

# ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

*An International Society for the Advancement of Economic  
Theory in its Relation to Statistics and Mathematics*

<http://www.econometricsociety.org/>

*Econometrica*, Vol. 83, No. 2 (March, 2015), 771–811

## NEARLY OPTIMAL TESTS WHEN A NUISANCE PARAMETER IS PRESENT UNDER THE NULL HYPOTHESIS

GRAHAM ELLIOTT

*University of California, San Diego, La Jolla, CA 92093, U.S.A.*

ULRICH K. MÜLLER

*Princeton University, Princeton, NJ 08544, U.S.A.*

MARK W. WATSON

*Princeton University, Princeton, NJ 08544, U.S.A. and NBER*

---

The copyright to this Article is held by the Econometric Society. It may be downloaded, printed and reproduced only for educational or research purposes, including use in course packs. No downloading or copying may be done for any commercial purpose without the explicit permission of the Econometric Society. For such commercial purposes contact the Office of the Econometric Society (contact information may be found at the website <http://www.econometricsociety.org> or in the back cover of *Econometrica*). This statement must be included on all copies of this Article that are made available electronically or in any other format.

---

## NEARLY OPTIMAL TESTS WHEN A NUISANCE PARAMETER IS PRESENT UNDER THE NULL HYPOTHESIS

BY GRAHAM ELLIOTT, ULRICH K. MÜLLER, AND MARK W. WATSON<sup>1</sup>

This paper considers nonstandard hypothesis testing problems that involve a nuisance parameter. We establish an upper bound on the weighted average power of all valid tests, and develop a numerical algorithm that determines a feasible test with power close to the bound. The approach is illustrated in six applications: inference about a linear regression coefficient when the sign of a control coefficient is known; small sample inference about the difference in means from two independent Gaussian samples from populations with potentially different variances; inference about the break date in structural break models with moderate break magnitude; predictability tests when the regressor is highly persistent; inference about an interval identified parameter; and inference about a linear regression coefficient when the necessity of a control is in doubt.

KEYWORDS: Least favorable distribution, composite hypothesis, maximin tests.

### 1. INTRODUCTION

CONSIDER A STATISTICAL HYPOTHESIS TEST concerning a parameter  $\theta = (\beta', \delta)'$ , where  $\beta$  is the parameter of interest and  $\delta$  is a nuisance parameter. Both the null and alternative are composite:

$$(1) \quad H_0: \beta = \beta_0, \quad \delta \in D \quad \text{against} \quad H_1: \beta \in B, \quad \delta \in D,$$

so that the null specifies the value of  $\beta$ , but not  $\delta$ .

A key example of a hypothesis testing problem with a nuisance parameter is the Gaussian shift experiment, where the single observation  $Y$  is drawn from

$$(2) \quad Y = \begin{pmatrix} Y_\beta \\ Y_\delta \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \beta \\ \delta \end{pmatrix}, \Sigma \right)$$

and the positive definite covariance matrix  $\Sigma$  is known. With an unrestricted nuisance parameter space  $D$ , there are good reasons for simply ignoring  $Y_\delta$ , even if  $\Sigma$  is not block-diagonal: For scalar  $\beta$ , the one-sided test of (1) based on  $Y_\beta$  is uniformly most powerful. In the two-sided problem, rejecting for large values of  $|Y_\beta - \beta_0|$  yields the uniformly most powerful unbiased test. These arguments can be generalized to vector valued  $\beta_0$  and unrestricted  $B$  by either imposing an appropriate rotational invariance for  $Y_\beta$ , by focusing on most

<sup>1</sup>This research was funded in part by NSF Grant SES-0751056 (Müller). The paper supersedes the corresponding sections of the previous working papers “Low-Frequency Robust Cointegration Testing” and “Pre and Post Break Parameter Inference” by the same set of authors. We thank the co-editor, two referees, Andriy Norets, Andres Santos, the participants of the AMES 2011 meeting at Korea University, of the 2011 Greater New York Metropolitan Econometrics Colloquium, and of a workshop at USC for helpful comments.

stringent tests, or by maximizing weighted average power on alternatives that are equally difficult to distinguish (see, for instance, Choi, Hall, and Schick (1996) and Lehmann and Romano (2005) for a comprehensive treatment and references).

These results are particularly significant because LeCam's Limits of Experiments Theory implies that inference about the parameter of a well-behaved parametric model is large sample equivalent to inference in a Gaussian shift experiment. See, for instance, Lehmann and Romano (2005) or van der Vaart (1998) for textbook introductions. As a consequence, the usual likelihood ratio, Wald, and score tests have a well-defined asymptotic optimality property also in the presence of a nuisance parameter.

These standard results only apply to the Gaussian shift experiment with unrestricted  $D$ , however. Outside this class, it is sometimes possible to derive powerful tests in the presence of nuisance parameters using specific techniques. One approach is to impose invariance constraints. For example, Dufour and King's (1991) and Elliott, Rothenberg, and Stock's (1996) optimal unit root tests impose translation invariance that eliminates the mean parameter. In many problems, however, invariance considerations only reduce the dimensionality of the nuisance parameter space. In the weak instrument problem with multiple instruments, for instance, rotational invariance reduces the effective nuisance parameter to the concentration parameter, a nonnegative scalar. What is more, even if an invariance transformation can be found such that the maximal invariant is pivotal under the null hypothesis, the restriction to invariant tests might not be natural. Imposing invariance can then rule out perfectly reasonable, more powerful procedures. We provide such an example in Section 5.2 below.

A second approach is to impose similarity, unbiasedness, or conditional unbiasedness. In particular, conditioning on a statistic that is sufficient for  $\delta$  ensures by construction that conditional distributions no longer depend on  $\delta$ . Depending on the exact problem, this allows the derivation of optimal tests in the class of all similar or conditionally unbiased tests, such as Moreira's (2003) CLR test for the weak instrument problem. The applicability of this approach, however, is quite problem specific. In addition, it is again possible that an exclusive focus on, say, similar tests rules out many reasonable and powerful tests a priori.<sup>2</sup>

In this paper, we adopt a well-known general solution to hypothesis tests in the presence of a nuisance parameter by integrating out the parameter  $\theta$  with

<sup>2</sup>In the Behrens–Fisher problem, Linnik (1966, 1968) and Salaevskii (1963) have shown that all similar tests have highly undesirable features, at least as long as the smaller sample has at least three observations. More recently, Andrews (2011) showed that similar tests have poor power in the context of moment inequality tests. However, it may sometimes be useful to impose similarity or other constraints on power functions; see Moreira and Moreira (2013).

respect to some probability distribution  $\Lambda$  under the null, and some probability distribution  $F$  under the alternative. The test statistic is then simply the likelihood ratio of the resulting integrated null and alternative densities. We treat  $F$  as given, so that the problem effectively reduces to testing against the point alternative of a hyper-model where  $\theta$  is drawn from  $F$ . In terms of the original composite alternative hypothesis,  $F$  represents the relative weights a researcher attaches to the power under various alternatives, so we seek tests that are optimal in the sense of maximizing weighted average power. The main concern of the paper is the probability distribution  $\Lambda$  under the null hypothesis, which has to be carefully matched to the problem and  $F$ . Technically, the distribution  $\Lambda$  that yields the optimal likelihood ratio test is known as the “least favorable distribution” (see [Lehmann and Romano \(2005\)](#) for details).

The least favorable approach is very general. Indeed, the standard results about the Gaussian location problem (2) reviewed above are obtained in this fashion. For nonstandard problems, however, it can be challenging to identify the least favorable distribution, and thus the efficient test. This is the problem that we address in this paper.

Our approach is based on the notion of an “approximate least favorable distribution” (ALFD), that we determine numerically. The ALFD plays two conceptually distinct roles: first, it yields an analytical upper bound on the weighted average power of *all* valid tests, and thus can be used to evaluate the optimality or near-optimality of extant tests. For example, [Andrews, Moreira, and Stock \(2008\)](#) showed that [Moreira’s \(2003\)](#) CLR test essentially achieves the power bound from an ALFD implying that the test is essentially optimal. Second, the test based on the likelihood ratio statistic with the null density integrated out with respect to the ALFD yields weighted average power close to the upper bound.

For most nonstandard testing problems, much of the parameter space essentially corresponds to a standard testing problem. For instance, in the weak instrument problem, a large concentration parameter essentially turns the problem into a standard one. We extend our approach so that tests switch to the “standard” test (with high probability) in this “almost standard” part of the parameter space. The weighting function for power and ALFD thus only need to be determined in the genuinely nonstandard part of the parameter space. A corresponding modification of the power bound result shows that the resulting tests are nearly weighted average power maximizing among all tests that have at least as much power as the standard test in the standard part of the parameter space. In our numerical work, we determine tests whose weighted average power is within 0.5 percentage points of the bound, and this is the sense in which the tests are nearly optimal.

The algorithm may be applied to solve for nearly optimal tests in a variety of contexts: small sample and asymptotic Limit of Experiment-type problems, time series and cross-section problems, nonstandard and Gaussian shift problems. Specifically, we consider six applications. First, we introduce a running

example to motivate our general approach that involves the Gaussian shift problem (2) with scalar  $\beta$  and  $\delta$ , where  $\delta$  is known to be nonnegative. This arises, for instance, in a regression context where the sign of the coefficient of one of the controls is known. Second, we consider the small sample problem of testing for the equality of means from two normal populations with unknown and possibly different variances, the so called “Behrens–Fisher problem.” While much is known about this well-studied problem (see [Kim and Cohen \(1998\)](#) for a survey), small sample optimal tests have not been developed, making the application of the algorithm an interesting exercise. Third, we consider inference about the break date in a time series model with a single structural change. In this problem,  $\delta$  is related to the size of the parameter break, where ruling out small breaks (as, e.g., [Bai \(1994, 1997\)](#) and much of the subsequent literature) may lead to substantially oversized tests (see [Elliott and Müller \(2007\)](#)). We compare our near-optimal test to the invariant tests developed in [Elliott and Müller \(2007\)](#), and find that the invariance restriction is costly in terms of power. The fourth example concerns inference in the predictive regression model with a highly persistent regressor. We compare our near-optimal tests to the tests derived by [Campbell and Yogo \(2006\)](#), and find that our tests have higher power for most alternatives. The fifth example considers nearly optimal inference about a set-identified parameter as in [Imbens and Manski \(2004\)](#), [Woutersen \(2006\)](#), and [Stoye \(2009\)](#). Finally, we consider a canonical model selection problem, where the parameter of interest  $\beta$  is the coefficient in a linear regression, and the necessity of including a particular control variable is in doubt. It is well understood that standard model selection procedures do not yield satisfactory inference for this problem—[Leeb and Pötscher \(2005\)](#) provided a succinct review and references. The application of our approach here yields a power bound for the performance of any uniformly valid procedure, as well as a corresponding test with power very close to the power bound.

The remainder of the paper is organized as follows. Section 2 formally states the problem, introduces the running example, presents the analytical power bound result, and uses the power bound to define the approximate least favorable distribution. This section also highlights the connection between the hypothesis testing problem and minimax decision rules as discussed, for instance, in [Blackwell and Girshick \(1954\)](#) and [Ferguson \(1967\)](#). Section 3 takes up the numerical problem of computing the ALFD, reviews the existing approaches of [Krafft and Witting \(1967\)](#), [Kempthorne \(1987\)](#), [Chamberlain \(2000\)](#), [Srikanthakumar and King \(2006\)](#), and [Chiburis \(2009\)](#), and proposes a simple and effective algorithm that we recommend for practical use. Section 4 discusses modifications of the ALFD tests so that they (essentially) coincide with known optimal tests in the standard region of the parameter space, and the corresponding modification of the power bound. Finally, Section 5 contains the results for the additional five examples. Appendix A contains additional details on the algorithm and the applications.

## 2. HYPOTHESIS TESTS WITH COMPOSITE NULL

2.1. *Statement of the Problem*

We observe a random element  $Y$  that takes values in the metric space  $\mathcal{Y}$ . The distribution of  $Y$  is parametric with parameter  $\theta \in \Theta \in \mathbb{R}^k$ , so that the probability density function is  $f_\theta(y)$  relative to some sigma-finite measure  $\nu$ . Based on this single observation, we seek to test the hypotheses

$$(3) \quad H_0: \theta \in \Theta_0 \quad \text{against} \quad H_1: \theta \in \Theta_1,$$

where  $\Theta_0 \cap \Theta_1 = \emptyset$  and  $\Theta_0$  is not a singleton, so that the null hypothesis is composite.

Tests of (3) are measurable functions  $\varphi: \mathcal{Y} \mapsto [0, 1]$ , where  $\varphi(y)$  indicates the rejection probability conditional on observing  $Y = y$ . Thus, a nonrandomized test has restricted range  $\{0, 1\}$ , and critical region  $\{y: \varphi(y) = 1\}$ . If  $\varphi(y) \in (0, 1)$  for some  $y \in \mathcal{Y}$ , then  $\varphi$  is a randomized test. In either case, the rejection probability of the test is equal to  $\int \varphi f_\theta d\nu$  for a given  $\theta \in \Theta$ , so that the size of the test is  $\sup_{\theta \in \Theta_0} \int \varphi f_\theta d\nu$ , and by definition, a level  $\alpha$  test has size smaller than or equal to  $\alpha$ .

In many problems, a composite null hypothesis arises due to the presence of a nuisance parameter. In a typical problem,  $\theta$  can be parameterized as  $\theta = (\beta', \delta)'$ , where  $\beta \in \mathbb{R}^{k_\beta}$  is the parameter of interest and  $\delta \in \mathbb{R}^{k_\delta}$  is a nuisance parameter. The hypothesis testing problem (3) then is equivalent to

$$(4) \quad H_0: \beta = \beta_0, \quad \delta \in D \quad \text{against} \quad H_1: \beta \in B, \quad \delta \in D,$$

where  $\beta_0 \notin B$ ,  $\Theta_0 = \{\theta = (\beta', \delta)': \beta = \beta_0, \delta \in D\}$ , and  $\Theta_1 = \{\theta = (\beta', \delta)': \beta \in B, \delta \in D\}$ .

One motivation for the single observation problem involving  $Y$  is a small sample parametric problem, where  $Y$  simply contains the  $n$  observations (or a lower dimensional sufficient statistic). Alternatively, the single observation problem may arise as the limiting problem in some asymptotic approximation, as we now discuss.

**RUNNING EXAMPLE:** To clarify ideas and help motivate our proposed testing procedures, we use the following example throughout the paper. (Related problems were considered by [Moon and Schorfheide \(2009\)](#) and [Andrews and Guggenberger \(2010\)](#).) Suppose we observe  $n$  observations from a parametric model with parameter  $(\gamma, \eta) \in \mathbb{R}^2$ . The hypothesis of interest is  $H_0: \gamma = \gamma_0$ , and it is known a priori that  $\eta \geq \eta_0$ . For instance,  $\gamma$  and  $\eta$  may correspond to regression coefficients, and it is known that the coefficient associated with the control variable is nonnegative. Let  $\beta = \sqrt{n}(\gamma - \gamma_0)$  and  $\delta = \sqrt{n}(\eta - \eta_0)$ . If the model is locally asymptotic normal at  $(\gamma, \eta) = (\gamma_0, \eta_0)$  at the usual parametric  $\sqrt{n}$  rate with nonsingular Fisher information matrix  $\Sigma^{-1}$ , then by Corollary 9.5

of van der Vaart (1998), the Limit Experiment local to  $(\gamma_0, \eta_0)$  concerns the bivariate normal observation

$$(5) \quad Y = \begin{pmatrix} Y_\beta \\ Y_\delta \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \beta \\ \delta \end{pmatrix}, \Sigma \right),$$

where  $\Sigma$  is known. The hypothesis testing problem concerning (5) is

$$(6) \quad H_0: \beta = 0, \quad \delta \geq 0 \quad \text{against} \quad H_1: \beta \in B, \quad \delta \geq 0,$$

where  $B = (0, \infty)$  and  $B = \mathbb{R} \setminus \{0\}$  correspond to one-sided and two-sided alternatives, respectively. It is clear that in either case, we can normalize  $\Sigma$  to be unity on the diagonal without loss of generality, so that the testing problem is only indexed by the correlation  $\rho \in (-1, 1)$ .

By the Asymptotic Representation Theorem (van der Vaart (1998, Theorem 9.3 and Theorem 15.1)), the local asymptotic rejection profile of any test in the original  $n$  observation problem (should it exist) can be matched by a test in the single observation problem (5). What is more, for any test of (5), it is typically straightforward to construct a corresponding test in the original parametric problem with the same asymptotic local power. Thus, the derivation of large sample tests with good local asymptotic power for the original problem reduces to the derivation of good tests for (5).

If the original parametric model concerns additional nuisance parameters, then the Limit Experiment (5) involves a larger dimensional normal variate. It is clear, however, that any valid test of the bivariate problem can still be applied, as the additional Gaussian observations in the Limit Experiment may simply be ignored (although additional arguments, such as invariance considerations, would be needed to argue for the optimality of such a procedure). A similar point applies in the presence of infinite dimensional additional nuisance parameters, that is, if the underlying model is semiparametric (see Choi, Hall, and Schick (1996) for details).

Finally, one could also rely on the approach developed by Müller (2011) to argue for the asymptotic reduction to the single observation problem (5). We omit details for brevity.

## 2.2. Weighted Average Power

The determination of a good test of (3) is difficult because both the null and the alternative are composite. A composite null requires that the test controls rejection probability over all values of  $\theta \in \Theta_0$ ; a composite alternative leads to the consideration of how the test's power varies over  $\theta \in \Theta_1$ . A standard approach for composite alternatives is to consider weighted average power as the scalar criterion to choose among tests

$$(7) \quad \text{WAP}(\varphi) = \int \left( \int \varphi f_\theta d\nu \right) dF(\theta),$$

where  $F$  is a probability measure with support in (the closure of)  $\Theta_1$ . The weighting function  $F$  describes the importance a researcher attaches to the ability of the test to reject under different alternatives. This approach underlies the optimality of Wald's (1943) statistics and has been employed in the influential work by Andrews and Ploberger (1994).

Since tests that maximize WAP equivalently maximize  $\int \varphi(\int f_\theta dF(\theta)) d\nu$  (where the interchange of the order of integration is allowed by Fubini's Theorem), efficient tests under the WAP criterion also maximize power against the single density  $g = \int f_\theta dF(\theta)$ . Thus, with a WAP criterion, the hypothesis testing problem (3) becomes one of finding a powerful test for the problem

$$(8) \quad \begin{aligned} H_0 &: \text{the density of } Y \text{ is } f_\theta, \quad \theta \in \Theta_0 \quad \text{against} \\ H_{1,F} &: \text{the density of } Y \text{ is } g = \int f_\theta dF(\theta), \end{aligned}$$

where the alternative  $H_{1,F}$  is simple. The power of a test under  $H_{1,F}$  is synonymous to weighted average power under the composite alternative  $H_1$  with weighting function  $F$ .

If a uniformly most powerful test exists, then it maximizes WAP for all choices of  $F$ , so that in this sense a focus on WAP is without loss of generality. In most problems, however, the choice of the weighting function  $F$  matters, as there is no uniformly most powerful test: there are many tests whose power functions cross, and one can reasonably disagree about the overall preferred test. We therefore offer no general remarks about  $F$ , but rather discuss our choices in the context of the running example and the particular testing problems analyzed in Section 5.

### 2.3. A Set of Power Bounds

Under the weighted average power criterion (7), the challenge is to derive a good test of a composite null against a simple alternative, that is a good test of (8). This subsection does not derive such a test directly, but rather provides a general set of bounds on the power of any level  $\alpha$  test. These bounds are useful both for constructing an approximately efficient test and for evaluating the efficiency of ad hoc tests.

Suppose the composite null hypothesis in (8) is replaced by the single hypothesis

$$H_{0,\Lambda} : \text{the density of } Y \text{ is } \int f_\theta d\Lambda(\theta),$$

where  $\Lambda$  is a probability distribution with support on  $\Theta_0$ . In general, the size  $\alpha$  Neyman–Pearson test of  $H_{0,\Lambda}$  against  $H_{1,F}$  is *not* a level  $\alpha$  test of  $H_0$  in (8), as its null rejection probability is equal to  $\alpha$  by definition only when  $Y$  is drawn from



the mixture distribution  $\int f_\theta d\Lambda(\theta)$  and does not satisfy the size constraint for the composite null  $H_0$ . Its properties are nevertheless helpful to bound the power of any level  $\alpha$  test of (8).

LEMMA 1: Let  $\varphi_\Lambda$  be the size  $\alpha$  test of  $H_{0,\Lambda}$  against  $H_{1,F}$  of the Neyman–Pearson form

$$(9) \quad \varphi_\Lambda(y) = \begin{cases} 1 & \text{if } g(y) > cv \int f_\theta(y) d\Lambda(\theta), \\ \kappa & \text{if } g(y) = cv \int f_\theta(y) d\Lambda(\theta), \\ 0 & \text{if } g(y) < cv \int f_\theta(y) d\Lambda(\theta), \end{cases}$$

for some  $cv \geq 0$  and  $0 \leq \kappa \leq 1$ . Then for any level  $\alpha$  test  $\varphi$  of  $H_0$  against  $H_{1,F}$ ,  $\int \varphi_\Lambda g d\nu \geq \int \varphi g d\nu$ .

PROOF: Since  $\varphi$  is a level  $\alpha$  test of  $H_0$ ,  $\int \varphi f_\theta d\nu \leq \alpha$  for all  $\theta \in \Theta_0$ . Therefore,  $\iint \varphi f_\theta d\nu d\Lambda(\theta) = \iint \varphi f_\theta d\Lambda(\theta) d\nu \leq \alpha$ , where the equality follows from Fubini’s Theorem, so that  $\varphi$  is also a level  $\alpha$  test of  $H_{0,\Lambda}$  against  $H_{1,F}$ . The result now follows from the Neyman–Pearson Lemma. Q.E.D.

Lemma 1 formalizes the intuitive result that replacing the composite null hypothesis  $H_0$  with the single mixture null hypothesis  $H_{0,\Lambda}$  can only simplify the testing problem in the sense of allowing for more powerful tests. Its appeal lies in the fact that the power of the test  $\varphi_\Lambda$  can be easily computed. Thus, Lemma 1 provides a set of explicit power bounds on the original problem, indexed by the distribution  $\Lambda$ .

RUNNING EXAMPLE—ctd: Suppose  $\rho = \text{corr}(Y_\beta, Y_\delta) = -1/2$  in the running example, and consider maximizing weighted average power for the degenerate distribution  $F$  that puts all mass at  $\theta_1 = (\beta, \delta)' = (1, 0)'$ . Further, choose  $\Lambda$  as a degenerate distribution with all its mass at  $\theta_0 = (0, 1)'$ . The likelihood ratio test  $\varphi_\Lambda$  of  $H_{0,\Lambda}$  against  $H_{1,F}$  then rejects for large values of  $Y_\beta - Y_\delta$ . Since  $Y_\beta - Y_\delta | H_{0,\Lambda} \sim \mathcal{N}(-1, 3)$ ,  $\varphi_\Lambda(y) = \mathbf{1}[y_\beta - y_\delta > 1.85]$ , where the critical value 1.85 is chosen to produce a rejection probability of 5% under  $H_{0,\Lambda}$ . Note that  $\varphi_\Lambda$  is not a valid 5% level test of  $H_0: \beta = 0, \delta \geq 0$ , since it has a rejection probability greater than 5% when  $\delta < 1$ . Under the alternative,  $Y_\beta - Y_\delta | H_{1,F} \sim \mathcal{N}(1, 3)$ , so that the power of  $\varphi_\Lambda$  is given by  $\int \varphi_\Lambda g d\nu = 0.31$ . While  $\varphi_\Lambda$  may not control size under  $H_0$ , Lemma 1 implies that any 5% level test of  $H_0: \beta = 0, \delta \geq 0$  against  $H_{1,F}$  has power that does not exceed 0.31.

Lemma 1 follows directly from the arguments leading to a standard result concerning tests with a composite null hypothesis; see, for instance, Theorem 3.8.1 of Lehmann and Romano (2005): A distribution  $\Lambda^\dagger$  is *least favorable* if the best level  $\alpha$  test of  $H_{0,\Lambda^\dagger}$  against the single alternative  $H_{1,F}$  is also of level

$\alpha$  in the testing problem with the composite null hypothesis  $H_0$  against  $H_{1,F}$ , so that—using the same reasoning as in the proof of Lemma 1—this test is also the best test of  $H_0$  against  $H_{1,F}$ . In contrast to this standard result, Lemma 1 is formulated without any restriction on the probability distribution  $\Lambda$ . This is useful because in many contexts, it is difficult to identify the least favorable distribution  $\Lambda^\dagger$ .

#### 2.4. Using the Power Bound to Gauge Potential Efficiency of ad hoc Tests

It is sometimes possible to construct an ad hoc test  $\varphi_{\text{ah}}$  of (3) that is known to be of level  $\alpha$ , even if the nuisance parameter space is high dimensional, but  $\varphi_{\text{ah}}$  has no optimality property by construction. The power bounds from Lemma 1 can then be used to check its efficiency: if the (weighted average) power of  $\varphi_{\text{ah}}$  is close to the power bound arising from some distribution  $\Lambda$ , then  $\varphi_{\text{ah}}$  is known to be close to optimal, as no substantially more powerful test exists. The check is partial, though, as a large difference between the power of  $\varphi_{\text{ah}}$  and the bound can arise either because  $\varphi_{\text{ah}}$  is inefficient, or because this specific  $\Lambda$  yields a bound far above the least upper bound.

For this strategy to work, one must try to guess a  $\Lambda$  that yields a low power bound. Intuitively, a low power bound arises if the density of  $Y$  under  $H_{0,\Lambda}$  is close to the density  $g$  under the alternative  $H_{1,F}$ . This may suggest a suitable choice of  $\Lambda$  directly. Alternatively, one can parameterize  $\Lambda$  in some suitable fashion, and numerically minimize some convenient distance between  $\int f_\theta d\Lambda(\theta)$  and  $g$ . For example, the testing problem of Müller and Watson (2013b) involves hypotheses about the covariance matrix of a mean zero multivariate normal, which under the null hypothesis is a function of a high dimensional nuisance parameter  $\delta \in D$ . With  $\Lambda = \Lambda_\delta$  restricted to put point mass at some  $\delta$ , one can use the Kullback–Leibler divergence between the null and alternative density as a convenient distance function, and use numerical methods to find  $\Lambda_\delta$ . In that application, the resulting power bound comes close to the power of a particular ad hoc test, which shows that the ad hoc test is close to efficient, and also that the power bound computed in this fashion is close to the least power bound. As a second example, Andrews, Moreira, and Stock (2008) showed that Moreira’s (2003) CLR test almost achieves the power bound in a weak instrument IV testing problem, and thus is nearly optimal in that context.

#### 2.5. Approximately Least Favorable Distributions

The least favorable distribution  $\Lambda^\dagger$  has the property that the size  $\alpha$  Neyman–Pearson test  $\varphi_{\Lambda^\dagger}$  of the simple hypothesis  $H_{0,\Lambda^\dagger}$  against  $H_{1,F}$  also yields a level  $\alpha$  test of the composite null hypothesis  $H_0$  against  $H_{1,F}$ . As noted above, for many problems it is difficult to analytically determine  $\Lambda^\dagger$ . A natural reaction is then to try to numerically approximate  $\Lambda^\dagger$ . In many problems, however, it

is nontrivial to approximate  $\Lambda^\dagger$  arbitrarily well, as its definition depends on the typically unbounded number of constraints  $\int \varphi_{\Lambda^\dagger} f_\theta d\nu \leq \alpha$  for all  $\theta \in \Theta_0$ . To ease the computational burden, it would be useful to be able to determine whether a potentially coarse approximation to  $\Lambda^\dagger$  is good enough in terms of generating a test with near optimal power.

Lemma 1 is very helpful in this regard. Specifically, consider the following definition of an approximate least favorable distribution (ALFD).

DEFINITION 1: An  $\varepsilon$ -ALFD is a probability distribution  $\Lambda^*$  on  $\Theta_0$  satisfying

- (i) the Neyman–Pearson test (9) with  $\Lambda = \Lambda^*$  and  $(\text{cv}, \kappa) = (\text{cv}^*, \kappa^*)$ ,  $\varphi_{\Lambda^*}$ , is of size  $\alpha$  under  $H_{0,\Lambda^*}$ , and has power  $\bar{\pi}$  against  $H_{1,F}$ ;
- (ii) there exists  $(\text{cv}^{*\varepsilon}, \kappa^{*\varepsilon})$  such that the test (9) with  $\Lambda = \Lambda^*$  and  $(\text{cv}, \kappa) = (\text{cv}^{*\varepsilon}, \kappa^{*\varepsilon})$ ,  $\varphi_{\Lambda^*}^\varepsilon$ , is of level  $\alpha$  under  $H_0$ , and has power of at least  $\bar{\pi} - \varepsilon$  against  $H_{1,F}$ .

Suppose that a suitable  $\varepsilon$ -ALFD can be identified, where  $\varepsilon$  is small. By (ii),  $\varphi_{\Lambda^*}^\varepsilon$  is a level  $\alpha$  test under  $H_0$ , and by (i), (ii), and Lemma 1, it has power that is within  $\varepsilon$  of the power bound. Thus  $\varphi_{\Lambda^*}^\varepsilon$  is a nearly optimal test of  $H_0$  against  $H_{1,F}$ .

Crucially, the demonstration of near-optimality of  $\varphi_{\Lambda^*}^\varepsilon$  only requires the rejection probability of  $\varphi_{\Lambda^*}^\varepsilon$  under  $H_0$  and the rejection probabilities of  $\varphi_{\Lambda^*}$  and  $\varphi_{\Lambda^*}^\varepsilon$  under  $H_{1,F}$ , respectively. Thus, the argument is *not* based on the notion that  $\Lambda^*$  is necessarily a good approximation to the actual least favorable distribution  $\Lambda^\dagger$  (should it exist) in some direct sense. Rather, any  $\Lambda^*$  that satisfies the two parts of Definition 1 yields a demonstrably nearly optimal test  $\varphi_{\Lambda^*}^\varepsilon$  of  $H_0$  against  $H_{1,F}$ .

## 2.6. A Decision Theoretic Interpretation of the ALFD

From a decision theoretic perspective, most powerful tests for composite hypotheses are related to minimax rules in the problem of deciding between  $H_0$  against  $H_{1,F}$  (Blackwell and Girshick (1954, Chapter 7.7)), and the test based on the ALFD is an approximate minimax decision rule.

To be precise, consider the hypothesis testing problem (8) as a decision problem where a false rejection of  $H_0$  induces a loss of 1, a false rejection of  $H_1$  induces a loss of  $\phi > 0$ , and correct decision has a loss of 0. The (frequentist) risk of the decision rule  $\varphi$  then is  $R_0(\varphi, \theta) = \int \varphi f_\theta d\nu$ ,  $\theta \in \Theta_0$  under  $H_0$  and  $R_1(\varphi) = \phi \int (1 - \varphi) g d\nu$  under  $H_{1,F}$ . The largest risk under  $H_0$ ,  $\alpha(\varphi) = \sup_{\theta \in \Theta_0} \int \varphi f_\theta d\nu$  is recognized as the size of the test  $\varphi$ , and the power of the test under  $H_{1,F}$ ,  $\pi(\varphi)$  equals  $1 - R_1(\varphi)/\phi$ . Nature's strategies consist of drawing  $Y$  from  $H_0$  with probability  $0 \leq q \leq 1$ , and from  $H_1$  with probability  $1 - q$  and, conditional on drawing from  $H_0$ , drawing  $\theta$  from  $\Lambda$ , a probability distribution with support in  $\Theta_0$ . An adversarial nature seeks to maximize expected risk, that is, to choose  $(q, \Lambda)$  that maximize  $r(\varphi, q, \Lambda) =$

$q \int R_0(\varphi, \theta) d\Lambda(\theta) + (1 - q)R_1(\varphi)$ . Note that for any  $(q, \Lambda)$ ,  $\inf_{\varphi} r(\varphi, q, \Lambda) \leq \inf_{\varphi} \max(\sup_{\theta \in \Theta_0} R_0(\varphi, \theta), R_1(\varphi)) = \bar{V}$ , the econometrician's minimax risk.

For any prior  $(q, \Lambda)$ , the posterior probabilities are  $p_{q,\Lambda}^0(y) = q \times \int f_{\theta}(y) d\Lambda(\theta)/f(y)$  under  $H_0$  and  $p_{q,\Lambda}^1(y) = (1 - q)g(y)/f(y)$  under  $H_{1,F}$ , where the marginal likelihood  $f(y)$  is given by  $f(y) = q \int f_{\theta}(y) d\Lambda(\theta) + (1 - q)g(y)$ . The posterior expected loss of decision  $\varphi$  equals  $\varphi(y)p_{q,\Lambda}^0(y) + \phi(1 - \varphi(y))p_{q,\Lambda}^1(y)$ , so that Bayes rules are of the form

$$\varphi^B(q, \Lambda) = \begin{cases} 1 & \text{if } g(y) > \frac{q}{\phi(1-q)} \int f_{\theta}(y) d\Lambda(\theta), \\ \kappa & \text{if } g(y) = \frac{q}{\phi(1-q)} \int f_{\theta}(y) d\Lambda(\theta), \quad 0 \leq \kappa \leq 1, \\ 0 & \text{if } g(y) < \frac{q}{\phi(1-q)} \int f_{\theta}(y) d\Lambda(\theta), \end{cases}$$

mimicking the form of the Neyman–Pearson test in (9). Minimizing posterior expected loss for each draw  $y$  yields a decision that minimizes prior weighted risk, that is,  $\inf_{\varphi} r(\varphi, q, \Lambda) = r(\varphi^B(q, \Lambda), q, \Lambda)$ .

Now consider an  $\varepsilon$ -ALFD  $\Lambda^*$  in the sense of Definition 1, and set  $\phi = \alpha/(1 - \bar{\pi})$ . Then  $R_1(\varphi_{\Lambda^*}) = \alpha$ , and since  $\varphi_{\Lambda^*}$  is of level  $\alpha$  under  $H_{0,\Lambda^*}$ ,  $\int R_0(\varphi_{\Lambda^*}, \theta) d\Lambda^*(\theta) = \alpha$ . Furthermore, let  $0 \leq q^* \leq 1$  solve  $\frac{q^*}{1-q^*} = \phi cv^*$ , so that  $\varphi_{\Lambda^*} = \varphi^B(q^*, \Lambda^*)$ . Thus  $\bar{V} \geq \inf_{\varphi} r(\varphi, q^*, \Lambda^*) = r(\varphi_{\Lambda^*}, q^*, \Lambda^*) = q^*\alpha + (1 - q^*)\alpha = \alpha$ . By definition of an  $\varepsilon$ -ALFD, the adjusted test  $\varphi_{\Lambda^*}^{\varepsilon}$  has size  $\alpha$  and power within  $\varepsilon$  of  $\bar{\pi}$ . Thus  $\sup_{\theta \in \Theta_0} R_0(\varphi_{\Lambda^*}^{\varepsilon}, \theta) \leq \alpha$  and  $R_1(\varphi_{\Lambda^*}^{\varepsilon}) \leq \alpha + \alpha\varepsilon/(1 - \bar{\pi})$ , so that the maximal risk of  $\varphi_{\Lambda^*}^{\varepsilon}$  exceeds the lower bound of  $\alpha$  by at most  $\alpha\varepsilon/(1 - \bar{\pi})$ . In this sense,  $\varphi_{\Lambda^*}^{\varepsilon}$  is an approximate minimax decision rule.

Minimax rules are inherently pessimistic, and they might be considered unattractive if they are rationalized by an unreasonable distribution for  $\theta$ . This can be assessed for a given test  $\varphi_{\Lambda^*}^{\varepsilon}$  derived with the algorithm developed in the next section by inspecting the ALFD  $\Lambda^*$ . From a Bayesian perspective, the ALFD might be used as a prior selection device, which guarantees attractive frequentist properties of Bayes rule  $\varphi_{\Lambda^*}^{\varepsilon} = \varphi^B(q^{*\varepsilon}, \Lambda^*)$ , where  $q^{*\varepsilon}$  solves  $\frac{q^{*\varepsilon}}{1-q^{*\varepsilon}} = \phi cv^{*\varepsilon}$ .

The exact least favorable distribution (should it exist) arises naturally in this decision theoretic perspective by considering nature's best strategy: the largest risk that nature can induce is  $\underline{V} = \sup_{q,\Lambda} \inf_{\varphi} r(\varphi, q, \Lambda)$ . If the maximin theorem holds (a sufficient condition is finiteness of  $\Theta_0$ ; see, for instance, Theorem 2.9.1 in Ferguson (1967)), then the game has value  $V = \bar{V} = \underline{V}$ . Furthermore, there exists a least favorable prior  $(q^{\dagger}, \Lambda^{\dagger})$  such that  $\varphi^{\dagger} = \varphi^B(q^{\dagger}, \Lambda^{\dagger})$  achieves the maximin risk  $V$ . Its level  $\alpha(\varphi^{\dagger}) = V = \sup_{\theta \in \Theta_0} \int \varphi^{\dagger} f_{\theta} d\nu$  is endogenously determined by the structure of the problem ( $f_{\theta}$  and  $g$ ) and the loss function parameter  $\phi$ . Also note that  $V = r(\varphi^{\dagger}, q^{\dagger}, \Lambda^{\dagger})$  implies that the least

favorable prior  $(q^\dagger, A^\dagger)$  has positive mass only on points that lead to maximum risk.

### 3. NUMERICAL DETERMINATION OF NEARLY OPTIMAL TESTS

We now discuss the suggested numerical algorithm to determine an  $\varepsilon$ -ALFD for small  $\varepsilon$ , and thus a nearly WAP maximizing test.

In a first step, it is useful to approximate  $H_0$  by a set of finitely many distributions for  $Y$ . A straightforward way to achieve this is to simply discretize  $\Theta_0$  into a finite number of values. As long as  $\Theta_0$  is compact, some continuity implies that controlling size on a fine enough grid on  $\Theta_0$ ,  $\{\theta_i\}_{i=1}^M \subset \Theta_0$  leads to at worst minimally oversized tests under the unrestricted  $H_0$ . Note that discretizing  $\Theta_0$  is only one way to discretize  $H_0$ ; alternatively, one can also specify a finite set of “base” distributions  $\Psi_i$ ,  $i = 1, \dots, M$ , each with support in  $\Theta_0$ , so that the  $M$  possible densities for  $Y$  under  $H_0$  are of the form  $f_i = \int f_\theta d\Psi_i(\theta)$ . This reduces to a grid on  $\Theta_0$  if the  $\Psi_i$  represent point masses. It is advantageous to use nondegenerate  $\Psi_i$ 's, which lead to smoother rejection regions for the same number  $M$ .

With  $H_0$  discretized,  $\Lambda$  is described by a point in the  $M$ -dimensional simplex  $\Lambda = (\lambda_1, \dots, \lambda_M)$ . The Neyman–Pearson test (9) is of the form<sup>3</sup>  $\varphi_\Lambda(y) = \mathbf{1}[g(y) > \text{cv}_\Lambda \sum_{i=1}^M \lambda_i f_i(y)]$ , with critical value  $\text{cv}_\Lambda$  determined by  $\int (\sum_{i=1}^M \lambda_i f_i) \varphi_\Lambda d\nu = \alpha$ . It is convenient to optimize in  $\mathbb{R}^M$  instead of the  $M$ -dimensional simplex, and to subsume the critical value  $\text{cv}_\Lambda$  into the weights. Thus, let  $\mu_i = \ln(\text{cv}_\Lambda \lambda_i) \in \mathbb{R}$  and  $\mu = (\mu_1, \dots, \mu_M)$ , so that  $\varphi_\mu = \mathbf{1}[g > \sum_{i=1}^M \exp(\mu_i) f_i]$ . Our suggested approach is to simply repeatedly adjust  $\mu$  as a function of the rejection probabilities of  $\varphi_\mu$ : start with some  $\mu^{(0)} \in \mathbb{R}^M$ , and for a suitable  $\omega > 0$ , set

$$(10) \quad \mu_j^{(i+1)} = \mu_j^{(i)} + \omega \left( \int \varphi_{\mu^{(i)}} f_j d\nu - \alpha \right)$$

for  $i = 1, \dots, O$ . An iteration of (10) increases the weights  $\exp(\mu_j)$  on those  $f_j$  that lead to overrejections relative to  $\alpha$ , and decreases the weights  $\exp(\mu_j)$  for those that lead to an underrejection. With sufficiently many iterations on (10), the implied  $\hat{\Lambda} \propto (\exp \mu_1^{(O)}, \dots, \exp \mu_M^{(O)})$  then serves as a candidate for an  $\varepsilon$ -ALFD  $\Lambda^*$  in the sense of Definition 1.<sup>4</sup> So it remains to check that the

<sup>3</sup>This holds at least as long as all convex combinations of  $g(Y)$  and  $f_i(Y)$ ,  $i = 1, \dots, M$  have an absolutely continuous distribution, which is the case in all applications we consider.

<sup>4</sup>For this argument,  $\hat{\Lambda}$  does not need to equal the exact LFD  $\Lambda_0^\dagger$  with  $\Theta$  restricted to the grid  $\Theta_0$ . At the same time, [Moreira and Moreira \(2013\)](#) proved that if one were to employ a sequence of finer and finer grids  $\Theta_{0,n}$  on a compact  $\Theta$ , then the power functions of the exact LDF tests  $\varphi_{\Lambda_n^\dagger}$  converge to the power function of the weighted average power maximizing test  $\varphi_{\Lambda^\dagger}$  on the original parameter space  $\Theta$ .

adjusted test  $\varphi_{\lambda}^{\varepsilon}$  controls size under the original null hypothesis. This may be determined by a fine grid search;<sup>5</sup> if that search reveals overrejections, then the algorithm is restarted with a finer discretization of  $H_0$  (larger  $M$ ). More comments on the efficient computation of  $\int \varphi_{\mu^{(i)}} f_j d\nu$  and a step-by-step description of the algorithm are in Appendix A.2.1.

We found that this algorithm reliably generates an  $\varepsilon$ -ALFD in all problems we considered. In our experience, the weights  $\exp(\mu_j^{(O)})$  generated by (10) are numerically fairly insensitive to the starting value  $\mu^{(0)}$ , the tuning parameter  $\omega$ , and the number of iterations  $O$  (as long as  $O$  is larger than, say, 300). Also, in simple problems where this is easy to assess (such as in the running example), different choices for the discretization of  $H_0$  end up yielding tests  $\varphi_{\lambda^*}^{\varepsilon}$  with very similar critical regions. The number of densities in  $H_0$  can be chosen fairly large (say, larger than 100) without much difficulty; in Müller and Watson (2013a), this algorithm was employed to determine an  $\varepsilon$ -ALFD in a problem involving a three-dimensional nuisance parameter, using a discretization with  $M = 204$  (strategically chosen) points.

The literature contains a number of alternative approaches to approximating least favorable distributions. One simple but powerful approach is to exploit the linearity of the rejection probability  $\int \varphi f_{\theta} d\nu$  as a function of  $\varphi$  to obtain a linear programming problem. See Krafft and Witting (1967) for a general discussion of the relationship between least favorable distributions and linear programming. For recent implementations, see Chiburis (2009) and, for more general linearly constrained tests, Moreira and Moreira (2013). A disadvantage is, however, that unless the sample space  $\mathcal{Y}$  can usefully be partitioned into a small number of regions, both the primal and dual of the linear program are of (potentially very) high dimension.

Kempthorne (1987) provided a numerical algorithm to determine maximin rules for general decision problems, which could be adopted to the testing problem described in Section 2.6. But the algorithm involves repeated maximizations of convex combinations of null rejection probabilities, both as a function of the weights, and as a function of null parameter values. A more attractive variant of determining maximin rules for discrete  $H_0$  was developed in Chamberlain (2000): If  $H_0$  consists of  $M$  possible distributions for  $Y$ , then the prior  $(q, \Lambda)$  can be represented as a point in the  $M + 1$ -dimensional simplex. Chamberlain noted that the Bayes risk  $r(\varphi, q, \Lambda)$  of the Bayes rule  $\varphi = \varphi^B(q, \Lambda)$  is a concave function on this simplex. The least favorable distribution  $(q^\dagger, \Lambda^\dagger)$  maximizes Bayes risk, so that it can be determined using concave programming. The level of the resulting maximin test is an endogenous function of the loss function parameter  $\phi$ , however, so that the least favorable distribution would have to be determined repeatedly to obtain a test for a given level  $\alpha$ .

<sup>5</sup>Elliott and Müller (2012) developed a technique for numerically checking size control without discretization of  $H_0$ .

A third approach is based on the observation that the least favorable prior  $\Lambda^\dagger$  puts all its mass on the values of  $\theta \in \Theta_0$  where  $\varphi_{\Lambda^\dagger}$  has rejection probability equal to  $\alpha$ . Thus, with  $H_0$  discretized, the problem of determining  $\Lambda^\dagger = (\lambda_1^\dagger, \dots, \lambda_M^\dagger)$  can be cast as the nonlinear complementarity problem (NCP)  $\lambda_i^\dagger \geq 0$ ,  $\alpha - \int \varphi_{\Lambda^\dagger} f_i d\nu \geq 0$  and  $\lambda_i^\dagger (\alpha - \int \varphi_{\Lambda^\dagger} f_i d\nu) = 0$ ,  $i = 1, \dots, M$ . See [Facchinei and Pang \(2003\)](#) for a comprehensive overview and references in the literature on algorithms for NCPs. One possibility is to cast the NCP into a constrained nonlinear optimization problem, such as  $\min_{\Lambda} \sum_i \mathbf{1}[\lambda_i > 0] \lambda_i^2 (\alpha - \int \varphi_{\Lambda} f_i d\nu)^2$  subject to  $\lambda_i \geq 0$  and  $\sum_i \lambda_i = 1$ . The algorithm of [Sriananthakumar and King \(2006\)](#) builds on this approach.

#### 4. SWITCHING TESTS

In most problems of interest, the nuisance parameter space is unbounded. This raises issues for the approach discussed so far under both the null and alternative hypothesis. Under the null hypothesis, discretizations of the null parameter space for noncompact  $\Theta_0$  are inherently coarse, complicating the algorithm discussed in the last section. Under the alternative, attempting to use a WAP criterion for a noncompact  $\Theta_1$  faces the problem that any integrable  $F$  puts almost all weight on a compact subset. Thus, any such  $F$  effectively puts very little weight on some large region of  $\Theta_1$ , leading to potentially poor power properties in that region.

To address this issue, note that for many nonstandard testing problems involving a nuisance parameter  $\delta \in \mathbb{R}$ , one can choose a parameterization in which the testing problem for large values of  $\delta$  essentially reduces to a standard problem. For example, in the weak instrument problem with concentration parameter  $\delta$ , a large  $\delta$  implies that the instruments are “almost” strong; inference problems involving a local-to-unity parameter  $\delta \geq 0$ , such as predictive regressions studied in [Cavanagh, Elliott, and Stock \(1995\)](#) and [Jansson and Moreira \(2006\)](#), essentially reduce to standard stationary time series problems as  $\delta \rightarrow \infty$ ; and similarly, in our running example, the problem becomes standard as  $\delta \rightarrow \infty$ .

It seems sensible, therefore, to consider tests that essentially reduce to the “standard best test” in the standard problem when  $\delta$  is large, and to employ the weighted average power criterion only to ensure near-optimality for small  $\delta$ . This has the additional advantage that size control only needs to be carefully checked in the nonstandard case of small  $\delta$ .

We proceed in three steps: First, we formally discuss the convergence to a standard problem. Second, we determine the nearly WAP maximizing tests in a class of tests that, by a functional form restriction, have nearly the same rejection properties as the standard best test for  $\delta$  large. Third, we examine whether the restriction to this class is costly in terms of weighted average power.



4.1. *Limiting Problem*

Consider the testing problem (4) under a reparameterization of  $\theta = (\beta', \delta')' \in \mathbb{R}^{k_\beta+1}$  in terms of  $\Delta, d \in \mathbb{R}$ , and  $b \in \mathbb{R}^{k_\beta}$  via  $\delta = r_\delta(\Delta, d)$  and  $\beta = r_\beta(\Delta, b)$ , where  $\Delta \rightarrow \infty$  implies  $\delta \rightarrow \infty$  for fixed  $d$ , and  $b = 0$  implies  $\beta = \beta_0$ . Think of  $\Delta$  as some approximate baseline value of  $\delta$ , and of  $d$  as the deviation of  $\delta$  from this baseline. Now construct a sequence of testing problems, indexed by  $n$ , by setting  $\Delta = \Delta_n$ , but for fixed  $h = (b, d) \in \mathbb{R}^{k_\beta+1}$ . Denote by  $f_{n,h}$  the density of  $Y$  in this parameterization (and define it arbitrarily if the implied  $\theta \notin \Theta_0 \cup \Theta_1$ ). Further, let  $X \in \mathcal{X}$  be a random element with density  $f_{X,h}$  relative to some dominating measure, and  $h \in \mathbb{R}^{k_\beta+1}$ . The following condition provides sufficient assumptions to ensure convergence as  $\Delta_n \rightarrow \infty$  to the experiment of observing the single random element  $X$ .<sup>6</sup>

CONDITION 1: (i) Suppose that  $f_{X,h_1}$  is absolutely continuous relative to  $f_{X,h_2}$ , for all  $h_1, h_2 \in \mathbb{R}^{k_\beta+1}$ .

(ii) For all sequences  $\Delta_n \rightarrow \infty$  and fixed finite subsets  $H \subset \mathbb{R}^{k_\beta+1}$ ,

$$(11) \quad \left\{ \frac{f_{n,h}(Y)}{f_{n,0}(Y)} \right\}_{h \in H} \Rightarrow \left\{ \frac{f_{X,h}(X)}{f_{X,0}(X)} \right\}_{h \in H},$$

where  $Y$  and  $X$  are distributed according to  $f_{n,0}$  and  $f_{X,0}$ , respectively.

The experiment involving  $X$  is typically much easier than that involving  $Y$ ; in most of our applications, it is simply an unrestricted Gaussian shift experiment (2). Denote by  $\varphi_S^{\text{lim}} : \mathcal{X} \mapsto [0, 1]$  the “standard best test” of level  $\alpha$  in the limiting problem of

$$(12) \quad H_0 : b = 0, \quad d \in \mathbb{R} \quad \text{against} \quad H_1 : b \neq 0, \quad d \in \mathbb{R}$$

based on the single observation  $X$ . Suppose further that there exists a test  $\varphi_S : \mathcal{Y} \mapsto [0, 1]$  in the original problem that has the same asymptotic rejection properties as  $\varphi_S^{\text{lim}}$  as  $\Delta_n \rightarrow \infty$  for all fixed values of  $h$ .

RUNNING EXAMPLE: Set  $r_\beta(\Delta_n, b) = b$  and  $r_\delta(\Delta_n, d) = \Delta_n + d$ . For any fixed  $h = (b, d) \in \mathbb{R}^2$  and all  $\Delta_n \geq -d$ ,

$$(13) \quad \log \frac{f_{n,h}(Y)}{f_{n,0}(Y)} = \begin{pmatrix} Y_\beta \\ Y_\delta - \Delta_n \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} b \\ d \end{pmatrix} - \frac{1}{2} \begin{pmatrix} b \\ d \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} b \\ d \end{pmatrix}.$$

<sup>6</sup>See the proof of Lemma 2 for details on how Condition 1 implies convergence of experiments.



Under  $h = 0$ , for any  $\Delta_n > 0$ ,  $(Y_\beta, Y_\delta - \Delta_n)' \sim \mathcal{N}(0, \Sigma)$ . One can therefore apply the reasoning in the proof of Theorem 9.4 in van der Vaart (1998) to show that (11) holds with

$$(14) \quad X = \begin{pmatrix} X_b \\ X_d \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ d \end{pmatrix}, \Sigma\right).$$

As discussed in the **Introduction**, the usual (admissible) test of  $H_0 : b = 0$ ,  $d \in \mathbb{R}$  versus  $H_1 : |b| > 0$ ,  $d \in \mathbb{R}$  in this limiting problem is of the form  $\varphi_S^{\text{lim}}(X) = \mathbf{1}[|X_b| > cv]$ . The power properties of this test are obtained in the original problem as  $\Delta_n \rightarrow \infty$  by the test  $\varphi_S(Y) = \mathbf{1}[|Y_\beta| > cv]$ .

**WEAK INSTRUMENTS EXAMPLE:** Consider inference in a linear regression with a single endogenous variable and a weak instrument. In Hillier’s (1990) and Chamberlain’s (2007) parameterization, the problem becomes inference about  $\beta \in (-\pi, \pi] = B$  based on the bivariate observation

$$(15) \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \delta \sin \beta \\ \delta \cos \beta \end{pmatrix}, I_2\right)$$

with  $\delta \geq 0$  measuring the strength of the instrument. Now consider the hypothesis test  $H_0 : \beta = \beta_0 < \pi$  against  $H_a : \beta \neq \beta_0$ , and set  $\delta = \Delta_n + d$  and  $\beta = \beta_0 + b\Delta_n^{-1}$ .<sup>7</sup> Then under  $h = 0$ , as  $\Delta_n \rightarrow \infty$ ,

$$\begin{aligned} \log \frac{f_{n,h}(Y)}{f_{n,0}(Y)} &= -\frac{1}{2} \left\| \begin{pmatrix} Y_1 - (\Delta_n + d) \sin(\beta_0 + b\Delta_n^{-1}) \\ Y_2 - (\Delta_n + d) \cos(\beta_0 + b\Delta_n^{-1}) \end{pmatrix} \right\|^2 \\ &\quad + \frac{1}{2} \left\| \begin{pmatrix} Y_1 - \Delta_n \sin \beta_0 \\ Y_2 - \Delta_n \cos \beta_0 \end{pmatrix} \right\|^2 \\ &\Rightarrow \begin{pmatrix} Z_b \\ Z_d \end{pmatrix}' \begin{pmatrix} b \\ d \end{pmatrix} - \frac{1}{2} \begin{pmatrix} b \\ d \end{pmatrix}' \begin{pmatrix} b \\ d \end{pmatrix}, \end{aligned}$$

with  $(Z_b, Z_d)' \sim \mathcal{N}(0, I_2)$ , since  $\Delta_n(\sin(\beta_0 + b\Delta_n^{-1}) - \Delta_n \sin \beta_0) \rightarrow b \cos \beta_0$ ,  $d \sin(\beta_0 + b\Delta_n^{-1}) \rightarrow d \sin \beta_0$ ,  $\Delta_n(\cos(\beta_0 + b\Delta_n^{-1}) - \cos \beta_0) \rightarrow -b \sin \beta_0$ ,  $d \cos(\beta_0 + b\Delta_n^{-1}) \rightarrow d \cos \beta_0$ , and  $(Y_1 - \Delta_n \sin \beta_0, Y_2 - \Delta_n \cos \beta_0) \sim \mathcal{N}(0, I_2)$ , so (11) holds with  $X$  as defined in (14) and  $\Sigma = I_2$  by Theorem 9.4 of van der Vaart (1998). The power properties of the standard test  $\varphi_S^{\text{lim}}(X) = \mathbf{1}[|X_b| > cv]$  are obtained in the original problem as  $\Delta_n \rightarrow \infty$  by the test  $\varphi_S(Y) = \mathbf{1}[\hat{\delta}|\hat{\beta} - \beta_0| > cv]$ , where  $(\hat{\beta}, \hat{\delta})$  is the MLE in (15).

<sup>7</sup>The restriction to  $\beta_0 < \pi$  is only for notational convenience; the result still goes through with  $\beta_0 = \pi$  by setting  $\beta = -\pi + b\Delta_n^{-1} \in B$  if  $b > 0$ .

The convergence of experiments implies that one cannot systematically outperform  $\varphi_S$  for large  $\delta$ , at least as long as  $\varphi_S^{\text{lim}}$  is an admissible test in the limiting problem, as the following lemma makes precise. Write  $E_{(b,\Delta)}[\cdot]$  for integration with respect to the density of  $Y$  in the parameterization  $\delta = r_\delta(\Delta, 0)$  (i.e.,  $d = 0$ ) and  $\beta = r_\beta(\Delta, b)$  defined above, and  $E_h[\cdot]$  for integration with respect to  $f_{X,h}$ .

LEMMA 2: Let  $\varphi_S^{\text{lim}}$  be a level  $\alpha$  test of (12) with rejection probability  $E_{(b,d)}[\varphi_S^{\text{lim}}(X)]$  that does not depend on  $d$ , and assume that  $\varphi_S^{\text{lim}}$  is admissible in the sense that  $E_{(b_1,d)}[\varphi^{\text{lim}}(X)] > E_{(b_1,d)}[\varphi_S^{\text{lim}}(X)]$  for some  $b_1 \neq 0$  and level  $\alpha$  test  $\varphi^{\text{lim}}$  of (12) implies existence of  $b_2 \neq 0$  where  $E_{(b_2,d)}[\varphi^{\text{lim}}(X)] < E_{(b_2,d)}[\varphi_S^{\text{lim}}(X)]$ . Under Condition 1, for any level  $\alpha$  test  $\varphi$  in the original problem (4) and any  $b_1 \neq 0$ ,

$$\limsup_{\Delta \rightarrow \infty} (E_{(b_1,\Delta)}[\varphi(Y)] - E_{(b_1,\Delta)}[\varphi_S(Y)]) > 0$$

implies the existence of  $b_2 \neq 0$  such that

$$\liminf_{\Delta \rightarrow \infty} (E_{(b_2,\Delta)}[\varphi(Y)] - E_{(b_2,\Delta)}[\varphi_S(Y)]) < 0.$$

#### 4.2. Switching to Standard Tests

The convergence to a “standard” experiment as  $\delta \rightarrow \infty$  via Condition 1 suggests that the part of the alternative parameter space  $\Theta_1$  with  $\delta$  large essentially corresponds to a standard testing problem. Accordingly, partition  $\Theta_1$  into a subset  $\Theta_{1,S}$  corresponding to  $\delta > \kappa_\delta$  for some large  $\kappa_\delta$ , and a subset  $\Theta_{1,N}$  that is “nonstandard”  $\delta \leq \kappa_\delta$ , so that

$$\Theta_1 = \Theta_{1,S} \cup \Theta_{1,N}.$$

The convergence to the experiment involving  $X$  discussed above implies that for large enough  $\kappa_\delta$ , the test  $\varphi_S$  has a rejection profile arbitrarily close to  $\varphi_S^{\text{lim}}$ . It thus makes sense to pick  $\kappa_\delta$  just large enough for this to become a good enough approximation. Furthermore, in light of Lemma 2, it is then sensible to restrict attention to tests that have (nearly) the same power as  $\varphi_S$  on  $\Theta_{1,S}$ , so the WAP criterion is only relevant for  $\Theta_{1,N}$ .

A straightforward strategy to achieve power that nearly equals the power of  $\varphi_S$  on  $\Theta_{1,S}$  is to consider tests of the following “switching” form:

$$(16) \quad \varphi_{N,S,\chi}(y) = \chi(y)\varphi_S(y) + (1 - \chi(y))\varphi_N(y),$$

where  $\varphi_S$  is the standard test from the limiting problem and  $\chi: \mathcal{Y} \mapsto [0, 1]$  is a switching function chosen so that  $\varphi_{N,S,\chi}(y) = \varphi_S(y)$  with probability close to 1 for all  $\theta \in \Theta_{1,S}$ . In our applications, we choose  $\chi$  of the form  $\chi(y) = \mathbf{1}[\hat{\delta} > \kappa_\chi]$ ,

for some  $\kappa_\chi$ . The value of  $\kappa_\chi$  is chosen sufficiently smaller than  $\kappa_\delta$  so that with probability very close to 1,  $\hat{\delta} > \kappa_\chi$  whenever  $\delta > \kappa_\delta$ .<sup>8</sup>

The function  $\varphi_N: \mathcal{Y} \mapsto [0, 1]$  is determined to maximize WAP relative to  $F$ . A straightforward generalization of last section’s results for  $\varepsilon$ -ALFD tests yields the approximately best weighted average test, say  $\varphi_{A^*,S,\chi}^\varepsilon(y)$ , for tests of the form (16). Details are provided in Appendix A. In the next subsection, we consider the efficiency of  $\varphi_{A^*,S,\chi}^\varepsilon$  within a wider class of tests than (16), but before that discussion it is useful to return to the running example.

**RUNNING EXAMPLE—ctd:** We choose  $\kappa_\delta = 9$ , so that  $\Theta_{1,N}$  consists of parameter values  $\theta = (\beta, \delta)$  with  $\beta \neq 0$  and  $0 \leq \delta \leq 9$ . The weighting function  $F$  is such that  $\delta$  is uniformly distributed on  $[0, 9]$ , and  $\beta$  takes on the values  $-2$  and  $2$  with equal probability. The two alternatives are thus treated symmetrically and the values  $\pm 2$  are chosen so that the test achieves approximately 50% power (following the rationale given in King (1987)). The switching function is specified as  $\chi(y) = \mathbf{1}[\hat{\delta} > 6]$ , where  $\hat{\delta} = y_\delta$ .

Figure 1 summarizes the results for the running example with  $\rho = 0.7$  for tests of level  $\alpha = 5\%$ . The ALFD was computed using  $\varepsilon = 0.005$ , so that the power of the nearly optimal tests differs from the power bound by less than 0.5 percentage points. The number of Monte Carlo draws under the null and alternative hypotheses in the algorithm are chosen so that Monte Carlo standard errors are approximately 0.1%. Panel A shows results for  $\varphi_{A^*,S,\chi}^\varepsilon$ , and for comparison, panel B shows results for the WAP maximizing test with the same weighting function  $F$  with support on  $\Theta_{1,N}$ , but without the constraint to the switching form (16).

The white and light gray band in the center of panel A.1 is the acceptance region of the nearly optimal test  $\varphi_{A^*,S,\chi}^\varepsilon$ , with the light gray indicating the acceptance region conditional on switching ( $|y_\beta| \leq 1.96$  and  $y_\delta \geq 6$ ). The dark shades show the critical region, with the darker shade indicating the critical region conditional on switching ( $|y_\beta| > 1.96$  and  $y_\delta \geq 6$ ). The critical region is seen to evolve smoothly as the test switches at  $y_\delta = 6$ , and essentially coincides with the standard test  $\varphi_S$  for values of  $y_\delta$  as small as  $y_\delta = 3$ . This suggests that other choices of  $\kappa_\delta$  and  $\kappa_\chi$  (and corresponding  $F$  with  $\delta$  uniform on  $[0, \kappa_\delta]$ ) would yield similar WAP maximizing tests; unreported results show that this is indeed the case. As  $y_\delta$  becomes negative, the critical region is approximately  $|y_\beta - \rho y_\delta| > 1.96\sqrt{1 - \rho^2}$ , which is recognized as the critical region of the uniformly best unbiased test for  $\delta = 0$  known.

Panel A.2 shows power (plotted as a function of  $\delta$ ) for selected values of  $\beta$ . The solid curves show the power of the nearly optimal test and the dashed

<sup>8</sup>A necessary condition for tests of the form (16) to control size is that the “pure switching test”  $\tilde{\varphi} = \chi\varphi_S$  (i.e., (16) with  $\varphi_N = 0$ ) is of level  $\alpha$ . Depending on the problem and choice of  $\kappa_\chi$ , this might require a slight modification of the most natural choice for  $\varphi_S$  to ensure actual (and not just approximate) size control of  $\varphi_S$  for all  $\delta > \kappa_\delta$ .

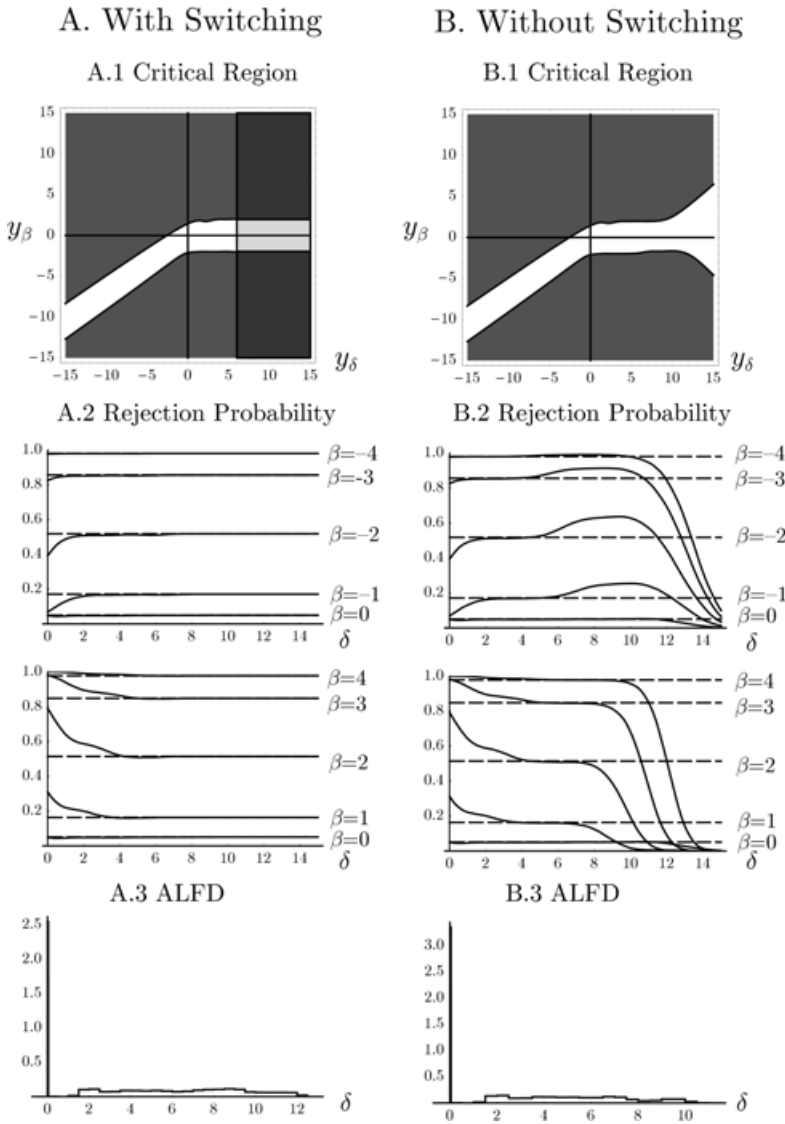


FIGURE 1.—Positive nuisance parameter. Darker shades for  $y_\delta \geq 6$  in panel A.1 indicate the part of the acceptance and critical region imposed by the switching rule. In panels A.2 and B.2, solid lines are the rejection probability of the nearly optimal tests  $\varphi_{\Lambda^*, S, \chi}^e$  (panel A) and  $\varphi_{\Lambda^*}^e$  (panel B), and dashed lines are for the usual test that ignores  $Y_\delta$ ,  $\varphi_S(y) = \mathbf{I}[|y_\beta| > 1.96]$ .

lines show the power of the standard test  $\varphi_S$ . The figures show that power is asymmetric in  $\beta$ , with substantially lower power for negative values of  $\beta$  when  $\delta$  is small; this is consistent with the critical region shown in panel A.1 where negative values of  $\beta$  and small values of  $\delta$  make it more likely that  $y$  falls in the lower left quadrant of panel A.1. Because weighted average power is computed for uniformly distributed  $\beta \in \{-2, 2\}$  and  $\delta \in [0, 9]$ , the optimal test maximizes the average of the power curves for  $\beta = -2$  and  $\beta = 2$  in panel A.2 over  $\delta \in [0, 9]$ . Weighted average power of  $\varphi_{A^*,S,\chi}^\varepsilon$  is higher than the power of  $\varphi_S$  for all pairs of values for  $\beta$  shown in the figure.

Panels B show corresponding results for the nearly optimal test  $\varphi_{A^*}^\varepsilon$  that does not impose switching to a standard test, computed using the algorithm as described in Section 3. Because  $F$  only places weight on values of  $\delta$  that are less than 9, this test sacrifices power for values of  $\delta > 9$  to achieve more power for values of  $\delta \leq 9$ . The differences between the power function for  $\varphi_{A^*,S,\chi}^\varepsilon$  (shown in panel A) and  $\varphi_{A^*}^\varepsilon$  (shown in panel B) highlight the attractiveness of switching to a standard test: it allows  $F$  to be chosen to yield high average power in the nonstandard portion of the parameter space (small values of  $\delta$ ) while maintaining good power properties in other regions.

Panels A.3 and B.3 show the ALFDs underlying the two tests, which are mixtures of uniform baseline densities  $f_i$  used in the calculations. We emphasize that the ALFDs are not direct approximations to the least favorable distributions, but rather are distributions that produce tests with nearly maximal weighted average power.

#### 4.3. Power Bounds Under Additional Constraints on Power

The test  $\varphi_{A^*,S,\chi}^\varepsilon$  (nearly) coincides with  $\varphi_S$  when  $\theta \in \Theta_{1,S}$ , and thus is (nearly) as powerful as the standard best test  $\varphi_S^{\text{lim}}$  in the limiting problem in that part of the parameter space; moreover,  $\varphi_{A^*,S,\chi}^\varepsilon$  comes close to maximizing WAP on  $\Theta_{1,N}$  among all tests of the form (16). A natural question is whether this class is restrictive, in the sense that there exist tests outside this class that have better WAP on  $\Theta_{1,N}$ . We investigate this in this section by computing a WAP upper bound for any test that satisfies the level constraint and achieves prespecified power on  $\Theta_{1,S}$ .

To begin, decompose the alternative into two hypotheses corresponding to  $\Theta_{1,N}$  and  $\Theta_{1,S}$ :

$$H_{1,N} : \theta \in \Theta_{1,N} \quad \text{and} \quad H_{1,S} : \theta \in \Theta_{1,S}.$$

Let  $\pi_S(\theta) = \int \chi \varphi_S f_\theta dv$  denote the rejection frequency for the test  $\tilde{\varphi} = \chi \varphi_S$ , where, as above,  $\chi$  is the switching function and  $\varphi_S$  is the standard test. The test  $\tilde{\varphi}$  is a useful benchmark for the performance of any test on  $\Theta_{1,S}$ , since it is a feasible level  $\alpha$  test of  $H_0$  against  $H_1$  with power  $\pi_S(\theta)$  under  $H_{1,S}$  that is very close to the power of the standard test  $\varphi_S$ , which in turn has power close to the

admissible test  $\varphi_S^{\text{lim}}$  in the limiting problem. Now, consider tests that (a) are of level  $\alpha$  under  $H_0$ , (b) maximize WAP under  $H_{1,N}$  relative to the weighting function  $F$  in (8), and (c) achieve power of at least  $\pi_S(\theta)$  under  $H_{1,S}$ .<sup>9</sup> Using an argument similar to that underlying Lemma 1, a bound on the power of level  $\alpha$  tests of  $H_0$  against  $H_{1,F}$  (cf. (8)) with power of at least  $\pi_S(\theta)$  for  $\theta \in \Theta_{1,S}$  can be constructed by considering tests that replace the composite hypotheses  $H_0$  and  $H_{1,S}$  with simple hypotheses involving mixtures. Thus, let  $\Lambda_0$  denote a distribution for  $\theta$  with support in  $\Theta_0$  and let  $\Lambda_1$  denote a distribution with support in  $\Theta_{1,S}$ , and consider the simple hypotheses

$$H_{0,\Lambda_0} : Y \text{ has density } f_{0,\Lambda_0} = \int f_\theta d\Lambda_0(\theta),$$

$$H_{1,\Lambda_1} : Y \text{ has density } f_{1,\Lambda_1} = \int f_\theta d\Lambda_1(\theta).$$

Let  $\bar{\pi}_S = \int \pi_S(\theta) d\Lambda_1(\theta)$  denote the weighted average of the power bound under  $H_{1,S}$ . The form of the best tests for these simple hypotheses is described by a generalized Neyman–Pearson Lemma.

LEMMA 3: *Suppose there exist  $cv_0 \geq 0, cv_1 \geq 0$ , and  $0 \leq \kappa \leq 1$  such that the test*

$$\varphi^{\text{NP}} = \begin{cases} 1 & \text{if } g + cv_1 f_{1,\Lambda_1} > cv_0 f_{0,\Lambda_0}, \\ \kappa & \text{if } g + cv_1 f_{1,\Lambda_1} = cv_0 f_{0,\Lambda_0}, \\ 0 & \text{if } g + cv_1 f_{1,\Lambda_1} < cv_0 f_{0,\Lambda_0}, \end{cases}$$

*satisfies  $\int \varphi^{\text{NP}} f_{0,\Lambda_0} d\nu = \alpha$ ,  $\int \varphi^{\text{NP}} f_{1,\Lambda_1} d\nu \geq \bar{\pi}_S$ , and  $cv_1(\int \varphi^{\text{NP}} f_{1,\Lambda_1} d\nu - \bar{\pi}_S) = 0$ . Then for any other test satisfying  $\int \varphi f_{0,\Lambda_0} d\nu \leq \alpha$  and  $\int \varphi f_{1,\Lambda_1} d\nu \geq \bar{\pi}_S$ ,  $\int \varphi^{\text{NP}} g d\nu \geq \int \varphi g d\nu$ .*

PROOF: If  $cv_1 = 0$ , the result follows from the Neyman–Pearson Lemma. For  $cv_1 > 0$ , by the definition of  $\varphi^{\text{NP}}$ ,  $\int (\varphi^{\text{NP}} - \varphi)(g + cv_1 f_{1,\Lambda_1} - cv_0 f_{0,\Lambda_0}) d\nu \geq 0$ . By assumption,  $\int (\varphi^{\text{NP}} - \varphi) f_{1,\Lambda_1} d\nu \leq 0$  and  $\int (\varphi^{\text{NP}} - \varphi) f_{0,\Lambda_0} d\nu \geq 0$ . The result now follows from  $cv_0 \geq 0$  and  $cv_1 \geq 0$ . Q.E.D.

As in Lemma 1, the power of the Neyman–Pearson test for the simple hypotheses provides an upper bound on power against  $H_{1,F}$  when  $H_0$  and  $H_{1,S}$  are composite.

LEMMA 4: *Let  $\varphi^{\text{NP}}$  denote the Neyman–Pearson test defined in Lemma 3, and let  $\varphi$  denote any test that satisfies (a)  $\sup_{\theta \in \Theta_0} \int \varphi f_\theta d\nu \leq \alpha$  and (b)  $\inf_{\theta \in \Theta_{1,S}} [\int \varphi f_\theta d\nu - \pi_S(\theta)] \geq 0$ . Then  $\int \varphi^{\text{NP}} g d\nu \geq \int \varphi g d\nu$ .*

<sup>9</sup>This is an example in the class of constraints considered by [Moreira and Moreira \(2013\)](#).

PROOF: Because  $\int \varphi f_{0,\Lambda_0} d\nu \leq \alpha$  and  $\int \varphi f_{1,\Lambda_1} d\nu \geq \int \pi_S(\theta) d\Lambda_1(\theta) = \bar{\pi}_S$ , the result follows from Lemma 3. *Q.E.D.*

Of course, to be a useful guide for gauging the efficiency of any proposed test, the power bound should be as small as possible. For given  $\alpha$  and power function  $\pi_S(\theta)$  of  $\tilde{\varphi} = \varphi_S \chi$ , the numerical algorithm from the last section can be modified to compute distributions  $\Lambda_0^*$  and  $\Lambda_1^*$  that approximately minimize the power bound. Details are provided in Appendix A. We use this algorithm to assess the efficiency of the level  $\alpha$  switching test  $\varphi_{\Lambda^*,S,\chi}^e$ , which satisfies the power constraint of Lemma 4 by construction, as  $\varphi_{\Lambda^*,S,\chi}^e(y) \geq \tilde{\varphi}(y)$  for all  $y$ .

RUNNING EXAMPLE—ctd: The power bound from Lemma 4 evaluated at  $\Lambda_0^*$  and  $\Lambda_1^*$  is 53.5%. Thus, there does not exist 5% level test with WAP larger than 53.5% and with at least as much power as the test  $\tilde{\varphi}(y) = \chi(y)\varphi_S(y) = \mathbf{1}[\hat{\delta} > 6]\mathbf{1}[|y_\beta| > 1.96]$  for alternatives with  $\delta \geq 9$ . Since  $\varphi_{\Lambda^*,S,\chi}$  of panel A of Figure 1 is strictly more powerful than  $\tilde{\varphi}$ , and it has WAP of 53.1%, it is thus also nearly optimal in the class of 5% level tests that satisfy this power constraint.

## 5. APPLICATIONS

In this section, we apply the algorithm outlined above (with the same parameters as in the running example) to construct nearly weighted average power maximizing 5% level tests for five nonstandard problems. In all of these problems, we set  $\varepsilon = 0.005$ , so that the ALFD test has power within 0.5% of the power bound for tests of the switching rule form (16) (if applicable). Unreported results show that the weighted average power of the resulting tests is also within 0.0065 of the upper bound on tests of arbitrary functional form under the power constraint described in Section 4.3 above.<sup>10</sup> Appendix B of the Supplemental Material (Elliott, Müller, and Watson (2015)) contains further details on the computations in each of the problems, and Appendix C of the Supplemental Material contains tables and Matlab programs to implement these nearly optimal tests for a wide range of parameter values.

### 5.1. The Behrens–Fisher Problem

Suppose we observe i.i.d. samples from two normal populations  $x_{1,i} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ ,  $i = 1, \dots, n_1$  and  $x_{2,i} \sim \mathcal{N}(\mu_2, \sigma_2^2)$ ,  $i = 1, \dots, n_2$ , where  $n_1, n_2 \geq 2$ . We are interested in testing  $H_0: \mu_1 = \mu_2$  without knowledge of  $\sigma_1^2$  and  $\sigma_2^2$ . This is the “Behrens–Fisher” problem, which has a long history in statistics. It also

<sup>10</sup>In some of these examples, we restrict attention to tests that satisfy a scale or location invariance property. Our ALFD test then comes close to maximizing weighted average power among all invariant tests.

arises as an asymptotic problem when comparing parameters across two potentially heterogeneous populations, where the information about each population is in the form of  $n_1$  and  $n_2$  homogeneous clusters to which a central limit theorem can be applied (see Ibragimov and Müller (2010, 2013)).

Let  $\bar{x}_j = n_j^{-1} \sum_{i=1}^{n_j} x_{j,i}$  and  $s_j^2 = \frac{1}{n_j-1} \sum_{i=1}^{n_j} (x_{j,i} - \bar{x}_j)^2$  be the sample means and variances for the two groups  $j = 1, 2$ , respectively. It is readily seen that the four-dimensional statistic  $(\bar{x}_1, \bar{x}_2, s_1, s_2)$  is sufficient for the four parameters  $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ . Imposing invariance to the transformations  $(\bar{x}_1, \bar{x}_2, s_1, s_2) \rightarrow (c\bar{x}_1 + m, c\bar{x}_2 + m, cs_1, cs_2)$  for  $m \in \mathbb{R}$  and  $c > 0$  further reduces the problem to the two-dimensional maximal invariant  $Y$

$$Y = (Y_\beta, Y_\delta) = \left( \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \log\left(\frac{s_1}{s_2}\right) \right)$$

whose density is derived in Appendix B of the Supplemental Material (cf. Linnik (1966) and Tsui and Weerahandi (1989)). Note that  $Y_\beta$  is the usual two-sample  $t$ -statistic which converges to  $\mathcal{N}(0, 1)$  under the null hypothesis as  $n_1, n_2 \rightarrow \infty$ . The distribution of  $Y$  only depends on the two parameters  $\beta = (\mu_1 - \mu_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$  and  $\delta = \log(\sigma_1/\sigma_2)$ , and the hypothesis problem becomes

$$(17) \quad H_0: \beta = 0, \quad \delta \in \mathbb{R} \quad \text{against} \quad H_1: \beta \neq 0, \quad \delta \in \mathbb{R}.$$

While the well-known two-sided test of Welch (1947) with “data-dependent degrees of freedom” approximately controls size for sample sizes as small as  $\min(n_1, n_2) = 5$  (Wang (1971) and Lee and Gurland (1975)), it is substantially oversized when  $\min(n_1, n_2) \leq 3$ ; moreover, its efficiency properties are unknown. Thus, we employ the algorithm described above to compute nearly optimal tests for  $n_1 \in \{2, 3\}$  and  $n_1 \leq n_2 \leq 12$ ; these are described in detail in the Supplemental Material. In the following, we focus on the two cases  $(n_1, n_2) \in \{(3, 3), (3, 6)\}$ .

To implement the algorithm, we choose  $F$  as uniform on  $\delta \in [-9, 9]$  and  $\beta = \{-3, 3\}$ . Appendix A shows that as  $|\delta| \rightarrow \infty$ , the experiment converges to a one sample normal mean problem (since one of the samples has negligible variance).<sup>11</sup> In the limiting problem, the standard test is simply the one sample  $t$ -test with  $n_1 - 1$  or  $n_2 - 1$  degrees of freedom, depending on the sign of  $\delta$ . Thus,  $\varphi_S(y) = \mathbf{1}[y_\delta > 0] \mathbf{1}[|y_\beta| > T_{n_1-1}(0.975)] + \mathbf{1}[y_\delta < 0] \mathbf{1}[|y_\beta| > T_{n_2-1}(0.975)]$ , where  $T_n(x)$  is the  $x$ th quantile of a Student- $t$  distribution with  $n$  degrees of freedom. We use the switching function  $\chi(y) = \mathbf{1}[|y_\delta| > 6]$ . We compare the power of the resulting  $\varphi_{A^*, S, \chi}^e$  test to the “conservative” test obtained by using the 0.975 quantile of a Student- $t$  distribution with degrees of freedom equal to  $\min(n_1, n_2) - 1$ , which is known to be of level  $\alpha$  (cf. Mickey and Brown (1966)).

<sup>11</sup>Strictly speaking, there are two limit experiments, one as  $\delta \rightarrow \infty$ , and one as  $\delta \rightarrow -\infty$ .



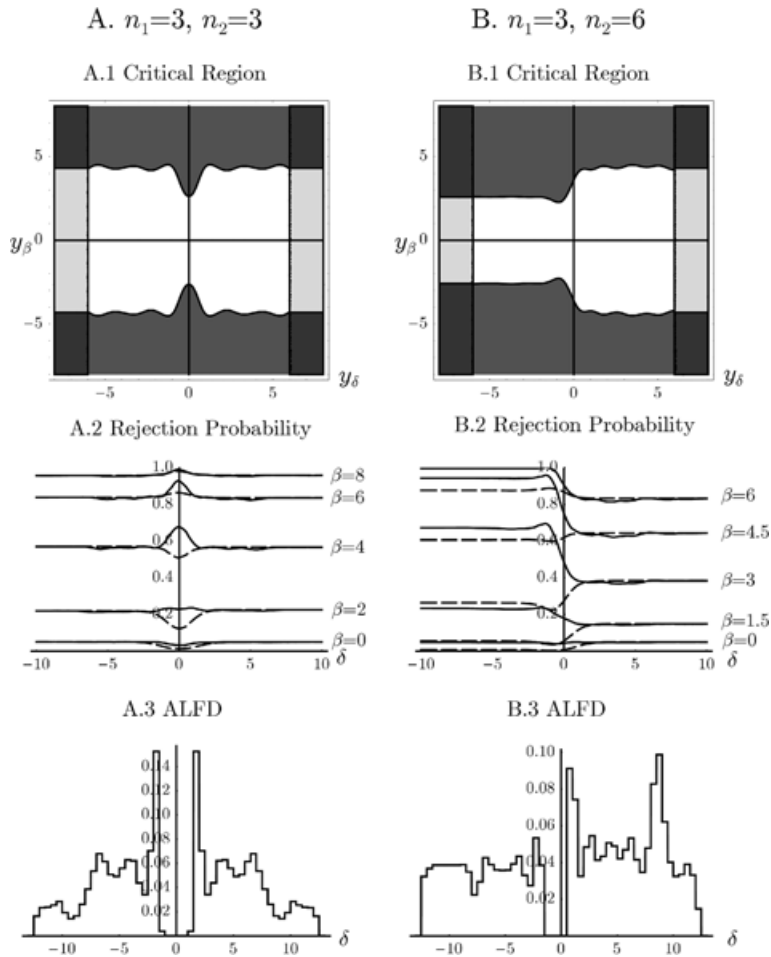


FIGURE 2.—Behrens-Fisher problem. Darker shades for  $|y_\delta| \geq 6$  in panels A.1 and B.1 indicate the part of the acceptance and critical region imposed by the switching rule. In panels A.2 and B.2, solid lines are the rejection probability of the nearly optimal test  $\varphi_{A^*, S, X}^\varepsilon$ , and dashed lines are for the usual  $t$ -test with critical value computed from the Student- $t$  distribution with  $n_1 - 1$  degrees of freedom.

Results are shown in Figure 2, where panel A shows results for  $(n_1, n_2) = (3, 3)$  and panel B shows results for  $(n_1, n_2) = (3, 6)$ . Looking first at panel A, the critical region transitions smoothly across the switching boundary. In the nonstandard part ( $|y_\delta| < 6$ ), the critical region is much like the critical region of the standard test  $\mathbf{1}[|y_\beta| > T_2(0.975)]$  for values of  $|y_\delta| > 2$ , but includes smaller values of  $|y_\beta|$  when  $y_\delta$  is close to zero. Evidently, small values of  $|y_\delta|$  suggest that the values of  $\sigma_1$  and  $\sigma_2$  are close, essentially yielding more degrees of freedom

for the null distribution of  $y_\beta$ . This feature of the critical region translates in the greater power for  $\varphi_{\Lambda^*, S, X}^\varepsilon$  than the conservative test when  $\delta$  is close to zero (see panel A.2). Panel B shows results when  $n_2$  is increased to  $n_2 = 6$ . Now, the critical region becomes “pinched” around  $y_\delta \approx -1$  apparently capturing a trade-off between a relatively small value of  $s_1$  and  $n_1$ . Panel B.2 shows a power function that is asymmetric in  $\delta$ , where the test has more power when the larger group has larger variance. Finally, the conservative test has a null rejection frequency substantially less than 5% when  $\delta < 0$  and weighted average power substantially below the nearly optimal test.

### 5.2. Inference About the Break Date in a Time Series Model

In this section, we consider tests for the break date  $\tau$  in the parameter of a time series model with  $T$  observations. A leading example is a one-time shift by the amount  $\eta$  of the value of a regression coefficient, as studied in Bai (1994, 1997). Bai’s asymptotic analysis focuses on breaks that are large relative to sampling uncertainty by imposing  $T^{1/2}|\eta| \rightarrow \infty$ . As discussed in Elliott and Müller (2007), this “large break” assumption may lead to unreliable inference in empirically relevant situations.

Under an alternative embedding for moderately sized breaks  $T^{1/2}\eta \rightarrow \delta \in \mathbb{R}$ , the parameter  $\delta$  becomes a nuisance parameter that remains relevant even asymptotically. As a motivating example, suppose the mean of a Gaussian time series shifts at some date  $\tau$  by the amount  $\eta$ ,

$$y_t = \mu + \mathbf{1}[t \geq \tau]\eta + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d. } \mathcal{N}(0, 1),$$

and the aim is to conduct inference about the break date  $\tau$ . As is standard in the structural break literature, assume that the break does not happen close to the beginning and end of the sample, that is, with  $\beta = \tau/T$ ,  $\beta \in B = [0.15, 0.85]$ . Restricting attention to translation invariant tests ( $\{y_t\} \rightarrow \{y_t + m\}$  for all  $m$ ) requires that tests are a function of the demeaned data  $y_t - \bar{y}$ . Partial summing the observations yields

$$(18) \quad T^{-1/2} \sum_{t=1}^{\lfloor sT \rfloor} (y_t - \bar{y}) \sim G(s) = W(s) - sW(1) - \delta(\min(\beta, s) - \beta s)$$

for  $s = j/T$  and integer  $1 \leq j \leq T$ , where  $W$  is a standard Wiener process. This suggests that asymptotically, the testing problem concerns the observation of the Gaussian process  $G$  on the unit interval, and the hypothesis of interest concerns the location  $\beta$  of the kink in its mean. Elliott and Müller (2014) formally showed that this is indeed the relevant asymptotic experiment for a moderate structural break in a well-behaved parametric time series model. By Gir-

sanov’s Theorem, the Radon–Nikodym derivative of the measure of  $G$  in (18) relative to the measure  $\nu$  of the standard Brownian Bridge, evaluated at  $G$ , is given by

$$(19) \quad f_\theta(G) = \exp\left[-\delta G(\beta) - \frac{1}{2}\delta^2\beta(1 - \beta)\right].$$

For  $|\delta| > 20$ , the discretization of the break date  $\beta$  becomes an important factor in this limiting problem, even with 1000 step approximations to Wiener processes. Since these discretization errors are likely to dominate the analysis with typical sample sizes for even larger  $\delta$ , we restrict attention to  $\delta \in \Delta = [-20, 20]$ , so that the hypotheses are

$$H_0: \beta = \beta_0, \quad \delta \in \Delta \quad \text{against} \quad H_1: \beta \neq \beta_0, \quad \delta \in \Delta.$$

To construct the ALFD test, we choose  $F$  so that  $\beta$  is uniform on  $B$  and  $\delta$  is  $\mathcal{N}(0, 100)$ , truncated to  $\Delta$ . Results are shown in Figure 3. Panel A shows results for  $\beta_0 = 0.2$ , where panel A.1 plots power as a function of  $\beta$  for five values of  $\delta$ ; panel B shows analogous results for  $\beta_0 = 0.4$ . (Since  $G$  is a continuous time stochastic process, the sample space is of infinite dimension, so it is not possible

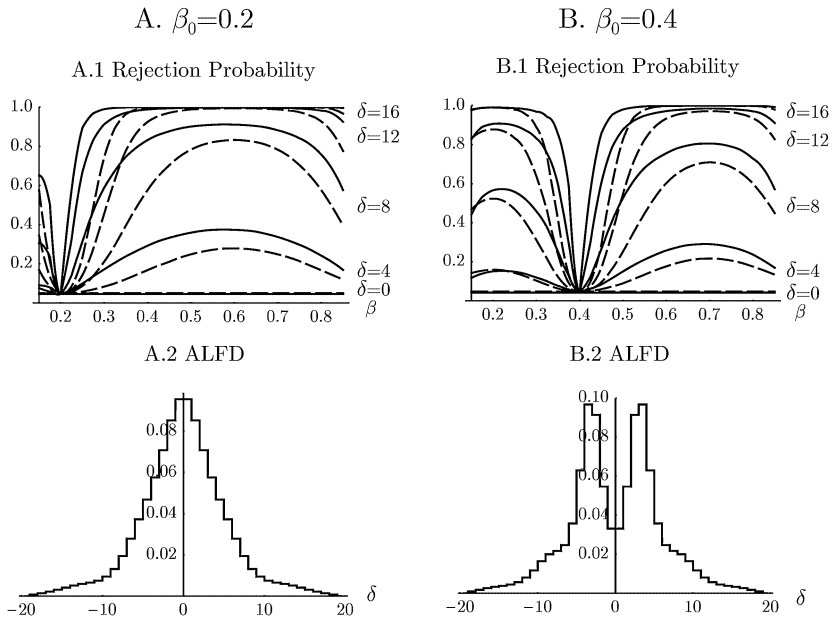


FIGURE 3.—Break date. In panels A.1 and B.1, solid lines are the rejection probability of the nearly optimal test  $\varphi_{A^*}^\epsilon$ , and dashed lines are for Elliott and Müller’s (2007) test that imposes an additional invariance.

to plot the critical region.) Rejection probabilities for a break at  $\beta_0 > 0.5$  are identical to those at  $1 - \beta_0$ .

Also shown in the figures are the corresponding power functions from the test derived in Elliott and Müller (2007) that imposes the additional invariance

$$(20) \quad G(s) \rightarrow G(s) + c(\min(\beta_0, s) - \beta_0 s) \quad \text{for all } c.$$

This invariance requirement eliminates the nuisance parameter  $\delta$  under the null, and thus leads to a similar test. But the transformation (20) is not natural under the alternative, leaving scope for reasonable and more powerful tests that are not invariant. Inspection of Figure 3 shows that the nearly optimal test  $\varphi_{A^*}^\varepsilon$  has indeed substantially larger power for most alternatives. Also, power is seen to be small when  $\beta$  is close to either  $\beta_0$  or the endpoints, as this implies a mean function close to what is specified under the null hypothesis.

### 5.3. Predictive Regression With a Local-to-Unity Regressor

A number of macroeconomic and finance applications concern the coefficient  $\gamma$  on a highly persistent regressor  $x_t$  in the model

$$(21) \quad \begin{aligned} y_t &= \mu + \gamma x_{t-1} + \varepsilon_{y,t}, \\ x_t &= r x_{t-1} + \varepsilon_{x,t}, \quad x_0 = 0, \end{aligned}$$

where  $E(\varepsilon_{y,t} | \{\varepsilon_{x,t-j}\}_{j=1}^{t-1}) = 0$ , so that the first equation is a predictive regression. The persistence in  $x_t$  is often modeled as a local-to-unity process (in the sense of Bobkoski (1983), Cavanagh (1985), Chan and Wei (1987), and Phillips (1987)) with  $r = r_T = 1 - \delta/T$ . Interest focuses on a particular value of  $\gamma$  given by  $H_0: \gamma = \gamma_0$  (where typically  $\gamma_0 = 0$ ). When the long-run covariance between  $\varepsilon_y$  and  $\varepsilon_x$  is nonzero, the usual  $t$ -test on  $\gamma$  is known to severely overreject unless  $\delta$  is very large.

After imposing invariance to translations of  $y_t$ ,  $\{y_t\} \rightarrow \{y_t + m\}$ , and an appropriate scaling by the (long-run) covariance matrix of  $(\varepsilon_{y,t}, \varepsilon_{x,t})'$ , the asymptotic inference problem concerns the likelihood ratio process  $f_\theta$  of a bivariate Gaussian continuous time process  $G$ ,

$$(22) \quad f_\theta(G) = K(G) \exp \left[ \beta Y_1 - \delta Y_2 - \frac{1}{2} \left( \beta + \frac{\rho}{\sqrt{1-\rho^2}} \delta \right)^2 Y_3 - \frac{1}{2} \delta^2 Y_4 \right],$$

where  $\beta$  is proportional to  $T(\gamma - \gamma_0)$ ,  $\rho \in (-1, 1)$  is the known (long-run) correlation between  $\varepsilon_{x,t}$  and  $\varepsilon_{y,t}$ ,  $\theta = (\beta, \delta)' \in \mathbb{R}^2$  is unknown, and the four-dimensional sufficient statistic  $Y = (Y_1, Y_2, Y_3, Y_4)$  has distribution

$$Y_1 = \int_0^1 W_{x,\delta}^\mu(s) dW_y(s) + \left( \beta + \frac{\rho}{\sqrt{1-\rho^2}} \delta \right) \int_0^1 W_{x,\delta}^\mu(s)^2 ds,$$

$$Y_2 = \int_0^1 W_{x,\delta}(s) dW_{x,\delta}(s) - \frac{\rho}{\sqrt{1-\rho^2}} Y_1,$$

$$Y_3 = \int_0^1 W_{x,\delta}^\mu(s)^2 ds, \quad Y_4 = \int_0^1 W_{x,\delta}(s)^2 ds,$$

where  $W_x$  and  $W_y$  are independent standard Wiener processes, and the Ornstein–Uhlenbeck process  $W_{x,\delta}$  solves  $dW_{x,\delta}(s) = -\delta W_{x,\delta}(s) ds + dW_x(s)$  with  $W_{x,\delta}(0) = 0$ , and  $W_{x,\delta}^\mu(s) = W_{x,\delta}(s) - \int_0^1 W_{x,\delta}(t) dt$  (cf. Jansson and Moreira (2006)).

Ruling out explosive roots,  $\delta < 0$ , the one-sided asymptotic inference problem is

$$(23) \quad H_0: \beta = 0, \quad \delta \geq 0 \quad \text{against} \quad H_1: \beta > 0, \quad \delta \geq 0.$$

While several methods have been developed that control size in (23) (leading examples include Cavanagh, Elliott, and Stock (1995) and Campbell and Yogo (2006)), there are fewer methods with demonstrable optimality. Stock and Watson (1996) numerically determined a weighted average power maximizing test within a parametric class of functions  $\mathbb{R}^4 \mapsto \{0, 1\}$ , and Jansson and Moreira (2006) derived the best conditionally unbiased tests of (23), conditional on the specific ancillary  $(Y_3, Y_4)$ . However, Jansson and Moreira (2006) reported that Campbell and Yogo's (2006) test has higher power for most alternatives. We therefore compare the one-sided ALFD test to this more powerful benchmark.

In Appendix A, we show that in a parameterization with  $\delta = \Delta_n - d\sqrt{2\Delta_n}$  and  $\beta = b\sqrt{2\Delta_n}/(1-\rho^2)$ , the experiment of observing  $G$  converges as  $\Delta_n \rightarrow \infty$  to the unrestricted two-dimensional Gaussian shift experiment. A natural way to obtain a test with the same asymptotic power function as  $\delta \rightarrow \infty$  is to rely on the usual maximum likelihood  $t$ -test (with observed information). From (22), the MLE is given by

$$\hat{\beta} = \frac{Y_1}{Y_3} - \frac{\rho}{\sqrt{1-\rho^2}} \hat{\delta}, \quad \hat{\delta} = -\frac{Y_2 + \frac{\rho}{\sqrt{1-\rho^2}} Y_1}{Y_4},$$

and the standard test becomes  $\varphi_S(Y) = \mathbf{1}[\hat{\beta}/\sqrt{Y_3^{-1} + \frac{\rho^2}{1-\rho^2} Y_4^{-1}} > cv_S]$ .

For numerical convenience, we use a discrete weighting function  $F$  with equal mass on 51 pairs of values  $(\beta, \delta)$ , where  $\delta \in \{0, 0.25^2, 0.5^2, \dots, 12.5^2\}$  and the corresponding values for  $\beta$  equal

$$(24) \quad \beta = b\sqrt{\frac{2\delta + 6}{1-\rho^2}} \quad \text{for} \quad \delta > 0$$

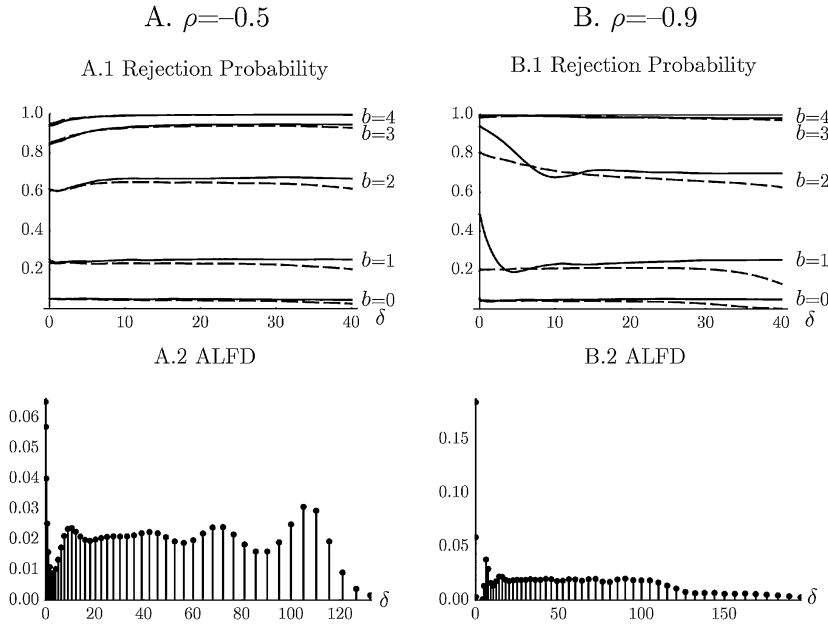


FIGURE 4.—Predictive regression with a local-to-unity regressor. In panels A.1 and B.1, solid lines are the rejection probability of the nearly optimal test  $\varphi_{A^*, S, \chi}^\varepsilon$ , and dashed lines are for Campbell and Yogo’s (2006) test. Since the latter is constructed under the assumption that  $\delta \leq 50$ , we only report its rejection probability for  $\delta \in [0, 40]$ . Alternatives are parameterized as in (24), with  $b \in \{0, 1, 2, 3, 4\}$ .

for  $b = 1.645$ . These alternatives are chosen so that power against each point in  $F$  is roughly 50%. (The spacing of the mass points for  $\delta$  and the choice of  $\beta$  correspond to an approximately uniform weighting over  $d$  and an alternative of 1.645 standard deviations for  $b$  in the limiting experiment.) We use the switching function  $\chi(Y) = \mathbf{1}[\hat{\delta} \geq 130]$ . The critical value  $cv_S$  in  $\varphi_S$  equals the usual 5% level value of 1.645 when  $\rho \geq 0$ , but we choose  $cv_S = 1.70$  when  $\rho < 0$ . This slight adjustment compensates for the heavier tail of the  $t$ -test statistic for moderate values of  $\delta$  and negative  $\rho$ .

Figure 4 compares the power of the resulting nearly optimal test to the test developed by Campbell and Yogo (2006) under the practically relevant values of  $\rho = -0.5$  and  $\rho = -0.9$ . Because the Campbell and Yogo (2006) test utilizes a confidence set for  $r$  with correct coverage only when  $r$  is close to unity (see Mikusheva (2007) and Phillips (2014)), Figure 4 plots power for  $\delta$  in the restricted range  $0 \leq \delta \leq 40$ , and over this range the optimal test nearly uniformly dominates the alternative test. For values of  $\delta > 40$  (so that  $r$  is much less than unity), the power of the Campbell and Yogo (2006) test falls dramatically, while the power of the optimal test increases slightly from its value with  $\delta = 40$ , and smoothly transitions to the power of  $\varphi_S$ . Unreported results show similar

results for positive values of  $\rho$ . In the Supplemental Material Appendix B, we also provide a comparison of our test with a modification of the Campbell and Yogo (2006) procedure that inverts Hansen's (1999) confidence intervals for  $r$ , which does not suffer from the uniformity issues mentioned above, and again find that our test dominates in terms of power.

#### 5.4. Testing the Value of a Set-Identified Parameter

The asymptotic problem introduced by Imbens and Manski (2004) and further studied by Woutersen (2006), Stoye (2009), and Hahn and Ridder (2011) involves a bivariate observation

$$Y = \begin{pmatrix} Y_l \\ Y_u \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu_l \\ \mu_u \end{pmatrix}, \begin{pmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{pmatrix} \right),$$

where  $\mu_l \leq \mu_u$ , and the elements  $\sigma_l, \sigma_u > 0$  and  $\rho \in (-1, 1)$  of the covariance matrix are known. The object of interest is  $\mu$ , which is only known to satisfy

$$(25) \quad \mu_l \leq \mu \leq \mu_u.$$

Without loss of generality, suppose we are interested in testing  $H_0: \mu = 0$  (the test of the general hypothesis  $\mu = \mu_0$  is reduced to this case by subtracting  $\mu_0$  from  $Y_l$  and  $Y_u$ ). Whilst under the null hypothesis the inequality (25) holds if and only if  $\mu_l/\sigma_l \leq 0 \leq \mu_u/\sigma_u$ , under the alternative the normalized means  $\mu_l/\sigma_l$  and  $\mu_u/\sigma_u$  may no longer satisfy the ordering  $\mu_l/\sigma_l \leq \mu_u/\sigma_u$ . It is thus not possible to reduce this problem to a single known nuisance parameter  $\rho$  without loss of generality. In the sequel, we demonstrate our approach when  $\sigma_l = \sigma_u = 1$  and various values of  $\rho$ .

It is useful to reparameterize  $(\mu_l, \mu_u)$  in terms of  $(\beta, \delta_L, \delta_P) \in \mathbb{R}^3$  as follows: let  $\delta_L = \mu_u - \mu_l$  be the length of the identified set  $[\mu_l, \mu_u]$ , let  $\beta = \mu_l$  if  $\mu_l > 0$ ,  $\beta = \mu_u$  if  $\mu_u < 0$ , and  $\beta = 0$  otherwise, and let  $\delta_P = -\mu_l$ , so that under the null hypothesis  $\delta_P$  describes the position of 0 in the identified set. In this parameterization, the hypothesis testing problem becomes

$$(26) \quad H_0: \beta = 0, \quad \delta_L \geq 0, \quad \delta_P \in [0, \delta_L] \quad \text{against} \quad H_1: \beta \neq 0, \quad \delta_L \geq 0.$$

Appendix A shows that the limiting problem as  $\delta_L \rightarrow \infty$  becomes a simple one-sided testing problem in the Gaussian shift experiment (2) with unrestricted nuisance parameter space. Thus, in this limiting problem, the standard test can be written as  $\varphi_S(y) = \mathbf{1}[y_l > 1.645 \text{ or } y_u < -1.645]$ . We switch to this standard test according to  $\chi(y) = \mathbf{1}[\hat{\delta}_L > 6]$ , where  $\hat{\delta}_L = Y_u - Y_l \sim \mathcal{N}(\delta_L, 2(1 - \rho))$ . The weighting function  $F$  is chosen to be uniform on  $\delta_L \in [0, 9]$ , with equal mass on the two points  $\beta \in \{-2, 2\}$ .

Note that (26) has a two-dimensional nuisance parameter under the null hypothesis, as neither the length  $\delta_L = \mu_u - \mu_l$  nor the distance  $\delta_P$  of  $\mu_l$  from

zero is specified under  $H_0$ . It is reasonable to guess, though, that the least favorable distribution only has mass at  $\delta_P \in \{0, \delta_L\}$ , so that one of the endpoints of the interval coincides with the hypothesized value of  $\mu$ . Further, the problem is symmetric in these two values for  $\delta_P$ . In the computation of the ALFD, we thus impose  $\delta_P \in \{0, \delta_L\}$  with equal probability, and then check that the resulting test  $\varphi_{A^*,S,\chi}^\varepsilon$  does indeed control size also for  $\delta_P \in (0, \delta_L)$ .

Figure 5 shows results for two values of  $\rho$ . Looking first at the critical regions, when  $y_u$  is sufficiently large (say  $y_u > 2$ ), the test rejects when  $y_l > 1.645$ ,

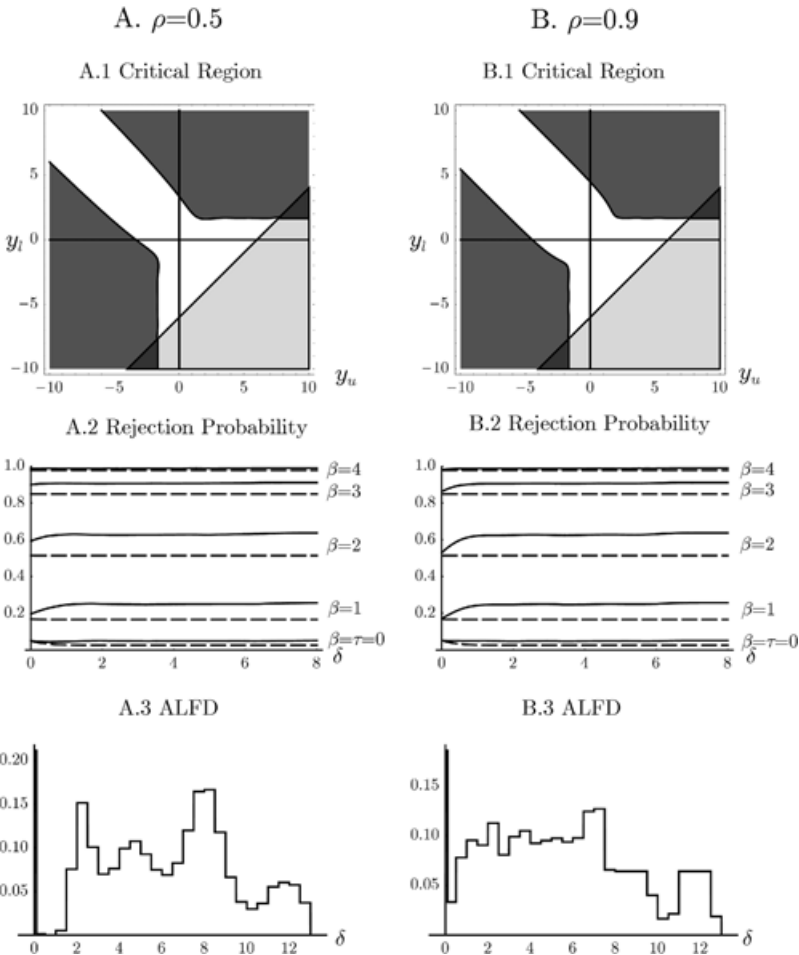


FIGURE 5.—Set-identified parameter. Darker shades for  $y_u + y_l \geq 6$  in panels A.1 and B1 indicate the part of the acceptance and critical region imposed by the switching rule. In panels A.2 and B.2, solid lines are the rejection probability of the nearly optimal tests  $\varphi_{A^*,S,\chi}^\varepsilon$ , and dashed lines are for Stoye’s (2009) test  $\varphi_{\text{Stoye}}(y) = \mathbf{1}[y_l > 1.96 \text{ or } y_u < -1.96]$ .



and similarly when  $y_l$  is sufficiently negative. The upper left-hand quadrant of the figures in panels A.1 and B.1 shows the behavior of the test when the observations are inverted relative to their mean values,  $y_l > y_u$ . In that case, the test rejects unless  $y_l + y_u$  is close to zero. Panels A.2 and B.2 compare the power of the ALFD test  $\varphi_{A^*,S,\chi}^e$  to the test  $\varphi_{\text{Stoye}}(y) = \mathbf{1}[y_l > 1.96 \text{ or } y_u < -1.96]$ , which is large sample equivalent to Stoye’s (2009) suggestion under local asymptotics. Note that this test has null rejection probability equal to 5% when  $\delta_L = 0$  and  $\delta_P \in \{0, \delta_L\}$ . Not surprisingly,  $\varphi_{A^*,S,\chi}^e$  dominates  $\varphi_{\text{Stoye}}$  when  $\delta_L$  is large, but it also has higher power when  $\delta_L$  is small and  $\rho = 0.5$  (because when  $\delta_L$  is small, the mean of  $Y_l$  and  $Y_u$  is more informative about  $\mu$  than either  $Y_l$  or  $Y_u$  unless  $\rho$  is close to 1).

### 5.5. Regressor Selection

As in Leeb and Pötscher (2005), consider the bivariate linear regression

$$(27) \quad y_i = \gamma x_i + \eta z_i + \varepsilon_i, \quad i = 1, \dots, n, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where  $\sigma^2$  is known. We are interested in testing  $H_0: \gamma = \gamma_0$ , and  $\eta$  is a nuisance parameter. Suppose there is substantial uncertainty whether the additional control  $z_i$  needs to be included in (27), that is,  $\eta = 0$  is deemed likely, but not certain. Denote by  $(\hat{\gamma}, \hat{\eta})$  the OLS estimators from the “long” regression of  $y_i$  on  $(x_i, z_i)$ . Let  $\beta = n^{1/2}(\gamma - \gamma_0)$ ,  $\delta = n^{1/2}\eta$ ,  $(Y_\beta, Y_\delta) = n^{1/2}(\hat{\gamma} - \gamma_0, \hat{\eta})$ , and for notational simplicity, assume that the regressors and  $\sigma^2$  have been scale normalized so that

$$(28) \quad Y = \begin{pmatrix} Y_\beta \\ Y_\delta \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \delta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where  $-\rho$  is the known sample correlation between  $x_i$  and  $z_i$ . Note that with the Gaussian assumption about  $\varepsilon_i$ ,  $Y$  is a sufficient statistic for the unknown parameters  $(\beta, \delta)$ .

For  $\delta = 0$  known, the statistic  $Y_\beta - \rho Y_\delta$  is more informative about  $\beta$  than  $Y_\beta$ . Intuitively,  $Y_\beta - \rho Y_\delta$  is the (rescaled) regression coefficient in the “short” regression of  $y_i$  on  $x_i$ , omitting  $z_i$ , while  $Y_\beta$  corresponds to the coefficient in the “long” regression. Ideally, one would like to let the data decide whether indeed  $\delta = 0$ , so that one can appropriately base inference on  $Y_\beta - \rho Y_\delta$ , or on  $Y_\beta$ . As reviewed by Leeb and Pötscher (2005), however, data-dependent model selection procedures do not perform uniformly well for all values of  $\eta$ , even in large samples, so that optimal inference is not obtained in this manner.

As one possible notion of optimality, suppose that we seek a test of  $H_0: \beta = 0$  that is as powerful as possible when  $\delta = 0$ , but under the constraint that the test controls size for all values of  $\delta \in \mathbb{R}$ . The idea is that we want to maximize power in the a priori likely case of  $\delta = 0$ , while at the same time controlling the null rejection probability even if  $\delta \neq 0$ .

Consider first the one-sided problem. With  $F$  degenerate at  $\beta_1 > 0$ , we obtain the hypothesis test

$$(29) \quad H_0: \beta = 0, \quad \delta \in \mathbb{R} \quad \text{against} \quad H_{1,F}: \beta = \beta_1, \quad \delta = 0.$$

Note that rejecting for large values of  $Y_\beta$  is the Neyman–Pearson test of  $H_{1,F}$  against the single null hypothesis  $H_0^s: (\beta, \delta) = (0, \delta_0)$ , where  $\delta_0 = -\rho\beta_1$ . Since any level  $\alpha$  test of (29) is also of level  $\alpha$  under  $H_0^s$ , the uniformly most powerful one-sided test of (29) thus rejects for large values of  $Y_\beta$  (cf. Proposition 15.2 in van der Vaart (1998)). Thus, as long as one insists on uniform size control, the question of best one-sided inference about  $\beta$  has a straightforward answer: simply rely on the coefficient estimate of the long regression.

Now consider the two-sided problem. It is known that rejecting for large values of  $|Y_\beta|$  yields the uniformly most powerful test among tests that are unbiased for all values of  $\delta \in \mathbb{R}$  (cf. problem 1 on page 226 of van der Vaart (1998)). But with a focus on power at the point  $\delta = 0$ , this might be considered a too restrictive class of tests. Thus, we consider the unconstrained problem of maximizing weighted average power in the hypothesis testing problem

$$(30) \quad H_0: \beta = 0, \quad \delta \in \mathbb{R} \quad \text{against} \quad H_1: \beta \neq 0, \quad \delta = 0$$

and choose a weighting function  $F$  that puts equal mass at the two points  $\{-2, 2\}$ . For large  $|Y_\delta|$ , we switch to the standard test  $\varphi_S(y) = \mathbf{1}[|y_\beta| > 1.96]$  via  $\chi(y) = \mathbf{1}[|y_\delta| > 6]$ . Unreported results show that imposing this switching rule leads to no discernible loss in power when  $\delta = 0$ . At the same time, this switching rule leads to much higher power when  $|\delta|$  is large.

By construction, the weighted average power at  $\delta = 0$  of  $\varphi_{A^*,S,\chi}^\varepsilon$  in Figure 6 is nearly the largest possible among all 5% valid tests. To get a more comprehensive view of the potential gains in power as a function of  $\rho$ , Figure 7 depicts the power bound, the power of  $\varphi_{A^*,S,\chi}^\varepsilon$ , and the power of  $\varphi_S$  against  $\delta = 0$  and  $\beta \in \{-2, 2\}$ . The experiment (28) becomes more informative about  $\beta$  as  $\rho$  increases, and correspondingly, the power bound is an increasing function of  $\rho$ .<sup>12</sup> It is striking, though, how flat the power bound becomes once  $\rho \geq 0.75$ . The gains in power at  $\delta = 0$  over the standard test  $\varphi_S$  are never larger than 12 percentage points, and the test  $\varphi_{A^*,S,\chi}^\varepsilon$  described in the Supplemental Material comes very close to fully exploiting the available information.

## 6. CONCLUSION

Many statistics have been explored as the basis for constructing valid tests in nonstandard econometric problems: maximum likelihood estimators, quasi

<sup>12</sup>Adding mean-zero Gaussian noise to  $Y_\delta$  and an appropriate rescaling yields an equivalent experiment with smaller  $|\rho|$ .

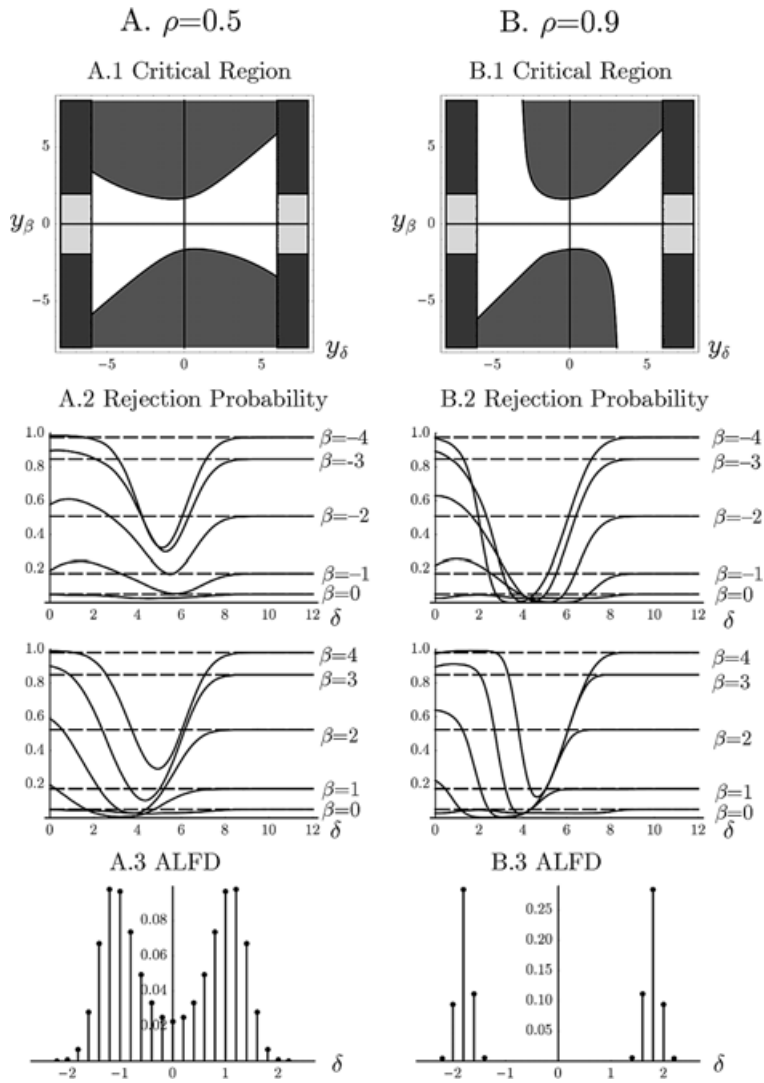


FIGURE 6.—Regressor selection. Darker shade for  $|y_\delta| \geq 6$  in panels A.1 and B.1 is the part of the critical region imposed by the switching rule. In panels A.2 and B.2, solid lines are the rejection probability of the nearly optimal tests  $\varphi_{A^*, S, X}^e$ , and dashed lines are for the usual test that ignores  $Y_\delta$ ,  $\varphi_S(y) = \mathbf{1}[|y_\beta| > 1.96]$ .

likelihood ratio statistics, moment based criterion functions, statistics involving pretests, etc. Similarly, appropriate critical values may be obtained by various methods: by searching for the largest appropriate quantile of the test statistic under the null hypothesis, by arguing for the validity of bootstrap or subsam-

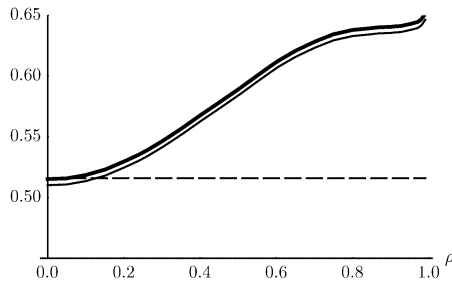


FIGURE 7.—Weighted average power in regressor selection problem as function of  $\rho$ . Thick solid line is power bound, thin solid line is power of 5% level test, and dashed line is power of usual test that ignores  $Y_\delta$ ,  $\varphi_S(y) = \mathbf{1}[|y_\beta| > 1.96]$ .

pling methods, or by refining these approaches with Bonferroni inequalities or pretests.

The numerical determination of an approximately least favorable null distribution suggested here is, in general, not substantially harder, either conceptually or computationally. And, in contrast to these other approaches, it delivers a test that is nearly optimal in the sense that it comes close to maximizing a weighted average power criterion. What is more, once the approximate least favorable distribution is determined, the test is straightforward to implement in practice, compared to, say, methods that require resampling. Also, the low upper bound on weighted average power implied by the approximately least favorable distribution can be used to argue for the near-optimality of a convenient test statistic, such as the quasi likelihood ratio statistic. For these reasons, the algorithm suggested here seems a practically useful approach to deal with nonstandard testing problems.

The six applications considered in this paper have a one-dimensional nuisance parameter, and Müller and Watson (2013b) and Müller (2014) discussed additional applications with a one-dimensional nuisance parameter. While a high-dimensional nuisance parameter generally poses numerical difficulties for the algorithm, Müller and Watson (2013a) successfully implemented the algorithm in a problem with a three-dimensional nuisance parameter. The relatively hardest part in the computations is the numerical check of size control, but unless the problem has special features, the difficulty of this step is common to all approaches to constructing nontrivial tests.

## APPENDIX A

### A.1. Proof of Lemma 2

For any  $h_0$ , we have, from (11),

$$\frac{f_{n,h_0}(Y)}{f_{n,0}(Y)} \Rightarrow \frac{f_{X,h_0}(X)}{f_{X,0}(X)},$$

so that  $f_{n,h_0}$  is contiguous to  $f_{n,0}$  by LeCam's first lemma. A general version of LeCam's third lemma (see, for instance, [van der Vaart \(2002\)](#)) thus implies that (11) also holds with  $Y$  and  $X$  distributed according to  $f_{n,h_0}$  and  $f_{X,h_0}$ , respectively. This establishes convergence of experiments via Definition 9.1 of [van der Vaart \(1998\)](#).

With  $\epsilon = \limsup_{\Delta \rightarrow \infty} (E_{(b_1, \Delta)}[\varphi(Y)] - E_{(b_1, \Delta)}[\varphi_S(Y)]) > 0$ , pick  $\Delta_n$  with  $\Delta_n \rightarrow \infty$  such that  $\lim_{n \rightarrow \infty} (E_{(b_1, \Delta_n)}[\varphi(Y)] - E_{(b_1, \Delta_n)}[\varphi_S(Y)]) = \epsilon$ . With this definition of  $\Delta_n$ ,  $E_{(b, \Delta_n)}[\cdot]$  corresponds to integration w.r.t.  $f_{n,h}$ , where  $h = (b, 0) \in \mathbb{R}^{k\beta+1}$ , denoted by  $E_{n,h}[\cdot]$ . By assumption about  $\varphi_S$  and  $\varphi_S^{\text{lim}}$ ,  $E_{n,h_1}[\varphi_S(Y)] \rightarrow E_{h_1}[\varphi_S^{\text{lim}}(X)]$ , where  $h_1 = (b_1, 0)$ . Thus,  $E_{n,h_1}[\varphi(Y)] \rightarrow E_{h_1}[\varphi_S^{\text{lim}}(X)] + \epsilon$ . Further, by contiguity, also  $E_{n,h}[\varphi(Y)]$  converges for all fixed  $h$ . By Theorem 15.1 of [van der Vaart \(1998\)](#), there exists a test  $\varphi^{\text{lim}}$  with a power function  $E_h[\varphi^{\text{lim}}(X)]$  equal to  $\lim_{n \rightarrow \infty} E_{n,h}[\varphi(Y)]$ . The test  $\varphi^{\text{lim}}$  is of level  $\alpha$ , as  $E_{(0, \delta_n)}[\varphi(Y)] \leq \alpha$  for all  $n$  implies  $\lim_{n \rightarrow \infty} E_{(0, \delta_n)}[\varphi(Y)] \leq \alpha$ . Further, since  $\varphi_S^{\text{lim}}$  is admissible with a power function that does not depend on  $d$ , and  $\varphi^{\text{lim}}$  has higher power than  $\varphi_S^{\text{lim}}$  at  $b = b_1$ , there must exist  $b_2$  such that with  $h_2 = (b_2, 0)$ ,  $E_{h_2}[\varphi^{\text{lim}}(X)] < E_{h_2}[\varphi_S^{\text{lim}}(X)]$ . The conclusion follows from  $E_{n,h_2}[\varphi(Y)] \rightarrow E_{h_2}[\varphi^{\text{lim}}(X)]$  and  $E_{n,h_2}[\varphi_S(Y)] \rightarrow E_{h_2}[\varphi_S^{\text{lim}}(X)]$ .

## A.2. Algorithm Details

### A.2.1. Details of the Algorithm Described in Section 3

Repeated application of (10) requires evaluation of the rejection probability  $\int \varphi_\mu f_j d\nu$ . This can be implemented straightforwardly by Monte Carlo integration, using  $N_0$  independent draws  $Y_{j,l}$  from  $f_j$ ,  $l = 1, \dots, N_0$ , for a total of  $M \cdot N_0$  independent draws. Thus, one simple estimator of  $\int \varphi_\mu f_j d\nu$  is  $N_0^{-1} \sum_{l=1}^{N_0} \mathbf{1}[g(Y_{j,l}) > \sum_{i=1}^M \exp(\mu_i) f_i(Y_{j,l})]$ . Note that the  $M \cdot N_0$  scalar random variables  $g(Y_{i,l})$  and  $M^2 \cdot N_0$  variables  $f_i(Y_{j,l})$  can be computed and stored once, prior to any fixed point iteration. But with  $f_i$  readily available, it makes sense to improve the estimator of  $\int \varphi_\mu f_j d\nu$  via importance sampling, that is, to use

$$(31) \quad \widehat{\text{RP}}_j(\mu) = (MN_0)^{-1} \sum_{k=1}^M \sum_{l=1}^{N_0} \frac{f_j(Y_{k,l})}{\bar{f}(Y_{k,l})} \mathbf{1} \left[ g(Y_{k,l}) > \sum_{i=1}^M \exp(\mu_i) f_i(Y_{k,l}) \right],$$

where  $\bar{f}(y) = M^{-1} \sum_{j=1}^M f_j(y)$ . This has the advantage that with a finer discretization of  $H_0$  (larger  $M$ ), one may decrease  $N_0$  for the same level of Monte Carlo accuracy, as the draws from neighboring densities become relatively more informative for the rejection probability under  $f_j$ .

The ultimate algorithm consists of the following eight steps:

1. For each  $k$ ,  $k = 1, \dots, M$ , generate  $N_0$  draws  $Y_{k,l}$ ,  $l = 1, \dots, N_0$ , with density  $f_k$ . The draws  $Y_{k,l}$  are independent across  $k$  and  $l$ .

2. Compute and store  $g(Y_{k,l})$ ,  $f_j(Y_{k,l})$ , and  $\bar{f}(Y_{k,l})$ ,  $j, k = 1, \dots, M$ ,  $l = 1, \dots, N_0$ .
3. Set  $\mu^{(0)} = (-2, \dots, -2) \in \mathbb{R}^M$ .
4. Compute  $\mu^{(i+1)}$  from  $\mu^{(i)}$  via  $\mu_j^{(i+1)} = \mu_j^{(i)} + \omega(\widehat{\text{RP}}_j(\mu^{(i)}) - \alpha)$  and  $\omega = 2$ , and repeat this step  $O = 600$  times. Denote the resulting element in the simplex by  $\hat{\Lambda}^* = (\hat{\lambda}_1^*, \dots, \hat{\lambda}_M^*)$ , where  $\hat{\lambda}_j^* = \exp(\mu_j^{(O)}) / \sum_{k=1}^M \exp(\mu_k^{(O)})$ .
5. Compute the number  $\text{cv}$  such that the test  $\varphi_{\hat{\Lambda}^*}$  is exactly of (Monte Carlo) level  $\alpha$  when  $Y$  is drawn from the mixture  $\sum_{i=1}^M \hat{\lambda}_i^* f_i$ , that is, solve  $\sum_{j=1}^M \hat{\lambda}_j^* (\widehat{\text{RP}}_j^*(\text{cv}) - \alpha) = 0$ , where  $\widehat{\text{RP}}_j^*(\text{cv})$  is the Monte Carlo estimate (31) of the rejection probability of the test  $\mathbf{1}[g > \text{cv} \sum_{i=1}^M \hat{\lambda}_i^* f_i]$  under  $f_j$ .
6. Compute the (estimate of) the power bound  $\bar{\pi}$  of  $\varphi_{\hat{\Lambda}^*} = \mathbf{1}[g > \text{cv} \sum_{i=1}^M \hat{\lambda}_i^* f_i]$  via  $N_1^{-1} \sum_{l=1}^{N_1} \varphi_{\hat{\Lambda}^*}(Y_l)$ , where  $Y_l$  are  $N_1$  i.i.d. draws of  $Y$  with density  $g$ .<sup>13</sup>
7. Compute the number  $\text{cv}^\varepsilon > \text{cv}$  such that the test  $\varphi_{\hat{\Lambda}^*}^\varepsilon = \mathbf{1}[g > \text{cv}^\varepsilon \sum_{i=1}^M \hat{\lambda}_i^* f_i]$  has (estimated) power  $\bar{\pi} - \varepsilon$ , using the same estimator as in Step 6.
8. Check size control of  $\varphi_{\hat{\Lambda}^*}^\varepsilon$  by evaluating its null rejection probabilities on a fine grid of  $H_0$ , using estimates of null rejection probabilities of the form (31). If  $\varphi_{\hat{\Lambda}^*}^\varepsilon$  is found to overreject, restart the algorithm at Step 1 with a finer discretization of  $H_0$  (larger  $M$ ). Otherwise,  $\hat{\Lambda}^*$  satisfies the definition of an  $\varepsilon$ -ALFD,  $\Lambda^* = \hat{\Lambda}^*$ .

With  $N_0 = 20,000$  and  $N_1 = 100,000$  (so that Monte Carlo standard deviations of  $\widehat{\text{RP}}_j(\mu)$  are about 0.1%) and  $M = 50$ , the algorithm takes about one minute on a modern PC, and it spends most of the time on Steps 2 and 8.

In some additional problems not discussed in this paper, we found that the iterations of Step 4 can be accelerated by introducing iteration-specific and adaptive step length factors  $\omega_j^{(i)}$ , so that Step 4 reads  $\mu_j^{(i+1)} = \mu_j^{(i)} + \omega_j^{(i)}(\widehat{\text{RP}}_j(\mu^{(i)}) - \alpha)$ , where  $\omega_j^{(i+1)}$  is made much smaller (say,  $\omega_j^{(i+1)} = 0.5\omega_j^{(i)}$ ) if the size control constraint switches sign in the last two iterations ( $\widehat{\text{RP}}_j(\mu^{(i)}) - \alpha)(\widehat{\text{RP}}_j(\mu^{(i-1)}) - \alpha) < 0$ , and  $\omega_j^{(i+1)}$  increases slowly otherwise (say,  $\omega_j^{(i+1)} = 1.03\omega_j^{(i)}$ ).

### A.2.2. Details of the Algorithm Used to Compute $\varepsilon$ -ALFD Switching Tests

First note the following result that extends Lemma 1 to tests of the switching form (16).

<sup>13</sup>It makes sense to choose  $N_1 > N_0$ , as the importance sampling in (31) leads to a relatively smaller Monte Carlo standard error.

LEMMA 5: For given  $\chi$  and  $\varphi_S$ , let  $SW$  be the set of tests of the form (16). Let  $\varphi_{\Lambda, S, \chi} \in SW$  be of size  $\alpha$  under  $H_{0, \Lambda}$  with  $\varphi_{\Lambda}$  of the Neyman–Pearson form

$$\varphi_{\Lambda} = \begin{cases} 1 & \text{if } g(y) > cv \int f_{\theta}(y) d\Lambda(\theta), \\ \kappa & \text{if } g(y) = cv \int f_{\theta}(y) d\Lambda(\theta), \\ 0 & \text{if } g(y) < cv \int f_{\theta}(y) d\Lambda(\theta), \end{cases}$$

for some  $cv \geq 0$  and  $0 \leq \kappa \leq 1$ . Then for any test  $\varphi \in SW$  that is of level  $\alpha$  under  $H_0$ ,  $\int \varphi_{\Lambda, S, \chi} g \, d\nu \geq \int \varphi g \, d\nu$ .

PROOF: Note that by construction,  $\int (\varphi_{\Lambda, S, \chi} - \varphi)(g - cv \int f_{\theta} d\Lambda(\theta)) \, d\nu \geq 0$ . Since  $\varphi$  is of level  $\alpha$  under  $H_0$ ,  $\int \varphi f_{\theta} \, d\nu \leq \alpha$  for all  $\theta \in \Theta_0$ . Therefore,  $\iint \varphi f_{\theta} \, d\nu \, d\Lambda(\theta) = \int \varphi (\int f_{\theta} d\Lambda(\theta)) \, d\nu \leq \alpha$ , where the equality follows from Fubini’s Theorem. Thus  $\int (\varphi_{\Lambda, S, \chi} - \varphi) (\int f_{\theta} d\Lambda(\theta)) \, d\nu \geq 0$ , and the result follows. Q.E.D.

Given this result, the only change to the algorithm of Section A.2.1 is the replacement of  $\varphi_{\Lambda}$  and  $\varphi_{\Lambda}^{\varepsilon}$  by  $\varphi_{\Lambda, S, \chi}$  and  $\varphi_{\Lambda, S, \chi}^{\varepsilon}$ , respectively. Specifically, the (estimated) rejection probability (31) now reads

$$\widehat{\text{RP}}_j(\mu) = (MN_0)^{-1} \sum_{k=1}^M \sum_{l=1}^{N_0} \frac{f_j(Y_{k,l})}{\bar{f}(Y_{k,l})} \left( \chi(Y_{k,l}) \varphi_S(Y_{k,l}) + (1 - \chi(Y_{k,l})) \mathbf{1} \left[ g(Y_{k,l}) > \sum_{i=1}^M \exp(\mu_i) f_i(Y_{k,l}) \right] \right),$$

and in Steps 6 and 7,  $\bar{\pi}$  and  $\bar{\pi} - \varepsilon$  are the (estimated) powers of the tests  $\varphi_{\hat{\Lambda}, S, \chi} = \chi \varphi_S + (1 - \chi) \mathbf{1}[g > cv \sum_{i=1}^M \hat{\lambda}_i f_i]$  and  $\varphi_{\hat{\Lambda}, S, \chi}^{\varepsilon} = \chi \varphi_S + (1 - \chi) \mathbf{1}[g > cv^{\varepsilon} \sum_{i=1}^M \hat{\lambda}_i f_i]$ , respectively. Correspondingly, the size check in Step 8 also concerns  $\varphi_{\hat{\Lambda}, S, \chi}^{\varepsilon}$ .

REFERENCES

ANDREWS, D. W. K. (2011): “Similar-on-the-Boundary Tests for Moment Inequalities Exist, but Have Poor Power,” Discussion Paper 1815, Cowles Foundation. [772]  
 ANDREWS, D. W. K., AND G. GUGGENBERGER (2010): “Asymptotic Size and a Problem With Subsampling and With the M out of N Bootstrap,” *Econometric Theory*, 26, 426–468. [775]  
 ANDREWS, D. W. K., AND W. PLOBERGER (1994): “Optimal Tests When a Nuisance Parameter Is Present Only Under the Alternative,” *Econometrica*, 62, 1383–1414. [777]

- ANDREWS, D. W. K., M. J. MOREIRA, AND J. H. STOCK (2008): "Efficient Two-Sided Nonsimilar Invariant Tests in IV Regression With Weak Instruments," *Journal of Econometrics*, 146, 241–254. [773,779]
- BAI, J. (1994): "Least Squares Estimation of a Shift in Linear Processes," *Journal of Time Series Analysis*, 15, 453–470. [774,795]
- (1997): "Estimation of a Change Point in Multiple Regressions," *Review of Economics and Statistics*, 79, 551–563. [774,795]
- BLACKWELL, D., AND M. A. GIRSHICK (1954): *Theory of Games and Statistical Decisions*. New York: Wiley. [774,780]
- BOBKOSKI, M. J. (1983): "Hypothesis Testing in Nonstationary Time Series," Ph.D. Thesis, Department of Statistics, University of Wisconsin. [797]
- CAMPBELL, J. Y., AND M. YOGO (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27–60. [774,798–800]
- CAVANAGH, C. L. (1985): "Roots Local to Unity," Working Paper, Harvard University. [797]
- CAVANAGH, C. L., G. ELLIOTT, AND J. H. STOCK (1995): "Inference in Models With Nearly Integrated Regressors," *Econometric Theory*, 11, 1131–1147. [784,798]
- CHAMBERLAIN, G. (2000): "Econometric Applications of Maximin Expected Utility," *Journal of Applied Econometrics*, 15, 625–644. [774,783]
- (2007): "Decision Theory Applied to an Instrumental Variables Model," *Econometrica*, 75 (3), 609–652. [786]
- CHAN, N. H., AND C. Z. WEI (1987): "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics*, 15, 1050–1063. [797]
- CHIBURIS, R. C. (2009): "Approximately Most Powerful Tests for Moment Inequalities," in "Essays on Treatment Effects and Moment Inequalities," Ph.D. Thesis, Department of Economics, Princeton University, Chapter 3. [774,783]
- CHOI, A., W. J. HALL, AND A. SCHICK (1996): "Asymptotically Uniformly Most Powerful Tests in Parametric and Semiparametric Models," *The Annals of Statistics*, 24, 841–861. [772,776]
- DUFOUR, J.-M., AND M. L. KING (1991): "Optimal Invariant Tests for the Autocorrelation Coefficient in Linear Regressions With Stationary or Nonstationary AR(1) Errors," *Journal of Econometrics*, 47, 115–143. [772]
- ELLIOTT, G., AND U. K. MÜLLER (2007): "Confidence Sets for the Date of a Single Break in Linear Time Series Regressions," *Journal of Econometrics*, 141, 1196–1218. [774,795–797]
- (2012): "Pre and Post Break Parameter Inference," Working Paper, Princeton University. [783]
- (2014): "Pre and Post Break Parameter Inference," *Journal of Econometrics*, 180, 141–157. [795]
- ELLIOTT, G., U. K. MÜLLER, AND M. W. WATSON (2015): "Supplement to 'Nearly Optimal Tests When a Nuisance Parameter Is Present Under the Null Hypothesis'," *Econometrica Supplemental Material*, 83, <http://dx.doi.org/10.3982/ECTA10535>. [792]
- ELLIOTT, G., T. J. ROTHENBERG, AND J. H. STOCK (1996): "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836. [772]
- FACCHINEI, F., AND J.-S. PANG (2003): *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vols. I and II. New York: Springer. [784]
- FERGUSON, T. S. (1967): *Mathematical Statistics—A Decision Theoretic Approach*. New York: Academic Press. [774,781]
- HAHN, J., AND G. RIDDER (2011): "A Dual Approach to Confidence Intervals for Partially Identified Parameters," Working Paper, UCLA. [800]
- HANSEN, B. E. (1999): "The Grid Bootstrap and the Autoregressive Model," *Review of Economics and Statistics*, 81, 594–607. [800]
- HILLIER, G. (1990): "On the Normalization of Structural Equations: Properties of Direction Estimators," *Econometrica*, 58, 1181–1194. [786]



- IBRAGIMOV, R., AND U. K. MÜLLER (2010): "T-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business & Economic Statistics*, 28, 453–468. [793]
- (2013): "Inference With Few Heterogeneous Clusters," Working Paper, Princeton University. [793]
- IMBENS, G., AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857. [774,800]
- JANSSON, M., AND M. J. MOREIRA (2006): "Optimal Inference in Regression Models With Nearly Integrated Regressors," *Econometrica*, 74, 681–714. [784,798]
- KEMPTHORNE, P. J. (1987): "Numerical Specification of Discrete Least Favorable Prior Distributions," *SIAM Journal on Scientific and Statistical Computing*, 8, 171–184. [774,783]
- KIM, S., AND A. S. COHEN (1998): "On the Behrens–Fisher Problem: A Review," *Journal of Educational and Behavioral Statistics*, 23, 356–377. [774]
- KING, M. L. (1987): "Towards a Theory of Point Optimal Testing," *Econometric Reviews*, 6, 169–218. [788]
- KRAFFT, O., AND H. WITTING (1967): "Optimale Tests und ungünstige Verteilungen," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 7, 289–302. [774,783]
- LEE, A. F. S., AND J. GURLAND (1975): "Size and Power of Tests for Equality of Means of Two Normal Populations With Unequal Variances," *Journal of the American Statistical Association*, 70, 933–941. [793]
- LEEB, H., AND B. M. PÖTSCHER (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59. [774,802]
- LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. New York: Springer. [772,773,778]
- LINNIK, Y. V. (1966): "Randomized Homogeneous Tests for the Behrens–Fisher Problem," *Selected Translations in Mathematical Statistics and Probability*, 6, 207–217. [772,793]
- (1968): *Statistical Problems With Nuisance Parameters*. New York: American Mathematical Society. [772]
- MICKEY, M. R., AND M. B. BROWN (1966): "Bounds on the Distribution Functions of the Behrens–Fisher Statistic," *The Annals of Mathematical Statistics*, 37, 639–642. [793]
- MIKUSHEVA, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75, 1411–1452. [799]
- MOON, H. R., AND F. SCHORFHEIDE (2009): "Estimation With Overidentifying Inequality Moment Conditions," *Journal of Econometrics*, 153, 136–154. [775]
- MOREIRA, H., AND M. J. MOREIRA (2013): "Contributions to the Theory of Optimal Tests," Working Paper, FGV/EPGE. [772,782,783,791]
- MOREIRA, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048. [772,773,779]
- MÜLLER, U. K. (2011): "Efficient Tests Under a Weak Convergence Assumption," *Econometrica*, 79, 395–435. [776]
- (2014): "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business & Economic Statistics*, 32, 311–322. [805]
- MÜLLER, U. K., AND M. W. WATSON (2013a): "Measuring Uncertainty About Long-Run Predictions," Working Paper, Princeton University. [783,805]
- (2013b): "Low-Frequency Robust Cointegration Testing," *Journal of Econometrics*, 174, 66–81. [779,805]
- PHILLIPS, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547. [797]
- (2014): "On Confidence Intervals for Autoregressive Roots and Predictive Regression," *Econometrica*, 82, 1177–1195. [799]
- SALAEVSKII, O. V. (1963): "On the Non-Existence of Regularly Varying Tests for the Behrens–Fisher Problem," *Soviet Mathematics–Doklady*, 4, 1043–1045. [772]

- SRIANANTHAKUMAR, S., AND M. L. KING (2006): "A New Approximate Point Optimal Test of a Composite Null Hypothesis," *Journal of Econometrics*, 130, 101–122. [774,784]
- STOCK, J. H., AND M. W. WATSON (1996): "Confidence Sets in Regressions With Highly Serially Correlated Regressors," Working Paper, Harvard University. [798]
- STOYE, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315. [774,800–802]
- TSUI, K.-W., AND S. WEERAHANDI (1989): "Generalized  $p$ -Values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters," *Journal of the American Statistical Association*, 84, 602–607. [793]
- VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge, U.K.: Cambridge University Press. [772,776,786,803,806]
- (2002): "The Statistical Work of Lucien Le Cam," *The Annals of Statistics*, 30, 631–682. [806]
- WALD, A. (1943): "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large," *Transactions of the American Mathematical Society*, 54, 426–482. [777]
- WANG, Y. Y. (1971): "Probabilities of the Type I Errors of the Welch Tests for the Behrens–Fisher Problem," *Journal of the American Statistical Association*, 66, 605–608. [793]
- WELCH, B. L. (1947): "The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28–35. [793]
- WOUTERSEN, T. (2006): "A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters," Unpublished Manuscript, Johns Hopkins University. [774,800]

*University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, U.S.A.; grelliott@ucsd.edu,*

*Dept. of Economics, Princeton University, Princeton, NJ 08544, U.S.A.; umueller@princeton.edu,*

*and*

*Dept. of Economics, Princeton University, Princeton, NJ 08544, U.S.A. and NBER; mwatson@princeton.edu.*

*Manuscript received January, 2012; final revision received September, 2014.*