From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2016

Inference in Regressions with Many Controls

Li, Chenchuan

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01w3763925c

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

INFERENCE IN REGRESSIONS WITH MANY CONTROLS

Chenchuan Li

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF ECONOMICS

[Adviser: Ulrich Mueller]

September 2016

© Copyright by Chenchuan Li, 2016. All rights reserved.

Contents

Abstract **iv**

Acknowledgements **vi**

I Inference in Regressions under a Sum-of-Squares Bound **1**

1 Introduction **1**

2 A Gaussian, Homoskedastic Model **4**

2.1 Preliminaries 4

2.2 Background: Benchmark Specifications for $\mathcal{D}_0$ and the Pitfalls of Consistent Model Selection 6

2.3 An L^2 Bound for d 8

2.4 Sufficient Statistics 9

2.5 Bivariate Tests 10

3 Asymptotic Inference in the Gaussian, Homoskedastic Model **12**

3.1 Preliminaries 12

3.2 Invariance 13

3.3 Asymptotic Framework and Main Results 14

3.3.1 Main Result: Limit Experiment 15

3.3.2 Main Result: Asymptotic Optimality of the Bivariate Test 17

3.4 Separation of Primary Controls 21

4 Robust Implementation under General Error Structures **22**

4.1 Preliminaries 22

4.2 Construction of the Robust Bivariate Test 24

4.3 Asymptotics of the Robust Bivariate Test 25

5 Comparisons with Post-Double-Selection **27**

5.1 Theory 27

5.2 Monte Carlo 28

6 Empirical Application **30**

6.1 Preliminaries 30

6.2 Applying the Robust Bivariate Test 32

7 Conclusion **33**

II Power Bounds in Regressions under Sparsity **35**

8 Introduction **35**

9 Setup and Background **36**

9.1 Preliminaries 36

9.2 Least-Favorable Distributions 38

10 Power Bounds for Symmetric W_{xz} **40**

10.1 An Analytical Least-Favorable Distribution for $N = 2$, $K = 2$, and $M = 1$ 41

10.1.1 First-Order Necessary Conditions 42

10.1.2 Second-Order Sufficient Conditions 44

10.2 Computing Power Bounds 44

11 Power Bounds for Randomly-Generated Designs **48**

11.1 Approximately-Sparse Designs 49

11.2 Designs with Varying Sparsity 50

12 Conclusion **52**

III Conditionally-Unbiased Asymptotic Variance Estimation **54**

13 Introduction **54**

14 Setup and Background **55**

15 Main Result: Extension of Cattaneo et al. (2015) to the Clustered Case **58**

15.1 Fixed-Effects 61

16 Monte Carlo Simulations **62**

17 Conclusion **64**

Appendix **67**

A Proofs, Background, and Figures from Part I **67**

A.1 Optimality, Weighted Average Power, and the Least-Favorable Distribution 68

A.1.1 Weighted Average Power Criterion 68

A.1.2 Least-Favorable Distribution 69

A.1.3 Application to Full Gaussian Baseline Model 71

A.2 Bivariate Tests 72

A.3 Asymptotics and the Limit of Experiments 75

A.3.1 Mathematical Lemmas 75

A.3.2 Invariance 76

A.3.3 Limit of Experiments 78

B Proofs from Part II **82**

B.1 Mathematical Lemmas 82

B.2 Proof of Theorem 10.2 83

B.2.1 The Least-Favorable Distribution under Sparsity 84

B.2.2 First-Order Conditions for Local Validity 85

B.2.3 Second-Order Conditions for Local Validity 87

C Proofs from Part III **88**

C.1 Proof of Theorem 15.3 88

C.2 Proof of Corollary 15.4 90

Abstract

In this thesis, we consider inference on a scalar coefficient of interest in a linear regression model with many potential control variables. Without any constraint on the control coefficients in a canonical model with Gaussian, homoskedastic errors, one cannot improve upon the standard t-test in the regression that includes all controls. In the following three chapters, we investigate *(i)* the possibility for inference to expand under constraints on the control coefficients and *(ii)* how to implement these improvements under general error structures.

In the first chapter, we impose a bound on the weighted sum-of-squared control coefficients, which amounts to a bound on the R^2 of controls. We develop a simple testing procedure to exploit this constraint, and we show that our procedure is, under asymptotics where the number of controls is a fraction of sample size, *(i)* of correct size under potential heteroskedasticity and clustered error structures and *(ii)* weighted-average-power maximizing under a sequence of bounds that shrinks to zero in the canonical model. We apply the new test to an empirical study of the relationship between crime and abortion by Donohue III and Levitt (2001), where we determine the marginal value of the R^2 bound which induces a significant result.

In the second chapter, we study how a sparsity assumption, which restricts the number of nonzero control coefficients, can admit improvements over the standard t-test in the canonical model. When the design satisfies a symmetry criterion, we are able to derive the infimum power bound against a point alternative over all valid tests under sparsity. For designs outside this family, we construct an algorithm to derive non-infimum power bounds over valid tests. For various sample sizes and designs, we find that when the R^2 of the regressor of interest on the controls does not exceed .9, the power gained from assuming that no more than 10% of control coefficients are non-zero is roughly equivalent to increasing the number of observations by less than four-fold and applying the standard t-test.

In the third chapter, we study the estimation of the asymptotic variance of linear statistics in the presence of many regressors. The natural and most popular estimator of variance under potential heteroskedasticity and clustering in a model with finitely-many regressors is an observation-weighted average of squared residuals, where consistency holds because the sampling uncertainty of coefficient

estimates becomes asymptotically-negligible. In a high-dimensional model where the number of regressors is of the same order as the number of observations, the regression coefficients can no longer be estimated with sufficient precision for this reasoning to apply. We construct a cluster-robust variance estimator which is *(i)* conditionally-unbiased in finite samples, *(ii)* consistent under some regularity assumptions on the sequence of regressor values, and *(iii)* invariant to the true data-generating coefficients of the linear model. Our estimator is an extension to the cluster case of a procedure by Cattaneo, Jansson, and Newey (2015) for potentially-heteroskedastic, independent errors.

Acknowledgements

With much gratitude to Ulrich Mueller, Paul Sally, and Larry Puetz, three outstanding mentors with whom I've been fortunate to study. A special thanks to Ulrich, whose unbounded energy and generosity always inspire me to put my best foot forward. I'm also indebted to faculty members Kirill Evdokimov, Bo Honore, Michal Kolesar, and Mark Watson, who, not without great strain, chiseled me into a better teacher and researcher. Lastly, a heartfelt thank you to the family and friends whose support made all of this possible.

This content of this thesis was greatly helped by comments from Diane Alexander, Kirill Evdokimov, Bo Honore, Moshe Katzwer, Shaowei Ke, Michal Kolesar, Inessa Liskovich, Ulrich Mueller, Chris Sims, Sharon Traiberman, Peter Van Tassel, and Mark Watson.

Part I

Inference in Regressions under a Sum-of-Squares Bound[1]

1 Introduction

In economics, linear regressions have encountered widespread use as tools for inference on causal effects. A feature of economic data, however, is that outcomes are rarely driven exclusively by the causal relationships of interest. In Donohue III and Levitt (2001), the running empirical example of this chapter, linear regressions are used to determine the consequences of an exogenous change in abortion rates on future crime rates, but the prevalence of crime is correlated with economic and social factors which can generate confounding bias when abortion rates are not randomly assigned. In general, disentangling a single causal channel from many confounding factors is central to empirical analysis, and it demands the use of control variables.

Typically, the set of controls under consideration can be quite expansive, consisting of primary variables identified by economic theory as well as a large number of "technical" terms — *e.g.* higher order and interaction effects — accounting for uncertainty in functional forms. Hence, a *long regression* incorporating all controls can induce overfitting, render inference powerless, and motivate the practitioner to narrow attention to subsets of controls, but the determination of an appropriate subset poses additional complications, and many popular approaches can lead to invalid inference (overrejection of the null hypothesis).

A common empirical strategy is to experiment with different controls, settle on a baseline specification, and proceed under that specification as given. Subsequently, robustness checks are conducted — whereby variables are added or removed and results re-examined in each iteration — to ensure that conclusions remain stable to perturbations from the baseline.[2] This informal procedure, though ubiquitous, is rarely accompanied by proper accounting of model selection uncertainty from a theoretical standpoint. In a practical but related matter, it can attract controversy when results

[1] Co-authored with Ulrich Mueller, Princeton University.
[2] This is the approach of Donohue III and Levitt (2001).

prove sensitive to the selection process.[3]

An alternative to *ad hoc* inspection is to apply a formal mechanism for choosing control variables, a task to which consistent model selection procedures — *e.g.* the Bayesian Information Criterion (Schwarz, 1978) and the SCAD estimator (Fan and Li, 2001) — appear naturally suited. Consistency, here, denotes that the correct model is revealed with near-certainty in large samples for fixed values of the control coefficients, and it is customarily invoked as theoretical justification to follow the pre-selection of controls with standard protocol conditional on knowing the relevant controls. However, as illustrated by Leeb and Pötscher (2005, 2008a), such inference is invalidated by the existence of small coefficients for which the corresponding control variables can simultaneously *(i)* evade consistent model selection and *(ii)* induce omitted variable bias of the same order of magnitude as the standard error of estimates, upon exclusion from the regression. We explore this in detail in example 2.5 of Section 2.2.

This drawback might impel a practitioner to explore other formal approaches, but a straightforward exercise demonstrates that if the control coefficients are left unrestricted, *no* valid procedure can outperform the long regression within the canonical Gaussian, homoskedastic model, a feature which suggests that any attempt to bypass the long regression must operate through a constraint on control coefficients.

Belloni et al. (2014) follow this prescription by imposing two "spasity" constraints: *(i)* at most a small fraction of control variables are allowed to have non-zero coefficients, and *(ii)* at most a small fraction of controls are allowed to be correlated with the regressor of interest. Under these assumptions, Belloni et al. (2014) show that a pair of LASSO regressions — L^1 penalized regressions[4] — comprise a valid variable-selection procedure. However, in economic applications where the effects of regressors are rarely exactly zero, sparsity may not be an appropriate assumption.[5] Moreover, sparsity leaves inference sensitive to linear recombinations of controls, as shown by the following example: consider a regression with a set of two covariates $\{age, period\}$, where it is typically without loss of generality to substitute $cohort = period - age$ for either *period* or *age* within the regression. This no longer holds if one imposes that at most one covariate has a nonzero coefficient, since each variable from $\{age, period, cohort\}$ represents a separate explanatory effect.

In contrast to sparsity, the constraint in this chapter is given by a bound on a weighted sum-of-squares (L^2 norm) of control coefficients, naturally related to a restriction on the R^2 of controls. In

[3]For instance, Foote and Goetz (2008) discover that results in Donohue III and Levitt (2001) are not robust to a particular set of modifications.

[4]Developed in Tibshirani (1996).

[5]For a discussion pertaining the use of sparsity in social science applications, see Gelman (2011).

essence, the bound acts as a device to curb omitted variable bias,[6] and under asymptotics where the number of regressors grows proportionally to the number of observations, it enables us to construct a test with the following properties:

1. Under potential heteroskedasticity and clustering of errors, it is asymptotically-valid, rejecting with correct size whenever the control coefficients obey the bound.

2. Within a Gaussian, homoskedastic model, it is valid in any sample size, and when the bound shrinks to zero at a certain rate, it is optimal in a weighted-average-power sense.[7]

We prove these claims in three theoretical sections. In Section 2, we define a finite-sample, Gaussian, homoskedastic framework. We show that consistent variable selection is invalid, and we observe that, absent coefficient constraints, the long regression is the optimal valid procedure. We subsequently define an L^2 bound on the control coefficients, discuss its R^2 interpretation, and construct tests based on a simple pair of OLS estimates — denoted the bivariate statistic — which are valid for arbitrary, given values of the bound. We observe that the mean of the bivariate statistic consists of two correlated components: the coefficient of interest and a measure of omitted variable bias. The least-favorable distribution technique of Elliott et al. (2015) then allows us to construct the test which makes best use of the bivariate statistic, in a weighted average power sense.

In Section 3, we extend the previous analysis to an asymptotic setting, where the number of controls becomes a fixed fraction of sample size. We find that as the L^2 bound shrinks to zero at a certain rate, the bivariate statistic absorbs all information from the experiment, becoming "asymptotically-sufficient." Hence, under these conditions, the "bivariate test" becomes optimal over all possible valid tests.

In Section 4, we extend this analysis to inference under potentially non-Gaussian, heteroskedastic, and clustered errors. We construct a robust bivariate test which is *(i)* valid under this general error structure and *(ii)* retains the optimality properties of Section 3 under a Gaussian, homoskedastic model.

Following the theoretical sections of this chapter, we compare our robust bivariate test to the LASSO-based Post-Double-Selection procedure of Belloni et al. (2014) in Section 5. A simple thought experiment demonstrates that both the bivariate test and Post-Double-Selection effectively constrain omitted variable bias. Subsequently, in a Monte Carlo exercise, we find that Post-Double-Selection fails to control size under certain conditions for which the robust bivariate test is valid.

[6] The generic notion of bounding the influence of a nuisance parameter for inference has been applied in Altonji et al. (2008) and Kline and Santos (2013).

[7] For a discussion of weighted average power, we refer readers to Andrews et al. (2011) and Elliott et al. (2015).

In Section 6, we apply the robust bivariate test to the empirical study of Donohue III and Levitt (2001). Specifically, we locate the margin at which our bound begins to induce a statistically-significant result, and we translate this marginal value into an R^2 assumption.

For $n \in \mathbb{N}$ and $\mu \in \mathbb{R}^N$, we take μ to be an $n \times 1$ column vector. We write μ_i, $i = 1, \ldots, n$ to denote the components of μ, and we write $\|\mu\|^2 = \sum_{i=1}^n \mu_i^2$ to denote the L^2 (Euclidean) norm of μ. Throughout the chapter, we let $\epsilon > 0$ denote an arbitrarily-small constant, and we let $\alpha \in (0, 1)$ denote a fixed level, unless otherwise specified.

2 A Gaussian, Homoskedastic Model

2.1 Preliminaries

In this section, we consider inference on β in a linear, homoskedastic model

$$y_i = x_i\beta + \sum_{k=1}^{K} z_{ik}\delta_k + e_i = x_i\beta + z_i'\delta + e_i, \tag{2.1}$$

where β is the scalar of interest, $\delta \in \mathbb{R}^K$ is a nuisance parameter, N denotes the number of observations, and $i = 1, \ldots, N$ denotes index. The observables are a dependent variable $y_i \in \mathbb{R}$, a regressor of interest $x_i \in \mathbb{R}$, and a vector of control variables $z_i \in \mathbb{R}^K$, and the error disturbance is e_i. Stacking variables in matrices, we write

$$\underset{(N\times 1)}{Y} = \underset{(N\times 1)}{X}\beta + \underset{(N\times K)}{Z}\delta + \underset{(N\times 1)}{e}.$$

Assumption 2.1. $e|X, Z \sim \mathcal{N}(0, I_N)$.

Within Section 2, we assume e is conditionally Gaussian and homoskedastic for tractibility. Section 4 relaxes assumptions on e to accommodate implementation under general data-generating processes, but the purpose of our stylized model is to provide context for evaluating statistical performance.

Assumption 2.2. *$W_{xz} = (X, Z)$ is a deterministic matrix.*

Assumption 2.2 establishes that the regressors are non-stochastic. In particular, inference

must be level-α valid conditional upon every realization of the design matrix W_{xz}, precluding the use of randomness in x_i and z_i to obtain size control.[8]

Assumption 2.3. $N^{-1}W_{xz}'W_{xz} = \begin{pmatrix} 1 & \rho' \\ \rho & I_K \end{pmatrix}$, *where* $\rho = N^{-1}Z'X \in \mathbb{R}^K$, *and* $\|\rho\|^2 \in (0,1)$.

Assumption 2.3 restricts the column space of $N^{-1/2}Z$ to be orthonormal and $N^{-1/2}X$ to have length 1. In this framework, $\rho \in \mathbb{R}^K$ represents the sample correlation between z_i and x_i, and $\|\rho\|^2$ is the R^2 of x_i on z_i. In what follows, our inferential procedure is constructed to be invariant to linear recombinations of the controls z_i, so Assumption 2.3 becomes without loss of generality.

The remainder of this chapter concerns inference on β. Suppose the null hypothesis is $\beta = 0$, and define local parameters $b = \sqrt{N}\beta$ and $d = \sqrt{N}\delta$. We consider level-α tests of

$$H_0 : b = 0,\, d \in \mathcal{D}_0 \quad vs. \quad H_1 : b \neq 0,\, d \in \mathbb{R}^K, \tag{2.2}$$

where $\mathcal{D}_0 \subseteq \mathbb{R}^K$ is a nuisance parameter set which constrains d under the null. In particular, we require tests to be *valid*, meaning the null rejection probability cannot exceed α under any $d \in \mathcal{D}_0$. In subsequent sections, we outline the proper treatment of d under the alternative.

Outline of Sections 2.2-2.5

In the remainder of Section 2, we examine the problem (2.2) in finite samples. Section 2.2 provides background on valid inference in two notable cases: when $\mathcal{D}_0 \subseteq \mathbb{R}^K$ is unrestricted and when $\mathcal{D}_0 = \{0\}$ is tightly restricted. We observe that in the former, consistent model-selection yields invalid inference, while the long regression of y_i on (x_i, z_i) is the optimal valid procedure. In the latter, the short regression of y_i on x_i alone is optimal.

In Section 2.3, we state the underlying assumption of this chapter: for a value $r_d > 0$, the nuisance parameter set $\mathcal{D}_0 = \{d : \|d\|^2 \leq r_d^2\}$ represents an L^2-bound on d, subsequently related to a bound on the R^2 of z_i.

In Section 2.4, we reparameterize $d \in \mathbb{R}^K$ using $(g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^K$, where $g_\rho = \rho' d \in \mathbb{R}$ is the component of d generating omitted variable bias and $g_\perp \in \mathbb{R}^K$ is the component unrelated to omitted variable bias. In the same fashion, we construct a sufficient statistic $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ which is the sample analog of $(b, g_\rho, g_\perp)$.

[8] Validity in a conditional framework implies unconditional validity, but the converse does not hold.

In Section 2.5, we construct a test based on a bivariate statistic $(\hat{b}, \hat{g}_\rho) \in \mathbb{R}^2$ which is valid for arbitrary values of the bound r_d.

2.2 Background: Benchmark Specifications for $\mathcal{D}_0$ and the Pitfalls of Consistent Model Selection

Definition 2.4. Let $(\hat{\beta}, \hat{\delta}) = (W_{xz}'W_{xz})^{-1}W_{xz}'Y$ be the OLS coefficients from the *long regression* of y_i on (x_i, z_i), and let $(\hat{b}, \hat{d}) = \sqrt{N}(\hat{\beta}, \hat{\delta})$ denote its local analog.

We begin by charaterizing inference in (2.2) when the nuisance parameter d is unrestricted, and $\mathcal{D}_0 = \mathbb{R}^K$. By default, a valid test is given by the long regression defined above, but as indicated in Section 1, large standard errors driven by high-dimensional z_i may render this unattractive. An alternative is to apply a consistent model-selection technique to determine relevant controls (an appropriate subvector of z_i) for inclusion in a final-stage regression. However, stemming from an observation by Leeb and Pötscher (2005), the following example demonstrates this approach to be invalid.

Example 2.5 (Invalidity of Consistent Variable Selection)**.** Suppose, for simplicity, that z_i is a scalar. Then the model reduces to $y_i = x_i\beta + z_i\delta + e_i$, where $e_i|\{x_i, z_i\}_{i=1}^N \sim_{iid} \mathcal{N}(0,1)$, and by Assumption 2.3, the regressors are normalized to satisfy $N^{-1}\sum_{i=1}^N x_i^2 = N^{-1}\sum_{i=1}^N z_i^2 = 1$ and $N^{-1}\sum_{i=1}^N x_i z_i = \rho \in (0,1)$.

Suppose a noisy mechanism $\hat{S} \in \{0,1\}$ indicates whether z_i should be included in the regression, and we employ standard inference to the model designated by $\hat{S}$. Hence, when $\hat{S} = 1$, we regress y_i on x_i controlling for z_i, and when $\hat{S} = 0$, we regress y_i on x_i alone. Because $\hat{S}$ is subject to error, it is possible, even when the truth is $\delta = \delta_0 \neq 0$, that $\hat{S} = 0$. In this case, the post-selection OLS coefficient $\hat{\beta}^{PS}$ satisfies $\hat{\beta}^{PS}|\{\hat{S} = 0\} = \beta + \rho\delta + N^{-1}\sum_{i=1}^N x_i e_i$, and the t-statistic is distributed $\sqrt{N}\frac{(\hat{\beta}^{PS} - \beta)}{SE(\hat{\beta}^{PS})}|\{\hat{S} = 0\} \sim \mathcal{N}(\sqrt{N}\rho\delta, 1)$, exhibiting omitted variable bias of $\sqrt{N}\rho\delta$.

If $\hat{S}$ is pointwise-consistent,[9] then it selects the true model with probability approaching 1 for any *fixed* value of the data-generating parameter $\delta_0 \in \mathbb{R}$. This is a promising indication that even when there is bias from selecting the wrong model, the overall procedure can be valid in large samples, where $\hat{S}$ is correct with near-certainty. However, the following shows this conjecture to be false.

[9]Formally, $Pr(\hat{S} = 1|\delta = \delta_0 \neq 0) \to 1$ and $Pr(\hat{S} = 0|\delta = \delta_0 = 0) \to 1$ as $N \to \infty$

Suppose $\beta = 0$, so the model satisfies the null. A valid procedure has correct size regardless of N and the true data-generating value of δ, so we apply this requirement to the specific case where $\delta = 1/\sqrt{N}$ for each N. Theoretical results show that when $\hat{S}$ is pointwise-consistent, it cannot distinguish between models generated by $\delta = 0$ and $\delta = 1/\sqrt{N}$ in large samples. Hence, with probability approaching 1, $\hat{S}$ mistakenly proclaims $\delta = 0$, and following incorrect selection by $\hat{S}$, the t-statistic is skewed by an omitted variable bias of $\sqrt{N}\rho\delta = \rho$, overrejects in arbitrarily large samples. As such, selection by $\hat{S}$ is invalid, driven by a data-generating coefficient $\delta = 1/\sqrt{N}$ too small to to be detected by consistent model-selection but large enough to nontrivially tamper with inference. ▶

In light of the issues highlighted by Example 2.5, some authors have considered approaches guided by inconsistent pre-selection of control variables.[10] However, a simple but instructive exercise shows that among valid procedures, a test based on $\hat{b}$ is: *(i)* the uniformly most powerful test against the one-sided alternative $H_1 : b > 0$, and *(ii)* the uniformly most powerful unbiased test against two-sided alternatives $H_1 : |b| > 0$. In this sense, exceeding the performance of $\hat{b}$ is impossible without constraints on d.

In general, when the nuisance parameter set shrinks, the set of valid tests expands, and the supremum of power over valid tests increases. When $\mathcal{D}_0 = \{0\}$ is the smallest possible set, the short regression coefficient $\hat{b}_{short} = \sqrt{N}(X'X)^{-1}X'Y$, formally defined below, outperforms any statistic against the one-sided alternative $H_1 : b > 0$, $d = 0$. Naturally, with *a priori* knowledge that controls are irrelevant, regressing y_i on x_i alone is optimal.

Definition 2.6. Let $\hat{\beta}_{short} = (X'X)^{-1}X'Y$ be the coefficient from the "short" regression of y_i on x_i alone, and let $\hat{b}_{short} = \sqrt{N}\hat{\beta}_{short}$. Observe that, equivalently, $\hat{\beta}_{short} = \hat{\beta} + \rho'\hat{\delta}$, and $\hat{b}_{short} = \hat{b} + \rho'\hat{d}$, with $(\hat{\beta}, \hat{\delta})$ and $(\hat{b}, \hat{d})$ as in Definition 2.4.

Other restrictions on $\mathcal{D}_0 \subset \mathbb{R}^K$ allow valid tests to span the entire range of power delimited by the long and short regressions. Sparse inference assumes this role by imposing an L^0 bound on d whereas we, instead, consider an L^2 bound.

[10] Leeb and Pötscher (2005, 2008a), among others.

2.3 An L^2 Bound for d

In Section 2.2, we established that smaller nuisance parameter sets admit more powerful valid tests, highlighting the tradeoff between power and size control with two benchmark examples. In this section, we place an L^2 bound on d, relate the L^2 bound to a restriction on the population R^2, and apply a change-of-variables to d to decompose its effect into two orthogonal components: one contributing to omitted variable bias, and one unrelated to omitted variable bias.

Assumption 2.7. *For given $r_d > 0$, let $\mathcal{D}_0 = \{d \in \mathbb{R}^K : \|d\|^2 \leq r_d^2\}$ be the nuisance parameter set under the null.*

The focus of this chapter is valid inference under Assumption 2.7. We argue that this is a natural constraint, because the regressors are deterministic in our model, so the contribution of z_i to y_i is fully determined by $Zd \in \mathbb{R}^N$. In conjunction with the normalization of Assumption 2.3, the magnitude of that contribution is $N\|d\|$, and the range of that contribution depends only on the column span of Z, so our model is invariant to the linear representation of controls.[11] This is a strict departure from sparsity, which favors setting individual components of d to 0, as discussed in Section 1.

The invariance established above elicits a natural R^2 interpretation. Under Assumption 2.2, the column space of $N^{-1/2}Z$ is orthonormal, so the population R^2 of z_i on y_i under $\beta = 0$ is:

$$R_{yz}^2 = \frac{\|Z\delta\|^2}{E[\|Y\|^2 | \beta = 0]} = \frac{N^{-1}\|d\|^2}{N^{-1}\|d\|^2 + 1}$$

Hence, by restricting d to satisfy $\|d\|^2 \leq r_d$ in Assumption 2.7, we implicitly impose the constraint $R_{yz}^2 \leq c_R^2 = \frac{N^{-1}r_d^2}{N^{-1}r_d^2+1}$, where the bound c_R^2 has a 1-to-1 correspondence with the bound r_d^2. Therefore, under homoskedasticity, it is equivalent to define the nuisance parameter set via the constraint $R_{yz}^2 \leq c_R^2$.

[11] We note that if $\mathcal{D}_0$ is initially given by a bound on $\|Zd\|^2$, the contribution of z_i towards the outcome y_i, then the normalization in Assumption 3.20 is without loss of generality. For instance, suppose we initially observe a matrix of controls Z_0 which is not normalized by default. Define $\Omega_{z,0} = N^{-1}Z_0'Z_0 \neq I_K$, and let $\Omega_{z,0}^{-1/2}$ be any matrix such that $(\Omega_{z,0}^{-1/2})'\Omega_{z,0}(\Omega_{z,0}^{-1/2}) = N \cdot I_K$, guaranteed to exist if Z_0 is full-rank. Define $Z = Z_0\Omega_{z,0}^{-1/2}$, so that $Z'Z = N \cdot I_K$, and Z_0 has the same column span as Z. Then defining $\mathcal{D}_0 = \left\{d \in \mathbb{R}^K : \|Z_0 d\|^2 \leq r_d^2\right\}$ as a constraint on the contribution of z_i is equivalent to replacing Z_0 with Z and defining $\mathcal{D}_0 = \left\{d : \|d\|^2 \leq r_d^2\right\}$.

2.4 Sufficient Statistics

Observe that it is without loss of generality to replace the observables Y, X, Z with a $K \times 1$ dimensional sufficient statistic $(\hat{b}, \hat{d}) \in \mathbb{R} \times \mathbb{R}^K$ — the OLS coefficients from the long regression — distributed as follows:

$$\begin{pmatrix} \hat{b} \\ \hat{d} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ d \end{pmatrix}, \begin{pmatrix} \frac{1}{1-\|\rho\|^2} & -\frac{\rho'}{1-\|\rho\|^2} \\ -\frac{\rho}{1-\|\rho\|^2} & (I_K - \rho\rho')^{-1} \end{pmatrix} \right).$$

The symmetry of $\mathcal{D}_0$ under Assumption 2.7 suggests a natural decomposition of the nuisance parameter d and observable $\hat{d}$.

Recall from Assumption 2.3 that the sample covariance/correlation between x_i and z_i is represented by the vector $\rho = N^{-1}Z'X$. Hence, the omitted variable bias from a regression of y_i on x_i, controlling for a subset $S \subseteq \{1, \ldots, K\}$ of variables $\{z_{ik}\}_{k \in S}$, is given by $N^{-1/2} \sum_{k \notin S} \rho_k d_k$. Observe that when d is constrained by the nuisance parameter set $\mathcal{D}_0 = \{d : \|d\|^2 \leq r_d^2\}$, the maximal amount of omitted variable bias is generated when all controls are excluded, and d is parallel to $\rho \in \mathbb{R}^K$. This suggests that the quantity $\rho' d$ plays an important role in maintaining size control over $\mathcal{D}_0$, motivating the change-of-variables below.

Definition 2.8. Let U be a $K \times (K-1)$ matrix such that $U'U = I_{K-1}$ and $U'\rho = 0$, implying $UU' = I - \rho\rho'$. Let $(\hat{g}_\rho, \hat{g}_\perp) = (\rho'\hat{d}, U'\hat{d})$ be a rotation of the observable $\hat{d}$, and let $(g_\rho, g_\perp) = (\rho' d, U' d)$ be a rotation of the parameter d. The full sufficient statistic $(\hat{b}, \hat{d})$ transforms into $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$, distributed

$$\begin{pmatrix} \hat{b} \\ \hat{g}_\rho \\ \hat{g}_\perp \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ g_\rho \\ g_\perp \end{pmatrix}, \begin{pmatrix} \frac{1}{(1-\|\rho\|^2)} & -\frac{\|\rho\|^2}{(1-\|\rho\|^2)} & 0 \\ -\frac{\|\rho\|^2}{(1-\|\rho\|^2)} & \frac{\|\rho\|^2}{(1-\|\rho\|^2)} & 0 \\ 0 & 0 & I_{K-1} \end{pmatrix} \right). \tag{2.3}$$

Remark 2.9. Observe that $(b, g_\rho, g_\perp)$ is the coefficient from the population regression of y_i on $(x_i, z_i'\rho\|\rho\|^{-2}, z_i'U)$, so the contribution of controls $z_i'd$ can be decomposed as follows: $z_i'd = (z_i'\rho\|\rho\|^{-2})g_\rho + (z_i'U)g_\perp$. Analogously, $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ is the OLS coefficient from the sample regression of y_i on $(x_i, z_i'\rho\|\rho\|^{-2}, z_i'U)$, and $z_i'\hat{d} = (z_i'\rho\|\rho\|^{-2})\hat{g}_\rho + (z_i'U)\hat{g}_\perp$. One implication of this decomposition is that d and $(g_\rho, g_\perp)$ are related by the equation $g_\rho^2\|\rho\|^{-2} + \|g_\perp\|^2 = \|d\|^2$, so the constraint $\|d\|^2 \leq r_d^2$ becomes $g_\rho^2\|\rho\|^{-2} + \|g_\perp\|^2 \leq r_d^2$.

Hereafter, we apply the change-of-variables from Definition 2.8. Formally, we proceed with

inference in:

$$H_0 : b = 0,\ (g_\rho, g_\perp) \in \mathcal{D}_0^g \quad vs. \quad H_1 : b \neq 0,\ (g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1}, \tag{2.4}$$

$$\mathcal{D}_0^g = \{(g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1} : g_\rho^2 \|\rho\|^{-2} + \|g_\perp\|^2 \leq r_d^2\}$$

where $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K-1}$ from (2.3) is the observable, and $\mathcal{D}_0^g$ substitutes for $\mathcal{D}_0$ as the nuisance parameter set under the null.

2.5 Bivariate Tests

One strategy to construct a valid test for (2.4) is to narrow attention from the full observable $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ to the first two components $(\hat{b}, \hat{g}_\rho)$.

Definition 2.10. The pair $(\hat{b}, \hat{g}_\rho) \in \mathbb{R}^2$ is denoted the *bivariate statistic*.

Because the distribution of the bivariate statistic does not depend on the parameter $g_\perp$, level-α validity of any test based on the bivariate statistic requires only that it control size for all possible values of g_ρ within the set $\mathcal{D}_0^\rho = \{g_\rho : g_\rho^2 \|\rho\|^{-2} \leq r_d^2\}$. Hence, the bivariate testing problem can be written:

$$H_0 : b = 0,\ g_\rho \in \mathcal{D}_0^\rho \quad vs. \quad H_1 : b \neq 0,\ g_\rho \in \mathbb{R} \tag{2.5}$$

where the observable is distributed $(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma)$ with Σ known. In the case of a Gaussian, homoskedastic model, the appropriate covariance matrix is $\Sigma = \Sigma^{hom}$, defined below.

Definition 2.11. $\Sigma^{hom} = \begin{pmatrix} \frac{1}{1-\|\rho\|^2} & -\frac{\|\rho\|^2}{1-\|\rho\|^2} \\ -\frac{\|\rho\|^2}{1-\|\rho\|^2} & \frac{\|\rho\|^2}{1-\|\rho\|^2} \end{pmatrix}$.

Note that under the alternative, the parameter $g_\perp$ disappears, because it is irrelevant to the distribution of $(\hat{b}, \hat{g}_\rho)$. Subsequently, the hypothesis testing problem (2.5) reduces to inference in the presence of a one-dimensional nuisance parameter g_ρ, and the least-favorable distribution technique of Elliott et al. (2015) facilitates the contruction of an optimal valid test.

Definition 2.12. Given a general covariance matrix Σ and L^2 bound r_d^2, let $\varphi^{biv}(\hat{b}, \hat{g}_\rho; \Sigma, r_d^2)$ be the optimal valid test for (2.5) based on the least-favorable distribution. Here, optimality denotes the

maximization of a weighted average of power, in the sense of Elliott et al. (2015). [12] We denote by φ^{biv} the *bivariate test*.

We note that the optimal test φ^{biv} essentially mirrors the standard test based on the long regression when *(i)* r_d is sufficiently large, *(ii)* $\hat{g}_\rho$ lies within the interior of $\mathcal{D}_0^\rho = \{g_\rho : g_\rho^2\|\rho\|^{-2} \leq r_d^2\}$, and *(iii)* $\hat{g}_\rho$ lies away from the boundary $\partial\mathcal{D}_0^\rho = \{-r_d\|\rho\|, r_d\|\rho\|\}$. This conforms to intuition, because the aforementioned conditions imply that $g_\rho \in \mathcal{D}_0^g$ with high likelihood, constituting a scenario in which the null restriction on g_ρ provides no additional information.

Definition 2.13. Given a general covariance matrix Σ and L^2 bound r_d^2, let $\varphi_{GL}^{biv}(\hat{b}, \hat{g}_\rho; \Sigma, r_d^2)$ be a generalized likelihood-ratio test for (2.5). Explicitly,

$$\varphi_{GL}^{biv}(\hat{b}, \hat{g}_\rho; \Sigma, r_d^2) = \frac{\phi(\hat{b}, \hat{g}_\rho; \bar{b}_1, \bar{g}_{1\rho}, \Sigma)}{\phi(\hat{b}, \hat{g}_\rho; 0, \bar{g}_{0\rho}, \Sigma)},$$

where $\phi(\hat{b}, \hat{g}_\rho; b, g_\rho, \Sigma) = (2\pi \det(\Sigma))^{-1/2} \exp[-\frac{1}{2} \cdot (\hat{b} - b, \hat{g}_\rho - g_\rho)'\Sigma^{-1}(\hat{b} - b, \hat{g}_\rho - g_\rho)]$ is the density of the distribution $(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma)$, and $(\bar{b}_1, \bar{g}_{1\rho})$ and $(0, \bar{g}_\rho)$ are the maximum likelihood estimates of (b, g_ρ) under the composite alternative and null, respectively.

Unreported results show that the optimal test φ^{biv} is well-approximated by a generalized likelihood-ratio statistic φ_{GL}^{biv}, defined above. Due to its computational simplicity, φ_{GL}^{biv} may be more attractive than φ^{biv}, and for this reason it is the procedure underlying the Monte Carlo exercises of Section 5.

Bivariate tests may be valid, but in finite samples tbey are suboptimal, even though $(\hat{b}, \hat{g}_\rho)$ and $\hat{g}_\perp$ are independent. The reason is straightforward: *(i)* the excluded statistic $\hat{g}_\perp$ is a signal of $g_\perp$, *(ii)* $g_\perp$ and g_ρ are linked by the constraint $g_\rho^2\|\rho\|^2 + \|g_\perp\|^2 \leq r_d^2$, and *(iii)* g_ρ and b are linked when $Cov[\hat{b}, \hat{g}_\rho] \neq 0$. It naturally follows that the signal to noise ratio in $\hat{g}_\perp$ gauges the deficiency of a bivariate test. When $\|g_\perp\|^2$ is too small to be reliably estimated, we expect bivariate tests to be nearly-optimal, but when $\|g_\perp\|^2$ is large, a test based on the full statistic $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ is more efficient. Ultimately, the form of the optimal procedure depends on the magnitude of $g_\perp$.

[12] For details on the contruction of φ^{biv}, which follows Elliott et al. (2015) closely, we refer readers to Section A.2 of the appendix.

3 Asymptotic Inference in the Gaussian, Homoskedastic Model

3.1 Preliminaries

In Section 2.5, we constructed a bivariate test which is valid for any value of r_d — the L^2 bound on the nuisance parameter — but remains suboptimal in finite samples, because it ignores information from the statistic $\hat{g}_\perp$. In this section, we reevaluate testing with the full observable $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$, and we characterize the optimal procedure as the number of observations N and number of regressors K diverge at the same rate.

Recall the full testing problem (2.4) , restated here:

$$H_0 : b = 0,\ (g_\rho, g_\perp) \in \mathcal{D}_0^g \quad vs. \quad H_1 : b \neq 0,\ (g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1} \tag{3.1}$$

where $\mathcal{D}_0^g = \{(g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1} :\ g_\rho^2 \|\rho\|^{-2} + \|g_\perp\|^2 \leq r_d^2\}$ is the nuisance parameter set, and the observable is $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$, distributed:

$$(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma^{hom}), \quad \hat{g}_\perp \sim \mathcal{N}(g_\perp, I_{K-1}), \quad (\hat{b}, \hat{g}_\rho) \perp \hat{g}_\perp \tag{3.2}$$

Outline of Sections 3.2-3.4

In Section 3.2, we impose an invariance assumption for the remainder of our analysis, reducing the observable to a triplet $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp) \in \mathbb{R}^2 \times \mathbb{R}_+$, where $\hat{r}_\perp = \|\hat{g}_\perp\|$. The relevant nuisance parameter reduces from $(g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1}$ to a pair $(g_\rho, r_\perp) \in \mathbb{R} \times \mathbb{R}_+$, where $r_\perp = \|g_\perp\|$ represents the magnitude of $g_\perp$, and the nuisance parameter restriction becomes $g_\rho^2 \|\rho\|^{-2} + r_\perp^2 \leq r_d^2$. We relate this invariance to testing against a certain class of alternatives.

In Section 3.3, we establish asymptotics in which $N \to \infty$, and the number of controls K grows proportionally with N. Using limit experiments, we derive the main results of this chapter: *(i)* when $r_\perp$ does not exceed the rate $N^{1/4}$, the informational content within $\hat{r}_\perp$ disappears, and the bivariate statistic becomes "asymptotically-sufficient," and *(ii)* when r_d does not exceed the rate $N^{1/4}$, the bivariate test becomes the best valid procedure in large samples.

In Section 3.4, we extend our procedure to accommodatee a set of primary controls, included by default, whose coefficients are not under the reach of the L^2 bound.

3.2 Invariance

In principle, it is possible to derive optimality results for (3.1) using the approach of Elliott et al. (2015), but the high-dimensional nature of $g_\perp \in \mathbb{R}^{K-1}$ renders this infeasible when K is large. Hence, prior to conducting asymptotics, we place an invariance assumption on the observable $\hat{g}_\perp$ which effectively reduces the nuisance parameter $g_\perp \in \mathbb{R}^{K-1}$ to the scalar $r_\perp = \|g_\perp\|$. Subsequently, we relate this assumption to a natural restriction on the structure of the alternative.

Consider the group of rotations $\{O : \mathbb{R}^{K-1} \to \mathbb{R}^{K-1} : \ \|Ov\| = \|v\| \text{ for all } v \in \mathbb{R}^{K-1}\}$. A maximal invariant to this group is the statistic $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$, distributed

$$(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma^{hom}), \quad \hat{r}_\perp^2 \sim \chi^2_{NC}(r_\perp^2, K-1), \quad (\hat{b}, \hat{g}_\rho) \perp \hat{r}_\perp, \tag{3.3}$$

where $\hat{r}_\perp = \|\hat{g}_\perp\|$, and χ^2_{NC} is the noncentral chi-squared distribution. Because the distribution of $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$ depends only on the parameters $(b, g_\rho, r_\perp) \in \mathbb{R}^2 \times \mathbb{R}_+$, restricting attention to rotationally-invariant procedures distills the full test (3.1) to:

$$H_0 : b = 0, \, (g_\rho, r_\perp) \in \mathcal{D}_0^{g,r} \quad vs. \quad H_1 : b \neq 0, \, (g_\rho, r_\perp) \in \mathbb{R} \times \mathbb{R}_+, \tag{3.4}$$

where $\mathcal{D}_0^{g,r} = \left\{(g_\rho, r_\perp) \in \mathbb{R}^2 : \, g_\rho^2 \|\rho\|^{-2} + r_\perp^2 \leq r_d^2\right\}$ is the nuisance parameter set.

Definition 3.1. $\hat{r}_\perp = \|\hat{g}_\perp\|$ and $r_\perp = \|g_\perp\|$.

In this case, the bivariate statistic $(\hat{b}, \hat{g}_\rho)$ remains independent from the excluded statistic $\hat{r}_\perp$, and the nuisance parameter constraint becomes $g_\rho^2 \|\rho\|^{-2} + r_\perp^2 \leq r_d^2$. Hence, the argument for the suboptimality of the bivariate test carries over from Section 2.5. Because $\hat{r}_\perp$ contains information about $r_\perp$ which can potentially be used to tighten the restriction on g_ρ, the bivariate test loses power to the optimal test based on the full statistic $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$, and the size of this gap is linked to the magnitude of $r_\perp$.

This invariance assumption can be interpreted as a restriction that the high-dimensional statistic $g_\perp \in \mathbb{R}^{K-1}$ is uniformly-distributed in all directions under the alternative. In light of the decomposition of controls in Section 2.4: $z_i'd = (z_i'\rho)g_\rho + (z_i'U)g_\perp$, alternatives from this class reflect indifference to the identity of control variables orthogonal to x_i, represented by $z_i'U$. In Section A.3.2 of the appendix, we formally show that test based on the maximal invariant $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$ is, in

the sense of maximizing any weighted average of power, the optimal procedure against this class.

Regarding rotational invariance as a sensible mechanism, we utilize the framework of (3.4) and, by default, take $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$ to be the observable and $(b, g_\rho, r_\perp)$ to be the unknown parameter in the remainder of this chapter.

3.3 Asymptotic Framework and Main Results

We let the number of observations N diverge and characterize the optimal procedure for (3.4) under assumptions on r_d and $r_\perp$. Henceforth, the L^2 bound r_{Nd} and all observables $\{\hat{b}_N, \hat{g}_{N\rho}, \hat{r}_{N\perp}, \rho_N, \Sigma_N^{hom}\}$ are indexed by N, and all limits are taken as $N \to \infty$.

Assumption 3.2. $K/N \to c_K \in (0, 1)$.

Assumption 3.3. $\|\rho_N\|^2 \to c_\rho \in (0, 1)$.

Definition 3.4. Let $\Sigma_A^{biv} = \begin{pmatrix} (1-c_\rho^2)^{-1} & -c_\rho^2(1-c_\rho^2)^{-1} \\ -c_\rho^2(1-c_\rho^2)^{-1} & c_\rho^2(1-c_\rho^2)^{-1} \end{pmatrix}$ be the limit of Σ_N^{hom}.

Assumption 3.2 replicates a high-dimensional environment by allowing the number of regressors to increase proportionally to N, relevant to empirical exercises with many controls. Because the design matrix W_{xz} is deterministic, the limiting fraction of regressors c_K is bounded by 1 for identification. Assumption 3.3 stipulates that the controls z_i do not become perfectly collinear with the regressor of interest x_i, ensuring the long regression retains nontrivial power under fixed local alternatives $b = b_1 \in \mathbb{R}$. In particular, note that the correlation $N^{-1}Z'X = \rho_N$ and covariance matrix Σ_N^{hom} are observables indexed by N which converge in the limit.

Definition 3.5. Let $\bar{r}_N$ be a sequence such that for some $\epsilon > 0$, $\bar{r}_N = O(N^{1/4-\epsilon})$.

We previously argued that the magnitude of the parameter $r_\perp$ regulates the amount of information contained the statistic $\hat{r}_{N\perp}$. This section contains the two main results of this chapter:

1. When $r_\perp$ is bounded by $\bar{r}_N$, the informational content in $\hat{r}_{N\perp}$ vanishes as N grows, so the bivariate statistic $(\hat{b}_N, \hat{g}_{N\rho})$ becomes *asymptotically-sufficient*.

2. Consider the testing problem (3.4). If $r_{Nd} \leq \bar{r}_N$, and we impose the L^2 bound on the nuisance

parameter: $(g_\rho, r_\perp) \in \mathcal{D}_0^g = \{g_\rho^2 \|\rho_N\|^{-2} + r_\perp^2 \leq r_{Nd}^2\}$ under *both* the null and alternative, then the bivariate test is the optimal procedure in the limit.

Our proof methodology is as follows: initially, we consider a thought experiment in which the parameter g_ρ is known to be some fixed distance away from an arbitrary baseline sequence $g_{N\rho}^*$. With this knowledge, we construct a benchmark asymptotic testing problem which bounds the limiting performance of any procedure whenever $r_{Nd}^2 \leq \bar{r}_N$ under both the null and alternative. Subsequently, we note that bivariate tests — which do not rely on knowledge of $g_{N\rho}^*$ — approach this performance ceiling whenever $r_{Nd}^2 \leq \bar{r}_N$.

3.3.1 Main Result: Limit Experiment

For the first main result, we construct a limit experiment to show that $(\hat{b}_N, \hat{g}_{N\rho})$ is asymptotically-sufficient when the magnitude of the parameter $r_\perp$ is bounded by $\bar{r}_N$. Informally, a *limit experiment* is a random variable A governed by a family of distributions which characterizes the asymptotic behavior of a sequence of observables drawn from a fixed collection of contiguous data-generating processes. *Contiguity* between two processes denotes sufficient proximity for their likelihood ratio process to be well-behaved and non-degenerate in the limit. For details, we refer readers to Van der Vaart (1998).

In our framework, $(\hat{b}_N, \hat{g}_{N\rho}, \hat{r}_{N\perp})$ is the finite-sample observable, and a collection of contiguous processes consists of a baseline sequence $(b_N^*, g_{N\rho}^*, r_{N\perp}^*)$ and a set of local deviations from the baseline. The restriction to local deviations is intrinsic to contiguity, because two processes which are not local would admit a single statistic consistent for different limiting values — behavior which no likelihood ratio could reconcile.

Example 3.6.

1. Suppose the baseline is $(b_N^*, g_{N\rho}^*, r_{N\perp}^*) = (0, 0, 0)$ for all N, and let $\mathcal{H} = \{(b, 0, 0) : b \in \mathbb{R}\}$ represent a collection of local deviations indexed by b. For any two values $b_0, b_1 \in \mathbb{R}$, the distribution of the observable $(\hat{b}_N, \hat{g}_{N\rho}, \hat{r}_{N\perp})$ under $(b, g_\rho, r_\perp) = (b_1, 0, 0)$ and $(b, g_\rho, r_\perp) = (b_0, 0, 0)$ admit a proper likelihood ratio in the limit. $\mathcal{H}$ represents a set of contiguous processes, and the limit experiment for $\mathcal{H}$ is a random variable A generated by a family of distributions $\{\mathcal{N}(b, (1 - c_\rho^2)^{-1}) : b \in \mathbb{R}\}$, where $(1 - c_\rho^2)^{-1}$ is the limit of $(1 - \|\rho_N\|^2)^{-1}$ (the finite-sample variance of $\hat{b}$). In this stylized example, the limit experiment summarizes all informational content from processes within $\mathcal{H}$.

2. Let $\mathcal{H} = \{(\beta\sqrt{N}, 0, 0) : \beta \in \mathbb{R}\}$ be a collection of non-local sequences of data-generating parameters indexed by β (recall that b is the local variant of β). For any two values $\beta_0, \beta_1 \in \mathbb{R}$, the statistic $N^{-1/2}\hat{b}$ can consistently distinguish between data generated by $(b, g_\rho, r_\perp) = (\beta_1\sqrt{N}, 0, 0)$ and $(b, g_\rho, r_\perp) = (\beta_0\sqrt{N}, 0, 0)$, so no likelihood ratio bridges the two processes. In turn, $\mathcal{H}$ does not admit a unifying limit experiment.▶

In Definitions 3.7-3.10, we fix the baseline sequence $(b^*_N, r^*_{N\perp}) = (0, 0)$, while the path for the baseline $g^*_{N\rho}$ is allowed to diverge. For ease of notation, we apply a change-of-variables to recenter $\hat{g}_{N\rho}$ and g_ρ around $g^*_{N\rho}$, and we subsequently establish a limit experiment for the new variables, implicitly summarizing the behavior of observables governed by data-generating parameters b local to 0, g_ρ local to $g^*_{N\rho}$, and $0 \leq r_{N\perp} \leq \bar{r}_N$. Finally, we specify a hyperparameter space $\mathcal{H}$, sample experiment $\mathcal{E}_N$, and limit experiment $\mathcal{E}$.

Definition 3.7. Let $g^*_{N\rho}$ be a baseline sequence for g_ρ, and denote by $(\hat{\gamma}_N, \gamma) = (\hat{g}_{N\rho} - g^*_{N\rho},\ g_\rho - g^*_{N\rho})$ the deviation of $(\hat{g}_{N\rho}, g_\rho)$ from $g^*_{N\rho}$. Then $(\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp})$ is distributed:

$$(\hat{b}_N, \hat{\gamma}_N)' \sim \mathcal{N}((b, \gamma)', \Sigma^{hom}_N), \quad \hat{r}^2_{N\perp} \sim \chi^2_{NC}(r^2_\perp, K-1), \quad (\hat{b}_N, \hat{\gamma}_N) \perp \hat{r}_{N\perp}. \tag{3.5}$$

Definition 3.8 (Hyperparameters)**.**

1. Fix sequences $r_{sN\perp}$, $s \in \{0, 1\}$ such that $r_{sN\perp} \in [0, \bar{r}_N]$ for all N.

2. Let $\mathcal{H} = \mathbb{R}\times\mathbb{R}\times\{0, 1\}$ be the hyperparameter space. Elements from $\mathcal{H}$ are denoted $h = (b, \gamma, s)$, where $s \in \{0, 1\}$ switches between data generated by $r_\perp = r_{sN\perp}$ in finite samples.

Remark 3.9. The first two components of h denote fixed local deviations $b \in \mathbb{R}$ and $\gamma \in \mathbb{R}$ from the baseline values $b^*_N = 0$ and $g^*_{N\rho}$. In the third component, it suffices for our derivations to use an indicator s which switches between two arbitrary $r_{0N\perp}$ and $r_{1N\perp}$, rather than to index all possible sequences $r_{N\perp}$ with a hyperparameter.

Definition 3.10 (Experiments)**.**

1. Suppose the *sample observable* $\hat{\theta}_N = (\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp})$ and *limit observable* $A = (A_b, A_\gamma)$ are random variables governed by the hyperparameter h. Conditional on $h = (b, \gamma, s)$, let $\hat{\theta}_N$ be

distributed

$$(\hat{b}_N, \hat{\gamma}_N)' \sim \mathcal{N}((b,\gamma)', \Sigma_N^{hom}), \quad \hat{r}^2_{N\perp} \sim \chi^2_{NC}(r^2_{sN\perp}, K-1), \quad (\hat{b}_N, \hat{\gamma}_N) \perp \hat{r}_{N\perp},$$

and let A be distributed $(A_b, A_\gamma)' \sim \mathcal{N}((b,\gamma)', \Sigma_A^{biv})$.

2. For each $N \in \mathbb{N}$, let P_{Nh} denote the distribution of A_N under h, and define the *sample experiment* to be the collection $\mathcal{E}_N = \{A_N, \mathcal{B}(\mathbb{R}^2 \times \mathbb{R}_+), \{P_{Nh}\}_{h\in\mathcal{H}}\}$, where $\mathcal{B}(\cdot)$ represents the Borel σ-algebra. Let P_h denote the distribution of A under h, and define the *limit experiment* to be the collection $\mathcal{E} = \{A, \mathcal{B}(\mathbb{R}^2), \{P_h\}_{h\in\mathcal{H}}\}$.

Theorem 3.11 (Convergence of Experiments)**.** *For every finite subset $I \subset \mathcal{H}$ and fixed $h_0 \in \mathcal{H}$,*

$$\left\{ \frac{dP_{N,h}}{dP_{N,h_0}}(\hat{\theta}_N) \right\}_{h\in I} \to_d \left\{ \frac{dP_h}{dP_{h_0}}(A) \right\}_{h\in I}$$

under $\hat{\theta}_N \sim P_{N,h_0}$.

Theorem 3.11 establishes the convergence of experiments. Essentially, it states that the ability to identify the hyperparameter $h = (b, \gamma, s)$ from the observable $\hat{\theta}_N = (\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp})$ is characterized by the limit experiment $\mathcal{E}$. But within $\mathcal{E}$ itself, the collection of distributions $\{P_h\}_{h\in\mathcal{H}}$ does not depend on s, so it becomes impossible for $\hat{\theta}_N$ to distinguish between two arbitrary sequences $r_{0N\perp} \leq \bar{r}_N$ and $r_{1N\perp} \leq \bar{r}_N$. Hence, the statistic $\hat{r}_{N\perp}$ — which serves as a signal of $r_\perp$ — is no longer useful, and naturally the Gaussian limit observable $A \sim \mathcal{N}((b,\gamma)', \Sigma_{A,biv})$ emulates the asymptotic behavior of the (baseline-centered) bivariate statistic $(\hat{b}_N, \hat{\gamma}_N) \sim \mathcal{N}((b,\gamma)', \Sigma_{biv})$. For this reason, we denote the bivariate statistic *asymptotically-sufficient* when $r_\perp \leq \bar{r}_N$.

3.3.2 Main Result: Asymptotic Optimality of the Bivariate Test

In this section, we impose the L^2 bound $g^2_\rho \|\rho\|^{-2} + r^2_\perp \leq r^2_{Nd}$ under *both* the null and alternative, and we impose the bound $r_{Nd} = O(N^{1/4-\epsilon})$, implicitly constraining the magnitude of the parameter $r_\perp$. Applying the first main result, we prove that a collection of four limit problems — corresponding to different baseline sequences $g^*_{N\rho}$ — benchmarks the performance of any sequence of valid tests. Subsequently, we note that the bivariate test φ^{biv} of Section 2.5 approaches the benchmark and is, hence, asymptotically-optimal in the weighted-average-power sense of Elliott et al. (2015).

Assumption 3.12. *Let r_{Nd} be a sequence such that $r_{Nd} = O(N^{1/4-\epsilon})$.*

Formally, recall the testing problem (3.4), and suppose the nuisance parameter $(g_\rho, r_\perp)$ now satisfies $(g_\rho, r_\perp) \in \mathcal{D}^{g,r} = \left\{(g_\rho, r_\perp) \in \mathbb{R}^2 : g_\rho^2 \|\rho\|^{-2} + r_\perp^2 \leq r_{Nd}^2\right\}$ under both H_1 and H_0. Conditional on the baseline sequence $g^*_{N\rho}$, we re-state (3.4) here with recentered variables substituted:

$$H_0 : b = 0,\ (\gamma, r_\perp) \in \mathcal{D}^{\gamma,r} \quad vs. \quad H_1 : b \neq 0,\ (\gamma, r_\perp) \in \mathcal{D}^{\gamma,r}. \tag{3.6}$$

In this context, $\mathcal{D}^{\gamma,r} = \{(\gamma, r_\perp) : (g^*_{N\rho} + \gamma, r_\perp) \in \mathcal{D}^{g,r}\} = \{(\gamma, r_\perp) : (g^*_{N\rho} + \gamma)^2 \|\rho\|^{-2} + r_\perp^2 \leq r_{Nd}^2)\}$ is a nuisance parameter set, and $(\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp})$ is the observable, distributed as in (3.5).

Subsequently, we construct *limit problems* which extract the essential components of (3.6) under Assumption 3.12, and we show that for any fixed baseline sequence $g^*_{N\rho}$, a sequence of valid finite-sample tests φ_N cannot outperform the optimal, valid test in the limit problem.

Definition 3.13. The bivariate limit testing problem is:

$$H_0 : b = 0,\ \gamma \in \Gamma \quad vs. \quad H_1 : b \neq 0,\ \gamma \in \mathbb{R}, \tag{3.7}$$

where $\Gamma \subseteq \mathbb{R}$ depends on the path for $g^*_{N\rho}$, and $A = (A_b, A_\gamma)' \sim \mathcal{N}((b, \gamma)', \Sigma_A^{biv})$ is the limit observable.

In order to compare testing performance, we evaluate rejection profiles, which requires us to define the power function π.

Definition 3.14. For a general observable $\hat{\theta}$, data-generating parameter θ, and randomized test $\varphi(\hat{\theta}) \in [0, 1]$, let $\pi(\hat{\theta}; \theta) = E_\theta[\varphi]$ denote the rejection probability of φ when $\hat{\theta}$ is generated by θ. Then $E_\theta[\varphi]$ represents size when θ is in the null, and $E_\theta[\varphi]$ represents power when θ is in the alternative.

Theorem 3.15 (Asymptotic Representation)**.** *Given Assumption 3.12, suppose in each case below that $\varphi_N(\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp})$ is a sequence of valid tests in (3.6) for which $\pi(\varphi_N; b, \gamma, r_\perp)$ converges whenever $(b, \gamma) \in \mathbb{R}^2$ and $r_\perp = 0$ are fixed. Let $r_{N\perp}$ be any sequence such that $r_{N\perp} \leq \bar{r}_N$.*

1. *Suppose $g^*_{N\rho} \in [-r_{Nd}\|\rho_N\|, r_{Nd}\|\rho_N\|]$ for all N, and $\min\{|g^*_{N\rho} - r_{Nd}\|\rho_N\||,\ |g^*_{N\rho} + r_{Nd}\|\rho_N\||\} \to \infty$. Setting $\Gamma = \mathbb{R}$, there exists a valid test φ in (3.7) such that $\lim \pi(\varphi_N; b, \gamma, r_{N\perp}) \leq \pi(\varphi; b, \gamma)$*

for each $(b, \gamma) \in \mathbb{R}^2$.

2. *Suppose* $|g^*_{N\rho} - r_{Nd}\|\rho_N\|| \to 0$. *Setting* $\Gamma = \mathbb{R}_-$, *there exists a valid test* φ *in (3.7) such that* $\lim \pi(\varphi_N; b, \gamma, r_{N\perp}) \leq \pi(\varphi; b, \gamma)$ *for each* $(b, \gamma) \in \mathbb{R}^2$.

3. *Suppose* $|g^*_{N\rho} + r_{Nd}\|\rho_N\|| \to 0$. *Setting* $\Gamma = \mathbb{R}_+$, *there exists a valid test* φ *in (3.7) such that* $\lim \pi(\varphi_N; b, \gamma, r_{N\perp}) \leq \pi(\varphi; b, \gamma)$ *for each* $(b, \gamma) \in \mathbb{R}^2$.

4. *Suppose* $g^*_{N\rho} \to 0$ *and* $r_{Nd} \to c_r c_\rho^{-1} \in \mathbb{R}$ *for* $c_r \geq 0$. *Setting* $\Gamma = [-c_r, c_r]$, *there exists a valid test* φ *in (3.7) such that* $\lim \pi(\varphi_N; b, \gamma, r_{N\perp}) \leq \pi(\varphi; b, \gamma)$ *for each* $(b, \gamma) \in \mathbb{R}^2$.

Proof. This proof follows that of Theorem 15.1 in Van der Vaart (1998) closely. Without loss of generality, consider the first case, and suppose φ_N is a sequence of valid tests in (3.6). Because all processes are Gaussian, this problem is sufficiently well-behaved to apply Prohorov's theorem, which states that there exists a subsequence N' and some (possibly-randomized) test φ for which $\varphi_{N'}$ converges in distribution under every fixed value of $(b, \gamma) \in \mathbb{R}^2$, where $r_\perp = 0$ is fixed.

Hence, the assumptions of the Asymptotic Representation Theorem of Van der Vaart (1998) are satisfied, and in conjunction with Theorem 3.11, it implies that there exists a (possibly-randomized) test statistic φ in the bivariate limit testing problem (3.7) such that for all $(b, \gamma) \in \mathbb{R}^2$, the distribution of φ is the limit distribution of $\varphi_{N'}$ with $r_\perp = 0$ along the subsequence N'. The consequences of this are twofold: (i) because $\varphi_{N'}$ is valid, $\pi(\varphi_N; 0, \gamma, 0) \leq \alpha$ as $N \to \infty$ for all fixed $\gamma \in \mathbb{R}$, so φ is a valid test in (3.7), and (ii) Le Cam's third lemma, in conjunction with Theorem 3.11, implies that for any sequence $r_{N\perp} \leq \bar{r}_N$,

$$\begin{aligned} \lim_{N' \to \infty} \pi(\varphi_{N'}; b, \gamma, r_{N\perp}) &= \lim_{N \to \infty} \pi(\varphi_N; b, \gamma, r_{N\perp}) \\ &= \lim_{N \to \infty} \pi(\varphi_N; b, \gamma, 0) \\ &= \pi(\varphi; b, \gamma) \end{aligned}$$

Then the power profile of the sequence φ_N is equivalent to the power profile of φ in the bivariate limit testing problem, which admits at least one valid test which weakly-dominates the limiting behavior of φ_N. □

Theorem 3.15 states that when r_{Nd} is bounded by $\bar{r}_N$, there exists an appropriate specification of Γ such that the rejection profile of any well-behaved sequence of valid tests φ_N can be replicated by a valid test within the bivariate limit problem (3.7). Ultimately, the form of Γ depends on the path

for $g^*_{N\rho}$. In the knife-edge case where $g^*_{N\rho}$ is close to the boundary of interval $[-r_{Nd}\|\rho_N\|, r_{Nd}\|\rho_N\|]$, Γ becomes one-sided (Cases 1 and 2), and when $g^*_{N\rho}$ is completely embedded within the interval, Γ spans both sides of $g^*_{N\rho}$ (Case 3). Forshadowed by results from Section 2.5, Case 3 reduces to a standard problem, where inference based on the long regression is the optimal unbiased procedure. When r_{Nd} does not diverge, Γ remains a compact interval (Case 4) and without loss of generality we assume the baseline value converges to 0.

In essence, enforcing $r_{Nd} \leq \bar{r}_N$ suppresses the magnitude of the parameter $r_\perp$, rendering $\hat{r}_{N\perp}$ uninformative and the bivariate statistic asymptotically sufficient. Hence, there exists a valid test in (3.7) whose rejection profile — in particular, whose power against *any* local alternative $(b_1, \gamma_1) \in \mathbb{R}^2$ — matches that of any limiting sequence of valid tests φ_N, so long as $r_{Nd} \leq \bar{r}_N$. In this sense, a sequence of valid tests for (3.6) cannot *outperform* the class of bivariate tests for (3.7), even when the baseline $g^*_{N\rho}$ is known. It remains to see whether the performance ceiling implied by the benchmark bivariate limit problem can be attained by sequences of finite-sample tests.

Recall the optimal finite-sample bivariate test $\varphi^{biv}(\hat{b}_N, \hat{g}_{N\rho}; \Sigma^{hom}_N, r^2_{Nd}) = \varphi^{biv}(\hat{b}_N, \hat{\gamma}_N + g^*_{N\rho}; \Sigma^{hom}_N, r^2_{Nd})$ constructed in Section 2.5. Applying Elliott et al. (2015) shows that in the limit, φ^{biv} — which, itself does not depend on $g^*_{N\rho}$ — closely approximates what is possible employing knowledge of $g^*_{N\rho}$ as $N \to \infty$.

Proposition 3.16. *Under Assumption 3.12, φ^{biv} becomes the asymptotically-optimal procedure for inference in*

$$H_0 : b = 0,\ (g_\rho, r_\perp) \in \mathcal{D}^{g,r} \quad vs. \quad H_1 : b \neq 0,\ (g_\rho, r_\perp) \in \mathcal{D}^{g,r}$$

$$\mathcal{D}^{g,r} = \left\{(g_\rho, r_\perp) \in \mathbb{R}^2 :\ g_\rho^2 \|\rho\|^{-2} + r_\perp^2 \leq r_{Nd}^2\right\},$$

in the weighted-average-power maximizing sense of Elliott et al. (2015).

Remark 3.17. In terms of the unscaled coefficient δ, Assumption 3.12 implies $\|\delta\| \leq N^{-1/4-\epsilon}$, so the contribution of control variables z_i must shrink to zero in the limit for optimality to hold. Depending on context, this may be unrealistic, but if a set of primary controls is identified and separated from z_i in a preliminary stage, then Assumption 3.12 might become more plausible. We detail this procedure in Section 3.4.

3.4 Separation of Primary Controls

The preceeding sections established that in the Gaussian, homoskedastic model, the bivariate test is valid for a given L^2 bound on δ, and it is optimal when the bound implies $\|\delta\| \leq N^{-1/4-\epsilon}$. However, when the controls represent factors of known economic importance or fixed effects, a restriction on their contribution may be inappropriate. In this section, we separate controls into primary and secondary groups, and we account for the primary controls with a preliminary regression.

Letting $\tilde{N}$ denote the number of observations and $i = 1, ..., \tilde{N}$ the observational index, suppose now that the full model is

$$\tilde{y}_i = \tilde{x}_i\beta + \tilde{z}_i'\delta + \tilde{v}_i'\eta + \tilde{e}_i, \tag{3.8}$$

where $\tilde{y}_i$ is a scalar dependent variable, $\tilde{z}_i \in \mathbb{R}^K$ is a vector of *secondary* control variables, $\tilde{v}_i \in \mathbb{R}^J$ is a vector of *primary* control variables (included in the regression by default), and $K < \tilde{N} - J$. Let $\tilde{z}_i$ assume the same role as z_i of Section 2 so that its contribution to $\tilde{y}_i$ falls under the scope of the L^2 bound. In contrast, $\tilde{v}_i$ is identified by the practitioner *a priori*, and we do not constrain its coefficient η.

As before, we condition on deteministic regressors and impose normality, and stacking observations in matrices, we write

$$\underset{(\tilde{N}\times 1)}{\tilde{Y}} = \underset{(\tilde{N}\times 1)}{\tilde{X}}\beta + \underset{(\tilde{N}\times K)}{\tilde{Z}}\delta + \underset{(\tilde{N}\times J)}{\tilde{V}}\eta + \underset{(\tilde{N}\times 1)}{\tilde{e}}. \tag{3.9}$$

Assumption 3.18. *$\tilde{W}_{xzv} = (\tilde{X}, \tilde{Z}, \tilde{V})$ is a deterministic matrix.*

Assumption 3.19. *$\tilde{e}_i|\{\tilde{x}_i, \tilde{z}_i, \tilde{v}_i\}_{i=1}^N \sim_{iid} \mathcal{N}(0,1)$. In matrix notation, $\tilde{e}|\tilde{W}_{xzv} \sim \mathcal{N}(0, I_{\tilde{N}})$.*

To remove the confounding effect of the nuisance parameter η, we take residuals from a preliminary regression of $\tilde{Y}, \tilde{X}, \tilde{Z}, \tilde{e}$ on $\tilde{V}$, in turn excluding η from any bound. Formally, we impose invariance to the group of transformations $\mathcal{V} = \{\tilde{Y} \mapsto \tilde{Y} + v,\ v \in \text{span}(\tilde{V})\}$.

Let $N = \tilde{N} - J$, and let Q_v be an $N \times \tilde{N}$ matrix such that $Q_v\tilde{V} = 0$ and $Q_v'Q_v = I_N$. Define $(Y, X, Z, e) = Q_v(\tilde{Y}, \tilde{X}, \tilde{Z}, \tilde{e})$, $\rho = N^{-1}Z'X$, and $W_{xz} = (X, Z)$. Then $Y = Q_v\tilde{Y}$ is a maximal invariant of $\mathcal{V}$ whose distribution is modeled by (3.10):

$$Y = X\beta + Z\delta + e, \quad e|W_{xz} \sim \mathcal{N}(0, I_N). \tag{3.10}$$

Without loss of generality, we proceed with inference in (3.10). As before, let $\beta = 0$ under the null and let $\sqrt{N}\beta = b$, $\sqrt{N}\delta = d$ be local parameters. For a given nuisance space $\mathcal{D}_0 \subseteq \mathbb{R}^K$ we test

$$H_0 : b = 0,\, d \in \mathcal{D}_0 \quad vs. \quad H_1 : b \neq 0. \tag{3.11}$$

Assumption 3.20. $N^{-1}W_{xz}'W_{xz} = \begin{pmatrix} 1 & \rho' \\ \rho & I_K \end{pmatrix}$.

Assumption 3.21. *For $r_d > 0$, let $\mathcal{D}_0 = \{d \in \mathbb{R}^K : \|d\|^2 \leq r_d^2\}$.*

As in Section 2, we normalize regressors and restrict the nuisance parameter with an L^2 bound on d. In this context $\|d\|^2 = N^{-1}\|Zd\|^2$ represents the partial contribution of secondary controls after projecting off primary controls, and it naturally relates to $R^2_{\tilde{y},\tilde{z}|\tilde{v}}$ the population partial R^2 of $\tilde{z}_i$ on $\tilde{y}_i$ conditional on $\tilde{v}_i$, via the following equation:

$$R^2_{\tilde{y},\tilde{z}|\tilde{v}} = \frac{\|Zd\|^2}{E[\|Y\|^2]} = \frac{N^{-1}\|d\|^2}{N^{-1}\|d\|^2 + 1}.$$

Hence, mirroring Section 2.3, the L^2 bound $\|d\|^2 \leq r_d^2$ is equivalent to the constraint $R^2_{\tilde{y},\tilde{z}|\tilde{v}} \leq c_R^2 = \frac{N^{-1}r_d^2}{N^{-1}r_d^2+1}$ when errors are homoskedastic.

By default, all results derived in Sections 3.3 carry over to Section 3.4, with an additional stipulation of $\mathcal{V}$-invariance. For concision, we do not state them in detail, but we reiterate that within this environment: *(i)* the bivariate statistic is valid, and *(ii)* it yields the asymptotically-optimal test when the L^2 bound is active under both the null and alternative, and the sequence r_{Nd} satisfies $r_{Nd} = O(N^{1/4-\epsilon})$. As noted in Remark 3.17, a shrinking bound on d may be less restrictive in this framework, where $\tilde{z}_i$ is secondary.

4 Robust Implementation under General Error Structures

4.1 Preliminaries

The preceeding sections characterize inference in a stylized model with Gaussian, homoskedastic errors, assumptions which may not be tenable in applications. In this section, we construct a procedure which is valid under more expansive conditions on the underlying data-generating process

while remaining nearly optimal in the sense of Section 3.3 when errors are, in fact, Gaussian and homoskedastic. Our procedure combines the bivariate test φ^{biv} with the high-dimensional, cluster-robust estimator of Li (2016).

Formally, consider the model

$$\tilde{y}_i = \tilde{x}_i\beta + \tilde{z}_i'\delta + \tilde{v}_i'\gamma + \tilde{e}_i, \quad i = 1, ..., N, \tag{4.1}$$

where N denotes the number of observations, $\tilde{y}_i$ is the dependent variable , $\tilde{x}_i$ is the regressor of interest, $\tilde{z}_i \in \mathbb{R}^K$ is a vector of secondary controls, and $\tilde{v}_i \in \mathbb{R}^J$ is a vector of primary controls. Stacking observations into matrices $\tilde{Y}, \tilde{X}, \tilde{Z}, \tilde{V}$, we write

$$\underset{(N\times 1)}{\tilde{Y}} = \underset{(N\times 1)}{\tilde{X}}\beta + \underset{(N\times K)}{\tilde{Z}}\delta + \underset{(N\times J)}{\tilde{V}}\eta + \underset{(N\times 1)}{\tilde{e}}. \tag{4.2}$$

As before, we condition on deterministic regressors $\tilde{W}_{xzv} = (\tilde{X}, \tilde{Z}, \tilde{V})$, but we do not normalize $\tilde{W}_{xzv}$, and we do not assume that errors are Gaussian. Instead, in Assumption 4.1, we impose the weaker condition of strict exogeneity.

Assumption 4.1. $E[\tilde{e}|\tilde{W}_{xzv}] = 0$

For a vector $\mu \in \mathbb{R}^N$ and $N \times N$ matrix M, let $\mu^p \in \mathbb{R}^n$ denote the p'th block of μ. To accommodate heteroskedasticity and clustering, let P denote the number of clusters, and suppose there are n observations within each cluster $p = 1, \ldots, P$, so that $nP = N$. Then we permit the covariance matrix to take the following form:

Assumption 4.2.

$$\Omega = E[\tilde{e}\tilde{e}'|\tilde{W}_{xzv}] = \begin{pmatrix} \Omega^{11} & 0 & \ldots & 0 \\ 0 & \Omega^{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Omega^{PP} \end{pmatrix},$$

where for each $p = 1, \ldots, P$, the $n \times 1$ vector $\tilde{e}^p = (\tilde{e}_{np-n+1}, \ldots, \tilde{e}_{np})'$ represents the p'th block of $\tilde{e}$, and $\Omega^{pp} = E[\tilde{e}^p(\tilde{e}^p)'|\tilde{W}_{xzv}]$ represents an $n \times n$ intra-cluster covariance matrix.

Prompted by the results of Section 3.4, we define observables following a preliminary regression on the primary controls $\tilde{V}$.

Definition 4.3.

1. Let $\tilde{M}_v = I_N - \tilde{V}(\tilde{V}'\tilde{V})^{-1}\tilde{V}'$, and define $(Y, X, Z) = (\tilde{M}_v\tilde{Y}, \tilde{M}_v\tilde{X}, \tilde{M}_v\tilde{Z})$ to be the residuals from preliminary regressions of all variables on $\tilde{V}$.

2. Let $\{y_i, x_i, z_i\}_{i=1}^N$ denote the individual observations within (Y, X, Z), and without loss of generality, normalize X so that $N^{-1}\|X\|^2 = 1$.

Suppose that for a given bound $r > 0$, the practitioner wishes to test

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta \neq 0, \tag{4.3}$$

restricting the L^2 contribution of controls to satisfy

$$\sum_{i=1}^{N}(z_i'\delta)^2 = \|Z\delta\|^2 \leq r^2 \tag{4.4}$$

under both the null and alternative. Letting $\hat{R}^2 = \frac{\|Z\delta\|^2}{\|Y\|^2}$ denote the in-sample partial R^2 of $\tilde{z}_i$ on $\tilde{y}_i$, the practitioner can relate the bound (4.4) to a partial R^2 assumption via the equation $\hat{R}^2 \leq r^2\|Y\|^{-2}$.

Below, we construct a heteroskedasticity and cluster-robust bivariate statistic φ^{rob} such that *(i)* φ^{rob} is uniformly-valid for (4.3) in large samples, and *(ii)* φ^{rob} is optimal when the true data-generating process conforms to the baseline model (3.8), and $r = O(N^{1/4-\epsilon})$ for some $\epsilon > 0$. We outline the full implementation of our inferential procedure below, and we detail its construction in Section 4.3.

4.2 Construction of the Robust Bivariate Test

1. The observations are $\{\tilde{y}_i, \tilde{x}_i, \tilde{z}_i, \tilde{v}_i\}_{i=1}^N$, with $\tilde{y}_i \in \mathbb{R}$, $\tilde{x}_i \in \mathbb{R}$, $\tilde{z}_i \in \mathbb{R}^K$, and $\tilde{v}_i \in \mathbb{R}^J$.

2. Accounting for $\tilde{v}_i$ with a preliminary regression, the observables become $\{y_i, x_i, z_i\}_{i=1}^N$, as in Definition 4.3.

3. Let $(Z'Z)^{-1}Z'X$ be the OLS coefficient from a regression of x_i on z_i, and let $\hat{X} = Z(Z'Z)^{-1}Z'X$ be the predicted value of x_i from that regression.

4. Define $\rho^2 = \|\hat{X}\|^2/\|X\|^2$ to be the sample R^2 of z_i on x_i.

5. Define $Z_{short} = \hat{X}/\rho^2$, and denote by $z_{i,short}$ the individual observations in Z_{short}.

6. Let $(\hat{\beta}, \hat{G}_\rho)$ be the OLS coefficient from a regression of y_i on $(x_i, z_{i,short})$, and define $(\hat{b}, \hat{g}_\rho) = \sqrt{N}(\hat{\beta}, \hat{G}_\rho)$ to be the local versions of that coefficient.[13]

7. Let $\hat{\Sigma}$ be the cluster-robust variance estimator of Li (2016).

8. For φ^{biv} from Section 2.5, define $\varphi^{rob} = \varphi^{biv}(\hat{b}, \hat{g}_\rho; \hat{\Sigma}, r_d^2)$ to be the robust Bivariate Test based on the least-favorable distribution.

9. For φ^{biv}_{GL} from Section 2.5, define $\varphi^{rob}_{GL} = \varphi^{biv}_{GL}(\hat{b}, \hat{g}_\rho; \hat{\Sigma}, r_d^2)$ to be the robust GLR Bivariate Test. Unreported results demonstrate that φ^{rob}_{GL} closely approximates φ^{rob} and, hence, may be used as an alternative to φ^{rob}.

Notation. In the remainder of Section 4, define ρ, $(\hat{\beta}, \hat{g}_\rho)$, $(\hat{\Sigma}, \Sigma)$ as above.

4.3 Asymptotics of the Robust Bivariate Test

Suppose, initially, that Σ and the Gaussianity of $(\hat{b}, \hat{g}_\rho)$ are given. Then the results of Section 2.5 imply that *(i)* the bivariate test $\varphi^{biv}(\hat{b}, \hat{g}_\rho; \Sigma, r^2)$ is valid for any value of r, and *(ii)* when the errors $\tilde{e}_i$ are homoskedastic, φ^{biv} is asymptotically-optimal under a sequence for r such that $r = O(N^{1/4-\epsilon})$. In practice, we take $(\hat{b}, \hat{g}_\rho)$ to be asymptotically-normal, and we plug in $\hat{\Sigma}$ for Σ.

Definition 4.4. Let $\varphi^{rob}(\hat{b}, \hat{g}_\rho) = \varphi^{biv}(\hat{b}, \hat{g}_\rho; \hat{\Sigma}, r_d^2)$, where $\hat{\Sigma}$ is the cluster-robust standard error of Li (2016).

In order for φ^{rob} to retain both aforementioned properties of the infeasible procedure φ^{biv} uniformly over sequences of the true data-generating coefficients, it remains to show that in a manner invariant to the values of such coefficients, $(\hat{b}, \hat{g}_\rho)$ is asymptotically-Gaussian, and $\hat{\Sigma}$ consistently estimates $\Sigma = E\left[(\hat{b}, \hat{g}_\rho)'(\hat{b}, \hat{g}_\rho)\right]$.[14] Because our framework conditions on the design matrix $\tilde{W}_{xzv}$, this reduces to checking deterministic properties of the regressors. Henceforth, all observables are indexed by N, all limits are taken as $N \to \infty$, and we suppress subscripts in N when context allows.

Assumption 4.5. *1.* $N \to \infty$, $K/N \to c_K \in (0,1)$, *and* $\|\rho\| \to_p c_\rho \in (0,1)$.

[13] We note that $\hat{\beta}$ is the OLS coefficient on $\tilde{x}_i$ from the long regression of $\tilde{y}_i$ onto $(\tilde{x}_i, \tilde{z}_i, \tilde{v}_i)$. Moreover, $\hat{\beta} + \hat{G}_\rho$ is the OLS coefficient on $\tilde{x}_i$ from the short regression of $\tilde{y}_i$ onto $(\tilde{x}_i, \tilde{v}_i)$.

[14] It is a straightforward exercise to show that the procedure of Section 4.2 is equivalent to the bivariate test of Section 2.5 when Σ^{hom} is substituted for $\hat{\Sigma}$ in step 8.

2. $\max_{i=1,\ldots,N} \mu_i^2 = o(\sqrt{N})$ *for* $\mu \in \{X, Z_{short}\}$.

3. $\max_{i=1,\ldots,N} \Omega_{ii} = O(1)$.

Theorem 4.6. *Suppose* $b_0, g_{0\rho}$ *are the true data-generating values of* b, g_ρ*, and let* $\Sigma = E[(\hat{b}, \hat{g}_\rho)'(\hat{b}, \hat{g}_\rho)]$. *Then under Assumption 4.5,*

$$\Sigma^{-1/2} \begin{pmatrix} \hat{b} - b_0 \\ \hat{g}_\rho - g_{0\rho} \end{pmatrix} \to_d \mathcal{N}(0, I_2),$$

uniformly over all sequences of data-generating coefficients.

Theorem 4.6 follows from the Lindeberg-Levy Central Limit Theorem. In particular, note that Assumption 4.5, Part 2 — stated in anticipation of Theorem 4.9 — is stronger than necessary for Theorem 4.6 to hold.

Because K is of the same order as N, the usual Huber-Eicker-White standard errors are not consistent for Σ. Hence, we adopt a method of Cattaneo et al. (2015) — extended to the clustered case in Li (2016) — to estimate Σ, and below we state assumptions ensuring consistency of our estimator. We remark that $\hat{\Sigma}$ exactly reduces to the estimator of Cattaneo et al. (2015) when there is no cluster structure.

Definition 4.7. For an $N \times N$ matrix M , let M^{pq} denote the $n \times n$ matrix representing the (p, q)'th block of M, where $p, q = 1, \ldots, P$. Let $M * M$ denote the $(Nn) \times (Nn)$ matrix

$$M * M = \begin{pmatrix} M^{11} \otimes M^{11} & M^{12} \otimes M^{12} & \ldots & M^{1P} \otimes M^{1P} \\ M^{21} \otimes M^{21} & M^{22} \otimes M^{22} & \ldots & M^{2P} \otimes M^{2P} \\ \vdots & \vdots & \ddots & \vdots \\ M^{P1} \otimes M^{P1} & M^{P2} \otimes M^{P2} & \ldots & M^{PP} \otimes M^{PP} \end{pmatrix},$$

Assumption 4.8. *Let* $\tilde{M}_{xzv} = I_N - \tilde{W}_{xzv}(\tilde{W}'_{xzv}\tilde{W}_{xzv})^{-1}\tilde{W}'_{xzv}$. *Then* $\sum_{i,j=1}^{Nn}[(\tilde{M}_{xzv} * \tilde{M}_{xzv})_{ij}^{-1}]^2 = O(N^2)$.

Theorem 4.9. *Define* $\hat{\Sigma}$ *to be the cluster-robust variance estimator of Li (2016). Then conditionally on the regressors, (i)* $E[\hat{\Sigma}] = \Sigma$*, and (ii) under Assumptions 4.5 and 4.8,* $Var(\hat{\Sigma} - \Sigma) \to 0$ *uniformly over all sequences of data-generating coefficients.*

Theorem 4.9 shows that under some regularity assumptions on the regressors, $\hat{\Sigma}$ is consistent when the squared sum of entries of the matrix $(\tilde{M}_{xzv} * \tilde{M}_{xzv})^{-1}$ is bounded in the limit. As in Cattaneo et al. (2015), there exist primitive conditions that ensure this constraint on $(\tilde{M}_{xzv} * \tilde{M}_{xzv})^{-1}$, but in applications where $(\tilde{M}_{xzv} * \tilde{M}_{xzv})^{-1}$ can simply be inspected, these conditions are uninstructive, so we omit their exposition. [15]

5 Comparisons with Post-Double-Selection

5.1 Theory

In this section, we examine the differences between the bivariate test and the Post-Double-Selection procedure of Belloni et al. (2014) with a theoretical exercise. Consider the linear model:

$$y_i = x_i\beta + z_i'\delta + e_i = x_i\beta + \sum_{k=1}^{K} z_{ik}\delta_k + e_i, \quad i = 1, \ldots, N,$$

where x_i is the regressor of interest, and $z_i = (z_{i1}, \ldots, z_{iK})'$ is the full vector of controls. As in the preceeding sections, suppose the number of regressors K is of the same order as N, and regressors are normalized so that $N^{-1}\sum_{i=1}^{N} x_i^2 = N^{-1}\sum_{i=1}^{N} z_{ik}^2$ and $N^{-1}\sum_{i=1}^{N} z_{ik}z_{ik'} = 0$ for all $k, k' = 1, \ldots, K$. In this framework, $\rho = N^{-1}\sum_{i=1}^{N} x_i z_i = (\rho_1, \ldots, \rho_K)'$ is a vector of sample correlations.

A convential procedure for inference on β is to select a subset of control variables $\mathcal{S} \subseteq \{1, \ldots K\}$ and regress y_i on $(x_i, \{z_{ik}\}_{k\in S})$. Then a regression of y_i on x_i controlling for $\{z_{ik}\}_{k\in\mathcal{S}}$ introduces omitted variable bias of $\omega = \sum_{k\in\mathcal{S}^c} \rho_k\delta_k$, where the set of excluded controls $\mathcal{S}^c$ is the complement of $\mathcal{S}$. Without constraints on (ρ, δ), this can lead to asymptotic size distortion, as in Example 2.5.

Suppose, instead, that one is able to identify *(i)* coefficients δ_k of order dominating $N^{-1/2}$ and *(ii)* correlations ρ_k of order dominating $N^{-1/2}$, following which a subset $\mathcal{S} \subseteq \{1, \ldots, K\}$ is designated to contain all controls z_{ik} identified by either *(i)* or *(ii)*. Then $\rho_k\delta_k = O(N^{-1})$ for all $k \in \mathcal{S}^c$, and the total omitted variable bias from a regression of y_i on x_i controlling for $\{z_{ik}\}_{k\in\mathcal{S}}$ is

$$\omega = \sum_{k\in\mathcal{S}^c} \rho_k\delta_k = \sum_{k\in\mathcal{S}^c} O(N^{-1}) = |\mathcal{S}^c| \cdot O(N^{-1}),$$

[15] Informal Monte Carlo results suggest that $(\tilde{M}_{xzv} * \tilde{M}_{xzv})^{-1}$ does not exceed the rate $O(N)$ under a wide array of data-generating processes.

where $|\mathcal{S}^c|$ represents the number of excluded controls. If $|S^c|$ has order $N^{1/2}$ or greater, then $\omega = |\mathcal{S}^c|O(N^{-1}) = O(N^{-1/2})$ is potentially large enough to induce non-negligible asymptotic size distortion. To limit the magnitude of ω, one can assume sparsity in $\delta = (\delta_1, \ldots, \delta_K)$, formally imposing the bound $\|\delta\|_0 = o(N^{1/2})$ to constrain the maximum number of nonzero coefficients by the rate $o(N^{1/2})$. With this assumption, the omitted variable bias from such a procedure cannot exceed $o(N^{1/2}) \cdot O(N^{-1}) = o(N^{-1/2})$, which is negligible to asymptotic testing.

This is the key insight of Belloni et al. (2014), who use LASSO regressions to identify correlations and coefficients of order dominating $N^{-1/2}$ to include in a final stage regression. They denote this a "Post-Double-Selection" procedure, where "double" refers to the fact that the correlation structure between x_i and z_i is taken into account, in addition to δ. To curb omitted variable bias, Belloni et al. (2014) assume sparsity of order $o(N^{1/2})$ in the coefficient δ.[16]

In contrast, the methodology of this chapter does not constrain the number of nonzero components of δ. Instead, we impose an L^2 bound on δ which effectively restricts $\omega = \rho'\delta$, and the following thought experiment exposes the main difference between the ensuing bivariate test and Post-Double-Selection.

Example 5.1. Fix $c_\rho \in (0, 1)$. Suppose the practitioner imposes the following L^2 bound: $\|d\|^2 \leq r_d^2 = 1 \Leftrightarrow \|\delta\| \leq N^{-1}$. If $\rho = c_\rho(N^{-1/2}, \ldots, N^{-1/2})'$, and the true data-generating parameter is $\delta = (N^{-1}, \ldots, N^{-1})'$, then the bound is satisfied, and the bivariate test is valid. However, because neither ρ nor δ has components dominating order $N^{-1/2}$, Post-Double-Selection has a tendency to excludes a nontrivial proportion of controls z_{ik}, $i = 1, \ldots, K$, admitting omitted variable bias of $\omega = \rho'\delta = c_\rho N^{-1/2}$ and failing to control size.▶

5.2 Monte Carlo

In this section, we use Monte Carlo simulations to demonstrate that the intuition outlined in Example 5.1 holds under full implementation in a model with heteroskedastic errors. In particular, we compare the power of the heteroskedasticity-robust versions of the Post-Double-Selection and Bivariate GLR procedures as the coefficients range from a sparse specification, where few coefficients are nonzero, to a uniform specification where all coefficients have the same, positive magnitude.

[16] Belloni et al. (2014) also assume sparsity in the population correlation between x_i and z_i. Our approach differs from that of Belloni et al. (2014) in that we do not allow for randomness in the regressors. Hence we impose a more restrictive definition of validity.

Data: In what follows, we set $N = 500$ and $K = 100$ to mimic conditions of the empirical study of Section 6. The observations are given by 2000 draws from the following data-generating process:[17]

$$\begin{aligned} z_{ik} &= \tilde{\varepsilon}_{ki} + \varepsilon_{ki}, \\ x_i &= \varepsilon_{xi} + c_{xz} N^{-1} \sum_{k=1}^{\lfloor \eta K \rfloor} \varepsilon_{ki}, \\ y_i &= x_i \beta + \sum_{k=1}^{K} z_{ik} \delta_{\eta k} + c_y e_i, \quad e | \{x_i, z_i\}_{i=1}^N \sim_{iid} \mathcal{N}(0, \Omega), \end{aligned}$$

where

1. $\eta \in (0, 1)$ is a parameter governing the sparsity of δ and of the correlation between x_i and z_i. Formally, $\delta_\eta = (\underbrace{1, \ldots, 1}_{\lfloor \eta K \rfloor}, 0, \ldots, 0)'$, and as such, small η corresponds to a sparse environment, while large η signifies the contrary.

2. c_{xz} and c_y are chosen so that, on average (and regardless of δ), *(i)* the in-sample R^2 of z_i on y_i is .043, and *(ii)* the in-sample R^2 of z_i on x_i is .5.

3. $\{\varepsilon_{ki}\}, \{\tilde{\varepsilon}_{ki}\}_{i,k}, \varepsilon_{xi}$ are *iid* $\mathcal{N}(0, 1)$ within and across observations $i = 1, \ldots, N$. Importantly, this implies that the covariance between x_i and z_{ik} is $N^{-1} c_{xz}$ for all $k = 1, \ldots, \lfloor \eta K \rfloor$, which, by design, does not fulfill the sparsity requirement of Belloni et al. (2014) for large η, allowing us to demonstrate the drawbacks of their approach.

4. To simulate heteroskedasticity, we let $\Omega = diag(\sigma_1^2, \ldots, \sigma_N^2)$, where $\sigma_1^2 = .5$, $\sigma_N^2 = 1.5$, and $\sigma_i^2 = \sigma_1^2 + (i-1) \cdot (\sigma_N^2 - \sigma_1^2)/(N-1)$ linearly-interpolates σ_1^2 and σ_N^2.

5. Under the null hypothesis, $\beta = 0$. Under the alternate hypothesis, we fix a single value $\beta = \beta_1 \in \mathbb{R}$ which facilitates power comparisons over all specifications of δ.

Bivariate Test: We implement the robust GLR Bivariate Test $\varphi_{GL}^{rob}(\hat{b}, \hat{g}_\rho; \hat{\Sigma}, r^2)$, which requires us to specify $r > 0$, the L^2 bound constraining the contribution of controls z_i. For purposes of demonstration, we choose $r = \|Z\delta\|$ to precisely meet the bound implied by δ, conditional on draws of z_i.[18] We remark that for r chosen higher, the bivariate test simply reduces to the long regression t-test, while for r chosen lower, the bivariate test trivially rejects with high probability under both the null and alternative.

[17] $\lfloor \cdot \rfloor$ denotes the floor, or greatest integer less than, function.
[18] Similar results hold when r is fixed to be the average value of $\|Z\delta\|$ over all draws.

Post-Double-Selection: We implement the Post-Double-Selection procedure of Belloni et al. (2014) to select control variables, following which we apply the standard t-test on the OLS estimate of β, accounting for potential heteroskedasticity throughout. This requires the choice of tuning parameters, for which we follow the prescription of Belloni et al. (2014) and set equal to $2.2\sqrt{N}\Phi^{-1}(1-\alpha/2)$,[19] where Φ denotes the distribution of a standard Gaussian.

Table 1: Monte Carlo Simulations for Level-5% Tests

	$\eta = 1$	$\eta = .8$	$\eta = .5$	$\eta = .25$	$\eta = .1$	$\eta = .09$	$\eta = .08$	$\eta = .05$	$\eta = .01$
PDS Size	.82	.80	.69	.67	.11	.085	.056	.051	.051
Bivariate Size	.04	.04	.04	.04	.04	.04	.05	.04	.05

In first row of Table 1, we display the size of the Post-Double-Selection procedure. It becomes evident that, as prefaced by Example 5.1, a LASSO-based selection technique which operates under a sparsity constraint exhibits poor size control when both the correlation between x_i and z_i and control coefficients δ become non-sparse. In the second row of Table 1, we display the size of the bivariate procedure, which is correct by design of the bivariate test.

6 Empirical Application

In this section, we apply the robust bivariate test to an empirical exercise from Belloni et al. (2014). In Section 6 of their chapter, Belloni et al. (2014) offer a comprehensive description of the exercise, which originates from a study conducted by Donohue III and Levitt (2001). We briefly review the study and Belloni et al. (2014)'s methodology before applying the robust bivariate statistic.

6.1 Preliminaries

Using a panel of 48 states from 1985 to 1997, Donohue III and Levitt (2001) estimate the causal effect of abortion rates on crime. As discussed in Belloni et al. (2014), an underlying mechanism for a negative causal link may be that access to abortion provides women with the flexibility to synchronize childbirth with periods of greater stability and wealth. Adopting notation from Belloni

[19] This is the same specification used in the empirical application of Belloni et al. (2014).

et al. (2014), let $c \in \{\text{violent, property, murder}\}$ denote one of three different measures of crime, let $p = 1, \ldots, 48$ index states, and let $t = 1, \ldots, 13$ index years. Donohue III and Levitt (2001) estimate the causal effect of abortion on crime in three linear specifications, one for each c. The model is

$$crime_{cpt} = abort_{cpt}\beta_c + cvar'_{cpt}\delta_c^{DL} + \eta_{cp}^{state} + \eta_{ct}^{year} + \varepsilon_{cpt} \tag{6.1}$$

where $crime_{cpt}$ is the contemporaneous crime rate, $abort_{cpt}$ is a weighted average of lagged abortion rates, $cvar_{cpt}$ is a vector of control variables, and $(\eta_{cp}^{state}, \eta_{ct}^{year})$ are state and year fixed-effects. In particular, $abort_{cpt}$ is the regressor of interest, and $cvar_{cpt}$, whose inclusion is designed to mitigate confounding effects, comprises 8 variables (with corresponding coefficients in δ_c^{DL}) in Donohue III and Levitt (2001)'s baseline specification. Ultimately the null hypothesis $H_0 : \beta_c = 0$ is rejected at the .001 level for all c, and Donohue III and Levitt (2001) justify their choice of controls by observing that $\hat{\beta}_c$ remains statistically-significant when certain variables are added to $cvar_{cpt}$.

The primary contention of Belloni et al. (2014) is that Donohue III and Levitt (2001), without motivating their selection of controls with a uniformly-valid procedure, admit possible size distortions. To rectify this, Belloni et al. (2014) expand the set of controls with new variables along with interactions and powers of existing variables in $cvar_{cpt}$. In total, they construct a set of $K = 130$ controls in addition to the 8 given by Donohue III and Levitt (2001).

We henceforth suppress the index c when possible, for clarity of exposition. To remove state fixed effects η_{cp}^{state}, Belloni et al. (2014) take first differences, formally transforming their model into

$$\tilde{y}_{pt} = \tilde{x}_{pt}\beta + \tilde{z}'_{pt}\delta + (\tilde{v}_{pt}^{DL})'\delta^{DL} + (\tilde{v}_{pt}^{FE})'\eta^{FE} + \tilde{e}_{pt}, \quad p = 1, \ldots, 48, \quad t = 1, \ldots 12 \tag{6.2}$$

where $\tilde{y}_{pt} = crime_{p,t+1} - crime_{pt}$ is the dependent variable, $\tilde{x}_{pt} = abort_{p,t+1} - abort_{pt}$ is the regressor of interest, $\tilde{z}_{pt} = \tilde{z}_{cpt}$ is the expanded vector of $K = 130$ controls by Belloni et al. (2014),[20] $\tilde{v}_{pt}^{DL} = cvar_{pt} - cvar_{p,t-1}$ is the differenced vector of 8 controls included by Donohue III and Levitt (2001), and $\tilde{v}_{pt}^{FE}$ subsumes a set of 12 time dummies with corresponding fixed effects η^{FE}. As expected, the presence of 138 total control variables, in addition to time fixed effects, renders the long regression which includes all controls nearly powerless. Subsequently, Belloni et al. (2014) apply their Post-Double-Selection technique to conduct inference on β,[21] from which they conclude that there is not enough evidence to reject the null hypothesis that $\beta = 0$ for any specification

[20] The identity of variables in $\tilde{z}_{cpt}$ depends on c.

[21] They account for time fixed effects with a preliminary regression on $\tilde{v}_{pt}$, and they account for variables in $\tilde{w}_{cpt}$ by including them in the regression as an "amelioration set," following Post-Double-Selection. For details, we refer readers to Belloni et al. (2014).

$c \in$ {violent, property, murder}. Outputs are summarized in Section 6, Table 2 of Belloni et al. (2014).

6.2 Applying the Robust Bivariate Test

As an alternative to the approach in Belloni et al. (2014), we use the robust bivariate statistic φ^{rob} of Section 4 to test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. Formally, we apply the procedure of Section 4.2 to the model (6.2), where we assume strict exogeneity: $E[\tilde{e}|\{\tilde{x}_{pt}, \tilde{z}_{pt}, \tilde{v}_{pt}^{DL}, \tilde{v}_{pt}^{FE}\}_{p,t}] = 0$ and allow for arbitrary correlation of the error term $\tilde{e}_{pt}$ within individual states $p = 1, \ldots, P$.

In this framework, we assume $\tilde{z}_{pt}$ — the 130 additional controls generated by Belloni et al. (2014) — are secondary, and we designate $\tilde{v}_{pt} = (\tilde{v}_{pt}^{DL}, \tilde{v}_{pt}^{FE})$ to be the primary controls, because we do not wish to restrain fixed effects or coefficients on variables selected by Donohue III and Levitt (2001). As in Section 4, we take a preliminary regression of observables $\{\tilde{y}_{pt}, \tilde{x}_{pt}, \tilde{z}_{pt}\}$ on $\tilde{v}_{pt}$, and we let $\{y_{pt}, x_{pt}, z_{pt}\}$ denote the residuals from such a regression.

For each value of $r > 0$, consider the bound $\sum_{p,t}(z_{pt}'\delta)^2 \leq r^2$ on a weighted sum of squares of $\delta = (\delta_1, \ldots, \delta_{130})'$. To lend interpretation to this bound, notice that $\hat{R}^2 = \frac{\sum_{p,t}(z_{pt}'\delta)^2}{\sum_{p,t} y_{pt}^2}$ is the in-sample partial R^2 of $\tilde{z}_i$ on $\tilde{y}_i$ controlling for $\tilde{v}_i$. Hence, setting $c_R^2 = r^2(\sum_{p,t} y_{pt}^2)^{-1}$ (which has a 1-to-1 correspondence to r), the sum of squares bound becomes equivalent to the constraint $\hat{R}^2 \leq c_R^2$,[22] and for each value c_R^2, we can apply the robust bivariate test φ^{rob} of Section 4 to determine whether the null hypothesis can be rejected under such a constraint.[23]

Table 2: Inference via the Bivariate Test

	Violent Crime	Property Crime	Murder
Long t-statistic	-1.720	-.958	.513
Short t-statistic	-2.11	-2.637	-.744
$\bar{c}_R^2$	.001	.005	N/A

[22] It is possible that r is large enough such that $c_R^2 > 1$, but it is easy to demonstrate that the value of c_R^2 at which φ^{rob} begins to reject lies within the range $[0, 1]$, provided the long regression which includes all covariates does not reject the null hypothesis.

[23] We note that in the *property* and *murder* specifications, a small-sample adjustment is made by excluding observations from single state, Delaware, whose inclusion compromises the estimation of standard errors, arising from insufficient variation in its corresponding observations.

In the first two rows of Table 2, we display — for each of three specifications $\{violent, property, murder\}$ — the t-statistics from *(i)* $\hat{\beta}$, the OLS estimate of β from the "long" regression of $\tilde{y}_{pt}$ on all controls $(\tilde{x}_{pt}, \tilde{z}_{pt}, \tilde{v}_{pt})$ and *(ii)* $\hat{\beta}_{short}$, the OLS estimate of β from the "short" regression of $\tilde{y}_{cpt}$ on $(\tilde{x}_{cpt}, \tilde{v}_{pt})$, excluding $\tilde{z}_{pt}$. In particular, we see that at the 5%-level, the long regression fails to reject in any specification, while the short regression rejects in all but the *murder* specification.[24] Hence, although $\hat{\beta}$ and $\hat{\beta}_{short}$ are both signals of β, the large standard error of $\hat{\beta}$ renders it impracticable, while one cannot rule out the possibility that the ostensibly significant t-statistic from $\hat{\beta}_{short}$ is driven by omitted variable bias.

The role of the bivariate test φ^{rob} is to bridge this gap by matching assumptions on c_R^2 with testing decisions. In the third row of Table 2, we display $\bar{c}_R^2$, the maximum value of c_R^2 such that φ^{rob} rejects at the 5%-level, for each of the three specifications $c \in \{\text{violent, property, murder}\}$.[25] The interpretation for $\bar{c}_R^2$ is as follows: whenever one assumes $c_R^2 < \bar{c}_R^2$ — so that the partial R^2 of potential controls identified by Belloni et al. (2014) is, in turn, bounded by $\bar{c}_R^2$ — one implicitly presumes that the magnitude of the omitted variable bias is too small to account for the magnitude of $\hat{\beta}_{short}$, so φ^{rob} rejects. In contrast, when $c_R^2 > \bar{c}_R^2$, one allows for the possibility that omitted variable bias is responsible for the statistical significance of $\hat{\beta}_{short}$, so φ^{rob} cannot reject. In the violent and property crime specifications, we observe that $\bar{c}_R^2$ is small, suggesting that one cannot adopt conclusions from Donohue III and Levitt (2001) without strong restrictions on the partial R^2 of the 130 potential controls identified by Belloni et al. (2014).

7 Conclusion

We have shown that in a Gaussian, homoskedastic setting, the bivariate statistic becomes asymptotically-sufficient when the component of δ irrelevant to omitted variable bias has magnitude no greater than $N^{-1/4-\epsilon}$. Consequently, when the magnitude of δ is, itself, constrained by the same rate, the bivariate test becomes optimal. We subsequently extended the bivariate test to accomodate heterskedasticity and clustering by identifying conditions under which the bivariate statistic is asymptotically-normal and constructing an estimator of its asymptotic covariance matrix.

We recognize that there are several substantial ways to build upon these results, which com-

[24] We recognize that in both Donohue III and Levitt (2001) and Belloni et al. (2014), the short regression t-statistics are greater than their counterparts in Table 2. This difference arises because we use standard errors that are different from White standard errors. We hope to bridge this discrepancy in future research.

[25] In the murder specification, there is no value of c_R^2 which leads to rejection, because the short regression does not reject.

prise only a first attempt at inference under an L^2 bound. Firstly, we note that although the bivariate statistic is valid in any sample, it discards potentially-useful information by ignoring the statistic representing OLS coefficients orthogonal to omitted variable bias, which provides a signal that can be used to nontrivially tighten any L^2 bound which does not shrink at the rate $N^{-1/4-\epsilon}$. While it may be straightforward to find an appropriate use of such information in a Gaussian, homoskedastic model, it remains to be seen how this information can be incorporated into a valid testing procedure under general error structures.

Secondly, we note that although applying the high-dimensional clustered standard errors of Li (2016) allows for consistent estimation of the asymptotic covariance of the bivariate statistic under fairly weak regularity conditions, there might exist consistent estimators with better finite-sample properties. Specifically, the standard errors proposed by both Li (2016) and Cattaneo et al. (2015) can lead to matrix estimates which are not positive-definite, and shrinkage techniques might be an appropriate remedy. Additionally, it might be possible to use the L^2 bound on δ to sharpen standard errors.

Thirdly, in this chapter, the link between the L^2 bound and R^2 in a setting with general error structures is based on the in-sample partial R^2, which is not a perfect signal of the population R^2. Hence, it remains to evaluate whether the L^2 bound can be related to a population R^2 restriction without affecting the uniform validity of the robust bivariate test.

Finally, we note that it may be possible to outperform the bivariate test by loosening the requirement of validity to incorporate randomness in the regressors. Specifically, this can take advantage of the fact that the population R^2 of x_i on z_i can be 0, even when the in-sample R^2 of x_i on z_i does not vanish under most draws.

Part II

Power Bounds in Regressions under Sparsity

8 Introduction

The problem of inference on a single coefficient in multiple linear regression has been widely studied. We refer to data generated by the following equation:

$$y_i = x_i\beta + z_{i1}\delta_1 + \cdots + z_{iK}\delta_K + e_i, \tag{8.1}$$

where y_i is the outcome, x_i is the regressor of interest, $, z_{i1}, \ldots, z_{iK}$ are control variables, and e_i is an orthogonal error. Various methods have been developed to deliver tractable and powerful inference on β in a many-controls setting where K is large. However, many of these methods have been demonstrated to lead to invalidity, or overrejection of the null hypothesis, when the control coefficients δ are left unconstrained (Leeb and Pötscher (2005, 2008a)).

More generally, it has been shown that without a constraint on δ, no valid procedure is more powerful than the standard t-test based on the full regression in (8.1) in a Gaussian, homoskedastic model. Hence, constructing a valid procedure to outperform the t-test calls for an assumption on δ, and an ostensibly natural candidate for this assumption is sparsity: a bound on the number of nonzero δ_k. In Belloni et al. (2014), for instance, asymptotic validity of the main technique — LASSO-based post-double selection (henceforth, LPDS) — relies on a "double-sparsity" condition.

The purpose of this chapter is to quantify the extent to which assuming sparsity can improve inference (conditional on the regressors) in a finite-sample setting. This exercise is complicated by the impracticability of establishing validity of specific procedures under sparsity, which grants (in a non-convex fashion) overwhelmingly many possibilities for the data-generating process (8.1).[26] In this chapter, rather than examining the power and size of individual procedures, we derive power bounds over valid tests by using the least-favorable distribution technique from Elliott et al. (2015). Throughout the derivation of these bounds, we assume a special case of (8.1) where *(i)* e_i is mean-zero, Gaussian, and homoskedastic, *(ii)* regressors x, z are fixed, and *(iii)* $\delta_1 = \cdots = \delta_K = 0$ under the alternative.

[26] For instance, Belloni et al. (2014) prove validity of LPDS under theoretical limiting behavior but remain silent about finite-sample designs and assumptions.

In Section 9, we provide background on the least-favorable distribution method. In Section 10, we derive an analytical result which allows us to compute the best possible power bound, given a certain class of symmetric designs in which the controls share the same pairwise correlation and are equally-correlated with the regressor of interest. Because sparsity produces higher gains when the regressor of interest is correlated with more controls, we consider this a best-case-scenario for a sparsity assumption, conditional on the R^2 of x on z. We find that, in general, the gains from sparsity depend heavily on the correlation between the controls, but when the population R^2 of x on z is .9 or less, such gains are universally minimal. In Section 11, we construct an algorithm to derive bounds for arbitrary designs, and we subsequently apply this algorithm to inference conditional on non-symmetric regressors. In Section 11.1, we generate examples of regressors according to the Monte Carlo exercise of Belloni et al. (2014) where the regressor of interest has approximately zero population correlation with the controls under repeated draws. In Section 11.2, we generate examples varying the number of controls with which the regressor of interest is correlated in population. In both cases, we find that the gains from sparsity are minimal for values of the population R^2 of x on z up to .9.

9 Setup and Background

9.1 Preliminaries

In the remainder of this chapter, we assume data is generated by the following:

$$y_i = x_i\beta + z_{i1}\delta_1 + \cdots + z_{iK}\delta_K + e_i, \quad e_i \sim_{iid} \mathcal{N}(0,1) \tag{9.1}$$

where $i = 1, \ldots, N$ denotes index, the number of observations N is finite, the scalar regressors $x_i, z_{i1}, \ldots, z_{iK}$ are assumed fixed, and the error disturbance e_i is Gaussian and homoskedastic. In obvious notation, we write $z_i = (z_{i1}, \ldots, z_{iK})'$ and $\delta = (\delta_1, \ldots, \delta_K)'$ for vectors and

$$\underset{(N\times 1)}{Y} = \underset{(N\times 1)}{X}\beta + \underset{(N\times K)}{Z}\delta + \underset{(N\times 1)}{e} \tag{9.2}$$

in matrix form.

Assumption 9.1.

1. *$W_{xz} = (X, Z)$ is a deterministic matrix.*

2. *$e \sim \mathcal{N}(0, I_N)$.*

We are interested in frequentist inference on β alone, so that $\delta_1, \ldots, \delta_K$ become nuisance parameters. Formally, we consider tests of

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta \neq 0, \tag{9.3}$$

assuming throughout this chapter that $\alpha \in (0, 1)$ is a fixed level. It has been demonstrated that without a restriction on δ, no valid test is more powerful than the standard t-test following a regression of y_i on x_i, z_i in (9.1) against one-sided altenatives $\beta > 0$, and no valid, unbiased test is more powerful against two-sided alternatives. This result implies that any nontrivial inferential procedure overrejects under certain values of δ, and it is the goal of this chapter to explicitly determine how much power can be gained by sacrificing size control over δ and assuming the sparsity restriction

$$\sum_{k=1}^{K} 1\{|\delta| > 0\} = \|\delta\|_0 \leq M \tag{9.4}$$

with $M \in \mathbb{N}$ representing the maximum number of controls with nonzero coefficients. We focus on the properties of tests which are valid under $\beta = 0$, for any value of δ satisfying (9.4), allowing overrejection only under δ for which (9.4) does not hold.

Sparsity is popular in statistical literature, where it serves to ensure that LASSO estimators — coefficients from regressions with L^1 penalties — have desirable asymptotic properties. Typically, the degree M is chosen to be asymptotically-negligible relative to the sample size N. When consistency of the LASSO estimator is the objective, it is generally enough to assume $M = o(\frac{N}{\log N})$. In Belloni et al. (2014), where uniformly-valid inference on β is the objective, $M = o(N^{1/2})$. Because such assumptions are centered around limiting behavior, it can be difficult to assess, for any specific procedure, which M (if any) ensures validity in finite samples.

In this chapter, we determine in a finite-sample setting whether Assumption 9.4 creates room for improvement over the standard t-test. In light of the aforementioned difficulties, we do not examine individual procedures, but instead we derive bounds on the power of any test of (9.3) valid under (9.4) using least-favorable distribution analysis, and we compare these bounds to the power of the standard t-test. Intuition suggests that the power gained from sparsity should be highest when confounding bias is large, making it difficult for the practitioner to distinguish between the effects

of the regressor of interest and controls. This happens when *(i)* the R^2 of x on z is high, *(ii)* x is correlated with many controls z_k, and *(iii)* the controls themselves are uncorrelated amongst each other. Within Sections 10-11, we formally confirm this intuition.

In Section 10, we consider design matrices W_{xz} which satisfy a symmetry condition allowing us to derive an analytical least-favorable distribution. Using this theoretical result, we compute the infimum bound on the power of all tests valid under (9.4) against the one-sided point alternative $\beta = \beta_1 > 0$, for various M. Subsequently, we translate these power bounds into a Pitman efficiency gain, which roughly captures the required increase in the number of observations in order for the standard t-test to have power equal to the computed infimum bound.

In Section 11, we generate random draws of the design matrix W_{xz}. Holding W_{xz} fixed for each draw, we apply an algorithm to determine a power bound on valid tests for (9.4) against point alternatives $\beta = \beta_1 > 0$ for various M. In general, the generated W_{xz} do not satisfy the symmetry conditions under which we derived analytical least-favorable distributions in Section 10, so the algorithm does not produce infimum bounds. Nevertheless, the derived bounds are informative, and we tabulate the implied efficiency gains.

9.2 Least-Favorable Distributions

The least-favorable distribution method, presented in Elliott et al. (2015), is a means of deriving power bounds in hypothesis testing problems involving nuisance parameters. For a detailed synopsis, we refer readers to Elliott et al. (2015) and Reinhardt (1961). In this section, we discuss the application of the least-favorable distribution to hypothesis testing in linear regressions under sparsity.

Definition 9.2. Power Bounds and least-favorable Distributions under Sparsity

1. Denote the null nuisance parameter set $\Delta_M = \{\delta : \|\delta\|_0 \leq M\}$. Observe that testing (9.3) under (9.4) can be recast as:

$$H_0 : \beta = 0,\, \delta \in \Delta_M \quad vs. \quad H_1 : \beta \neq 0,\, \delta \in \Delta_M \tag{9.5}$$

2. A weighted alternative is a distribution $\Lambda_1(\beta, \delta)$ over $(\beta, \delta) \in \mathbb{R} \times \Delta_M$.

3. Let $\phi(Y, W_{xz}; \beta, \delta)$ denote the likelihood of (9.2) under β, δ. For a given weighted alternative

Λ_1, a least-favorable distribution is a distribution $\Lambda_{LFD}(\beta, \delta)$ over $(\beta, \delta) \in \{0\} \times \Delta_M$ such that the likelihood ratio test

$$\varphi_{LFD}(Y, W_{xz}) = 1\left\{ \frac{\int \phi(Y, W_{xz}; \beta, \delta)\, d\Lambda_1(\beta, \delta)}{\int \phi(Y, W_{xz}; \beta, \delta)\, d\Lambda_{LFD}(\beta, \delta)} > cv_\alpha \right\}$$

(i) rejects with probability equal to α over the support of Λ_{LFD} and *(ii)* rejects with probability no more than α under any $(\beta, \delta) \in \{0\} \times \Delta_M$.

4. For a given weighted alternative Λ_1 and least-favorable distribution Λ_{LFD}, let $\pi_{LFD}(\Lambda_1, \Lambda_{LFD}) = \int \varphi_{LFD}(Y, W_{xz}) \cdot \left(\int \phi(Y, W_{xz}; \beta, \delta)\, d\Lambda_1(\beta, \delta) \right)\, dY$ denote the weighted average power of φ_{LFD}. By Lemma 1 of Elliott et al. (2015), π_{LFD} is a bound on the power of any valid test over $(\beta, \delta) \in \{0\} \times \Delta_M$. Because φ_{LFD} is valid over $(\beta, \delta) \in \{0\} \times \Delta_M$, π_{LFD} is the infimum over all such bounds.

5. In some cases, the derivation or computation of a least-favorable distribution is not possible. Suppose, then, that $\Lambda_0(\beta, \delta)$ is an arbitrary distribution over $(\beta, \delta) \in \{0\} \times \Delta_M$, not necessarily least-favorable. Then the power of the likelihood ratio test

$$\varphi_0(Y, W_{xz}) = 1\left\{ \frac{\int \phi(Y, W_{xz}; \beta, \delta)\, d\Lambda_1(\beta, \delta)}{\int \phi(Y, W_{xz}; \beta, \delta)\, d\Lambda_{LFD}(\beta, \delta)} > cv_\alpha \right\}$$

is an upper bound on the power of any valid test over $(\beta, \delta) \in \{0\} \times \Delta_M$. That is, $\pi_0(\Lambda_1, \Lambda_0) := \int \varphi_0(Y, W_{xz}) \cdot \left(\int \phi(Y, W_{xz}; \beta, \delta)\, d\Lambda_0(\beta, \delta) \right)\, d(Y, W_{xz}) \geq \pi_{LFD}$. Qualitatively, if Λ_{LFD} is well-approximated by Λ_0, then π_{LFD} should be well-approximated by π_0.

Hence, any null distribution Λ_0 generates a power bound on valid tests, and, in principle, finding a least-favorable distribution Λ_{LFD} yields the infimum power bound π_{LFD}. In Section 10, we analytically derive Λ_{LFD} and π_{LFD} in the case where W_{xz} exhibits a certain symmetry.

In general, however, it may be difficult to analytically obtain least-favorable distributions. The Approximately least-favorable Distribution (ALFD) algorithm of Elliott et al. (2015) is a means to computationally solve for φ_{LFD} and π_{LFD}, but it requires a grid approximation of Δ_M. Unfortunately, it is typically infeasible to apply the ALFD algorithm directly to the problem (9.5), because implementing a sufficiently-fine grid for Δ_M becomes too computationally-intensive. Hence, in Section 11 where designs W_{xz} are randomly drawn, we employ candidate distributions $\Lambda_0(\beta, \delta)$ which are not least-favorable but nonetheless generate power bounds over valid tests, as outlined in Definition 9.2, part 5.

10 Power Bounds for Symmetric W_{xz}

As discussed in Section 9.2, the least-favorable distribution can take on a complicated form for a general weighted alternative Λ_1 and design matrix W_{xz}. In this section, we consider the alternative which places all weight on fixed $\beta = \beta_1 > 0$, $\delta = 0$, and we constrain our attention to a class of symmetric designs indexed by two scalar parameters: *(i)* ρ_{zz}, the pairwise correlation between controls (held constant over different pairs) and *(ii)* R^2_{xz}, the sample R^2 of x on z. Under these conditions, we prove that the least-favorable distribution is a uniform mixture over points in the null parameter space $\{0\} \times \Delta_M$, identified up to scale. Computationally, we approximate the scale, yielding infimum power bounds for arbitrary pairs (ρ_{zz}, R^2_{xz}).

Begin by defining the desired symmetry:

Assumption 10.1. *For each pair $(\rho_{zz}, R^2_{xz}) \in [0,1]^2$, let $Z^{sym}(\rho_{zz})$ be an $N \times K$ matrix and $X^{sym}(\rho_{zz}, R^2_{xz})$ be an $N \times 1$ vector such that:*

1. *The pairwise sample correlation between any two columns in $Z^{sym}(\rho_{zz})$ is ρ_{zz}.*
2. *The projection of $X^{sym}(\rho_{zz}, R^2_{xz})$ on $Z^{sym}(\rho_{zz})$ is a scalar multiple of the average of the columns of $Z^{sym}(\rho_{zz})$.*
3. *The sample R^2 of X^{sym} on Z^{sym} is R^2_{xz}.*

$X^{sym}(\rho_{zz}, R^2_{xz})$ represents the regressor of interest and $Z^{sym}(\rho_{zz})$ represents the controls, in columns. We omit the notation (ρ_{zz}, R^2_{xz}) and write X^{sym}, Z^{sym}, and $W^{sym}_{xz} = (X^{sym}, Z^{sym})$ when context allows.

Recall the test (9.5), where $M \leq K$ represents the true number of potentially non-zero coefficients, and consider the weighted alternative Λ_1 which places all mass on the point $\beta = \beta_1$ and $\delta = 0$, with $\beta_1 > 0$ fixed. The test can be re-written:

$$H_0 : \beta = 0, \delta \in \Delta_M \quad vs. \quad H_1 : \beta \neq \beta_1 > 0, \delta = 0. \tag{10.1}$$

Under the conditions stated in Assumption 10.1, there is a simple form for the least-favorable distribution in (10.1), presented in Theorem 10.2. Intuitively, this relies on the rotational symmetry of the design matrix , which suggests a least-favorable distribution which is symmetric in each "arm" — span of M columns from Z^{sym} — of the nuisance parameter space Δ_M.

Theorem 10.2. *Fix $(\rho_{zz}, R^2_{xz}) \in [0,1]^2$, and suppose the design matrix $W^{sym} = (X^{sym}, Z^{sym})$ satisfies Assumption 10.1, where without loss of generality the columns of Z^{sym} have length 1. For each subset $S \subseteq \{1, \ldots, K\}$ of size M, let Z^{sym}_S denote the matrix consisting of columns of Z^{sym} indexed by the elements in S, and let $\delta_S = ((Z^{sym}_S)' Z^{sym}_S)^{-1} (Z^{sym}_S)' X^{sym}$. Then there exists a constant $\gamma \geq 0$ for which the least-favorable distribution in (10.1) places uniform weight on the $\frac{K!}{M!(K-M)!}$ points:*

$$\{(\beta, \delta) = (0, d) \,:\, d = \gamma \delta_S,\, S \subseteq \{1, \ldots, K\},\, |S| = M\}$$

We outline the main ideas behind the proof with a simple example below, while the proof of the general case is a straightforward, but tedious extension located in the appendix. Essentially, the proof relies on two facts: *(i)* a symmetric test valid over an "arm" of Δ_M is, in turn, valid over all $\delta \in \Delta_M$, and *(ii)* rejection regions for likelihood ratio tests of Gaussian distributions are convex, so that a test which maximizes size within a small neighborhood of a point δ also maximizes size over the arm of Δ_M which contains δ. Section 10.1 demonstrates this in a setting where $K = 2$ and $M = 1$.

10.1 An Analytical Least-Favorable Distribution for $N = 2$, $K = 2$, and $M = 1$

Let $N = 2$, $K = 2$, and $M = 1$, and suppose $X = (1, 1)'$ and $Z = (z_1, z_2)$ where $z_1 = (1, 0)'$, $z_2 = (0, 1)'$. In terms of notation from Assumption 10.1, $Z = Z^{sym}(0)$ and $X = X^{sym}(0, 1)$. Letting $\mu_1 := (1, 1)'$, the testing problem (10.1) can be re-framed:

$$H_0 : \mu \in \{\text{span}(z_1) \cup \text{span}(z_2)\} \quad vs. \quad H_1 : \mu = \mu_1$$

where we observe $Y \sim \mathcal{N}(\mu, I_2)$. Geometrically, the null parameter space consists of the union of the horizontal and vertical axes in $\mathbb{R}^2$.

Suppose we guess an analytical least-favorable distribution, indexed by a scalar $t \geq 0$, which places mass $1/2$ on each of the points $\mu = tz_1$ and $\mu = tz_2$. Let the corresponding likelihood-ratio test under this guess be denoted $\varphi(Y, t)$. Then

$$\varphi(Y, t) = 1\left\{g\left(\|Y - tz_1\|^2 - \|Y - \mu_1\|^2\right) + g\left(\|Y - tz_2\|^2 - \|Y - \mu_1\|^2\right) < cv_\alpha(t)\right\}$$

where $g(r) = \exp(-r/2)$, and $cv_\alpha(t)$ endogenously takes on a value such that $\varphi(Y,t)$ rejects with probability α when $\mu \in \{tz_1, tz_2\}$. We conjecture that there exists $t^* \in (0,2)$ for which $\varphi(Y,t^*)$ is level-α valid over all $Y \sim \mathcal{N}(\mu, I_2)$ with $\mu \in \{\text{span}(z_1) \cup \text{span}(z_2)\}$. It suffices, by symmetry, to check that there exists $t^* \in (0,2)$ for which $\varphi(Y,t^*)$ is level-α valid under $Y \sim \mathcal{N}(tz_1, I_2)$, for all $t \in \mathbb{R}$.

For notational convenience, let $E_\mu[\cdot]$ denote the expectation taken under $Y \sim \mathcal{N}(\mu, I_2)$. The rejection probability of $\varphi(Y,t)$ under $Y \sim \mathcal{N}(\mu, I_2)$ can be written $E_\mu[\varphi(Y,t)] = \int \varphi(y,t) \cdot g(\|y-\mu\|^2)dy$. It remains to check that for some $t^* \in (0,2)$, $E_{tz_1}[\varphi(Y,t^*)]$ is maximized at $t = t^*$.

10.1.1 First-Order Necessary Conditions

A necessary condition for $E_{tz_1}[\varphi(Y,t^*)]$ to be maximized at $t = t^* \in (0,2)$ is

$$0 = \frac{\partial}{\partial t}\left(E_{tz_1}[\varphi(Y,t^*)]\right)\Big|_{t=t^*} = E_{t^*z_1}\left[\varphi(Y,t^*) \cdot ((Y - t^*z_1)'z_1)\right].$$

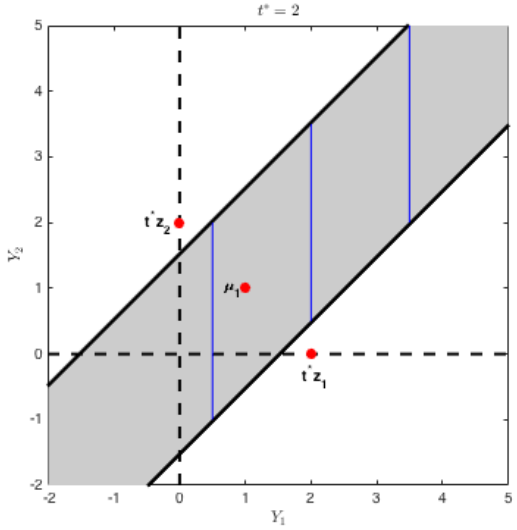
We claim that there exists at least one value of $t^* \in (0,2)$ for which $E_{t^*z_1}[\varphi(Y,t^*) \cdot ((Y-t^*z_1)'z_1)] = 0$. This follows from the fact that $E_{t^*z_1}[\varphi(Y,t^*) \cdot ((Y-t^*z_1)'z_1)]$ is continuous in t^* (a standard exercise), $E_{t^*z_1}[\varphi(Y,t^*) \cdot ((Y-t^*z_1)'z_1)] < 0$ when $t^* = 2$, and $E_{t^*z_1}[\varphi(Y,t^*) \cdot ((Y-t^*z_1)'z_1)] > 0$ when $t^* = 0$. The latter statements are shown below.

1. First consider $t^* = 2$. Then $t^*z_1 = (2,0)$, $\mu_1 = (1,1)$, and $t^*z_2 = (0,2)$ all lie on the same line, so $\varphi(\cdot, t^*)$ simply rejects for small values of $|(Y-\mu_1)'(z_2 - z_1)|$. Hence, for some scalar $c \geq 0$, the rejection region can be re-written: $\{Y : \varphi(Y,t^*) = 1\} = \{(Y_1, Y_2) : Y_2 \leq Y_1 + c \cap Y_2 \geq Y_1 - c\}$, where c has a 1-to-1 relationship with the critical value $cv_\alpha(t^*)$. It follows immediately that

$$\begin{aligned}
&E_{t^*z_1}[\varphi(Y,t^*) \cdot ((Y-t^*z_1)'z_1)] \\
&\propto \int\int \varphi(y,t^*) \cdot ((y-t^*z_1)'z_1) \cdot \exp(-\frac{1}{2}(y_1-t^*)^2) \cdot \exp(-\frac{1}{2}y_2^2)\, dy_1 dy_2 \\
&\propto \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(y_1-t^*)^2) \cdot (y_1 - t^*) \cdot \left(\int_{y_1-c}^{y_1+c} \exp(-\frac{1}{2}y_2^2)\, dy_2\right) dy_1 \\
&= \int_{-\infty}^{\infty} \exp(-\frac{1}{2}(y_1-t^*)^2) \cdot (y_1 - t^*) \cdot (\Phi(y_1+c) - \Phi(y_1-c))\, dy_1
\end{aligned}$$

$$= \int_0^{\infty} \exp(-\frac{1}{2}\tilde{y}_1^2) \cdot \tilde{y}_1 \cdot \underbrace{(\Phi(t^*+\tilde{y}_1+c) - \Phi(t^*+\tilde{y}_1-c) - \Phi(t^*-\tilde{y}_1+c) + \Phi(t^*-\tilde{y}_1-c))}_{\text{negative for all } \tilde{y}_1 > 0} d\tilde{y}_1 < 0$$

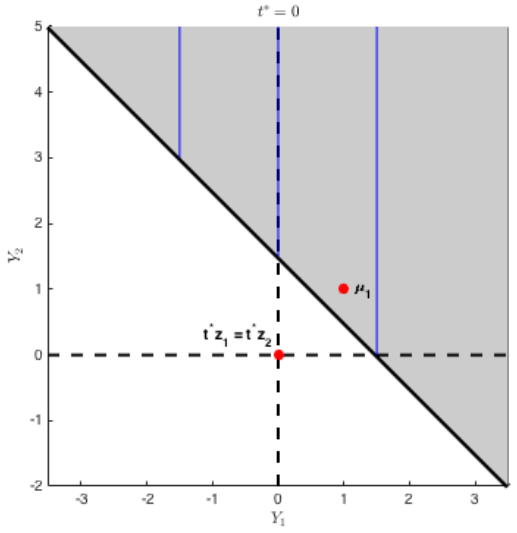
Figure 10.1: Rejection Regions for $\varphi(Y, t^*)$, $t^* = 2$



Grey denotes the rejection region $\{Y : \varphi(Y, t^*) = 1\}$, and the blue lines are isoquants for $\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)$. For each positive quantity of $\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)$ to the right of t^*z_1, there is a corresponding negative quantity of $\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)$ to the left, while clearly the density of $X \sim \mathcal{N}(t^*z_1, I_2)$ places more mass on the segment to the left, which is located closer to t^*z_1. Hence, $E_{t^*z_1}[\varphi(X, t^*) \cdot ((X - t^*z_1)'z_1)] < 0$.

2. Now consider $t^* = 0$. Then $t^*z_1 = t^*z_2 = (0, 0)$, and $\varphi(\cdot, t^*)$ rejects for large values of $Y'\mu_1$. It follows from a similar analysis as in part 1 that $E_{t^*z_1}[\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)] > 0$.

Figure 10.2: Rejection Regions for $\varphi(Y, t^*)$, $t^* = 0$



Grey denotes the rejection region $\{Y : \varphi(Y, t^*) = 1\}$, and the blue lines are isoquants for $\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)$. For each positive quantity of $\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)$ to the right of t^*z_1, there is a corresponding negative quantity of $\varphi(Y, t^*) \cdot ((Y - t^*z_1)'z_1)$ to the left, while clearly the density of $X \sim \mathcal{N}(t^*z_1, I_2)$ places more mass on the segment to the right, which is located closer to t^*z_1. Hence, $E_{t^*z_1}[\varphi(X, t^*) \cdot ((X - t^*z_1)'z_1)] > 0$.

10.1.2 Second-Order Sufficient Conditions

In Section 10.1.1, we showed that there exists $t^* \in (0,2)$ for which $0 = \frac{\partial}{\partial t}\left(E_{tz_1}[\varphi(Y,t^*)]\right)\Big|_{t=t^*}$. If, in addition, $\frac{\partial^2}{\partial t^2}\left(E_{tz_1}[\varphi(Y,t^*)]\right)\Big|_{t=t^*} < 0$, then $E_{tz_1}\left[\varphi(Y,t^*)\right]$ is maximized within a neighborhood of $t = t^*$.

Begin by observing that the rejection region $\{Y : \varphi(Y,t^*) = 1\}$ is convex. In turn, a mathematical result (Lemma B.2 in the appendix) implies

$$E_{t^*z_1}\left[(Y_1 - E[Y_1 \,|\, \varphi(Y,t^*) = 1])^2 \,|\, \varphi(Y,t^*) = 1\right] < E_{t^*z_1}\left[(Y_1 - E[Y_1])^2\right] = 1$$

But from part 1, $E_{t^*z_1}\left[\varphi(Y,t^*) \cdot ((Y - t^*z_1)'z_1)\right] = 0$, which implies $E[Y_1 \,|\, \varphi(Y,t^*) = 1] = t^*z_1$. Combining,

$$E_{t^*z_1}\left[(Y_1 - E[Y_1 \,|\, \varphi(Y,t^*) = 1])^2 \,|\, \varphi(Y,t^*) = 1\right] = \frac{E_{t^*z_1}\left[\varphi(Y,t^*) \cdot (Y_1 - t^*)^2\right]}{E_{t^*z_1}\left[\varphi(Y,t^*)\right]} < 1$$

$$\Rightarrow \frac{\partial^2}{\partial t^2}\left(E_{tz_1}[\varphi(Y,t^*)]\right)\Big|_{t=t^*} = E_{t^*z_1}\left[\varphi(Y,t^*) \cdot \left((Y_1 - t^*)^2 - 1\right)\right] < 0.$$

Hence, $E_{tz_1}\left[\varphi(Y,t^*)\right]$ is locally-maximized at $t = t^*$. Because the rejection region $\{Y : \varphi(Y,t^*) = 1\}$ is convex, a mathematical result (Lemma B.3 in the appendix) implies that $E_{tz_1}\left[\varphi(Y,t^*)\right]$ is also globally-maximized at $t = t^*$.

10.2 Computing Power Bounds

In Theorem 10.2, we determined the form of an analytical least-favorable distribution for the test (10.1):

$$H_0 : \beta = 0,\, \delta \in \Delta_M \quad vs. \quad H_1 : \beta \neq \beta_1 > 0,\, \delta = 0$$

for a class of symmetric design matrices satisfying Assumption 10.1. In Section 10.2, we use this to compute infimum power bounds over all valid tests of (10.1) under various N, K, M and various specifications for the parameter (ρ_{zz}, R^2_{xz}) which governs these designs.

We begin by examining the likelihood ratio test implied by the least-favorable distribution.

Corollary 10.3. *Consider the testing problem (10.1), and let Λ_{LFD} be the least-favorable distribution from Theorem 10.2. The associated likelihood ratio test is given by:*

$$\varphi_{LFD}(Y, W_{xz}) = 1\left\{ \sum_{S\subset\{1,\dots,K\},\,|S|=M} \exp(-\frac{1}{2}(\|Y - Z^{sym}\delta_S\gamma\|^2 - \|Y - X^{sym}\beta_1\|^2)) < cv_\alpha^{-1} \right\}$$

for some scalar $\gamma \in [0, \frac{K}{M}]$. Furthermore, the power of φ_{LFD} under Λ_1,

$$\pi_{LFD} = \int \varphi_{LFD}(Y, W_{xz}^{sym}) \cdot \left(\int \phi(Y, W_{xz}^{sym}; \beta, \delta)\, d\Lambda_1(\beta, \delta) \right) dY,$$

is the infimum power bound over valid tests of (10.1).

In principle, computing π_{LFD} from Corollary 10.3 yields the desired bounds. However, this requires one to compute *(i)* the correct critical value cv_α, *(ii)* the constant γ which scales δ_S in Theorem 10.2, and *(iii)* a sum over all subsets $S \subset \{1, \dots, K\}$ of size M. Following task *(iii)*, tasks *(i)* and *(ii)* are straightforward, but *(iii)* is difficult to implement even for modest K, M. For this reason, we approximate φ_{LFD} with a tractable version $\tilde{\varphi}_{LFD}$, which uses a randomly drawn subpopulation of subsets $S \subset \{1, \dots, K\}$ in its argument. Subsequently, we derive approximations to π_{LFD} by evaluating the power of $\tilde{\varphi}_{LFD}$.

Algorithm 10.4. *Construction of $\tilde{\varphi}_{LFD}$ and $\tilde{\pi}_{LFD}$:*

1. *For fixed n, draw random subsets $\{S_1, \dots, S_n\}$ of size M from $\{1, \dots, K\}$.*
2. *Let $\tilde{\varphi}_{LFD}(Y, W_{xz}) = 1\left\{\sum_{j=1}^n \exp(-\frac{1}{2}(\|Y - Z^{sym}\delta_{S_j}\|^2 - \|Y - X^{sym}\beta_1\|^2)) < cv_\alpha^{-1}\right\}$.*[27]
3. *Let $\tilde{\pi}_{LFD} = \int \tilde{\varphi}_{LFD}(Y, W_{xz}^{sym}) \cdot \left(\int \phi(Y, W_{xz}^{sym}; \beta, \delta)\, d\Lambda_1(\beta, \delta) \right) dY$ be the approximation to the infimum power bound π_{LFD}.*

In Tables 3-5, we compute $\tilde{\pi}_{LFD}$ — approximate infimum power bounds — for the testing problem (10.1) with design matrix $W_{xz}^{sym}(\rho_{zz}, R_{xz}^2)$ under stated values of $N, \frac{K}{N}, \frac{M}{K}, \rho_{zz}$, and R_{xz}^2. We choose to express the number of regressors and sparsity index in terms of ratios K/N and M/K rather than absolute quantities K and M. Recall that ρ_{zz} represents the pairwise correlation between control variables, held constant over all pairs, and R_{xz}^2 is the sample R^2 of the regressor of interest on the controls. We translate $\tilde{\pi}_{LFD}$ into the Pitman efficiency gain, formally defined as follows:

[27] Unreported results show that the Monte Carlo error from such an approximation is suitably small, for $n \geq 10^5$.

Definition 10.5. Consider the testing problem (10.1) with fixed alternative $\beta_1 > 0$ and design matrix W_{xz}, and define $(\hat{\beta}, \hat{\delta}')' = (W_{xz}'W_{xz})^{-1}W_{xz}'Y$. For any value of power $\pi \in [\alpha, 1)$, there exists a unique $q > 0$ such that the standard test based on the statistic $\hat{\beta} \sim \mathcal{N}(\beta, (q \cdot W_{xz}'W_{xz})^{-1}_{1,1})$ has power π under H_1. Informally, this is the increase in the number of observations necessary for the standard test to have power π. Define the efficiency gain function to be $EG(\pi, W_{xz}) = q$. To ease notation, we simply write EG when context is clear.

We point out several features from Tables 3-5 (all statements made ceteris peribus):

1. The efficiency gain varies strongly with the parameters governing the design.

2. The efficiency gain is highest when the degree of sparsity M is low, because stronger assumptions on δ lead to more powerful inference.

3. The efficiency gain increases with R^2_{xz}. Naturally, inference is impeded when the controls are highly correlated with the regressor of interest, as happens when R^2_{xz} is large.

4. The efficiency gain decreases with ρ_{zz}. Informally, one is better able to account for a control z_k when it is highly-correlated with other controls. For instance, if one were to conduct inference by inspecting $\hat{\beta}$ in a regression of y on x controlling for a subset of $\{z_1, \ldots, z_K\}$, then the omitted variable bias from excluding a single, influential control z_k diminishes as the included controls better-approximate z_k.

5. Holding $\frac{M}{K}$ and $\frac{K}{N}$ fixed, the efficiency gain decreases with N. This coincides with traditional assumptions from LASSO procedures that that the sparsity index is an asymptotically-negligible fraction of sample size.

Table 3: Efficiency Gains for Symmetric Designs: $R^2_{xz} = .98$

N	K/N	M/K	ρ_{zz}	EG
100	.2	.1	0	31
100	.2	.1	.1	16
100	.2	.1	.2	8.4
100	.2	.5	0	5.8
100	.2	.5	.1	1.2
100	.2	.5	.2	1.2
100	.5	.1	0	25
100	.5	.1	.1	4.4
100	.5	.1	.2	2.0
100	.5	.5	0	3.2
100	.5	.5	.1	1.2
100	.5	.5	.2	1.0

N	K/N	M/K	ρ_{zz}	EG
500	.2	.1	0	21
500	.2	.1	.1	2.3
500	.2	.1	.2	1.0
500	.2	.5	0	2.0
500	.2	.5	.1	1.2
500	.2	.5	.2	1.0
500	.5	.1	0	13
500	.5	.1	.1	1.2
500	.5	.1	.2	1.0
500	.5	.5	0	1.2
500	.5	.5	.1	1.2
500	.5	.5	.2	1.0

Table 4: Efficiency Gains for Symmetric Designs: $R^2_{xz} = .95$

N	K/N	M/K	ρ_{zz}	EG
100	.2	.1	0	10
100	.2	.1	.1	4.8
100	.2	.1	.2	2.9
100	.2	.5	0	2.0
100	.2	.5	.1	1.2
100	.2	.5	.2	1.0
100	.5	.1	0	7.8
100	.5	.1	.1	1.7
100	.5	.1	.2	1.2
100	.5	.5	0	1.2
100	.5	.5	.1	1.0
100	.5	.5	.2	1.0

N	K/N	M/K	ρ_{zz}	EG
500	.2	.1	0	6.3
500	.2	.1	.1	1.2
500	.2	.1	.2	1.2
500	.2	.5	0	1.2
500	.2	.5	.1	1.0
500	.2	.5	.2	1.2
500	.5	.1	0	4.4
500	.5	.1	.1	1.0
500	.5	.1	.2	1.0
500	.5	.5	0	1.2
500	.5	.5	.1	1.0
500	.5	.5	.2	1.0

Table 5: Efficiency Gains for Symmetric Designs: $R^2_{xz} = .9$

N	K/N	M/K	ρ_{zz}	EG	N	K/N	M/K	ρ_{zz}	EG
100	.2	.1	0	4.8	500	.2	.1	0	2.9
100	.2	.1	.1	2.0	500	.2	.1	.1	1.2
100	.2	.1	.2	1.7	500	.2	.1	.2	1.0
100	.2	.5	0	1.2	500	.2	.5	0	1.4
100	.2	.5	.1	1.0	500	.2	.5	.1	1.0
100	.2	.5	.2	1.0	500	.2	.5	.2	1.0
100	.5	.1	0	3.6	500	.5	.1	0	2.3
100	.5	.1	.1	1.2	500	.5	.1	.1	1.0
100	.5	.1	.2	1.2	500	.5	.1	.2	1.0
100	.5	.5	0	1.2	500	.5	.5	0	1.0
100	.5	.5	.1	1.0	500	.5	.5	.1	1.0
100	.5	.5	.2	1.0	500	.5	.5	.2	1.0

11 Power Bounds for Randomly-Generated Designs

In this section, we consider designs W_{xz} which do not accommodate the derivation of analytical least-favorable distributions. We randomly generate these designs according to the regression equation

$$x_i = z_{i1}\eta_1 + \cdots + z_{iK}\eta_K + v_i = z_i'\eta + v_i, \quad z_i \sim \mathcal{N}(0, \Sigma_z), \quad v_i \sim \mathcal{N}(0, \sigma_{vi}) \tag{11.1}$$

where $z_i \perp v_i$, and observations are independent across i. Conditional on each design, we apply the method suggested in Definition 9.2, part 5 to compute a power bound over valid tests of (10.1). This requires us to construct candidate distributions $\Lambda_0(\beta, \delta)$ over $\{0\} \times \Delta_M$, which we tailor to the data-generating process for W_{xz}.

In Section 11.1, we replicate the Monte Carlo study of Belloni et al. (2014), where η takes on an "approximately-sparse" form, and find that efficiency gains from sparsity that are small. In Section 11.2, we depart from Belloni et al. (2014) and allow the number of nonzero components (degree of sparsity) in η to vary, resulting in a wide range for the efficiency gain. In both cases, we apply a candidate distribution Λ_0 which equally-weights a number of points, chosen so that the controls have large explanatory power on the regressor of interest under the null. Formally, we apply Algorithm 11.1 below:

Algorithm 11.1. *The computational demands of the algorithm are determined by two exogenously-*

chosen parameters $(p, n) \in \mathbb{N}^2$ *and a set* $C \subset \mathbb{R}_+$.

1. *Fix* $n \in \mathbb{N}$,[28] *and draw random subsets* $\{S_1, \ldots, S_n\}$ *of size* M *from* $\{1, \ldots, K\}$.

2. *For each* S_j, $j = 1, \ldots, n$, *compute* $R^2(x, S_j)$, *the sample* R^2 *of* x *on the subset of* M *controls* $\{z_k : k \in S_j\}$. *Without loss of generality, order* S_j *so that* $R^2(x, S_j) \geq R^2(x, S_{j'})$ *whenever* $j \geq j'$.

3. *Fix* $p \in \mathbb{N}$,[29] *and for fixed* $c \in \mathbb{R}$, *define* $\Delta(c) = \{\delta \in \mathbb{R}^K : \delta = c(Z'_{S_j} Z_{S_j})^{-1} Z'_{S_j} Y,\ j = 1, \ldots, p\}$, *where* Z_{S_j} *is the* $N \times M$ *matrix consisting of columns of* Z *indexed by the elements in* S_j.

4. *Fix* $C \subset \mathbb{N}$, *and for each* $c \in C$, [30] *find the least-favorable distribution* $\Lambda_0(\beta, \delta)$ *(either theoretically or via the ALFD algorithm) in the testing problem*

$$H_0 : \beta = 0,\ \delta \in \Delta(c) \quad vs. \quad H_1 : \beta \neq \beta_1 > 0,\ \delta = 0,$$

   and let $\tilde{\pi}_c$ *be the induced power bound.*

5. *Let* $\tilde{\pi} = \min_{c \in [0,2]} \tilde{\pi}_c$ *be the lowest such power bound, and let the associated efficiency gain be* $RG(\tilde{\pi}, W_{xz})$.

11.1 Approximately-Sparse Designs

In the Monte Carlo study of Belloni et al. (2014), regressors are generated via Equation (11.1), with $\eta_k = c_\eta / k^2$ and c_η chosen to tune the population R^2 of z on x. Belloni et al. (2014) denote this an "approximately-sparse" representation, because although x is correlated with all controls z_k, this correlation diminishes quickly as k increases. Moreover, we assume that the controls $\{z_{i1}, \ldots, z_{iK}\}$ are governed by an $AR(1)$ process with persistence .5. In Table 6, we fix $N = 500$ and $K = 200$ and replicate the approximately-sparse design of Belloni et al. (2014), drawing different instances of W_{xz} according to Equation (11.1) and the following specifications:

Condition 11.2. Specifications for Approximately-Sparse Designs

1. $\sigma^2_{vi} = 1$ for all $i = 1, \ldots, N$.

[28] $n = 10^3$ in Section 11.1, and $n = 10^4$ in Section 11.2.
[29] $p = 1$ in Section 11.1, and $p = 10^2$ in Section 11.2.
[30] $C = \{1\}$ in Section 11.1, and C is a grid on $[0, 2]$ spaced by 0.05 in Section 11.2.

2. $Cov(z_{ik}, z_{ik'}) = .5^{|k-k'|}$ for all $1 \leq k, k' \leq K$.

3. $\eta_k = \frac{c_\eta}{k^2}$, $1 \leq k \leq K$, where c_η solves $\bar{R}^2_{xz} = \frac{\eta' \Sigma_z \eta}{\eta' \Sigma_z \eta + \sigma^2_{vi}}$ for exogenously-chosen $\bar{R}^2_{xz} \in [0, 1]$.

Hence, Table 6 is indexed by the free parameters: *(i)* $\bar{R}^2_{xz}$, which governs the process from which W_{xz} is drawn, and *(ii)* $\frac{M}{K}$, the index of the sparsity assumption in (10.1), expressed as a ratio. We allow $R^2_{xz} \in \{0, .2, .4, .6, .8, .9\}$ (covering the cases in Belloni et al. (2014)) and $\frac{M}{K} \in \{.01, .1\}$ to deliver strong sparsity assumptions. For each pair $(\frac{M}{K}, \bar{R}^2_{xz})$, we draw 100 samples of W_{xz} and, for each draw, construct a power bound $\tilde{\pi}$ over valid tests of (10.1) according to Algorithm 11.1 with $n = 10^3$, $p = 1$, and $C = \{1\}$.

Hence, we generate 100 values of the efficiency gain EG for each pair $(\frac{M}{K}, \bar{R}^2_{xz})$, and within each entry of Table 6, we report the inter-quartile range. We see that even with a sparsity assumption which imposes that only 1% of the control coefficients are non-zero, the third quartile efficiency gain does not exceed 2. We conclude that power gains from assuming sparsity in δ are relatively small when the underlying relationship between x and z is approximately-sparse. This result is intuitive, in the sense that when the regressor of interest has very low correlation with many of the control variables, imposing zero coefficients on those variables does little to aid inference.

Table 6: Efficiency Gains for Approximately-Sparse, Random Designs

M/K	$\bar{R}^2_{xz}$	EG, IQR	M/K	$\bar{R}^2_{xz}$	EG, IQR
.01	0	[1.56,1.66]	.1	0	[1.49,1.66]
.01	.2	[1.64,1.74]	.1	.2	[1.54,1.61]
.01	.4	[1.66,1.82]	.1	.4	[1.54,1.64]
.01	.6	[1.74,1.93]	.1	.6	[1.54,1.64]
.01	.8	[1.99,2.19]	.1	.8	[1.59,1.72]
.01	.9	[2.56,2.89]	.1	.9	[1.74,1.90]

11.2 Designs with Varying Sparsity

In this section, we allow the degree of sparsity between x and z, denoted M_{xz}, to vary among our randomly-generated designs. When M_{xz} is large, x is correlated with many z_k, and we expect large gains from assuming sparsity in δ. Moreover, we assume that the data-generating process for the controls $\{z_{i1}, \ldots, z_{iK}\}$ is an $AR(1)$ process as in Section 11.1, but we allow the persistence — denoted $\bar{\rho}_{zz}$ — to take on values 0 and .5. When $\bar{\rho}_{zz}$ increases, each particular control has less incremental

influence on the outcome, suppressing the gains from assuming its coefficient is zero.

In Table 7, we fix $N = 500$ and $K = 200$ and draw different instances of W_{xz} according to Equation (11.1) and the following specifications:

Condition 11.3. Specifications for Designs with Varying Sparsity

1. $\sigma_{vi}^2 = 1$ for all $i = 1, \ldots, N$.
2. $Cov(z_{ik}, z_{ik'}) = \rho_{zz}^{|k-k'|}$ for all $1 \leq k, k' \leq K$, where $\rho_{zz} \in [0, 1]$ is chosen exogenously.
3. $\eta = c_\eta(\underbrace{1, \ldots, 1}_{M_{xz}}, 0, \ldots, 0)'$, where c_η solves $\bar{R}_{xz}^2 = \frac{\eta' \Sigma_z \eta}{\eta' \Sigma_z \eta + \sum_{i=1}^N \sigma_{vi}^2} = .9$, and $M_{xz} \in \{1, \ldots, K\}$ is chosen exogenously.

Hence, Table 7 is indexed by the free parameters: *(i)* (M_{xz}, ρ_{zz}), which govern the process from which W_{xz} is drawn, and *(ii)* $\frac{M}{K}$, the index of the sparsity assumption in (10.1), expressed as a ratio. For each pair (M_{xz}, ρ_{zz}) listed in Table 7, we draw a *single* instance of W_{xz} and, for each draw, construct a power bound $\tilde{\pi}$ over valid tests of (10.1) according to Algorithm 11.1 with $n = 10^4$, $p = 100$, and $C = \{0, .05, \ldots, 1.95, 2\}$.[31]

We see from Table 7 that the efficiency gain increases with $\frac{M_{xz}}{K}$ and ρ_{zz}, as predicted. Moreover, the efficiency gain drops below 2 in all specifications once the sparsity index $\frac{M}{K}$ reaches .1. Hence, even when the relationship between x and z is highly non-sparse, assuming sparsity in δ does not yield large gains for inference. Because this analysis relies on power bounds derived from the candidate null distribution from Algorithm 11.1 whose support consists of only 100 points, which is not necessarily a good approximation of the least-favorable distribution, we note that one can potentially obtain substantially smaller values for the efficiency gain.

Table 7: Efficiency Gains for Varying-Sparsity, Random Designs

M/K	M_{xz}/K	ρ_{zz}	EG	M/K	M_{xz}/K	ρ_{zz}	EG
.02	.7	0	6.3	.02	1	0	6.3
.05	.7	0	4.4	.05	1	0	4.4
.1	.7	0	2.9	.1	1	0	3.6
.02	.7	.5	4.8	.02	1	.5	5.3
.05	.7	.5	2.9	.05	1	.5	3.2
.1	.7	.5	1.7	.1	1	.5	1.7

[31] We draw a single instance only, because the algorithm is computationally-intensive.

12 Conclusion

In this chapter, we applied several methods to determine how the degree of sparsity M in the hypothesis testing problem (9.5) affects inferential power, using the performance of the standard t-test on the coefficient β with all controls $z_1, \ldots, z_K$ present in the regression as a baseline. For given M, we computed the power bound over valid tests of (9.5) and the corresponding efficiency gain — the increase in the number of observations necessary for the the standard t-test to have power equal to the bound.

In Section 10, we derived a theoretical result pertaining to designs where *(i)* pairwise correlations between control variables are constant and *(ii)* the average of the control variables is the best predictor of the regressor of interest from the linear span of controls. Conditional on these designs, we showed that the null least-favorable distribution against point-alternatives $\beta = \beta_1 > 0, \delta = 0$ places uniform weight on values of δ arising from least-squares regression of y on x and all subsets of M controls from $\{z_1, \ldots, z_K\}$, appropriately-scaled.[32] For future research, we note that the form of this least-favorable distribution suggests a nearly-optimal test for (9.5): evaluate the statistic which totals the residual sums-of-squares from linear regressions of y on x and all subsets of controls $\{z_1, \ldots, z_K\}$, taken M at a time. For tractability, emulate Algorithm 10.4 and use a random selection of subsets of M controls.

Subsequently, in Section 10.2, we computed infimum power bounds with the previously-developed theoretical tools, and we discovered that the efficiency gain depends heavily on the mutual correlation between controls as well as the R^2 of the regressor of interest on the controls. In samples of 100 and 500 observations, we observed that large efficiency gains — up to 30 — are possible when the R^2 is .98 and the mutual correlation is 0, while the gains decrease dramatically — to less than 4 — if the R^2 drops to .9 or correlation rises to .1.

In Section 11, we generated designs according to Equation (11.1) and derived valid, non-infimum power bounds. We determined in Section 11.1 that under the specifications from the Monte Carlo exercises of Belloni et al. (2014) where the regressor of interest and controls have an approximately-sparse relationship, the efficiency gains are small, never exceeding 2. Subsequently, we determined in Section 11.2 that even when the underlying relationship between the regressor of interest and controls is allowed to be non-sparse, the gains from sparsity are moderate when the

[32] We conjecture that Theorem 10.2 of Section 10 can be extended to accommodate general designs, for which the least-favorable distribution might also be a discrete mixture over finite-many points, but we leave this to future research.

sparsity assumption is set to $M = .1K$, never exceeding 4.

In aggregate, the results of Sections 10 and 11 suggest that assuming sparsity in finite samples is greatly beneficial only if one believes that less than 10% or less of the control variables have non-zero coefficients, or the control variables explain more than 90% of the variation in the regressor of interest. We recognize that our methodology relies on two stylized assumptions: *(i)* Gaussian, homoskedastic errors and *(ii)* $\delta = 0$ under the alternative. The first is a canonical assumption for statistical models, and we hesitate to alter it without justification for a particular error structure which procures qualitatively-different results. It is possible, in principle, to relax the second assumption, but this is equivalent to re-defining the regressor of interest x and implementing $\delta = 0$, and we conjecture that the various designs we considered are natural specifications for x. We leave these investigations to future research.

Part III

Conditionally-Unbiased Asymptotic Variance Estimation

13 Introduction

Ordinary least squares (OLS) is widely used for inference on coefficients in a linear model. Primitive assumptions typically ensure that OLS statistics are consistent and asymptotically-Gaussian, with mean equal to some function of the parameter(s) of interest. To deal with the unknown scale of the OLS statistic, it is common practice to replace its true asymptotic variance with a consistent estimator and invoke Slutsky's Lemma to ensure convergence to the desired asymptotic distribution.

There is a large literature on consistent estimation of the asymptotic variance. In a seminal paper, White (1980) constructs an estimator which accommodates potentially-heteroskedastic, independent errors. This result can be extended to the case where observations have arbitrary dependence within specified clusters, if it is assumed that the number of independent clusters approaches infinity. These methods exploit the following features: *(i)* the asymptotic variance of the OLS estimate is an observation-weighted average of error variances, *(ii)*replacing the error variances with realized squared regression errors in the weighted average induces the same probability limit, and *(iii)* OLS residuals are sufficiently-precise proxies of the true errors that their substitution in the weighted average results in asymptotically-negligible sampling uncertainty.

As discussed in Cattaneo et al. (2015), however, such methods can be unreliable when the number of covariates is large. The tendency for OLS is to overfit, causing the average of squared residuals to underestimate the average of true squared errors and invalidating feature *(iii)*. As an alternative to White (1980), Cattaneo et al. (2015) construct a heteroskedasticity-consistent estimator for the variance which allows the number of covariates to be a non-trivial fraction of the sample size. As in White (1980), the estimator of Cattaneo et al. (2015) is a quadratic form in the vector of OLS residuals, but the sandwich matrix is retooled to compensate for degrees of freedom eliminated by first-stage OLS. In deriving the consistency of their estimator, Cattaneo et al. (2015) assume that errors are independent across observations.

Cattaneo et al. (2015) and others[33] argue that many-covariate asymptotics are appropriate when conducting inference on a treatment with many possible confounding effects. Such is the case in Donohue III and Levitt (2001), who study the relationship between crime and abortion rates in 48 states over 13 years. Although Donohue III and Levitt (2001) only include 9 covariates along with fixed-effects in their main specification, subsequent investigations such as Belloni et al. (2014) have considered 276 other covariates consisting of higher-order transformations and interactions of variables from Donohue III and Levitt (2001). With only 576 observations, such circumstances lend themselves to the asymptotics of Cattaneo et al. (2015), but the estimator therein requires observations to be independent, while the data from Donohue III and Levitt (2001) exhibits panel structure, where it is typically assumed that observations can be correlated across time but independent across clusters/states.

In this chapter, we generalize the estimator of Cattaneo et al. (2015) to allow for error clustering. As a departure from Cattaneo et al. (2015), we assume a conditionally-unbiased model and constrain the limiting behavior of sequences of regressors rather than their underlying data-generating process. In Section 14, we review the methodology of Cattaneo et al. (2015) and discuss an alternative derivation of their variance estimator. In Section 15, we construct the cluster-robust version of this estimator, derive its consistency, and discuss its supporting assumptions. In particular, we focus on one assumption stipulating that the Hadamard square of the projection matrix which forms OLS residuals is invertible, and the sum of elements of that inverse is bounded by a rate proportional to N^2, where N is the number of independent clusters. In Section 16, we conduct Monte Carlo simulations to determine whether this assumption holds for various data-generating processes.

14 Setup and Background

We assume data is generated by the following equation

$$y_i = x_{i,1}\beta_1 + \cdots + x_{i,K}\beta_K + e_i, \quad E[e|X] = 0, \quad Var[e|X] = \Sigma_X \tag{14.1}$$

where $i = 1, \ldots, N$ is the observational index, y_i is the outcome, $x_{i,1}, \ldots, x_{i,K}$ are deterministic covariates, e_i is the (unobserved) error term, and in obvious notation $Y = X\beta + e$, $\beta = (\beta_1, \ldots, \beta_K)'$

[33] See Belloni et al. (2014), for instance.

denote stacked variables. In this chapter, K is allowed to vary with N, and we estimate the conditional variance of linear statistics allowing heteroskedasticity and clustering in e. All limits are taken as $N \to \infty$, unless otherwise specified.

Definition 14.1. For a matrix A, let $T(A) = A'e$.

A finite linear combination of OLS statistics can be written in the form $N^{-1/2}A'Y$, where A is an $N \times P$ matrix which may depend on X (with P fixed). It is typical to conduct inference on the mean $E[N^{-1/2}A'Y] = E[N^{-1/2}A'X]\beta$ — a transformation of β — by inspecting the re-centered statistic

$$N^{1/2}(N^{-1/2}A'Y - E[N^{-1/2}A'X]\beta) = A'e$$

and imposing sufficient regularity for $T(A) = A'e$ to be asymptotically-Gaussian. For this reason, the remainder of this chapter focuses on estimating the conditional variance of linear forms $T(A)$ when Σ_X is diagonal and block-diagonal. In Example 14.3, we outline a few cases which conform to this framework.

Definition 14.2.

1. Let the standard OLS estimator be denoted $\hat{\beta} = (X'X)^{-1}X'Y$.
2. Let $M = I - X(X'X)^{-1}X'$ and $\hat{e} = Y - X\hat{\beta} = MY$.
3. Let X_{-1} denote X absent its first column, and let $M_{-1} = I_N - X_{-1}(X'_{-1}X_{-1})^{-1}X'_{-1}$.

Example 14.3. Linear Statistics:

1. Suppose K is fixed with N, and consider inference on the full coefficient β via the statistic $N^{1/2}(\hat{\beta} - \beta) = N^{1/2}(X'X)^{-1}X'e$. In our notation, $N^{1/2}(\hat{\beta} - \beta) = T(A)$ with $A = N^{1/2}(X'X)^{-1}X'$ and $P = K$.
2. Consider Li and Mueller (2016), where regressors X are assumed fixed, $\frac{K}{N} \to c \in (0, 1)$, and one conducts inference on β_1 via the bivariate statistic $T_{biv} = N^{1/2}(\hat{\beta}_1 - \beta_1, \hat{\beta}_1^{short} - \beta_1^{short})'$, where $\hat{\beta}_1^{short} = (X_1'X_1)^{-1}X_1'Y$, and $\beta_1^{short} = E[\hat{\beta}^{short}] = (X_1'X_1)^{-1}X_1'X\beta$. We can write $T_{biv} = T(A)$, where $P = 2$, and $A = (A_1, A_2)$ with $A_1 = N^{1/2}(X_1'M_{-1}X_1)^{-1}X_1'M_{-1}$ and $A_2 = N^{1/2}(X_1'X_1)^{-1}X_1'$.▶

When K is assumed fixed, the estimation error in $\hat{\beta}$ is small enough that under weak regularity conditions, a White-type estimator whose accuracy depends on the degree to which residuals $\hat{e} = Y - X\hat{\beta}$ well-approximate e consistently estimates $\Omega_X = Var[T(A)|X]$. As discussed in Cattaneo et al. (2015), this no longer holds if $\frac{K}{N} \to c \in (0,1)$, because even small estimation errors in each component $\hat{\beta}_k$ can induce non-trivial biases when aggregated over all $k = 1, \ldots, K$. We demonstrate this below.

Example 14.4. Suppose that *(i)* independent of X, e_i are jointly independent with variance σ_i^2, so that $\Sigma_X = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$, *(ii)* $A_i = 1$ for all i, so the objective is to estimate the average error variance $\Omega = Var[T(A)] = N^{-1}\sum_{i=1}^N \sigma_i^2$, and *(iii)* $N^{-1}\sum_{i=1}^N x_i x_i' = O_p(1)$.

1. The White estimator $\hat{\Omega}^{white} = \sum_{i=1}^N \hat{e}_i^2$ is the sample analog of Ω, and when K is fixed with N, its consistency follows:

$$\hat{\Omega}^{white} = N^{-1}\sum_{i=1}^N \hat{e}_i = N^{-1}\sum_{i=1}^N (e_i + x_i'(\hat{\beta} - \beta))^2 + o_p(1)$$

$$= N^{-1}\sum_{i=1}^N e_i^2 + O_p(\underbrace{\|\hat{\beta} - \beta\|}_{o_p(1)} + \|\hat{\beta} - \beta\|^2) + o_p(1) = \underbrace{N^{-1}\sum_{i=1}^N E[e_i^2]}_{\Omega} + o_p(1).$$

2. Consider the case $\frac{K}{N} \to c \in (0,1)$, and suppose $e|X \sim \mathcal{N}(0, I_N)$, so that $\sigma_i^2 = 1$ for all i. Observe that because M is a rank $N-K$ projection matrix for any realization of the regressors, $\hat{\Omega}^{white} = N^{-1}e'Me \sim N^{-1} \cdot \chi^2_{N-K}$, which implies $\hat{\Omega}^{white} \to_p 1 - c \neq \text{plim}\Omega = 1$. Hence, $\hat{\Omega}^{white}$ is asymptotically-biased, because K — the number of degrees of freedom removed by the projection matrix M — is a nontrivial fraction of N. In this example, the bias correction should be $\text{plim}(\Omega - \hat{\Omega}^{white}) = c$, but in general, its form depends on A.▶

Cattaneo et al. (2015) propose an alternative to the White estimator, presented in Definition 14.5, for the case when $\frac{K}{N} \to c \in (0,1)$ and the data is generated under Equation (14.1) with heteroskedastic, independent errors. As a slight departure from the setup of this chapter, Cattaneo et al. (2015) do not assume exact strict exogeneity in the underlying model, but the mechanics behind their estimator carry over to our analysis. We briefly discuss it following Definition 14.5.

Definition 14.5. Let $\odot$ denote the Hadarmard or elementwise matrix product, and let $\hat{\Omega}^o = N^{-1}\hat{e}'S^o\hat{e}$ with $S^o = \text{diag}\left((M \odot M)^{-1}\text{vec}(A \odot A)\right)$.

In the setup of Cattaneo et al. (2015), it is assumed that $\Sigma_X = \text{diag}(\sigma_1^2, \ldots, \sigma_K^2)$. To motivate their methodology, consider using a quadratic form $\hat{\Omega}_S = N^{-1}\hat{e}'S\hat{e}$, where S is a general $N \times N$ matrix, to estimate Ω_X. Observe that the White estimator, which can be written $\hat{\Omega}^{white} = N^{-1}\hat{e}'S^{white}\hat{e}$ with $S^{white} = \text{diag}(A_1^2, \ldots, A_N^2)$, belongs to this class, but recall that it is asymptotically-biased under $K/N \to c \in (0,1)$ asymptotics. For conditional unbiasedness, it must be true that $E[\hat{e}'S\hat{e}|X] = Tr(MSM\Sigma_X) = \sum_{i=1}^N \sigma_i^2 A_i^2$ for all possible $\Sigma_X = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$, which holds if and only if for all $\sigma_1^2, \ldots, \sigma_N^2$,

$$Tr(MSM\Sigma_X) = \sum_{i=1}^N \sigma_i^2 \cdot (\sum_{j=1}^N M_{ij}^2 S_{jj}^2) = \sum_{i=1}^N \sigma_i^2 A_i^2$$

$$\Leftrightarrow (\sum_{j=1}^N M_{ij}^2 S_{jj}^2) = A_i^2 \quad \forall\, i = 1, \ldots, N$$

$$\Leftrightarrow (M \odot M) \cdot (S_{11}^2, \ldots, S_{NN}^2)' = (A_1^2, \ldots, A_N^2)'$$

$$\Leftrightarrow (S_{11}^2, \ldots, S_{NN}^2)' = (M \odot M)^{-1}(A_1^2, \ldots, A_N^2)'. \tag{14.2}$$

Equation (14.2) stipulates that the diagonal elements of S must match the vector $(M \odot M)^{-1}(A_1^2, \ldots, A_N^2)'$. This is true for $\hat{\Omega}^o$ — in which $S = S^o$ is simply the appropriate diagonal matrix — and it can be shown that when e is Gaussian and homoskedastic, setting $S = S^o$ minimizes the variance of $\hat{\Omega}_S$.

Implicitly, the method of Cattaneo et al. (2015) depends on the invertibility of $M \odot M$, which is not guaranteed for all projection matrices M.[34] The existence and magnitude of $(M \odot M)^{-1}$ — as measured by the sum of its elements — is also closely related to the consistency of $\hat{\Omega}^o$. Because these conditions are special cases of those in Section 15, we postpone their discussion.

15 Main Result: Extension of Cattaneo et al. (2015) to the Clustered Case

Cattaneo et al. (2015) construct an estimator for the variance of $T(A)$ under $\frac{K}{N} \to c \in [0,1)$ asymptotics, assuming error terms are independent across observations. As discussed in Section 13, the independence assumption is not compelling in empirical studies such as Donohue III and Levitt (2001), where data is potentially-correlated within clusters. In this section, we construct the analog

[34] Cattaneo et al. (2015) show that a sufficient condition for this to hold is for $M \odot M$ to be diagonally-dominant, but in this chapter, we simply take the invertibility as given and do not provide primitive assumptions.

of the estimator of Cattaneo et al. (2015) for balanced panels.

Suppose the data is structured as follows

$$y_{ij} = x_{ij,1}\beta_1 + \cdots + x_{ij,K}\beta_K + e_{ij} = x_{ij}'\beta + e_{ij}, \quad E[e_{ij}|\{x_{ij}\}_{j=1}^J] = 0 \tag{15.1}$$

where $i = 1, \ldots, N$ indexes clusters, $j = 1, \ldots, J$ indexes within-cluster observations, and J is fixed with N. Stacking observations, we write $Y = X\beta + e$, where $y^i = (y_{i1}, \ldots, y_{iJ})'$, $x_k^i = (x_{i1,k}, \ldots, x_{iJ,k})'$, $e^i = (e_{i1}, \ldots, e_{iJ})'$, and

$$Y = \begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix}, \quad X = \begin{pmatrix} x_1^1 & \cdots & x_K^1 \\ \vdots & \ddots & \vdots \\ x_1^N & \cdots & x_K^N \end{pmatrix}, \quad e = \begin{pmatrix} e^1 \\ \vdots \\ e^N \end{pmatrix}.$$

Assume that $e^i \perp e^{\tilde{i}}$ whenever $i \neq \tilde{i}$, so that N represents the total number of independent clusters. As a departure from Cattaneo et al. (2015), suppose for all i that $Var[e^i|X] = E[e^i(e^i)'|X] = \Sigma_X^i$ where Σ_X^i is a general covariance matrix which does not depend on X, allowing errors to be correlated within clusters. Define $\Sigma_X = E[ee'|X] = \text{diag}(\Sigma_X^1, \ldots, \Sigma_X^N)$ to be the conditional variance of the full vector of errors.

As in Section 14, we analyze the asymptotic properties of linear forms $A'e$, where A are $NJ \times P$ matrices which can vary with N, and P is finite. The conditional variance of $A'e$ can be written $\Omega_X = Var[A'e|X] = A'\Sigma_X A$, and under mild regularity conditions on X's behavior, $\Omega_X^{-1/2} T(A) = \Omega_X^{-1/2} A'e \to_d \mathcal{N}(0, I_P)$. Just as in Cattaneo et al. (2015), we estimate Ω_X with a quadratic form in the regression residuals. Let the conditionally-unbiased asymptotic variance estimator be a $P \times P$ matrix denoted $\hat{\Omega}^*$ with entries as in the following definition:

Definition 15.1. For fixed $1 \leq p, q \leq P$, let $\hat{\Omega}_{p,q}^* = \hat{e}' S^* \hat{e}$, where $S^* = \text{diag}(S^{*1}, \ldots, S^{*N})$ is a block-diagonal matrix, and S^{*i} are $J \times J$ matrices such that

$$\begin{pmatrix} \text{vec}(S^{*1}) \\ \vdots \\ \text{vec}(S^{*N}) \end{pmatrix} = (M * M)^{-1} v,$$

with v and $M * M$ defined below.

1. For $1 \leq p \leq P$, let $A_p \in \mathbb{R}^{NJ}$ denote the p'th column of A, and for $1 \leq i \leq N$ let $A_p^i =$

$(A_{p,N(i-1)+1}, \ldots, A_{p,N(i-1)+J}) \in \mathbb{R}^J$ denote the i'th block of A_p.

2. For $1 \leq p, q \leq P$, let $v \in \mathbb{R}^{NJ}$ denote the vector

$$v = \begin{pmatrix} \text{vec}(A_p^1(A_q^1)') \\ \text{vec}(A_p^2(A_q^2)') \\ \vdots \\ \text{vec}(A_p^N(A_q^N)') \end{pmatrix}.$$

3. For $1 \leq i, \tilde{i} \leq N$, let the $J \times J$ matrix $M^{i\tilde{i}}$ denote the $(i, \tilde{i})$ block of M, so that

$$M = \begin{pmatrix} M^{11} & M^{12} & \cdots & M^{1N} \\ M^{21} & M^{22} & \cdots & M^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M^{N1} & M^{N2} & \cdots & M^{NN} \end{pmatrix}.$$

4. Let $M * M$ denote the $NJ \times NJ$ matrix (assumed non-singular)

$$M * M = \begin{pmatrix} M^{11} \otimes M^{11} & M^{12} \otimes M^{12} & \cdots & M^{1N} \otimes M^{1N} \\ M^{21} \otimes M^{21} & M^{22} \otimes M^{22} & \cdots & M^{2N} \otimes M^{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M^{N1} \otimes M^{N1} & M^{N2} \otimes M^{N2} & \cdots & M^{NN} \otimes M^{NN} \end{pmatrix}.$$

Definition 15.2. Let $\hat{\Omega}^*$ be the $P \times P$ matrix whose elements are $\hat{\Omega}^*_{p,q}$ for $1 \leq p, q \leq P$.

The estimator $\hat{\Omega}^*$ is an extension of the estimator $\hat{\Omega}^o$ of Cattaneo et al. (2015), and when $J = 1$, $M * M = M \odot M$ and $\hat{\Omega}^* = \hat{\Omega}^o$. Like $\hat{\Omega}^o$, $\hat{\Omega}^*$ only depends on X and $\hat{e}$ and is, thus, invariant to translations of the outcome Y the linear span of regressors X. Conditional on the invertibility of $M * M$, sufficient conditions for the consistency of $\hat{\Omega}^*$ mirror those for $\hat{\Omega}^o$, with $M * M$ taking the place of $M \odot M$. This is formally stated in Theorem 15.3 and Corollary 15.4 below.

Theorem 15.3. *1.* $E[\hat{\Omega}^*|X] = \Omega_X$.

2. *Suppose* $\sup_X \max_{i=1,\ldots,N} Var[e_i|X] = O(1)$. *Then* $Var[\hat{\Omega}^*|X] = N^{-2}O(C_X)$, *with*

$$C_X = \max_{r=1,\ldots,NJ} A_r^2 \tilde{A}_r^2 \cdot \left(\sum_{r,t=1}^{NJ} \kappa_{r,t}^2 \right)^{1/2}$$

and $\kappa = (M * M)^{-1}$.

Corollary 15.4. *Suppose* $\sup_X \max_{i=1,\ldots,N} Var[e_i|X] = O(1)$. *Then* $C_X = o_p(N^2)$ *implies* $\hat{\Omega}^* - \Omega_X \to_p 0$.

Note that in Corollary 15.4, consistency does not explicitly constrain the rate at which K grows. Rather, embedded within the condition $C_X = o_p(N^2)$ is that the sum-of-squared elements in $(M * M)^{-1}$ does not diverge too quickly, the propensity of which decreases in K. We conjecture that under $\frac{K}{N} \to c \in (0, 1)$ asymptotics and relatively weak assumptions on the data-generating process underlying the regressors, *(i)* $\max_{r=1,\ldots,NJ} A_r^2 \tilde{A}_r^2$ is bounded by the rate $O_p(N^\alpha)$ for arbitrary $\alpha > 0$, and *(ii)* $\left(\sum_{r,t=1}^{NJ} \kappa_{r,t}^2 \right)^{1/2} = o_p(N^{2-\epsilon})$ for some sufficiently-small $\epsilon > 0$. We do not attempt to prove these conjectures, but in Section 16, we evaluate whether $C_X = o_p(N^2)$ can plausibly hold in Monte Carlo simulations.

15.1 Fixed-Effects

An important caveat of our methodology is that whenever there exists a single regressor which serves as an indicator for two or fewer clusters — as when x contains fixed-effect dummies for each cluster i — $M * M$ is not invertible and, thus, $\hat{\Omega}^*$ is not defined.

To deal with fixed-effects, consider the model $Y = X\beta + V\eta + e$, where V is an $NJ \times N$ matrix representing fixed-effect dummies in columns (formally, $V_{rs} = 1\{(s-1)J < r \leq sJ\}$ for all $r = 1, \ldots, NJ$ and $s = 1, \ldots, N$), and the coefficient $\eta \in \mathbb{R}^J$ collects corresponding levels. One can subsequently specify a full-rank, block-diagonal $N(J-1) \times NJ$ matrix

$$D = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_N \end{pmatrix}$$

in which each D_i, $i = 1, \ldots, N$ is a full-rank $(J-1) \times J$ matrix which satisfies $D_i(1, \ldots, 1)' = 0$ (in Section 16, for instance, D is the matrix which takes first-differences within each cluster). The purpose of D is to de-mean within-cluster observations.

Because the error term e is conditionally-unbiased, so too is $\tilde{e} := De$. It suffices to proceed within the model

$$\tilde{Y} = \tilde{X}\beta + \tilde{e}, \quad E[\tilde{e}|X] = 0, \quad Var[\tilde{e}|X] = D\Sigma_X D' \tag{15.2}$$

where $(\tilde{Y}, \tilde{X}, \tilde{e}) := D(Y, X, e)$. For various sequences of $N(J-1) \times P$ matrices $\tilde{A}$, linear statistics $\tilde{A}'\tilde{Y}$ provide inference for the quantity $E[N^{1/2}\tilde{A}'\tilde{X}]\beta$.[35] Because Equation (15.2) contains no fixed-effect, it becomes feasible again to estimate the asymptotic variance of linear forms.

16 Monte Carlo Simulations

In Section 15, we outlined, in Corollary 15.4, a sufficient condition for consistency of a cluster-robust estimator of asymptotic variance. In this section, we determine the extent to which this condition holds in a Monte Carlo exercise, assuming a linear model:

$$y_{ij} = \underbrace{x_{ij,1}\beta_1 + \cdots + x_{ij,K}\beta_K}_{x_{ij}'\beta} + \eta_i + e_{ij} \tag{16.1}$$

where $i = 1, \ldots, N$ indexes clusters, $j = 1, \ldots, J$ indexes within-cluster observations, η_i are non-random fixed effects, and regressors x_{ij} are drawn from Gaussian and t-distributions. Letting Ω be the asymptotic variance of the OLS estimator $\hat{\beta}_1$ in a first-differenced version of (16.1), we use simulations to find the error of $\hat{\Omega}^*$ (from Definition 15.1) in estimating Ω. As in Section 15.1, we write $Y = X\beta + V\eta + e$ to denote the stacked equation, with fixed-effect dummies in V, and we generate data according to Condition 16.1.

Condition 16.1. Suppose $\{y_{ij}, x_{ij}, e_{ij}\}$ are independent across clusters i, and suppose $x^i = (x_{i,1}, \ldots, x_{i,J})'$ is independent from $e^i = (e_{i,1}, \ldots, e_{i,J})$ for all i. Letting $\Sigma_{AR}(\rho)$ denote the $J \times J$ variance matrix of an AR(1) with persistence ρ (whose $(j, \tilde{j})$'th element is $\rho^{|j-\tilde{j}|}$), we impose the following:

1. e^i has covariance $\Sigma_{AR}(0.1)$.

[35] We recognize that not every parameter of interest $E[A'X]\beta$ in the original model (15.1) can be expressed as a parameter $E[\tilde{A}'\tilde{X}]\beta$ in the transformed model (15.2). Regardless, this framework is expansive enough to accommodate any linear combination of β in either model.

2. $(x_{i1,k}, \ldots, x_{iJ,k})' \in \mathbb{R}^J$ has covariance $\Sigma_{AR}(0.5)$ for each (i,k).

3. $(x_{i1,k}, \ldots, x_{iJ,k})$ are mutually independent over $k = 1, \ldots, K$.

4. In Design 1, we take x^i, e^i to be multivariate Gaussian, and in Design 2, we take x^i, e^i to be multivariate student-t with $d.f.$ 5.

To deal with fixed-effects η_i, we evaluate a first-differenced model $\tilde{y}_{i,j} = \tilde{x}_{i,j,1}\beta_1 + \cdots + \tilde{x}_{i,jK}\beta_K + \tilde{e}_{i,j}$ with $\tilde{y}_{ij} = y_{i+1,j} - y_{i,j}$, $\tilde{x}_{i,j} = x_{i+1,j} - x_{i,j}$, and $\tilde{e}_{i,j} = e_{i+1,j} - e_{i,j}$. In the context of Section 15.1 we apply an $N(J-1) \times NJ$ first-difference matrix

$$D = \begin{pmatrix} \bar{D} & 0 & \cdots & 0 \\ 0 & \bar{D} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{D} \end{pmatrix}, \quad \bar{D} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(J-1)\times J}$$

to yield $\tilde{Y} = \tilde{X}\beta + \tilde{e},\ E[\tilde{e}|\tilde{X}] = 0$, where $(\tilde{Y}, \tilde{X}, \tilde{e}) := D(Y, X, e)$.

Let $\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$ denote the usual OLS statistic within the differenced model, so that $\hat{\beta}_1$ is conditionally-unbiased for β_1. Defining $\tilde{M}_{-1} = I_{N(J-1)} - \tilde{X}_{-1}(\tilde{X}'_{-1}\tilde{X}_{-1})^{-1}\tilde{X}'_{-1}$ with $\tilde{X}_{-1} = (\tilde{X}_2, \ldots, \tilde{X}_K)$, observe that the Frisch-Waugh theorem implies $\hat{\beta}_1 = (\tilde{X}'\tilde{M}_{-1}\tilde{X})^{-1}\tilde{X}'\tilde{M}_{-1}\tilde{Y} = \beta_1 + (\tilde{X}'\tilde{M}_{-1}\tilde{X})^{-1}\tilde{X}'\tilde{M}_{-1}\tilde{e}$. The inferential statistic we analyze is $N^{1/2}(\hat{\beta}_1 - \beta_1) = A'\tilde{e}$ with $A = N^{1/2}\tilde{M}_{-1}\tilde{X}(\tilde{X}'\tilde{M}_{-1}\tilde{X})^{-1}$, and its variance can be written $\Omega_X = A'\tilde{\Sigma}A$ with $\tilde{\Sigma} = E[\tilde{e}\tilde{e}'] = D \cdot \text{diag}\,(\Sigma^1, \ldots, \Sigma^N) \cdot D'$.

The estimator $\hat{\Omega}^*$ of Section 15 is unbiased for Ω_X, and in the following Monte Carlo simulations we tabulate *(i)* its deviation from Ω_X, as a ratio, and *(ii)* the magnitude of the parameter C_X from Theorem 15.3 (whose boundedness constitutes a sufficient condition for $\hat{\Omega}^*$ to be consistent). Formally, for $N \in \{50, 100, 250, 500\}$, $J = 4$, and $K = N$, we generate 1000 instances of the observables according to Designs 1 and 2 described above. In each instance, we record $\hat{\Omega}^*$, Ω_X, and C_X. In the tables below we present inter-quartile ranges of $\hat{\Omega}^*/\Omega_X$ and $N^{-2}C_X$ over the 1000 draws.

Table 8: Monte Carlo Error of $\hat{\Omega}^*$

Design 1:

N	IQR of $\hat{\Omega}^*/\Omega_X$	IQR of $N^{-2}C_X$
50	$[.69, 1.21]$	$[4.9 \times 10^{-4}, 7.2 \times 10^{-4}]$
100	$[.82, 1.18]$	$[1.0 \times 10^{-4}, 1.4 \times 10^{-4}]$
250	$[.87, 1.11]$	$[1.2 \times 10^{-5}, 1.6 \times 10^{-5}]$
500	$[.91, 1.08]$	$[2.3 \times 10^{-6}, 3.0 \cdot 10^{-6}]$

Design 2:

N	IQR of $\hat{\Omega}^*/\Omega_X$	IQR of $N^{-2}C_X$
50	$[.63, 1.24]$	$[4.6 \times 10^{-4}, 7.4 \times 10^{-4}]$
100	$[.76, 1.21]$	$[9.4 \times 10^{-5}, 1.5 \times 10^{-4}]$
250	$[.85, 1.12]$	$[1.2 \times 10^{-5}, 2.0 \times 10^{-5}]$
500	$[.88, 1.09]$	$[2.7 \times 10^{-6}, 5.0 \times 10^{-6}]$

Table 8 shows that as N increases, the error of estimation in $\hat{\Omega}^*$ decreases as a proportion of Ω_X. In line with this observation, the constant C_X appears to shrink at the required rate $o_p(N^2)$ to guarantee consistency of $\hat{\Omega}^*$.

17 Conclusion

In this chapter, we constructed $\hat{\Omega}^*$, a cluster-robust version of the estimator of Cattaneo et al. (2015), to estimate Ω_X, the conditional asymptotic variance of a general linear statistic (noting that in a model with cluster fixed-effects, one must first-difference before applying $\hat{\Omega}^*$). We showed that a sufficient condition for the consistency of $\hat{\Omega}^*$ is for C_X, as defined in Theorem 15.3, to satisfy the rate $o_p(N^2)$. We conjectured that this holds under mild regularity assumptions and asymptotics where the number of regressors is a non-negligible fraction of sample size N. Finally, we demonstrated in simulations that, indeed, $N^{-2}C_X$ appears to shrink and $\hat{\Omega}^*$ appears to be consistent under reasonable data-generating processes. For future research, one might attempt to formally prove this conjecture.

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2008): "Using Selection on Observed Variables to Assess Bias from Unobservables when Evaluating Swan-Ganz Catheterization," *American Economic Review: Papers and Proceedings*, 98, 345–350.

Andrews, D., X. Cheng, and P. Guggenberger (2011): "Generic Results for Establishing the Asymptotic Size of Confidence Sets and Tests," *Cowles Foundation Discussion Paper No. 1813*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81.

Cattaneo, M., M. Jansson, and W. Newey (2015): "Heteroskedastic Consistent Standard Errors with Many Covariates," *Working Paper*.

Donohue III, J. J. and S. D. Levitt (2001): "The Impact of Legalized Abortion on Crime," *The Quarterly Journal of Economics*, 116, 379–420.

Elliott, G., U. Mueller, and M. Watson (2015): "Nearly Optimal Tests when a Nuisance Parameter is Present Under the Null Hypothesis," *Econometrica*, 83, 771–881.

Fan, J. and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348?1360.

Foote, C. L. and C. F. Goetz (2008): "The Impact of Legalized Abortion on Crime: Comment," *The Quarterly Journal of Economics*, 123, 425–440.

Gardner, R. J. and A. Zvavitch (2010): "Gaussian Brunn-Minkowski Inequalities," *Transactions of the American Mathematical Society*, 362, 5333–5353.

Gelman, A. (2011): "Review Essay: Causality and Statistical Learning," *American Journal of Sociology*, 117, 955–966.

Haberman, S. J. (1980): "A Note on the Conditional Moments of a Multivariate Normal Distribution Confined to a Convex Set," *Journal of Multivariate Analysis*, 10, 398–404.

Kline, P. and A. Santos (2013): "Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure," *Quantitative Economics*, 4, 231–267.

Leeb, H. and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2008a): "Can one Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338–376.

Li, C. (2016): "Conditionally-Unbiased Asymptotic Variance Estimation," *Working Paper*.

Li, C. and U. Mueller (2016): "Inference in Regressions with Many Controls under a Sum-of-Squares Bound," *Working Paper*.

Olver, F. W. (1997): *Asymptotics and Special Functions*, A K Peters/CRC Press; 2nd edition.

Reinhardt, H. E. (1961): "The Use of Least Favorable Distributions in Testing Composite Hypotheses," *The Annals of Mathematical Statistics*, 32, 1034–1041.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

Van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.

Appendix

A Proofs, Background, and Figures from Part I

Unless otherwise-specified, all notation throughout Section A is adopted from Part 1. Suppose X is a deterministic $N \times 1$ matrix, Z is a deterministic $N \times K$ matrix. With parameters $\beta \in \mathbb{R}$ and $\delta \in \mathbb{R}^K$, we consider a Gaussian, homoskedastic model

$$Y = X\beta + Z\delta + e, \quad e|X, Z \sim \mathcal{N}(0, I_n).$$

Recall the following definitions:

1. $W_{xz} = (X, Z)$, and$N^{-1}W_{xz}'W_{xz} = \begin{pmatrix} 1 & \rho' \\ \rho & I_K \end{pmatrix}$, where $\rho = N^{-1}Z'X \in \mathbb{R}^K$, and $\|\rho\|^2 \in (0, 1)$.
2. $(\hat{\beta}, \hat{\delta}')' = (W_{xz}'W_{xz})^{-1}W_{xz}'Y$.
3. $(\hat{b}, \hat{d}) = \sqrt{N}(\hat{\beta}, \hat{\delta})$ and $(b, d) = \sqrt{N}(\beta, \delta)$.
4. $\hat{g}_\rho = \rho'\hat{d}$ and $g_\rho = \rho' d$.
5. $\hat{g}_\perp = U'\hat{d}$ and $g_\perp = U'd$, where U is a $K \times (K-1)$ matrix such that $N^{-1}U'U = I_K$ and $U'\rho = 0$.

Then $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ is distributed

$$\begin{pmatrix} \hat{b} \\ \hat{g}_\rho \\ \hat{g}_\perp \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ g_\rho \\ g_\perp \end{pmatrix}, \begin{pmatrix} \frac{1}{(1-\|\rho\|^2)} & -\frac{\|\rho\|^2}{(1-\|\rho\|^2)} & 0 \\ -\frac{\|\rho\|^2}{(1-\|\rho\|^2)} & \frac{\|\rho\|^2}{(1-\|\rho\|^2)} & 0 \\ 0 & 0 & I_{K-1} \end{pmatrix} \right),$$

and applying a change of variables,

$$\begin{pmatrix} \hat{b} \\ \hat{b} + \hat{g}_\rho\|\rho\|^{-2} \\ \hat{g}_\perp \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ b + g_\rho\|\rho\|^{-2} \\ g_\perp \end{pmatrix}, \begin{pmatrix} (1-\|\rho\|^2)^{-1} & 0 & 0 \\ 0 & \|\rho\|^{-2} & 0 \\ 0 & 0 & I_{K-1} \end{pmatrix} \right),$$

The likelihood of the observables is

$$L(\hat{b}, \hat{g}_\rho, \hat{g}_\perp; b, g_\rho, g_\perp) = \mathcal{C}_N \cdot \exp\left(-\frac{1}{2}\left[(\hat{b}-b)^2(1-\|\rho\|^2) + (\hat{b}+\hat{g}_\rho\|\rho\|^{-2} - b - g_\rho\|\rho\|^{-2})^2\|\rho\|^2 + \|\hat{g}_\perp\|^2\right]\right) \cdot \cdots \cdot \exp\left(\hat{g}_\perp' g_\perp - \frac{1}{2}\|g_\perp\|^2\right),$$

where $\mathcal{C}_N = (2\pi)^{-(K+1)/2}(\rho\sqrt{1-\|\rho\|^2})^{-1}$ is the constant of integration.

A.1 Optimality, Weighted Average Power, and the Least-Favorable Distribution

When testing a point null against a point alternative, the optimal procedure is given by the Neyman-Pearson likelihood ratio test. In our model, the nuisance parameter d - equivalently, $(g_\rho, g_\perp)$ - takes on a continuum of values, so direct application of the likelihood ratio test would require fixing $d = d_0$ under the null and $d = d_1$ under the alternative. This is problematic, however, because the subsequent procedure is not guaranteed to be level-α valid over the entire nuisance parameter set $\mathcal{D}_0$, and it potentially sacrifices power against points $d \neq d_1$ under the alternative.

In place of a simple likelihood ratio test, we apply the weighted average power criterion and the least-favorable distribution to conduct inference in the presence of a nuisance parameter.

A.1.1 Weighted Average Power Criterion

Consider the general hypothesis test of an unknown parameter μ belonging to a composite null $\mathcal{M}_0 \subseteq \mathbb{R}^n$ and alternative $\mathcal{M}_1 \subseteq \mathbb{R}^n$

$$H_0 : \mu \in \mathcal{M}_0 \quad vs. \quad H_1 : \mu \in \mathcal{M}_1, \tag{A.1}$$

where the observable $\mathcal{X}$ has density $f(\cdot;\mu)$ for given $\mu \in \mathbb{R}^n$. To address the complexities introduced by a composite alternative, we integrate over μ as follows.

Definition A.1. A *weighted average power criterion* (weighted alternative) for (A.1) is a distribution Λ_1 in μ with support in the closure of $\mathcal{M}_1$. We use the shorthand $H_0 : \mu \in \mathcal{M}_0$ vs. $H_{1,w} : \mu \sim \Lambda_1$ to denote the test:

$$H_0 : \mu \in \mathcal{M}_0 \quad vs. \quad H_{1,w} : \mathcal{X} \text{ has density } f_{\Lambda_1}(\cdot) = \int f(\cdot;\mu) d\Lambda_1(\mu), \tag{A.2}$$

A weighted average power criterion Λ_1 is simply a mixture of the unknown parameter $\mu \in \mathcal{M}_1$ under the alternative. Testing against the criterion - as in (A.2) - ensures that rather than focusing power against a single point $\mu = \mu_1 \in \mathcal{M}_1$, we hedge against all possibilities $\mu \in \mathcal{M}_1$ with importance governed by Λ_1. To gauge the degree to which a test $\varphi(\mathcal{X})$ meets this criterion, we evaluate its *weighted average power*:

Definition A.2. The Λ_1-*weighted average power* of a test $\varphi(\mathcal{X})$, denoted $\pi(\varphi; \Lambda_1)$, is the rejection probability of φ under $H_{1,w}$. Formally, $\pi(\varphi; \Lambda_1) = \int \left(\int \varphi f(x; \mu) d\Lambda_1(\mu) \right) dx$.

Observe that changing the order of integration yields $\pi(\varphi; \Lambda_1) = \int \left(\int \varphi f(x; \mu) dx \right) d\Lambda_1(\mu)$, where $\pi_\mu(\varphi) = \int \varphi f(x; \mu) dx$ is the power of φ against a single point $\mu \in \mathcal{M}_1$. Hence, $\pi(\varphi; \Lambda_1)$ is also the Λ_1-weighted average of all power against individual alternatives $\mu \in \mathcal{M}_1$. In this sense, maximization of weighted average power offers an expansive notion of optimality.

A.1.2 Least-Favorable Distribution

Given a weighted alternative Λ_1, consider the problem

$$H_0 : \mu \in \mathcal{M}_0 \quad vs. \quad H_{1,w} : \mu \sim \Lambda_1, \tag{A.3}$$

where, as in Definition A.1, the observable $\mathcal{X}$ has density $f(\cdot; \mu)$, $\mu \in \mathcal{M}_0$ under H_0, and $\mathcal{X}$ has density $f_{\Lambda_1}(\cdot) = \int f(\cdot; \mu) d_{\Lambda_1}(\mu)$ under $H_{1,w}$. Suppose we seek to construct a level-α valid test $\varphi(\mathcal{X})$ admitting the greatest weighted average power $\pi(\varphi; \Lambda_1)$.

Consider weighting $\mu \in \mathcal{M}_0$ under the null as follows: for a distribution Λ_0 on the closure of $\mathcal{M}_0$, we test

$$\begin{aligned} H_{0,w} : \mathcal{X} \text{ has density } f_{\Lambda_0}(\cdot) &= \int f(\cdot; \mu) d\Lambda_0(\mu) \\ vs. \quad H_{1,w} : \mathcal{X} \text{ has density } f_{\Lambda_1}(\cdot) &= \int f(\cdot; \mu) d\Lambda_1(\mu), \end{aligned} \tag{A.4}$$

Observe that the optimal test in (A.4) - equivalently, the test which maximizes weighted average power $\pi(\varphi, \Lambda_1)$ in (A.4) - is given by the Neyman-Pearson likelihood ratio statistic:

$$\varphi_{LR}(\cdot; \Lambda_1, \Lambda_0) = 1\{LR(\cdot) > cv_\alpha\} + \kappa 1\{LR(\cdot) = cv_\alpha\}, \quad LR(\mathcal{X}) = f_{\Lambda_1}(\mathcal{X}) / f_{\Lambda_0}(\mathcal{X}),$$

where $cv_{alpha} \in \mathbb{R}_+$ and $\kappa \in [0, 1]$ are chosen so that φ_{LR} has size $\int \left(\int \varphi_{LR}(x; \Lambda_1, \Lambda_0) \cdot f(x, \mu) d\Lambda_0(\mu) \right) dx =$

α under $H_{0,w}$.

Importantly, observe that any test φ which controls size over all $\mu \in \mathcal{M}_1$ also controls size under $H_{0,w}$ in (A.4). By the construction of φ_{LR}, this implies $\pi(\varphi, \Lambda_1) \leq \pi(\varphi_{LR}, \Lambda_1)$ for such φ, so the weighted average power of φ_{LR} is actually an upper bound on the weighted average power of any valid test of (A.3). When $\varphi_{LR}(\cdot; \Lambda_1, \Lambda_0)$ is, itself, level-α valid for (A.3), we define Λ_0 to be a least-favorable distribution.

Definition A.3. A *Least-Favorable Distribution* (LFD) for (A.3) is a distribution Λ_0^{LFD} over $\mathcal{M}_0$ such that $\varphi_{LR}(\cdot; \Lambda_1, \Lambda_0^{LFD})$ is level-α valid for (A.3). In such instances, let $\varphi_{LFD}(\cdot; \Lambda_1)$ denote $\varphi_{LR}(\cdot; \Lambda_1, \Lambda_0^{LFD})$.

The significance of a least-favorable distribution Λ_0^{LFD} is that $\varphi_{LFD} = \varphi_{LR}(\cdot; \Lambda_1, \Lambda_0^{LFD})$ is the valid test of (A.3) admitting the greatest Λ_1-weighted average power. In this sense, φ_{LFD} is the optimal valid test of (A.3). Mild regularity conditions on $\mathcal{M}_0$, $\mathcal{M}_1$, and $f(\cdot, \mu)$ guarantee the existence of a least-favorable distribution Λ_0^{LFD}. In special cases, such as Example A.4 below, Λ_0^{LFD} has an explicit form.

Example A.4. Consider testing

$$H_0 : \mu \in \mathcal{M}_0 \quad vs. \quad H_{1,w} : \mu = \mu_1 \tag{A.5}$$

based on the observable $\mathcal{X} \sim N(\mu, I_n)$. If $\mathcal{M}_0$ is a convex set, then the least-favorable distribution Λ_0 for (A.5) is the point mass on $\mu_0^* = \arg\min_{\mu_0 \in \mathcal{M}_0} \|\mu_1 - \mu_0\|$, the projection of μ_1 on $\mathcal{M}_0$. The optimal test φ_{LFD} maximizes power against $\mu = \mu_1$ and rejects for large values of $(\mathcal{X} - \mu_0^*)'(\mu_1 - \mu_0^*)$, the projection of $\mathcal{X} - \mu_0^*$ on $\mu_1 - \mu_0^*$.

In general, the least-favorable distribution is not necessarily a simple point mass as in Example A.4, and its form is dependent on Λ_1, the weighted average power criterion. To compute the least-favorable distribution for general $\mathcal{M}_0$, $\mathcal{M}_1$, and Λ_1, we use the ALFD algorithm of Elliott et al. (2015).

A.1.3 Application to Full Gaussian Baseline Model

We are interested in testing:

$$H_0 : b = 0,\, (g_\rho, g_\perp) \in \mathcal{D}_0^g \quad vs. \quad H_1 : b \neq 0,\, (g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1} \tag{A.6}$$

where $\mathcal{D}_0^g = \{(g_\rho, g_\perp) \in \mathbb{R} \times \mathbb{R}^{K-1} :\, g_\rho^2 \|\rho\|^{-2} + \|g_\perp\|^2 \leq r_d^2\}$, and the observable is

$$\begin{pmatrix} \hat{b} \\ \hat{g}_\rho \\ \hat{g}_\perp \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ g_\rho \\ g_\perp \end{pmatrix}, \begin{pmatrix} \frac{1}{(1-\|\rho\|^2)} & -\frac{\|\rho\|^2}{(1-\|\rho\|^2)} & 0 \\ -\frac{\|\rho\|^2}{(1-\|\rho\|^2)} & \frac{\|\rho\|^2}{(1-\|\rho\|^2)} & 0 \\ 0 & 0 & I_{K-1} \end{pmatrix} \right)$$

In terms of notation from Section A.1.1, $(b, g_\rho, g_\perp)$ corresponds to the unknown parameter μ, and $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ corresponds to the observable $\mathcal{X}$. The domain of $(b, g_\rho, g_\perp)$ under the null is $\{0\} \times \mathcal{D}_0^g$, which corresponds to $\mathcal{M}_0$, and the domain of $(b, g_\rho, g_\perp)$ under the alternative is $\mathbb{R}^{K+1}$, which corresponds to $\mathcal{M}_1$. Consider a weighted average power criterion Λ_1 in this setting which has compact support over $(b, g_\rho, g_\perp) \in \mathbb{R}^{K+1}$. Because $\mathcal{M}_0$ is compact, and Gaussian distributions are well-behaved, an LFD exists and has support over $(b, g_\rho, g_\perp) \in \mathcal{M}_0$.

Lemma A.5. *Suppose $r_d = \infty$, so that $\mathcal{D}_0^g = \mathbb{R}^K$. The uniformly-most-powerful one-sided test of (A.6) is based on the long regression estimate $\hat{b}$. The uniformly-most-powerful two-sided test of (A.6) which is unbiased is based on the long regression estimate $\hat{b}$.*

Proof. Because $\hat{g}_\perp$ is independent of $(\hat{b}, \hat{g}_\rho)$, and the nuisance parameters $(g_\rho, g_\perp)$ are unconstrained, it is enough to consider tests which are functions of $(\hat{b}, \hat{g}_\rho)$ alone, which is a sufficient statistic. The problem reduces to testing

$$H_0 : b = 0,\, g_\rho \in \mathbb{R} \quad vs. \quad H_1 : b \neq 0,\, g_\rho \in \mathbb{R}$$

with the observable $(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma^{hom})$. The remainder follows from Proposition 15.2 and Problem 1, pg. 226 of Van der Vaart (1998). □

A.2 Bivariate Tests

Consider the bivariate testing problem

$$H_0 : b = 0,\ g_\rho \in \mathcal{D}_0^\rho \quad vs. \quad H_1 : b \neq 0,\ g_\rho \in \mathbb{R} \tag{A.7}$$

where $(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma)$ is the observable, $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ is a general covariance matrix, and $\mathcal{D}_0^\rho = \{g_\rho : g_\rho^2 \|\rho\|^{-2} \leq r_d^2\} = [-r_d\|\rho\|, r_d\|\rho\|]$. To test (A.7), we consider

$$H_0 : b = 0,\ g_\rho \in \mathcal{D}_0^\rho \quad vs. \quad H_1 : (b, g_\rho) \sim \Lambda_1, \tag{A.8}$$

where Λ_1 is distributed such that: *(i)* the marginal of Λ_1 in b places mass .5 on each $b \in \{-2\sigma_1^{-1}, 2\sigma_1^{-1}\}$, and *(ii)* conditional on b, Λ_1 is uniform in g_ρ on $\mathcal{D}_0^\rho$.

Definition A.6. Let $\varphi^{biv}(\hat{b}, \hat{g}_\rho; \Sigma, r_d^2)$ be the test based on the LFD in the problem (A.8).

Applying the ALFD algorithm, we calculate the full optimal test for (A.8) to construct $\varphi^{biv}(\hat{b}, \hat{g}_\rho; \Sigma, r_d^2)$. Below, we let $\Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$ and plot rejection profiles for φ^{biv} for several specifications of r_d and σ_{12}. The rejection regions are indicated by the grey areas, with $\hat{\beta}$ on the vertical axis and $\hat{g}_\rho$ on the horizontal.

Figure A.1: Rejection Region for $\hat{\varphi}^{biv}$ ($r_d = 1$ and $\sigma_{12} = -.8$)

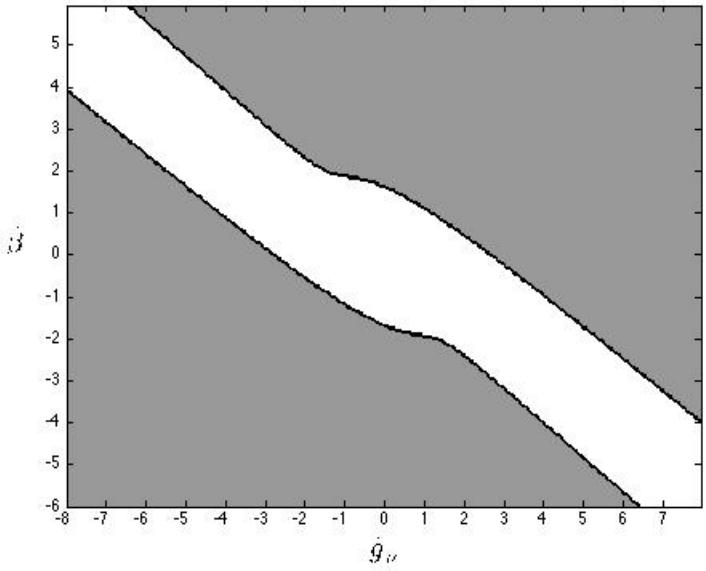


Figure A.2: Rejection Region for $\hat{\varphi}^{biv}$ ($r_d = 4$ and $\sigma_{12} = -.8$)

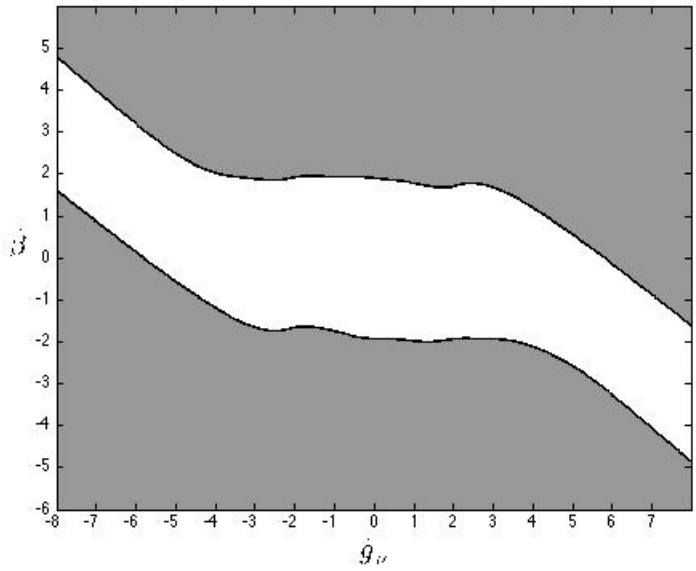


Figure A.3: Rejection Region for $\hat{\varphi}^{biv}$ ($r_d = 4$ and $\sigma_{12} = -.5$)

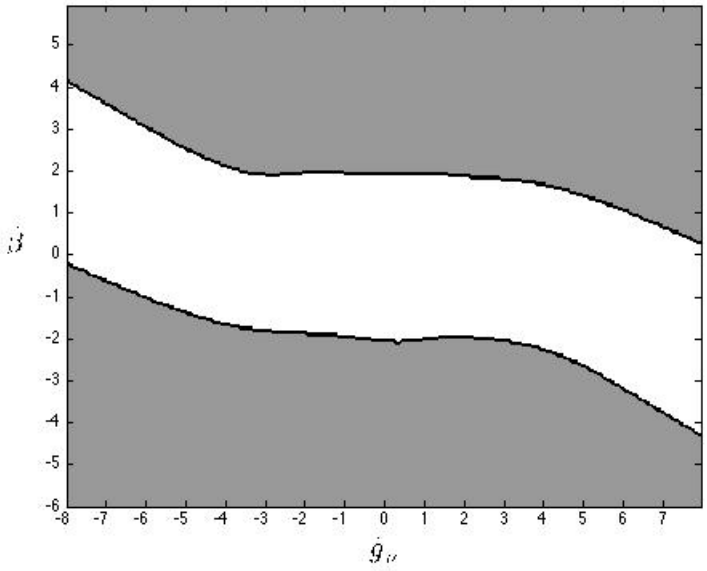
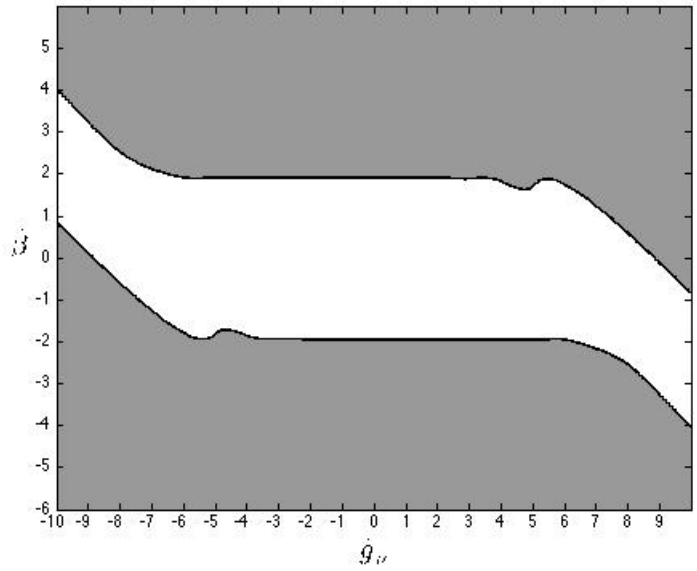


Figure A.4: Rejection Region for $\hat{\varphi}^{biv}$ ($r_d = 7$ and $\sigma_{12} = -.8$)



As we can see from the above figures, φ^{biv} reduces to the standard t-test which rejects when $|\hat{b}| \geq 1.96$ when $\hat{g}_\rho$ lies within the interior of $\mathcal{D}_0^g$. The intuition for φ^{biv} is given by the following, which considers a homoskedastic model where $\Sigma = \Sigma^{hom}$:

1. The short regression estimate $\hat{b}^{short}$ can be re-written $\hat{b}^{short} = \hat{b} + \hat{g}_\rho$, where

$$\begin{pmatrix} \hat{b}^{short} \\ \hat{g}_\rho \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b + g_\rho \\ g_\rho \end{pmatrix}, \begin{pmatrix} \|\rho\|^{-2} & 0 \\ 0 & \frac{\|\rho\|^2}{(1-\|\rho\|^2)} \end{pmatrix} \right)$$

   Observe that g_ρ is the omitted variable bias of $\hat{b}^{short}$, and $\hat{g}_\rho$ is unbiased for g_ρ.

2. When one observes $\hat{g}_\rho$ within the interior of $\mathcal{D}_0^\rho$, the nuisance parameter restriction provides no additional information, so the best estimate of bias is given by $\hat{g}_\rho$. To "bias-correct" $\hat{b}^{short}$, one simply subtracts $\hat{b}^{short} - \hat{g}_\rho$, since the two components $\hat{b}^{short}$ and $\hat{g}_\rho$ are independent. This yields $\hat{b}$, the long-regression estimate.

3. When one observes $\hat{g}_\rho$ which does not lie in $\mathcal{D}_0^\rho$, the nuisance parameter restriction binds, and the best estimate of bias is given by $sign\left(\hat{g}_\rho\right) \cdot \max(r_d\|\rho\|, \|\hat{g}_\rho\|)$. To "bias-correct" $\hat{b}^{short}$, one subtracts $\hat{b}^{short} - sign\left(\hat{g}_\rho\right) \cdot \min(r_d\|\rho\|, \|\hat{g}_\rho\|) = \hat{b}^{short} - sign\left(\hat{g}_\rho\right) r_d\|\rho\|$.

4. Putting together steps 2 and 3, the overall procedure essentially checks whether the long regression coefficient $\hat{b}$ lies within the range of the bias-corrected short regression coefficient: $\hat{b}^{short} \pm r_d\|\rho\|$. If so, then the bound provides no additional information, and the long regression estimate $\hat{b}$ is the best statistic for inference. If not, then the point in the range $\hat{b}^{short} \pm r_d\|\rho\|$ which lies closest to the long regression estimate is a corrected

A.3 Asymptotics and the Limit of Experiments

A.3.1 Mathematical Lemmas

Definition A.7.

1. Let $\mathcal{I}_n(\cdot)$ denote the Type-1 modified Bessel function of order $n > 0$.

2. For $r \in \mathbb{R}_+$ and $n \in \mathbb{N}$, let $\mathcal{S}_n(r) = \{v \in \mathbb{R}^n : \|v\| = r\}$ denote the sphere of radius $r > 0$.

Lemma A.8. (Bessel Approximation) *For $n \in \mathbb{N}$ and $y \in \mathbb{R}$, $\mathcal{I}_n(ny) = \frac{y^n \cdot \exp\left(n(\sqrt{1+y^2} - \log(1+\sqrt{1+y^2}))\right)}{\sqrt{2\pi n} \cdot (1+y^2)^{1/4}} \cdot (1+\eta)$ where $|\eta| < (8n)^{-1}$*

Proof. By a result in Olver (1997),

$$\mathcal{I}_n(ny) \quad = \quad \frac{y^n \cdot \exp\left(n(\sqrt{1+y^2} - \log(1+\sqrt{1+y^2}))\right)}{\sqrt{2\pi n} \cdot (1+y^2)^{1/4}} \cdot (1+\eta(n,y))$$

where $|\eta(n,y)| \leq 2\exp(\frac{2V_1(y)}{n}) \cdot \frac{V_1(y)}{n}$, and$V_1(y)$ is the total variation of the function $f(p) = (3p - 5p^3)/24$, $f : \mathbb{C} \to \mathbb{C}$, from $p = 1$ to $p = (1+y^2)^{-1/2}$ in the complex plane over piecewise-differentiable paths $\mathcal{P} : [0,1] \to \mathbb{C}$ such that $\text{Re}(\xi(\mathcal{P}(\cdot))) : [0,1] \to \mathbb{R} \cup \{-\infty, \infty\}$ is monotonic, with $\xi(p) = p^{-1} + \ln(p^{-2}-1)/2 - \ln(1+p^{-1})$. Note that $\partial_p \xi = (p^2(p^2-1))^{-1} < 0$ for $p \in (0,1)$.

Hence, $\text{Re}(\xi(\mathcal{P}))$ is monotonic if $\mathcal{P}(s) = 1 + s((1+y^2)^{-1/2} - 1)$, so without loss of generality $V_1(y)$ is the total variation of f on the real interval $[(1+y^2)^{-1/2}, 1]$. Then $(1+y^2)^{-1/2} \in (0,1]$ for all $y \in \mathbb{R}$, so that $|V_1(y)| \leq 1/8$, and $|\eta(n,y)| \leq (8n)^{-1}$ □

Lemma A.9. *Suppose $\alpha < 1/2$, $\gamma_n \leq n^\alpha$, and $u, \tilde{u} \sim_{iid} \text{Unif}(S_n(1))$. Then $E[\exp(\gamma_n u'\tilde{u})] \to 1$ as $n \to \infty$*

Proof. Note that for any $\gamma > 0$, $E[\exp(\gamma u'\tilde{u})] = \frac{2^{n/2-1}\mathcal{I}_{n/2-1}(\gamma)\Gamma(n/2)}{\gamma^{n/2-1}}$, and $\partial_\gamma \left(\frac{\mathcal{I}_{n/2-1}(\gamma)}{\gamma^{n/2-1}}\right) = \frac{\mathcal{I}_{n/2}(\gamma)}{\gamma^{n/2-1}} > 0$, so $\partial_\gamma E[\exp(\gamma u'\tilde{u})] > 0$. It remains to show $E[\exp(n^\alpha u'\tilde{u})] \to 1$.

Using the Bessel Approximation Lemma and Stirling's Approximation for $\Gamma(\cdot)$,

$$\begin{aligned}
& \frac{2^{n/2-1}\mathcal{I}_{n/2-1}(n^\alpha)\Gamma(n/2)}{n^{\alpha(n/2-1)}} \\
= \quad & \frac{(n^\alpha(n/2-1)^{-1})^{n/2-1} \cdot \Gamma(n/2) \cdot 2^{n/2-1}}{\sqrt{2\pi(n/2-1)} \cdot (1+(n^\alpha(n/2-1)^{-1})^2)^{1/4} \cdot n^{\alpha(n/2-1)}} \cdot (1+o(1)) \cdot \\
& \quad ... \cdot \exp\left((n/2-1) \cdot (\sqrt{1+(n^\alpha(n/2-1)^{-1})^2} - \log(1+\sqrt{1+(n^\alpha(n/2-1)^{-1})^2}))\right) \\
= \quad & ((n/2-1)^{-1})^{n/2-1} \cdot \exp\left((n/2-1) \cdot (1 - \log(2) + O(n^{2(\alpha-1)}))\right) \cdot \\
& \quad ... \cdot ((n/2-1)e^{-1})^{n/2-1} \cdot 2^{n/2-1} \cdot (1+o(1)) \\
\to \quad & 1
\end{aligned}$$

□

A.3.2 Invariance

In this section, we show that imposing the invariance of $\hat{g}_\perp$ to the group of orthogonal transformations $\{O : \mathbb{R}^{K-1} \to \mathbb{R}^{K-1} : \|Ov\| = \|v\| \text{ for all } v \in \mathbb{R}^{K-1}\}$ is, without loss of generality, the optimal procedure when maximizing weighted average power against the class of weightings over $(b, g_\rho, g_\perp)$

which are rotationally-symmetric in $g_\perp$.

Definition A.10. Let $r_\perp = \|g_\perp\|$ and $\hat{r}_\perp = \|\hat{g}_\perp\|$.

First, let the scalar $r_\perp \in \mathbb{R}_+$ replace $g_\perp \in \mathbb{R}^{K-1}$ as a free nuisance parameter. Consider drawing $g_\perp$ uniformly from $\mathcal{S}_{K-1}(r_\perp) = \{v \in \mathbb{R}^{K-1} : \|v\|^2 = r_\perp^2\}$, the sphere of radius $r_\perp \geq 0$. It follows from Lemma A.11 below that the observables $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ are then distributed:

$$(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma_{biv}), \quad \hat{g}_\perp | g_\perp \sim \mathcal{N}(g_\perp, I_{K-1}), \quad g_\perp \sim \Psi(\cdot; \mathcal{S}_{K-1}(r_\perp)) \tag{A.9}$$

where $(\hat{g}_\perp, g_\perp) \perp (\hat{b}, \hat{g}_\rho)$, and $\Psi(\cdot, \mathcal{S})$ denotes the uniform distribution over $\mathcal{S}$. Observe that under $g_\perp \sim \Psi(\cdot; \mathcal{S}_{K-1}(r_\perp))$, the parameters $(b, g_\rho, r_\perp)$ alone identify the distribution of $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$, whose likelihood is a function of the three scalars $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$ alone.

Lemma A.11. *The density of $(\hat{b}, \hat{g}_\rho, \hat{g}_\perp)$ in (A.9) can be written*

$$f(\hat{b}, \hat{g}_\rho, \hat{g}_\perp; b, g_\rho, r_\perp) \propto \exp[-1/2 \cdot ((\hat{b}, \hat{g}_\rho)\Sigma_{biv}^{-1}(\hat{b}, \hat{g}_\rho)' + \hat{r}_\perp^2 + r_\perp^2)] \cdot \mathcal{I}_{(K-1)/2}(\hat{r}_\perp r_\perp)$$

In particular, f is a function of the observable $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$, whose distribution under $(b, g_\rho, r_\perp)$ is given by:

$$(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma^{hom}), \quad \hat{r}_\perp^2 \sim \chi^2_{NC}(r_\perp^2, K-1), \quad \hat{r}_\perp \perp (\hat{b}, \hat{g}_\rho)$$

Proof. It is clear that conditional on $g_\perp$ drawn from $\mathcal{S}_{K-1}(r_\perp)$, the distribution of $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$ under $(b, g_\rho, r_\perp)$ is given by:

$$(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma^{hom}), \quad \hat{r}_\perp^2 \sim \chi^2_{NC}(r_\perp^2, K-1), \quad \hat{r}_\perp \perp (\hat{b}, \hat{g}_\rho)$$

so this holds unconditionally when $g_\perp \sim \Psi(\cdot, \mathcal{S}_{K-1}(r_\perp))$. The likelihood can be written:

$$\begin{aligned} f(\hat{b}, \hat{g}_\rho, \hat{g}_\perp; b, g_\rho, r_\perp) &= \int L(\hat{b}, \hat{g}_\rho, \hat{g}_\perp; b, g_\rho, g_\perp)\, d\Psi(g_\perp, \mathcal{S}_{K-1}(r_\perp)) \\ &\propto \exp[-1/2 \cdot ((\hat{b}, \hat{g}_\rho)\Sigma_{biv}^{-1}(\hat{b}, \hat{g}_\rho)' + \hat{r}_\perp^2 + r_\perp^2)] \cdot \int \exp(\hat{g}_\perp' g_\perp)\, d\Psi(g_\perp, \mathcal{S}_{K-1}(r_\perp)) \\ &= \exp[-1/2 \cdot ((\hat{b}, \hat{g}_\rho)\Sigma_{biv}^{-1}(\hat{b}, \hat{g}_\rho)' + \hat{r}_\perp^2 + r_\perp^2)] \cdot \mathcal{I}_{(K-1)/2}(\hat{r}_\perp r_\perp) \end{aligned}$$

□

Theorem A.12. *For fixed $r > 0$ and compactly-supported distribution Λ_1 over $\mathbb{R}^2 \times \mathbb{R}_+$, suppose the observable is*

$$(\hat{b}, \hat{g}_\rho)' \sim \mathcal{N}((b, g_\rho)', \Sigma_{biv}), \quad \hat{g}_\perp | g_\perp \sim \mathcal{N}(g_\perp, I_{K-1})$$

and consider testing

$$H_0 : b = 0,\ (g_\rho, g_\perp) \in \mathcal{D}_0^g \quad vs. \quad H_1 : (b, g_\rho, r_\perp) \sim \Lambda_1$$

where $\mathcal{D}_0^g = \{(g_\rho, g_\perp) \in \mathbb{R}^K : g_\rho^2 \|\rho\|^{-2} + \|g_\perp\|^2 \leq r^2\}$ under the null, and $g_\perp \sim \Psi(\cdot; \mathcal{S}_{K-1}(r_\perp))$ and $(\hat{g}_\perp, g_\perp) \perp (\hat{b}, \hat{g}_\rho)$ under the alternative. Then the test which maximizes WAP in this problem is a function of $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$.

Proof. Consider searching for the LFD. Suppose that under H_0, we substitute $r_\perp$ for $g_\perp$ as a free parameter, set $g_\perp | r_\perp \sim \Psi(\mathcal{S}_{K-1}(r_\perp))$, and artificially restrict the search to distributions Λ_0 governed by mixtures over $(g_\rho, r_\perp) \in \mathcal{D}_0^{g,r} = \left\{(g_\rho, r_\perp) \in \mathbb{R}^2 : g_\rho^2 \|\rho\|^{-2} + r_\perp^2 \leq r^2\right\}$. Then Λ_0 induces a test which is a function of $(\hat{b}, \hat{g}_\rho, \hat{r}_\perp)$, whose distribution is pinned down by the triplet $(b, g_\rho, r_\perp)$. Because the test is valid over the restricted nuisance parameter set $\mathcal{D}_0^{g,r}$, Λ_0 is also the LFD under an unrestricted search. □

A.3.3 Limit of Experiments

In this section we prove the limit of experiments results within Section 3.3. Ultimately, Lemma A.16 of this section directly implies Theorem 3.11 of Section 3.3.1 and Proposition 3.16 of Section 3.3.2. Henceforth, all observables $(\hat{b}_N, \hat{g}_{N\rho}, \hat{g}_{N\perp}, \hat{\gamma}_N, \hat{r}_{N\perp}, \rho_N)$ depend on N.

Assumption A.13. *$K/N \to c_K \in (0, 1)$ as $N \to \infty$.*

Recall that $g_{N\rho}^*$ is a baseline sequence for g_ρ, and denote by $(\hat{\gamma}_N, \gamma) = (\hat{g}_{N\rho} - g_{N\rho}^*,\ g_\rho - g_{N\rho}^*)$ the deviation of $(\hat{g}_{N\rho}, g_\rho)$ from $g_{N\rho}^*$.

Definition A.14. 1. For parameters $(b, \gamma, r_\perp) \in \mathbb{R}^2 \times \mathbb{R}_+$, suppose $\hat{\theta}_N = (\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp})'$ is distributed:

$$(\hat{b}_N, \hat{\gamma}_N)' \sim \mathcal{N}((b, \gamma)', \Sigma_N^{hom}), \quad \hat{r}_{N\perp}^2 \sim \chi_{NC}^2(r_\perp^2, K-1), \quad (\hat{b}_N, \hat{\gamma}_N) \perp \hat{r}_{N\perp}. \tag{A.10}$$

2. For parameters $(b, \gamma) \in \mathbb{R}^2$, suppose $A = (A_b, A_\gamma)'$ is distributed $(A_b, A_\gamma)' \sim \mathcal{N}((b, \gamma)', \Sigma_A^{biv})$.

Definition A.15 (Hyperparameters for Lemma A.16)**.**

1. Let $\mathcal{C}_{b,\gamma} \subset \mathbb{R}^2$ be a fixed, compact set, and let $\mathcal{C}_s \in \mathbb{N}$, neither of which is dependent on N. For each $N \in \mathbb{N}$ and $s \in \{1, \ldots, \mathcal{C}_s\}$, let Λ_{Ns} be a distribution over $(b, \gamma, r_\perp) \in \mathbb{R}^2 \times \mathbb{R}_+$ with support in $\mathcal{C}_{b,\gamma} \times [0, \bar{r}_N]$, and suppose the marginal of Λ_{Ns} in (b, γ) — denoted by $\Lambda_s^{b,\gamma}$ — does not depend on N.

2. Let $\mathcal{H}_{lem} = \mathbb{R}^2 \times \{0, 1, \ldots, c\}$ be the hyperparameter space for Lemma A.16.

3. Suppose $h = (b, \gamma, s) \in \mathcal{H}_{lem}$. If $s = 0$, let $\hat{\theta}_N | h = (\hat{b}_N, \hat{\gamma}_N, \hat{r}_{N\perp}) | h$ be distributed according to Definition A.14 with $(b, \gamma, r_\perp) = h$. If $s > 0$, let $\hat{\theta}_N$ be distributed according to Definition A.14 with $(b, \gamma, r_\perp) \sim \Lambda_{Ns}$. Let $P_{N,h}$ denote the distribution of $\hat{\theta}_N | h$.

4. Suppose $h = (b, \gamma, s) \in \mathcal{H}_{lem}$. If $s = 0$, let $A|h = (A_b, A_\gamma)|h$ be distributed according to Definition A.14, and if $s > 0$, let $(b, \gamma) \sim \Lambda_s^{b,\gamma}$. Let P_h denote the distribution of $A|h$.

Lemma A.16 (Limit of Experiments, General Framework)**.** *For every finite subset $I \subset \mathcal{H}_{lem}$ and fixed $h_0 \in \mathcal{H}_{lem}$,*

$$\left\{ \frac{dP_{N,h}}{dP_{N,h_0}}(\hat{\theta}_N) \right\}_{h \in I} \to_d \left\{ \frac{dP_h}{dP_{h_0}}(A) \right\}_{h \in I}$$

under $\hat{\theta}_N \sim P_{N,h_0}$.

Proof. For any two $h, h' \in \mathcal{H}_{lem}$, the ratio of densities $\frac{dP_{N,h'}}{dP_{N,h}}(\hat{\theta}_N)$ can be re-written $\frac{dP_{N,h'}}{dP_{N,h}} = \frac{dP_{N,h'}}{dP_{N,h_0}} \cdot \frac{dP_{N,h_0}}{dP_{N,h}}$. Hence, it suffices for the theorem to show that the following joint convergence holds

$$\left\{ \frac{dP_{N,h}}{dP_{N,h_0}}(\hat{\theta}_N) \right\}_{h \in I} \to_d \left\{ \frac{dP_h}{dP_{h_0}}(A) \right\}_{h \in I}$$

when $h_0 \in \mathcal{H}_{lem}$ is fixed at $h_0 = 0$.

To that end, fix arbitrary $h = (b, \gamma, s) \in \mathcal{H}_{lem}$ such that $s = 1$. We show that $\frac{dP_{N,h}}{dP_{N,h_0}}(\hat{\theta}_N) \to_d \frac{dP_h}{dP_{h_0}}(A)$, where $h_0 = 0$ is the true data-generating hyperparameter underlying $\hat{\theta}_N$ and A. Begin by

writing

$$\begin{aligned}\frac{dP_{N,h}}{dP_{N,h_0}}(\hat{\theta}_N) &= \frac{\int f(\hat{\theta}_N; b, \gamma, r_\perp)\, d\Lambda_{N1}(b, \gamma, r_\perp)}{f(\hat{\theta}_N; 0)} \\ &= \int \frac{f(\hat{\theta}_N; b, \gamma, r_\perp)}{f(\hat{\theta}_N; 0)}\, d\Lambda_{N1}(b, \gamma, r_\perp) \\ &= \int \frac{f(\hat{\theta}_N; b, \gamma, 0)}{f(\hat{\theta}_N; 0)}\, d\Lambda_{N1}^{b,\gamma} + \int \frac{f(\hat{\theta}_N; b, \gamma, r_\perp) - f(\hat{\theta}_N; b, \gamma, 0)}{f(\hat{\theta}_N; 0)}\, d\Lambda_{N1}(b, \gamma, r_\perp)\end{aligned}$$

Because $\int \frac{f(\hat{\theta}_N;b,\gamma,0)}{f(\hat{\theta}_N;0)}\, d\Lambda_{N1}^{b,\gamma} \to_d \frac{dP_h}{dP_{h_0}}(A)$ (a convergence which continues to hold jointly over finitely-many $h \in \mathcal{H}_{lem}$) it suffices to show that $\int \frac{f(\hat{\theta}_N;b,\gamma,r_\perp)-f(\hat{\theta}_N;b,\gamma,0)}{f(\hat{\theta}_N;0)}\, d\Lambda_{N1}^{b,\gamma} \to_d 0$.

Proceed by writing

$$\begin{aligned}&\left| \int \frac{f(\hat{\theta}_N; b, \gamma, r_\perp) - f(\hat{\theta}_N; b, \gamma, 0)}{f(\hat{\theta}_N; 0)}\, d\Lambda_{N1}(b, \gamma, r_\perp) \right| \\ = \; & E_{b,\gamma,r_\perp}\left[\frac{\exp\left(-\frac{1}{2}\left[(\hat{b}_N - b)^2(1 - \|\rho_N\|^2)^{-1} + (\hat{b}_N - b + \hat{\gamma}_N\|\rho_N\|^{-2} - \gamma\|\rho_N\|^{-2})^2\|\rho_N\|^{-2}\right]\right)}{\exp\left(-\frac{1}{2}\left[\hat{b}_N^2(1 - \|\rho_N\|^2)^{-1} + (\hat{b}_N + \hat{\gamma}_N\|\rho_N\|^{-2})^2\|\rho_N\|^{-2}\right]\right)} \right. \\ & \qquad \left. \ldots \cdot \left| \frac{\psi(\hat{r}_{N\perp}; r_\perp)}{\psi(\hat{r}_{N\perp}; 0)} - 1 \right| \right] \\ \leq \; & E_{b,\gamma,r_\perp}\left[\frac{\exp\left(-\left[(\hat{b}_N - b)^2(1 - \|\rho_N\|^2)^{-1} + (\hat{b}_N - b + \hat{\gamma}_N\|\rho_N\|^{-2} - \gamma\|\rho_N\|^{-2})^2\|\rho_N\|^{-2}\right]\right)}{\exp\left(-\left[\hat{b}_N^2(1 - \|\rho_N\|^2)^{-1} + (\hat{b}_N + \hat{\gamma}_N\|\rho_N\|^{-2})^2\|\rho_N\|^{-2}\right]\right)} \right]^{1/2} \cdot \\ & \qquad \ldots \cdot E_{b,\gamma,r_\perp}\left[\left(\frac{\psi(\hat{r}_{N\perp}; r_\perp)}{\psi(\hat{r}_{N\perp}; 0)} - 1 \right)^2 \right]^{1/2}\end{aligned}$$

where *(i)* the outer expectation $E_{b,\gamma,r_\perp}$ is taken with respect to $(b, \gamma, r_\perp) \sim \Lambda_{N1}$, *(ii)* $\psi(\hat{r}_{N\perp}; r_\perp)$ is the density of $\hat{r}_{N\perp} \sim \chi^2_{NC}(r_\perp^2, K-1)$, and *(iii)* the last line follows from the Cauchy-Schwarz inequality. Because $\hat{b}_N, \hat{\gamma}_N$ are well-behaved in the limit,

$$E_{b,\gamma,r_\perp}\left[\frac{\exp\left(-\left[(\hat{b}_N - b)^2(1 - \|\rho_N\|^2)^{-1} + (\hat{b}_N - b + \hat{\gamma}_N\|\rho_N\|^{-2} - \gamma\|\rho_N\|^{-2})^2\|\rho_N\|^{-2}\right]\right)}{\exp\left(-\left[\hat{b}_N^2(1 - \|\rho_N\|^2)^{-1} + (\hat{b}_N + \hat{\gamma}_N\|\rho_N\|^{-2})^2\|\rho_N\|^{-2}\right]\right)} \right]$$

converges weakly to a well-behaved limiting distribution, so it remains to show $E_{b,\gamma,r_\perp}\left[\left|\frac{\psi(\hat{r}_{N\perp};r_\perp)}{\psi(\hat{r}_{N\perp};0)} - 1\right|\right] \to_d 0$.

It then suffices, by Markov's inequality, to show

$$E_{\hat{g}_{N\perp}}\left[E_{b,\gamma,r_\perp}\left[\left(\frac{\psi(\hat{r}_{N\perp}; r_\perp)}{\psi(\hat{r}_{N\perp}; 0)} - 1 \right)^2 \right] \right] \to_d 0 \tag{A.11}$$

where $E_{\hat{g}_{N\perp}}$ denotes the expectation over $\hat{g}_{N\perp}$. Write $\hat{r}_{N\perp} = \|\hat{g}_{N\perp}\|^2$, where $\hat{g}_{N\perp} \sim \mathcal{N}(0, I_{K-1})$, so that

$$\begin{aligned}\frac{\psi(\hat{r}_{N\perp}; r_\perp)}{\psi(\hat{r}_{N\perp}; 0)} &= \int \exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2\right) d\Psi(u, \mathcal{S}_{K-1}(1)) \\ &= E_u\left[\exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2/2\right)\right]\end{aligned}$$

where E_u denotes taking the expectation with respect to $u \sim \Psi(\cdot, \mathcal{S}_{K-1}(1))$ distributed uniformly on the unit-sphere, with $\hat{g}_{N\perp} \perp u$. Hence, it suffices to show

$$E_{\hat{g}_{N\perp}}\left[E_{b,\gamma,r_\perp}\left[\left(E_u\left[\exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2/2\right)\right] - 1\right)^2\right]\right] \to 0$$

Observe that

$$\begin{aligned}& E_{\hat{g}_{N\perp}}\left[E_{b,\gamma,r_\perp}\left[\left(E_u\left[\exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2/2\right)\right] - 1\right)^2\right]\right] \\ = \;& E_{b,\gamma,r_\perp}\left[E_{\hat{g}_{N\perp}}\left[E_u\left[\exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2/2\right)\right]^2 - 2E_u\left[\exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2/2\right)\right] + 1\right]\right] \\ = \;& E_{b,\gamma,r_\perp}\left[E_{\hat{g}_{N\perp}}\left[E_{u,\tilde{u}}\left[\exp\left(\hat{g}'_{N\perp} u r_\perp + \hat{g}'_{N\perp}\tilde{u} r_\perp - r_\perp^2\right)\right] - 2E_u\left[\exp\left(\hat{g}'_{N\perp} u r_\perp - r_\perp^2/2\right)\right] + 1\right]\right]\end{aligned}$$

where $\tilde{u}, u \sim_{iid} \Psi(\cdot, \mathcal{S}_{K-1}(1))$ are *iid* uniform over the sphere.

It remains to observe that:

$$\begin{aligned}& E_{\hat{g}_{N\perp}}\left[E_{u,\tilde{u}}\left[\exp\left(\hat{g}'_{N\perp} u r_\perp + \hat{g}'_{N\perp}\tilde{u} r_\perp - r_\perp^2\right)\right]\right] \\ = \;& E_{\hat{g}_{N\perp}}\left[E_{u,\tilde{u}}\left[\exp\left(\hat{g}'_{N\perp}(u+\tilde{u}) r_\perp - r_\perp^2\right)\right]\right] \\ = \;& E_{u,\tilde{u}}\left[E_{\hat{g}_{N\perp}}\left[\exp\left(\hat{g}'_{N\perp}(u+\tilde{u}) r_\perp - r_\perp^2\right)\right]\right] \\ = \;& E_{u,\tilde{u}}\left[E_{\hat{g}_{N\perp}}\left[\exp\left(\|u+\tilde{u}\|^2 r_\perp^2/2 - r_\perp^2\right)\right]\right] \\ = \;& E_{u,\tilde{u}}\left[\exp\left(u'\tilde{u} r_\perp^2\right)\right]\end{aligned}$$

and by Jensen's inequality, $E_{u,\tilde{u}}\left[\exp\left(u'\tilde{u} r_\perp^2\right)\right] \geq \exp\left(E_{u,\tilde{u}}\left[u'\tilde{u} r_\perp^2\right]\right) = 1$.

By Lemma A.9, $E_{u,\tilde{u}}\left[\exp\left(u'\tilde{u}r_\perp^2\right)\right] \leq E_{u,\tilde{u}}\left[\exp\left(u'\tilde{u}\bar{r}_N^2\right)\right]$ whenever $r_\perp^2 \leq \bar{r}_N^2$, and

$$\begin{aligned} & E_{b,\gamma,r_\perp}\left[E_{\hat{g}_{N\perp}}\left[E_{u,\tilde{u}}\left[\exp\left(\hat{g}'_{N\perp}ur_\perp + \hat{g}'_{N\perp}\tilde{u}r_\perp - r_\perp^2\right)\right]\right]\right] \\ = \; & E_{b,\gamma,r_\perp}\left[E_{u,\tilde{u}}\left[\exp\left(u'\tilde{u}r_\perp^2\right)\right]\right] \\ \leq \; & E_{u,\tilde{u}}\left[\exp\left(u'\tilde{u}\bar{r}_N^2\right)\right] \\ \to \; & 1 \end{aligned}$$

Hence, $\sup_{r_\perp \in [0,\bar{r}_N]} E_{\hat{g}_{N\perp}}\left[E_{u,\tilde{u}}\left[\exp\left(\hat{g}'_{N\perp}ur_\perp + \hat{g}'_{N\perp}\tilde{u}r_\perp - r_\perp^2\right)\right]\right] \to_d 1$ which implies

$$E_{b,\gamma,r_\perp}\left[E_{\hat{g}_{N\perp}}\left[E_{u,\tilde{u}}\left[\exp\left(\hat{g}'_{N\perp}ur_\perp + \hat{g}'_{N\perp}\tilde{u}r_\perp - r_\perp^2\right)\right]\right]\right] \to_d 1$$

The same reasoning shows $E_{b,\gamma,r_\perp}\left[E_{\hat{g}_{N\perp}}\left[E_u\left[\exp\left(\hat{g}'_{N\perp}ur_\perp - r_\perp^2/2\right)\right]\right]\right] \to 1$, so the expression (A.11) is proved. □

B Proofs from Part II

Unless otherwise-specified, all notation throughout Section B is adopted from Part 2. Recall that $g(r) = \exp(-\frac{r}{2})$, and $E_\mu[f(Y)]$ denotes the expectation taken under $Y \sim \mathcal{N}(\mu, I_N)$ for an arbitrary function $f : \mathbb{R}^N \to \mathbb{R}$.

B.1 Mathematical Lemmas

Lemma B.1. *For fixed $N \geq 1$, $cv \in \mathbb{R}$, and $\mu_a, \mu_1, \ldots, \mu_n \in \mathbb{R}^N$, the region*

$$\{y \in \mathbb{R}^N : \sum_{i=1}^n g(\|y - \mu_i\|^2 - \|y - \mu_a\|^2) \leq cv\}$$

is convex.

Proof. Let $h(y) = \sum_{i=1}^n g\left(\|vs + y - \mu_i\|^2 - \|vs + y - \mu_a\|^2\right)$. It is enough to show that for any $y_1, y_2 \in \mathbb{R}^N$ and $s \in (0,1)$, $h(sy_1 + (1-s)y_2) < sh(y_1) + (1-s)h(y_2)$. It suffices that for any fixed

$y, v \in \mathbb{R}^N$, the function $h(y+vs)$ is convex, shown as follows:

$$\begin{aligned}\frac{\partial}{\partial s}(h(y+vs)) &= \sum_{i=1}^{n} g\left(\|vs+y-\mu_i\|^2 - \|vs+y-\mu_a\|^2\right) \cdot \\ &\qquad (-\frac{1}{2}) \cdot \left(2s\|v\|^2 + 2v'(y-\mu_i) - 2s\|v\|^2 - 2v'(y-\mu_a)\right) \\ &= \sum_{i=1}^{n} g\left(\|vs+y-\mu_i\|^2 - \|vs+y-\mu_a\|^2\right) \cdot (v'(\mu_i - \mu_a))\end{aligned}$$

$$\Rightarrow \frac{\partial^2}{\partial s^2}(h(y+vs)) = \sum_{i=1}^{n} g\left(\|vs+y-\mu_i\|^2 - \|vs+y-\mu_a\|^2\right) \cdot (v'(\mu_i - \mu_a))^2 \geq 0$$

□

Lemma B.2. *(Haberman (1980)) Suppose that $Y \sim \mathcal{N}(\mu, \Sigma)$, $C \subset \mathbb{R}^N$ is a strictly-convex set, and $v \in \mathbb{R}^N$ is a fixed vector. Then for $s \geq 1$,*

$$E\left[(v'(Y - E[Y \mid Y \in C]))^s\right] < E\left[(v'(Y-\mu))^s\right].$$

Lemma B.3. *Suppose that the rejection region of the test $\varphi(Y)$ is convex. Then for any linear subspace $\mathcal{W} \subseteq \mathbb{R}^N$, there exists a unique local maximum of $E_\mu[\varphi(Y)]$ over $\mu \in \mathcal{W}$.*

Proof. It suffices to fix arbitrary $\mu \in \mathcal{W}$ and show that there exists a unique local maximum of $h(r) = E_{r\mu}[\varphi(Y)]$ in $r \in \mathbb{R}$. Suppose, by contradiction, that there exists $r_1 \neq r_2$ such that both are local maxima of $h(\cdot)$. By Gardner and Zvavitch (2010), a Brunn-Minkowski inequality for Gaussian distributions implies that for any $s \in (0,1)$,

$$E_{(sr_1+(1-s)r_2)\mu}[\varphi(Y)] \geq E_{r_1\mu}[\varphi(Y)]^s \cdot E_{r_2\mu}[\varphi(Y)]^{1-s} \geq \min\left\{E_{r_1\mu}[\varphi(Y)], E_{r_2\mu}[\varphi(Y)]\right\}.$$

Hence, for either $s \to 0$ or $s \to 1$, one arrives at a contradiction of local maximality. □

B.2 Proof of Theorem 10.2

Fix $N \geq 1$, $K < N$, $M \leq K$, and $a \in (0,1)$.

Definition B.4.

1. Suppose V is an $N \times K$ matrix with *(i)* $V_{i,i} = 1$ for all $i \leq k$, *(ii)* $V_{i,j}$ for all $i > k$, and *(iii)*

$V_{i,j} = a$ for $i \leq k$ and $i \neq j$.

2. For a subset $S \subset \{1, \ldots, K\}$, let V_S denote the submatrix of V with columns indexed by elements in S (in increasing order, without loss of generality), and let $\Sigma_M = \{S \subset \{1, \ldots, K\}, |S| = M\}$.

3. Let $\mu_a \in \mathbb{R}^N$ be any vector with 1's in the first K coordinates.

B.2.1 The Least-Favorable Distribution under Sparsity

Consider the testing problem

$$H_a : \mu = \mu_a \quad vs. \quad H_0 : \mu \in \{\text{span}(V_S) \,:\, S \in \Sigma_M\} \tag{B.1}$$

where the observation is $Y \sim \mathcal{N}(\mu, I_N)$. This corresponds to testing whether the mean of Y lies in the "sparse" M-dimensional column subspace of V. It suffices for Theorem 10.2 to find the least-favorable distribution in the above problem, since any set of regressors (X^{sym}, Z^{sym}) as in Assumption 10.1 can be rotated via an orthogonal transformation into (μ_a, V).

For each $S \in \Sigma_M$, let $\tilde{\mu}^S$ be the projection of μ^a on V_S. Given the structure of V, we know that there exists a scalar b (which does not depend on q) such that $\{\tilde{\mu}^S\}_{S \in \Sigma_M} = \{b\mu^S \in \mathbb{R}^N \,:\, S \in \Sigma_M\}$, where

$$\mu_i^S = \begin{cases} a + \frac{1-a}{M} & i \in S \\ a & i \notin S,\, i \leq K \\ 0 & i > K \end{cases}.$$

It remains to show that for some $t^* > 0$, the least-favorable distribution in (B.1) places equal mass on each of the points $\{t^*\mu^S\}_{S \in \Sigma_M}$. Define the following likelihood ratio test:

$$\varphi(Y, t) = 1\left\{ \sum_{S \in \Sigma_M} g(\|Y - t\mu^S\|^2 - \|Y - \mu_a\|^2) < cv(t) \right\}$$

where $cv(t)$ endogenously takes on a value such that φ is level-α valid under $Y \sim \mathcal{N}(t\mu^S, I_N)$ for all $S \in \Sigma_M$.

We begin by deriving a necesary first-order condition for $\varphi(Y, t^*)$ to be valid and show that this must be satisfied for some value of $t^* \in (0, \bar{t})$, where $\bar{t}$ is the unique value such that the points $\mu_a, \{\bar{t}\mu^S\}_{S \in \Sigma_M}$ lie in the same hyperplane.[36] We then derive a necessary second-order condition,

[36] It is trivial to show the existence/uniqueness of $\bar{t}$.

and we invoke the mathematical lemmas to show that such a condition is, in fact, sufficient.

B.2.2 First-Order Conditions for Local Validity

For any $\mu \in \mathbb{R}^N$, the size of $\varphi(Y,t)$ under $Y \sim \mathcal{N}(\mu, I_N)$ is

$$E_\mu\left[\varphi(Y,t)\right] = \int \varphi(y,t) \cdot g(\|y-\mu\|^2)\, dy.$$

For arbitrary fixed $t \geq 0$, $S \in \Sigma_M$, and $\bar{\mu} \in V_S$, suppose $E_{\bar{\mu}}[\varphi(Y,t)] \geq E_\mu[\varphi(Y,t)]$ for all $\mu \in V_{\bar{S}}$ within an open neighborhood of $\bar{\mu}$. Write V^i to denote the i'th column of V. The implied first-order necessary condition is

$$\frac{\partial}{\partial r}\left(\int \varphi(y,t) \cdot g(\|y-\bar{\mu}+rV^i\|^2)\, dy\right)\Bigg|_{r=0} \propto E_{\bar{\mu}}\left[\varphi(Y,t) \cdot ((Y-\bar{\mu})'V^i)\right] = 0 \quad \text{for all } i \in S. \tag{B.2}$$

The following lemma shows that this condition is satisfied for some $t^* \in (0,\bar{t})$.

Lemma B.5. *There exists $t^* \in (0,\bar{t})$ such that for each $S \in \Sigma_M$, $E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t^*) \cdot ((Y-t^*\mu^S)'V^i)\right] = 0$ for all $i \in S$.*

Proof. Let $\bar{S} = \{1,\ldots,M\}$. By symmetry, it suffices to show this for the case where $S = \bar{S}$. For each $\sigma = (\sigma_1,\sigma_2) \subset \bar{S}$ of size 2 (where it is assumed that $\sigma_1 < \sigma_2$), define $v^\sigma \in \mathbb{R}^N$ such that $v^\sigma_i = 1$ for $i = \sigma_1$, $v^\sigma_i = -1$ for $i = \sigma_2$, and $v^\sigma_i = 0$ for $i \notin \sigma$. The set of vectors $\{v^\sigma : \sigma \subset \bar{S}, |\sigma| = 2\}$ span the linear subspace of $V_{\bar{S}}$ orthogonal to $\mu^{\bar{S}}$. Hence, to show (B.2) for $S = \bar{S}$, it suffices to show that there exists $t^* > 0$ such that: *(i)* $E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t^*) \cdot ((Y-\bar{\mu})'v^\sigma)\right] = 0$ for all $\sigma \subset \bar{S}$, $|\sigma| = 2$, and *(ii)* $E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t^*) \cdot ((Y-\bar{\mu})'\mu^{\bar{S}})\right] = 0$. Note, however, that the symmetry of the problem implies

$$\sum_{S \in \Sigma_M} g(\|\mu^{\bar{S}} + rv^\sigma - t\mu^S\|^2 - \|\mu^{\bar{S}} - \mu_a\|^2) = \sum_{S \in \Sigma_M} g(\|\mu^{\bar{S}} - rv^\sigma - t\mu^S\|^2 - \|\mu^{\bar{S}} - \mu_a\|^2)$$

$$\Rightarrow \varphi(\mu^{\bar{S}} + rv^\sigma, t) = \varphi(\mu^{\bar{S}} - rv^\sigma, t)$$

for all $r,t \geq 0$ and $\sigma \subset \bar{S}$ of size 2. As a consequence, point *(i)* above is satisfied, and it remains to show point *(ii)*.

Define $h(t) = E_{t\mu^{\bar{S}}}\left[\varphi(Y,t) \cdot ((Y-t\mu^{\bar{S}})'\mu^{\bar{S}})\right]$. The following shows $h(\bar{t}) < 0 < h(0)$ and, hence, by the continuity of $h(\cdot)$, the existence of t^* such that $h(t^*) = 0$. Define an orthonormal transformation O such that $O\mu^{\bar{S}} = (\|\mu^{\bar{S}}\|, 0, \ldots, 0)'$ and $O\mu_a = (\frac{\mu_a'\mu^{\bar{S}}}{\|\mu^{\bar{S}}\|}, \|\mu_a - \frac{\mu_a'\mu^{\bar{S}}}{\|\mu^{\bar{S}}\|}\|, 0, \ldots, 0)$.

1. We show $h(\bar{t}) < 0$. Because the points $\{\bar{t}\mu^S\}_{S\in\Sigma_M}$ all lie within the hyperplane $\{Y : (Y - \mu_a)'\mu_a = 0\}$, the region $R_{\bar{t}} = \{Y : \varphi(Y, \bar{t}) = 1\} \subseteq \mathbb{R}^N$ is invariant to translations by multiples of μ_a, so that $R_{\bar{t}} = \{Y : Y = r\mu_a + \Delta, r \in \mathbb{R}\}$ with $\Delta = \{Y : \varphi(Y, 0) = 1, Y'\mu^{\bar{S}} = 0\}$. Then $OR = \{Y : Y = (y_1, \kappa y_1, \ldots, 0)' + O\Delta, y_1 \in \mathbb{R}, \kappa = \frac{\mu_a'\mu^{\bar{S}}}{\|\mu_a\|\mu^{\bar{S}}\| - \mu_a'\mu^{\bar{S}}\|}\}$, observe that $O\Delta$ lies on the hyperplane $\{Y : Y_1 = 0\}$, and write $O\Delta = \{0\} \times \tilde{\Delta}$ with $\tilde{\Delta} \subset \mathbb{R}^{N-1}$. For some positive constant C_N which depends only on N,

$$
\begin{aligned}
& E_{\bar{t}\mu^{\bar{S}}}\left[\varphi(Y,\bar{t})\cdot((Y-\bar{t}\mu^{\bar{S}})'\mu^{\bar{S}})\right] \\
= \; & E_{\bar{t}\mu^{\bar{S}}}\left[\varphi(OY,\bar{t})\cdot((OY-\bar{t}O\mu^{\bar{S}})'O\mu^{\bar{S}})\right] \\
= \; & E_{\bar{t}O\mu^{\bar{S}}}\left[\varphi(Y,\bar{t})\cdot((Y-\bar{t}O\mu^{\bar{S}})'O\mu^{\bar{S}})\right] \\
= \; & C_N\cdot\int\varphi(y,\bar{t})\cdot((y-\bar{t}O\mu^{\bar{S}})'O\mu^{\bar{S}})\cdot g(\|y-\bar{t}O\mu^{\bar{S}}\|^2)\,dy \\
= \; & C_N\cdot\int_{(y_1,\kappa y_1,\ldots,0)'+O\Delta}((y-\bar{t}O\mu^{\bar{S}})'O\mu^{\bar{S}})\cdot g(\|y-\bar{t}O\mu^{\bar{S}}\|^2)\,dy \\
= \; & C_N\cdot\int_{-\infty}^{\infty}\left(\int_{(\kappa y_1,\ldots,0)'+\tilde{\Delta}}((y-\bar{t}O\mu^{\bar{S}})'O\mu^{\bar{S}})\cdot g(\|y-\bar{t}O\mu^{\bar{S}}\|^2)\,dy_2\ldots dy_N\right)dy_1 \\
= \; & C_N\cdot\int_{-\infty}^{\infty}g((y_1-\bar{t}\|\mu^{\bar{S}}\|)^2)\cdot(y_1-\bar{t}\|\mu^{\bar{S}}\|)\cdot\left(\int_{(\kappa y_1,\ldots,0)'+\tilde{\Delta}}g(y_2^2)\cdot g(\sum_{i=3}^N y_i^2)\,dy_2\ldots dy_N\right)dy_1 \\
= \; & C_N\cdot\int_{-\infty}^{\infty}g(\tilde{y}_1^2)\cdot\tilde{y}_1\cdot\left(\int_{\tilde{\Delta}}g[(\kappa(\tilde{y}_1+\bar{t}\|\mu^{\bar{S}}\|)+\tilde{y}_2)^2]\cdot g(\sum_{i=3}^N y_i^2)\,d\tilde{y}_2dy_3\ldots dy_N\right)d\tilde{y}_1 \\
= \; & C_N\cdot\int_0^{\infty}g(\tilde{y}_1^2)\tilde{y}_1\int_{\tilde{\Delta}}\left(g[(\kappa(\tilde{y}_1+\bar{t}\|\mu^{\bar{S}}\|)+\tilde{y}_2)^2]-g[(\kappa(-\tilde{y}_1+\bar{t}\|\mu^{\bar{S}}\|)+\tilde{y}_2)^2]\right)g(\sum_{i=3}^N y_i^2) \\
& \ldots d\tilde{y}_2dy_3\ldots dy_Nd\tilde{y}_1 \\
< \; & 0
\end{aligned}
$$

   where the last inequality follows, because for all $0 < y < y'$ and $\tilde{y}_2$, $g((y'+\tilde{y}_2)^2) - g((y+\tilde{y}_2)^2) < 0$.

2. Observe that $R_0 = \{Y : \varphi(Y, 0) = 1\} = \{Y : Y'\mu_a \geq c\}$ for some constant c. Because the components of $\mu^{\bar{S}}$ are non-negative, $h(0) = E_0\left[\varphi(Y,0)\cdot(Y'\mu^{\bar{S}})\right] > 0$ follows from a similar argument as in part 1.

□

B.2.3 Second-Order Conditions for Local Validity

From the results of Section B.2.2, there exists $t^* \in (0, \bar{t})$ such that for each $S \in \Sigma_M$, $E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t^*)\cdot((Y-t^*\mu^S)'V^i)\right] = 0$ for all $i \in S$. In order for $E_{t^*\mu^S}[\varphi(Y,t^*)] \geq E_\mu[\varphi(Y,t^*)]$ for all $\mu \in V_S$ within an open neighborhood of $t^*\mu^S$, the $M \times M$ Hessian matrix formed by the second-derivatives

$$H_{i,j} = \left(\frac{\partial^2}{\partial r_i \partial r_j}\int \varphi(y,t)\cdot g(\|y - t^*\mu^S + \sum_{s\in S} r_s V^s\|^2)\,dy\right)\Bigg|_{r_s=0\,\forall\, s\in S}, \quad i,j \in S$$

must be negative-definite. By symmetry, the off-diagonal elements $H_{i,j}$, $i \neq j$ are the same, so it suffices to show that the diagonal elements $H_{i,i}$ are negative for all $i \in S$.

By symmetry, it suffices to show this for $S = \bar{S}$. Observe

$$\begin{aligned}
H_{1,1} &= \left(\frac{\partial}{\partial r_1}\int \varphi(y,t)\cdot g(\|y - t^*\mu^{\bar{S}} + r_1V^1\|^2)\cdot(-\frac{1}{2}(2(y-t^*\mu^{\bar{S}})'V^1 + 2r_1\|V^1\|^2))\,dy\right)\Bigg|_{r_1=0}\\
&= \int \varphi(y,t)\cdot g(\|y-t^*\mu^{\bar{S}}\|^2)\cdot\left(((y-t^*\mu^{\bar{S}})'V^1)^2 - \|V^1\|^2\right)dy\\
&= \|V^1\|^2\int \varphi(y,t)\cdot g(\|y-t^*\mu^{\bar{S}}\|^2)\cdot\left(\left((y-t^*\mu^{\bar{S}})'\frac{V^1}{\|V^1\|}\right)^2 - 1\right)dy\\
&= \|V^1\|^2 E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t)\cdot\left(((Y-t^*\mu^{\bar{S}})'\frac{V^1}{\|V^1\|})^2 - 1\right)\right]dy.
\end{aligned}$$

Notice that $E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t^*)\cdot((Y-t^*\mu^S)'V^1)\right] = 0$ implies $E_{t^*\mu^{\bar{S}}}\left[\frac{Y'V^1}{\|V^1\|}\,|\,\varphi(Y,t^*)=1\right] = \frac{(t^*\mu^{\bar{S}})'V^1}{\|V^1\|}$. By Lemmas B.1 and B.2,

$$E_{t^*\mu^{\bar{S}}}\left[((Y-t^*\mu^{\bar{S}})'\frac{V^1}{\|V^1\|})^2\,\middle|\,\varphi(Y,t^*)=1\right] = \frac{E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t^*)\cdot((Y-t^*\mu^{\bar{S}})'\frac{V^1}{\|V^1\|})^2\right]}{E_{t^*\mu^{\bar{S}}}[\varphi(Y,t^*)]}$$

$$< E_{t^*\mu^{\bar{S}}}\left[((Y-t^*\mu^{\bar{S}})'\frac{V^1}{\|V^1\|})^2\right] = 1$$

$$\Rightarrow \|V^1\|^2 E_{t^*\mu^{\bar{S}}}\left[\varphi(Y,t)\cdot\left(((Y-t^*\mu^{\bar{S}})'\frac{V^1}{\|V^1\|})^2 - 1\right)\right] = H_{1,1} < 0.$$

By symmetry, this argument shows $H_{i,i} < 0$ for all $i \in \bar{S}$.

Hence, $\mu = t^*\mu^S$ is a local minimum for $E_\mu[\varphi(Y,t^*)]$ within V_S for all $S \in \Sigma_M$. By Lemma B.3, $\mu = t^*\mu^S$ is the global minimum for $E_\mu[\varphi(Y,t^*)]$ within V_S for all $S \in \Sigma_M$, and the theorem is proved.

C Proofs from Part III

Unless otherwise-specified, all notation throughout Section C is adopted from Part 3.

C.1 Proof of Theorem 15.3

For ease of notation, fix a determinstic sequence of $NJ \times K$ matrices X, and suppress dependence on X. Let $e = ((e^1)', \ldots, (e^N)')' \in \mathbb{R}^{NJ}$ denote the error term, and assume $E[e] = 0$. Suppose $\Sigma = E[ee'] = \text{diag}(\Sigma^{1,1}, \ldots, \Sigma^{N,N})$, where $\Sigma^{i,i}$ is a $J \times J$ covariance matrix for each $i = 1, \ldots, N$. For $A, \tilde{A} \in \mathbb{R}^{NJ}$, write $A = ((A^1)', \ldots, (A^N)')'$ and $\tilde{A} = ((\tilde{A}^1)', \ldots, (\tilde{A}^N)')'$, with $A^i, \tilde{A}^i \in \mathbb{R}^J$ for all $i = 1, \ldots, N$. Unless otherwise specified, for an $NJ \times NJ$ matrix W, let $W^{i,\tilde{i}}$ denote the $J \times J$ matrix forming $(i, \tilde{i})$'th block of W. Let $\mathbf{1}$ denote a vector of ones. All limits are taken as $N \to \infty$.

Define $\hat{\omega} = \hat{e}' S^* \hat{e}$, where $S^* = \text{diag}(S^{*1}, \ldots, S^{*N})$ is a block-diagonal matrix, and S^{*i} are $J \times J$ matrices such that $(\text{vec}(S^{*1})', \ldots, \text{vec}(S^{*N}))' = \kappa \left(\text{vec}(A^1(\tilde{A}^1)')', \ldots, \text{vec}(A^N(\tilde{A}^N)')'\right)'$ for $\kappa = (M * M)^{-1}$. It suffices for the proof to derive $E[\hat{\omega}]$ and $Var[\hat{\omega}]$.

1. To show $E[\hat{\omega}] = \omega$, observe

$$\omega = \sum_{i=1}^{N} (A^i)' \Sigma^i \tilde{A}^i = \sum_{i=1}^{N} \mathbf{1}' (\text{vec}(A^i(\tilde{A}^i)') \odot \text{vec}(\Sigma^i))$$

$$E[\hat{\omega}] = \sum_{i=1}^{N} \text{Tr}((MS^*M)^{i,i} \Sigma^i) = \sum_{i=1}^{N} \mathbf{1}' ((MS^*M)^{i,i} \odot \text{vec}(\Sigma^i))$$

where for all $i = 1, \ldots, N$,

$$\begin{pmatrix} \text{vec}(S^{*1}) \\ \vdots \\ \text{vec}(S^{*N}) \end{pmatrix} = \kappa \begin{pmatrix} \text{vec}(A^1(\tilde{A}^1)') \\ \vdots \\ \text{vec}(A^N(\tilde{A}^N)') \end{pmatrix}$$

$$\Rightarrow (M * M) \begin{pmatrix} \text{vec}(S^{*1}) \\ \vdots \\ \text{vec}(S^{*N}) \end{pmatrix} = \begin{pmatrix} \text{vec}(A^1(\tilde{A}^1)') \\ \vdots \\ \text{vec}(A^N(\tilde{A}^N)') \end{pmatrix}$$

$$\Rightarrow vec(A^i(\tilde{A}^i)') = \sum_{j=1}^{N} (M^{i,j} \otimes M^{i,j}) \text{vec}(S^{*i})$$

$$\Rightarrow A^i(\tilde{A}^i)' = \sum_{j=1}^{N} M^{i,j} S^{*j} M^{j,i} = (MS^*M)^{i,i}.$$

2. For the variance of $\hat{\omega}$, observe

$$\begin{aligned}
\hat{\omega} &= \sum_{i=1}^{N} (\hat{e}^i)' S^{*i} \hat{e}^i \\
&= \sum_{i=1}^{N} \mathrm{Tr}(S^{*i} \hat{e}^i (\hat{e}^i)') \\
&= \sum_{i=1}^{N} \mathrm{Tr}(\sum_{j,h=1}^{N} S^{*i} M^{i,j} e^j (e^h)' (M^{i,h})') \\
&= \sum_{i=1}^{N} \sum_{j,h=1}^{N} \mathrm{Tr}(M^{h,i} S^{*i} M^{i,j} e^j (e^h)') \\
&= \sum_{j,h=1}^{N} \sum_{i=1}^{N} \mathrm{Tr}(e^j (e^h)' M^{h,i} S^{*i} M^{i,j}) \\
&= \sum_{j,h=1}^{N} \mathrm{Tr}(e^j (e^h)' (\sum_{i=1}^{N} M^{h,i} S^{*i} M^{i,j}))
\end{aligned}$$

Let $\Gamma^{j,h} = \sum_{i=1}^{N} M^{h,i} S^{*i} M^{i,j}$, which implies

$$\begin{aligned}
\mathrm{vec}(\Gamma^{j,h}) &= \sum_{i=1}^{N} (M^{j,i} \otimes M^{h,i}) \mathrm{vec}(S^{*i}) \\
&= \sum_{i=1}^{N} (M^{j,i} \otimes M^{h,i}) \left(\sum_{\bar{i}=1}^{N} \kappa^{i,\bar{i}} \mathrm{vec}(A^{\bar{i}} (\tilde{A}^{\bar{i}})') \right) \\
&= \sum_{\bar{i}=1}^{N} \left(\sum_{i=1}^{N} (M^{j,i} \otimes M^{h,i}) \kappa^{i,\bar{i}} \right) \mathrm{vec}(A^{\bar{i}} (\tilde{A}^{\bar{i}})').
\end{aligned}$$

Suppose $\sup_X \max_{i=1,\ldots,N} Var[e_i] = O(1)$. Then

$$\begin{aligned}
Var[\hat{\omega}] &= Var[\sum_{j,h=1}^{N} \mathrm{Tr}(e^j (e^h)' \Gamma^{j,h})] \\
&= \sum_{j,h=1}^{N} Var[\mathrm{Tr}(e^j (e^h)' \Gamma^{j,h})] \\
&= O(1) \sum_{j,h=1}^{N} \sum_{m,n=1}^{J} (\Gamma^{j,h}_{m,n})^2.
\end{aligned}$$

For each (j,h), write

$$\begin{aligned}
&\sum_{m,n=1}^{J}(\Gamma_{m,n}^{j,h})^2\\
=\;&\|\mathrm{vec}(\Gamma^{j,h})\|^2\\
=\;&\left(\sum_{\bar{i}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'\left(\sum_{i=1}^{N}(M^{j,i}\otimes M^{h,i})\kappa^{i,\bar{i}}\right)'\right)\left(\sum_{\bar{i}=1}^{N}\left(\sum_{i=1}^{N}(M^{j,i}\otimes M^{h,i})\kappa^{i,\bar{i}}\right)\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}}))\right)\\
=\;&\sum_{\bar{i},\bar{j}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'\left(\sum_{i=1}^{N}(M^{j,i}\otimes M^{h,i})\kappa^{i,\bar{i}}\right)'\left(\sum_{i=1}^{N}(M^{j,i}\otimes M^{h,i})\kappa^{i,\bar{j}}\right)\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&\sum_{\bar{i},\bar{j}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'\left(\sum_{i,\bar{h}=1}^{N}((M^{j,i}\otimes M^{h,i})\kappa^{i,\bar{i}})'((M^{j,\bar{h}}\otimes M^{h,\bar{h}})\kappa^{\bar{h},\bar{j}})\right)\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&\sum_{\bar{i},\bar{j},i,\bar{h}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'(\kappa^{i,\bar{i}})'(M^{j,i}\otimes M^{h,i})'(M^{j,\bar{h}}\otimes M^{h,\bar{h}})\kappa^{\bar{h},\bar{j}}\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&\sum_{\bar{i},\bar{j},i,\bar{h}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'(\kappa^{i,\bar{i}})'(M^{i,j}\otimes M^{i,h})(M^{j,\bar{h}}\otimes M^{h,\bar{h}})\kappa^{\bar{h},\bar{j}}\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&\sum_{\bar{i},\bar{j},i,\bar{h}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'(\kappa^{i,\bar{i}})'((M^{i,j}M^{j,\bar{h}})\otimes(M^{i,h}M^{h,\bar{h}}))\kappa^{\bar{h},\bar{j}}\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})').
\end{aligned}$$

from which the conclusion follows:

$$\begin{aligned}
&\sum_{j,h=1}^{N}\sum_{m,n=1}^{J}(\Gamma_{m,n}^{j,h})^2\\
=\;&\sum_{j,h=1}^{N}\sum_{\bar{i},\bar{j},i,\bar{h}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'(\kappa^{i,\bar{i}})'((\sum_{j=1}^{N}M^{i,j}M^{j,\bar{h}})\otimes(\sum_{h=1}^{N}M^{i,h}M^{h,\bar{h}}))\kappa^{\bar{h},\bar{j}}\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&\sum_{\bar{i},\bar{j},i,\bar{h}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'(\kappa^{i,\bar{i}})'(M^{i,\bar{h}}\otimes M^{i,\bar{h}})\kappa^{\bar{h},\bar{j}}\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&\sum_{\bar{i},\bar{j}=1}^{N}\mathrm{vec}(A^{\bar{i}}(\tilde{A}^{\bar{i}})')'\kappa^{\bar{i},\bar{j}}\mathrm{vec}(A^{\bar{j}}(\tilde{A}^{\bar{j}})')\\
=\;&O(1)\cdot\max_{r=1,\ldots,NJ}A_r^2\tilde{A}_r^2\cdot\left(\sum_{r,\bar{r}=1}^{NJ}\kappa_{r,\bar{r}}^2\right)^{1/2}
\end{aligned}$$

C.2 Proof of Corollary 15.4

Fix $L>0$ such that $Var[\hat{\Omega}^*|X]\leq N^{-2}LC_X$ for all X. For arbitrary fixed $\delta,\epsilon\in(0,1)$, there exists $\bar{N}$ such that $Pr(\mathcal{E})>(1-\delta)^{1/2}$ for all $N\geq\bar{N}$, where $\mathcal{E}=\{|N^{-2}C_X|<(1-(1-\delta)^{1/2})\epsilon^2L^{-1}\}$.

Markov's inequality implies $Pr((\hat{\Omega}^* - \Omega_X)^2 > \epsilon^2 \,|\, \mathcal{E}) \leq N^{-2} L C_X \epsilon^{-2} < 1 - (1-\delta)^{1/2}$ and, hence, $Pr(|\hat{\Omega}^* - \Omega_X| < \epsilon \,|\, \mathcal{E}) > (1-\delta)^{1/2}$ for all $N \geq \bar{N}$. Combining,

$$Pr(|\hat{\Omega}^* - \Omega_X| < \epsilon) = Pr(\mathcal{E}) \cdot Pr(|\hat{\Omega}^* - \Omega_X| < \epsilon \,|\, \mathcal{E}) > 1 - \delta$$

for all $N \geq \bar{N}$.