests that the price we pay for a fully nonasymptotic confidence sequence is conservativeness. However, although our confidence sequences are asymptotic, the null hypothesis simulations of section 8 demonstrate that our Type-I error remains calibrated at the nominal α level, while exhibiting faster rejections under the alternative hypothesis.