Nearly Optimal Tests when a Nuisance Parameter is Present Under the Null Hypothesis*

Graham Elliott
UCSD

Ulrich K. Müller
Princeton University

Mark W. Watson
Princeton University
and NBER

November 2011

Abstract

This paper considers nonstandard hypothesis testing problems that involve a nuisance parameter. We establish a 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.

JEL classification: C12; C21; C22

Keywords: Least favorable distribution, composite hypothesis, maximin tests

*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 participants of the AMES 2011 meeting at Korea University and of a workshop at USC for helpful comments.

1 Introduction

This paper considers statistical hypothesis tests concerning a parameter $\theta = (\beta', \delta')'$ where β is a parameter of interest and δ is a nuisance parameter. Both the null and alternative are composite

$$H_0 : \beta = \beta_0, \delta \in \Delta \quad \text{against} \quad H_1 : \beta \in B, \delta \in \Delta \tag{1}$$

so that the null specifies the value of β, but not δ.

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

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

and the positive definite covariance matrix Σ is known. With an unrestricted nuisance parameter space Δ, there are good reasons for simply ignoring Y_δ, even if Σ is not block-diagonal: For scalar β, the one-sided test (1) based on Y_β 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 β_0 and unrestricted B by either imposing an appropriate rotational invariance for Y_β, by focussing on most 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 Δ, however. Outside this class it is sometimes possible to deal with 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 below.

A second approach is to impose similarity, unbiasedness or conditional unbiasedness. In particular, conditioning on a statistic that is sufficient for δ ensures by construction that conditional distributions no longer depend on δ. 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.[1]

A general solution to hypothesis tests in the presence of a nuisance parameter is obtained by integrating out the parameter θ with respect to some probability distribution under the null and alternative, respectively. The test statistic is then simply the likelihood ratio of the resulting integrated null and alternative densities. In this approach, the probability distribution under the alternative can be freely chosen and represents the relative weights a researcher attaches to the power under various alternatives. The resulting test is then optimal in the sense of maximizing weighted average power. The probability distribution over δ under the null hypothesis, in contrast, has to be carefully matched to the problem and weighting function. Technically, the null distribution 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 extremely 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), which we determine numerically. The ALFD plays two conceptually distinct roles: on the one hand, it yields an analytical upper bound on the weighted average power of *all* valid tests. On the other hand, 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

[1] 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 the smaller sample has at least three observations. More recently, Andrews (2011) shows that similar tests in a moment inequality model exist, but have poor power.

upper bound. The approach can be extended to tests that switch to a given "standard" test (with high probability) in particular regions of the nuisance 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 efficient 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 β and δ, where δ is known to be non-negative. 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. The third 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. Fourth, we consider inference about the break date in a time series model with a single structural change. In this problem δ is related to the size of the parameter break, and ruling out small breaks (as, for example Bai (1994, 1997) and much of the subsequent literature) may lead to substantially over-sized 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 fifth example considers near 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 β 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) contains 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 bound.

In all of these applications, δ is one dimensional. Generally, high dimensional δ pose substantial numerical difficulties for the algorithm. In two companion papers, Müller and Watson (2009) and Elliott and Müller (2009), we successfully apply the algorithm to two

additional time series problems with a two-dimensional nuisance parameter.

The hypothesis testing problem (1) can be recast in the form of a general decision problem, and we do so in the concluding Section 6 below. In that form, the analytical result on the upper bound for the power of any valid test is a consequence of the well-known Minimax Theorem of classical decision theory (see, for instance, Ferguson (1967), or Chamberlain (2000) for a recent application). In this general form, the analogous bound plays a prominent role in the numerical determination of the least favorable prior distribution suggested by Kempthorne (1987). Specialized to the hypothesis testing problem, Andrews, Moreira, and Stock (2008) discuss and employ the upper bound on power in the context of inference with weak instruments. Srianantha kumar and King (2006) numerically determine tests for a composite null hypothesis in the form of the generalized Neyman-Pearson Lemma, and apply it to specification tests for Gaussian time series models. Finally, Chiburis (2009) provides a numerical approach to determine the optimal critical region directly. This has the advantage of yielding a linear programming problem, but the disadvantage that feasibility requires a low dimensional sample space (whereas for our approach, only the dimension of the nuisance parameter space is relevant).

The remainder of the paper is organized as follows. Section 2 formally states the problem, introduces the running example and states the analytical power bound result. Section 3 describes the algorithm to determine the ALFD, and Section 4 extends the approach to tests that are constrained by a switching rule to a given standard test. Section 5 contains the results for the additional five examples. We conclude with a particular decision theoretic justification for the least favorable distribution, and offer a few remarks about the resulting AFLD tests and Bayes factors. The Appendix contains additional details on the algorithm, and the six 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 S. 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 ν. Based on this single observation, we seek to test the hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1 \tag{3}$$

where $\Theta_0 \cap \Theta_1 = \varnothing$ and Θ_0 is not a singleton, so that the null hypothesis is composite.

Tests of (3) are measurable functions $\varphi : S \mapsto [0,1]$, where $\varphi(y)$ indicates the rejection probability conditional on observing $Y = y$. Thus, a non-randomized test has restricted range $\{0,1\}$, and CR$= \{y : \varphi(y) = 1\}$ is its critical region. If $\varphi(y) \in (0,1)$ for some $y \in S$, then φ 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 α test has size smaller or equal to α.

In many problems, a composite null hypothesis arises due to the presence of a nuisance parameter. In a typical problem, θ can be parametrized 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

$$H_0 : \beta = \beta_0,\, \delta \in \Delta \quad \text{against} \quad H_1 : \beta \in B,\, \delta \in \Delta \tag{4}$$

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

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. Suppose we observe n observations from a parametric model with parameter $a = (b, d) \in \mathbb{R}^2$. The hypothesis of interest is $H_0 : b = b_0$, and it is known a priori that $d \geq d_0$. For instance, b and d may correspond to regression coefficients, and it is known that the marginal effect of the control variable is non-negative. Let $\beta = \sqrt{n}(b - b_0)$ and $\delta = \sqrt{n}(d - d_0)$. If the model is locally asymptotic normal at $(b, d) = (b_0, d_0)$ at the usual parametric $\sqrt{n}$ rate with non-singular Fisher information matrix Σ^{-1}, then by Corollary 9.5 of van der Vaart (1998), the Limit Experiment local to (b_0, d_0) concerns the bivariate normal observation

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

where Σ is known. The hypothesis testing problem concerning (5) is

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

where $B = (0, \infty)$ and $B = \mathbb{R} \backslash \{0\}$ correspond to one-sided and two-sided alternatives, respectively. It clear that in either case, we can normalize Σ 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)), the local asymptotic rejection profile of any test in the original n observation problem can be matched by a test in the single observation problem (5). What is more, for any test of (5), one can 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 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 typically, both the null and the alternative hypotheses are composite, so that one cannot directly appeal to the Neyman Pearson Lemma. In this subsection, we consider the complication that arises from a composite nature of the alternative, i.e. if Θ_1 is not a singleton.

A standard solution to this problem is to consider weighted average power as the scalar criterion to choose among tests

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

where F is a probability measure with support on (the closure of) Θ_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 $\int \varphi \left(\int f_\theta dF(\theta) \right) d\nu$ (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 $h = \int f_\theta dF(\theta)$. Thus, with a WAP criterion, the hypothesis testing problem (3) effectively becomes

$$H_0 : \text{the density of } Y \text{ is } f_\theta,\ \theta \in \Theta_0 \qquad \text{against} \qquad H_{1,F} : \text{the density of } Y \text{ is } h = \int f_\theta dF(\theta) \tag{8}$$

and under the simple alternative hypothesis $H_{1,F}$, the density h of Y is a mixture of f_θ, with mixing weights F. 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 discuss the choice of F in more detail in Section 4.

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 good tests of (8). This subsection does not derive such tests directly, but rather provides a general set of bounds on the power of any level α test. These bounds are useful both for constructing efficient tests 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 Λ is a probability distribution with support on Θ_0. In general, the size α Neyman-Pearson test of $H_{0,\Lambda}$ against $H_{1,F}$ is *not* a level α test of H_0 in (8), as its null rejection probability is equal to α 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 α test of (8).

Lemma 1 *Let φ_Λ be the size α test of $H_{0,\Lambda}$ against $H_{1,F}$ of the Neyman-Pearson form*

$$\varphi_\Lambda(y) = \begin{cases} 1 \text{ if } h(y) > \text{cv} \int f_\theta(y) d\Lambda(\theta) \\ p \text{ if } h(y) = \text{cv} \int f_\theta(y) d\Lambda(\theta) \\ 0 \text{ if } h(y) < \text{cv} \int f_\theta(y) d\Lambda(\theta) \end{cases} \tag{9}$$

for some $\text{cv} \geq 0$ *and* $p \in [0,1]$*. Then for any level* α *test* φ *of* H_0 *against* $H_{1,F}$*,* $\int \varphi_\Lambda h d\nu \geq \int \varphi h d\nu$*.*

Proof. *Since* φ *is a level* α *test of* H_0*,* $\int \varphi f_\theta d\nu \leq \alpha$ *for all* $\theta \in \Theta$*. Therefore,* $\int\int \varphi f_\theta d\nu d\Lambda(\theta) = \int\int \varphi f_\theta d\Lambda(\theta) d\nu \leq \alpha$*, where the equality follows from Fubini's Theorem, so that* φ *is also a level* α *test of* $H_{0,\Lambda}$ *against* $H_{1,F}$*. The result now follows from the Neyman-Pearson Lemma.* ■

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 φ_Λ can be easily computed. Thus, Lemma 1 provides a set of explicit power bounds on the original problem, indexed by the distribution Λ.

Running example, ctd: Suppose $\rho = 1/2$ 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 Λ degenerate with all mass at $\theta_0 = (0,1)'$. The likelihood ratio test φ_Λ 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 φ_Λ 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, $X | H_{1,F} \sim \mathcal{N}(1,3)$, so that the power of φ_Λ is given by $\int \varphi_\Lambda h d\nu = 0.31$. While φ_Λ 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 can usefully be thought of as generalizing a standard result concerning tests with a composite null hypothesis; see, for instance, Theorem 3.8.1 of Lehmann and Romano (2005): A distribution Λ^{**} is *least favorable* if the best level α test of $H_{0,\Lambda^{**}}$ against the single alternative $H_{1,F}$ is also of level α 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 Λ. This is useful because in many contexts, it is difficult to identify the least favorable distribution Λ^{**} (and it may not even exist).

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

It is sometimes possible to construct an *ad hoc* test φ_{ah} of (3) that is known to be of level α, even if the nuisance parameter space is high dimensional, but φ_{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 φ_{ah} is close to the power bound arising from some distribution Λ, then φ_{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 φ_{ah} and the bound can arise either because φ_{ah} is inefficient, or because this specific Λ yields a bound far above the least upper bound.

For this strategy to work, one must try to guess a Λ 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 h under the alternative $H_{1,F}$. This may suggest a suitable choice of Λ directly. Alternatively, one can parametrize Λ in some suitable fashion, and numerically minimize some convenient distance between $\int f_\theta d\Lambda(\theta)$ and h. For example, the testing problem of Müller and Watson (2009) 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 $\delta \in \Delta$. With $\Lambda = \Lambda_\delta$ restricted to put point mass at some δ, one can use the Kullback-Leibler divergence between the null and alternative density as a convenient distance function, and use numerical methods to find Λ_δ. 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) show 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 Λ^{**} has the property that the level α Neyman-Pearson test $\varphi_{\Lambda^{**}}$ of the simple hypothesis $H_{0,\Lambda^{**}}$ against $H_{1,F}$ also yields a level α test of the composite null hypothesis H_0 against $H_{1,F}$. As noted above, for many problems it is very difficult to analytically determine Λ^{**}. A natural reaction is then to try to numerically approximate Λ^{**}. In general, though, such an approach lacks a criterion to determine when a candidate approximation is "good enough".

Lemma 1 is very useful in this regard. Specifically, consider the following definition of an

approximate least favorable distribution (ALFD).

Definition 1 *An* ε-ALFD *is a probability distribution* Λ^* *on* Θ_0 *satisfying*

(i) the Neyman-Pearson test (9) with $\Lambda = \Lambda^*$ *and* $(\mathrm{cv}, p) = (\mathrm{cv}^*, p^*)$, φ_{Λ^*}, *is of size* α *under* H_{0,Λ^*}, *and has power* $\bar{\pi}$ *against* $H_{1,F}$*;*

(ii) there exists $(\mathrm{cv}^{*\varepsilon}, p^{*\varepsilon})$ *such that the test (9) with* $\Lambda = \Lambda^*$ *and* $(\mathrm{cv}, p) = (\mathrm{cv}^{*\varepsilon}, p^{*\varepsilon})$, $\varphi^{\varepsilon}_{\Lambda^*}$, *is of level* α *under* H_0, *and has power of at least* $\bar{\pi} - \varepsilon$ *against* $H_{1,F}$.

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

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

3 Numerical Determination of the ALFD

We now discuss a numerical algorithm that determines an approximately least favorable distribution Λ^*. The basic idea is to discretize the problem by specifying a finite set of base distributions Ψ_i on Θ_0, $i = 1, \cdots, M$. The ALFD Λ^* is then constructed as a mixture of these M base distributions. An important special case for the base distributions Ψ_i are point masses, but in many problems, it makes sense to impose some continuity and to use non-degenerate Ψ_i's.

Any test that is of level α under the composite hypothesis H_0 must have rejection probability of at most α when Y is drawn from $f_i = \int f_\theta d\Psi_i(\theta)$, $i = 1, \cdots, M$. Let J_N be a subset of N of the M baseline indices, $J \subset \{1, 2, \cdots, M\}$, and consider first the simpler problem where it is known that Y is drawn from f_i, $i \in J_N$ under the null. In this restricted set-up, the least favorable distribution is described by an N dimensional multinomial distribution P^*_N, that is by a point in the N dimensional simplex $p_i \geq 0$, $\sum_{i \in J_N} p_i = 1$. Relative to this restricted null, the best test, say φ^*_N, is of the form $\varphi^*_N(y) = \mathbf{1}[h(y) > \mathrm{cv}_N \sum_{i \in J_N} p^*_i f_i(y)]$, and can thus be characterized by the parameters (cv_N, P^*_N).[2] By Theorem 3.8.1 of Lehmann

[2] We restrict attention to problems where no randomization is necessary.

and Romano (2005), (cv_N, P_N^*) yield a test with two key properties: (i) $\int \varphi_N^* f_i d\nu \leq \alpha$ for $i \in J_N$ and (ii) $\int \varphi_N^* f_i d\nu < \alpha$ only if $p_i^* = 0$. When N is small (say, $N \leq 15$), numerical methods can be used to find (cv_N, P_N^*) that satisfy these two properties. By construction, the test φ_N^* has rejection probability of at most α when Y is drawn from f_i, $i \in J_N$. The algorithm now seeks to identify a set J_N so that the corresponding test $\varphi_N^{\varepsilon *}$ with slightly larger critical value $\text{cv}^{*\varepsilon}$ has null rejection probability below α under H_0. The number and type of base distributions M that are required for this to be possible depends on the continuity of the rejection probability of $\varphi_N^{\varepsilon *}$ under the null as a function of θ. In general, with a large set of point-mass like base distributions Ψ_i that evenly cover Θ_0, the largest size under f_i, $i = 1, \cdots, M$ is also close to the largest size under H_0. For computational reasons, though, it makes sense to pick M as small as possible.

Concretely, the algorithm consists of the following steps:

1. Pick an initial set of M base distributions. Set $N = 1$, and initialize $J_N = \{1\}$.

2. Determine the least favorable distribution P_N^* and critical value cv_N when it is known that Y is drawn from f_i, $i \in J_N$ by finding (cv_N, P_N^*) that satisfy (i) $\int \varphi_N^* f_i d\nu \leq \alpha$ for $i \in J_N$ and (ii) $\int \varphi_N^* f_i d\nu < \alpha$ only if $p_i^* = 0$. If some of the p_i^* are zero, then the corresponding elements of J_N are dropped, and N reduced accordingly.

3. Determine the power bound $\bar{\pi} = \int \varphi_N^* h d\nu$ of the test $\varphi_N^*(y) = \mathbf{1}[h(y) > \text{cv}_N \sum_{i \in J_N} p_i^* f_i(y)]$.

4. Determine the critical value $\text{cv}_N^\varepsilon > \text{cv}_N$ so that the test $\varphi_N^{\varepsilon *}(y) = \mathbf{1}[h(y) > \text{cv}_N^\varepsilon \sum_{i=1}^N p_i^* f_i(y)]$ has power $\int \varphi_N^{\varepsilon *} h d\nu = \bar{\pi}_N - \varepsilon$.

5. Compute the rejection probability of $\varphi_N^{\varepsilon *}$ under f_i, $i = 1, \cdots, M$. If the rejection probability exceeds α under some f_i, add this i to J_N, and go to step 2.

6. Check if $\varphi_N^{\varepsilon *}$ is of level α under H_0. If it is not, add more concentrated base distributions, and go to step 5.

Appendix A contains details on the implementation of the various steps. Importantly, Lemma 1 implies that the power bound computed from *any* distribution Λ is valid, so that the algorithm achieves an ε-ALFD with an approximate solution in Step 2 as long as rejection frequencies in Steps 3, 4 and 6 are accurately computed.[3]

[3] Elliott and Müller (2009) develop a technique for checking size control of a given test for *all* values of $\delta \in \Delta$, and not just on a fine grid, so that the usual Monte Carlo error is the only remaining numerical approximation. For simplicity, we do not pursue this here, and instead work with a grid.

4 Switching to Standard Tests

Nearly Standard Problem in Part of Parameter Space. For many nonstandard testing problems involving a nuisance parameter $\delta \in \mathbb{R}^{k_\delta}$, one can choose a parameterization in which the problem for large values of $||\delta||$ essentially reduces to a standard problem. For example, in the weak instrument problem with concentration coefficient δ, a large $|\delta|$ implies that the instruments are "almost" strong. Also, 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 $\delta \to \infty$. This is well-recognized in practice, as non-standard asymptotic arguments are only invoked if small to moderate values of δ are considered plausible. This plausibility can usually be gauged by considering an estimator $\hat{\delta}$ of δ: if $||\hat{\delta}||$ is large, then the standard mode of inference is applied, whereas small realizations of $||\hat{\delta}||$ are followed by inference using the non-standard asymptotic embedding. Staiger and Stock's (1997) rule of thumb to revert to standard TSLS inference if the first stage F statistic is larger than 10 is a prominent example of this approach.

In this section, we formalize this notion of switching to a standard test, and describe how to numerically determine a (nearly) efficient test conditional on such switching. We focus on tests of the form

$$\varphi_{D,S,\chi}(y) = (1 - \chi(y))\varphi_D(y) + \chi(y)\varphi_S(y) \tag{10}$$

where $\chi \mapsto \{0,1\}$ is a "switching rule" (such as $\chi(y) = \mathbf{1}[||\hat{\delta}|| > K]$), φ_S is a "Standard" test and φ_D is the test for the "Difficult" part of the parameter space. With χ and φ_S given, a restriction to tests of the form (10) can be viewed as a constraint on the critical region of the overall test $\varphi_{D,S,\chi}$: different functions $\varphi_D \mapsto \{0,1\}$ only affect the critical region in the subset $\{y : \chi(y) = 0\}$.

Running example, ctd: With $\hat{\delta} = Y_\delta$, a realization of $\hat{\delta} > 6$ is extremely unlikely when $\delta \leq 0$. Thus, when $\hat{\delta} > 6$, knowledge of the constraint $\delta \geq 0$ is not valuable, and one might as well switch to the usual two-sided t-test $\varphi_S(y) = \mathbf{1}[|y_\beta| > 1.96]$ via $\chi(y) = \mathbf{1}[\hat{\delta} > 6]$. With these φ_S and χ taken as given, the remaining task is to determine the critical region for realizations where $Y_\delta \leq 6$. ▲

Choice of Weighting Function F. In several of the applications we consider, we implement switching to a standard test. The (conditional) weighted average power optimality then only assesses the quality of the overall test for the genuinely nonstandard test component. Accordingly our choice of weighting function is guided by three considerations: First, it makes sense to put weight on alternatives in a way that yields a smooth transition

to the critical region of the standard test in the case of a switch. Typically, this can be achieved by an approximately uniform weighting along the nuisance parameter dimension in its natural parametrization. Second, for two-sided problems where φ_S is symmetric, we treat both directions symmetrically also in terms of the weighting function F, even if this yields an asymmetric power function of the WAP maximizing test. Finally, as argued by King (1988), it makes sense to focus on alternatives where good tests achieve power of roughly 50%. This ensures that the power function is tangent to the power envelope at that point, which will yield an overall good test also for more and less distant alternatives in many well behaved problems. For two-sided problems with asymmetric power, the 50% rule is applied to the simple average of the two directions.

Running example, ctd: In the example (where we switch to the standard test when $\hat{\delta} > 6$) these considerations lead us to construct weighted average power using an F in which δ is uniformly distributed on $[0, 8]$, and β takes on the values -2 and 2 with equal probability. ▲

Power Bound under Switching. The following Lemma is the straightforward generalization of the power bound Lemma 1 above to tests of the form (10).

Lemma 2 *For given χ and φ_S, let SW be the set of tests of the form (10). Let $\varphi_{\Lambda,S,\chi} \in SW$ be of size α under $H_{0,\Lambda}$ with φ_Λ of the Neyman-Pearson form*

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

for some $\text{cv} \geq 0$ *and* $p \in [0, 1]$. *Then for any test $\varphi \in SW$ that is of level α under H_0, $\int \varphi_{\Lambda,S,\chi} h d\nu \geq \int \varphi h d\nu$.*

Proof. Note that by construction, $\int (\varphi_{\Lambda,S,\chi} - \varphi)(h - \text{cv} \int f_\theta d\Lambda(\theta)) d\nu \geq 0$. Since φ is of level α under H_0, $\int \varphi f_\theta d\nu \leq \alpha$ for all $\theta \in \Theta$. Therefore, $\int \int \varphi f_\theta d\nu d\Lambda(\theta) = \int \varphi (\int f_\theta d\Lambda(\theta)) d\nu \leq \alpha$, 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. ■

Just like Lemma 1, Lemma 2 provides an explicit form of tests whose power constitutes an upper bound for any test of the form (2). Note, however, that particular choices for φ_S and χ can induce SW to be empty: reconsider the weak instrument problem, for instance, where χ is Staiger-Stock's rule of thumb and φ_S the TSLS-based test of nominal level α. Since under weak instruments, the rejection probability of φ_S asymptotes to α from above as $\delta \to \infty$, $\varphi_{D,S,\chi}$ will over-reject for large finite values of δ, even with $\varphi_D = 0$ in (10). In

such a case, one can either rely on a slightly larger critical value for the standard test, or use a critical value in φ_S that is an appropriate function of $\hat{\delta}$ (so that φ_S is only approximately equal to the standard test).

Incorporating Switching into the Algorithm. The algorithm for numerically determining an approximately least favorable Λ^*, conditional on the switching rule to the standard test, is a minor modification of the algorithm described in Section 3 above. Specifically, recall that without the switching, tests were of the form $\varphi_N^*(y) = \mathbf{1}[h(y) > \text{cv}_N \sum_{i \in J_N} p_i^* f_i(y)]$. To implement the switching in this convenient structure, replace h in the definition of φ_N^* by

$$\tilde{h}(y) = \begin{cases} \infty & \text{if } \chi(y)\varphi_S(y) = 1 \\ 0 & \text{if } \chi(y)(1 - \varphi_S(y)) = 1 \\ h(y) & \text{otherwise} \end{cases} \tag{11}$$

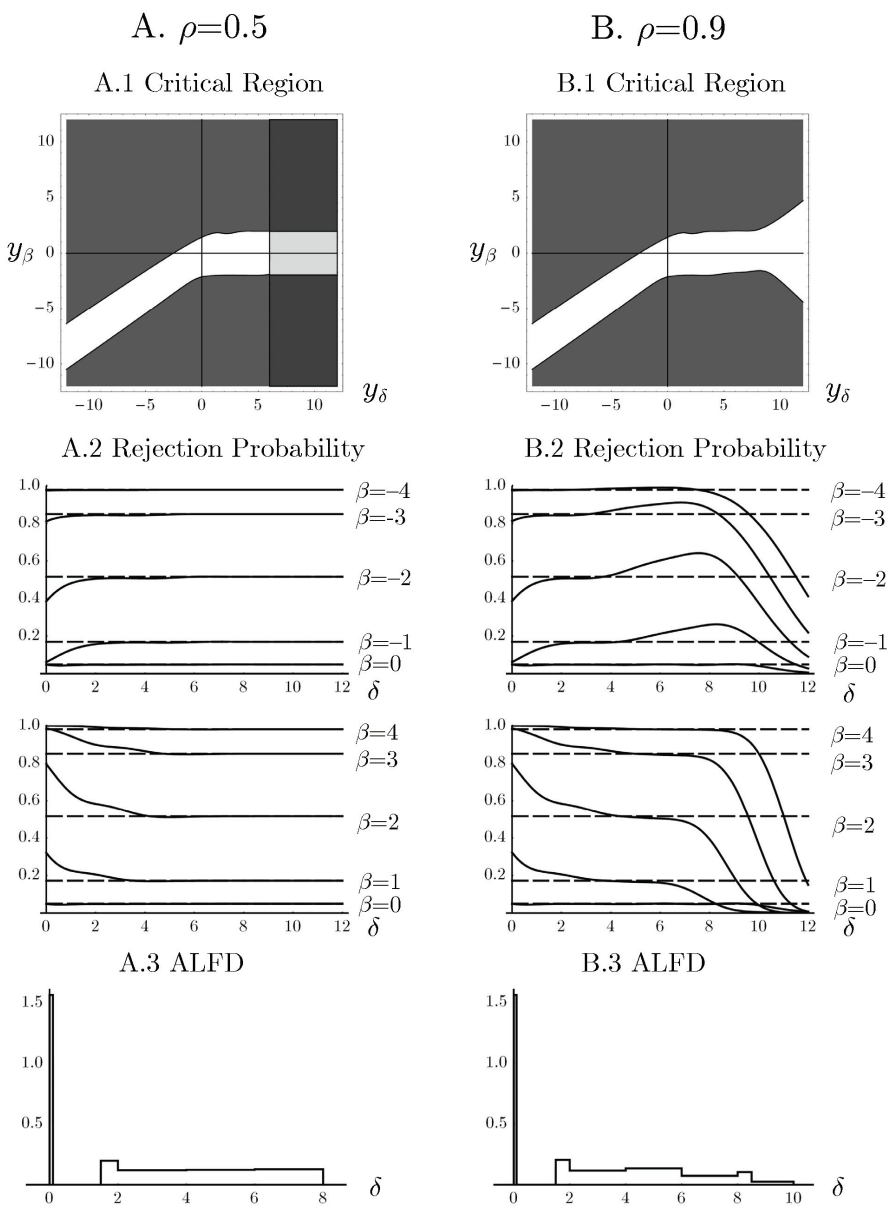
so that now $\varphi_N^*(y) = \mathbf{1}[\tilde{h}(y) > \text{cv}_N \sum_{i \in J_N} p_i^* f_i(y)]$. Intuitively, the replacement of h with $\tilde{h}$ ensures that in the case of a switch, $\chi(y) = 1$, the ratio of the (modified) "density" $\tilde{h}$ to the null density is such that the Neyman-Pearson test rejects if and only if the standard test φ_S does. By Lemma 2, the power bound $\bar{\pi} = \int \varphi_N^* h d\nu$ of step 3 of the algorithm then provides an upper bound on the power of all tests of the form (10).

Numerical Results for Running Example. Figure 1 summarizes the results for the running example with $\rho = 0.7$ for tests of level $\alpha = 5\%$. As discussed above, weighted average power was computed using an F that puts uniform weight on $\delta \in [0, 8]$ and $\beta \in \{-2, 2\}$. The AFLD 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. Rejection frequencies were approximated using 100,000 Monte Carlo draws. Panel A shows results for the test that switches to the standard test $\varphi_S(y) = 1[|y_\beta| > 1.96]$ if $y_\delta \geq 6$. For comparison, panel B shows results for the test that does not switch.

The white and light gray band in the center of panel A.1 is the acceptance region of the nearly optimal test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$, 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 φ_S for values of y_δ as small as $y_\delta = 1$. As y_δ becomes negative the critical region is approximately $|y_\beta - \rho y_\delta| > 1.96(1 - \rho^2)^{1/2}$, which is recognized as the critical region of the best test under the assumption $\delta = 0$.

Panel A.2 shows power (plotted as a function of δ) for selected values of β. The solid curves show the power of the nearly optimal test and the dashed lines shows the power of the

Figure 1: Positive Nuisance Parameter



Notes: 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 $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ (panel A) and $\varphi^{\varepsilon}_{\Lambda^*}$ (panel B), and dashed lines are for the usual test that ignores Y_δ, $\varphi_S(y) = 1[|y_\beta| > 1.96]$.

standard test φ_S. The figures show that power is asymmetric in β, with substantially lower power for negative values of β when δ is small; this is is consistent with the critical region shown in panel A.1 where negative values of β and small values of δ 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 $\delta \in [0, 8]$, the optimal test maximizes the average of the power curves for $\beta = -2$ and $\beta = 2$ in A.3 over $\delta \in [0, 8]$. Weighted average power of $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ is higher than the power of φ_S for all pairs of values for β shown in the figure.

Panels B show corresponding results for the nearly optimal test $\varphi^{\varepsilon}_{\Lambda^*}$ that does not impose switching to a standard test, computed using the algorithm in Section 3 without the modification (11). Because F only places weight on values of δ that are less than 8, this test sacrifices power for values of $\delta > 8$ to achieve more power for values of $\delta \leq 8$. The differences between the power function for $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ (shown in panel A) and $\varphi^{\varepsilon}_{\Lambda^*}$ (shown in panel B) highlights the attractiveness of switching to a standard test: it allows F to be chosen to yield high average power in the difficult portion of the parameter space (small values of δ) 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.

Overall Optimality. By construction, the algorithm with $\tilde{h}$ in place of h determines a test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi} \in SW$ whose weighted average power is within ε of the power bound *conditional* on the switching rule. It can also be interesting to compare the power of the resulting test with the unconditional power bound obtained from the original algorithm as described in Section 3. Specifically, if the standard test φ_S has some optimality property for $||\delta||$ large, then one can potentially make the stronger claim that $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ is (i) is close to maximizing weighted average power for small $||\delta||$, and (ii) inherits the optimality property of φ_S for $||\delta||$ large.

Running example, ctd: Consider the running example with one-sided alternative $\beta > 0$ and the standard test $\varphi_S(y) = \mathbf{1}[y_\beta > 1.645]$. It is straightforward to show that this test is the uniformly best test against alternatives $A = \{(\beta_1, \delta_1) : \delta_1 - \rho\beta_1 \geq 0\}$, and is thus the best test when δ is large.[4] But, for δ large, the power of $\varphi^{\varepsilon *}_N$ with $\chi(y) = \mathbf{1}[y_\delta > 6]$ is almost identical to the power of φ_S, since $\chi(Y) = 1$ with very high probability. For instance, for $\delta \geq 10$, $P(\chi(Y) = 1) = P(\mathcal{N}(\delta, 1) \geq 6) > 0.9999$. Thus, for all $\delta_1 \geq 10$, the power of

[4]Note that $\varphi_S(y)$ is the Neyman-Pearson test for the two single hypotheses $H_0^s : \beta = 0, \delta = \delta_1 - \rho\beta_1$ versus $H_1^s : \beta = \beta_1, \delta = \delta_1$.

$\varphi_N^{\varepsilon *}$ is within 0.01 percentage points of the power of φ_S. Furthermore, an application of the algorithm yields that for $\rho = 0.9$, the unconstrained power bound relative to a weighting function F that is uniform on $\delta \in [0, 10]$ and $\beta = 2$ is $\bar{\pi} = 0.675$, and the weighted average power of the 5% level switching test 0.669. Thus, the switching test is within 0.6 percentage points of maximizing weighted average power relative to F that places all of its weight on small values of δ, and at the same time, for large δ, it has essentially the same power as the uniformly best test. ▲

5 Applications

In this section we apply the algorithm outlined above to construct the AFLD optimal test for five non-standard problems. In all of these problems we set $\varepsilon = 0.005$ (so that the ALFD test is with 0.5% of the power bound), and approximate rejection probabilities using 100,000 Monte Carlo draws. Appendix B contains further details on the computations in each of the problems.

Most problems considered in this paper are indexed by a known, or at least consistently estimable parameter (ρ in the running example). Appendix C contains tables that describe a nearly optimal test for all practically relevant values of this parameter. Readers interested in applications of the tests derived in this paper should thus consult Appendix B after having read the relevant subsection below.

5.1 The Behrens-Fisher Problem

Suppose we observe i.i.d. samples from two normal populations $x_{1,i} \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $i = 1, \cdots, n_1$ and $x_{2,i} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $i = 1, \cdots, n_2$, where $2 \leq n_1 \leq n_2$. We are interested in testing $H_0 : \mu_1 = \mu_2$ without knowledge of σ_1^2 and σ_2^2. This is the "Behrens-Fisher" problem, which has a long history in statistics.

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 mean 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 $(\bar{x}_1, \bar{x}_2, s_1, s_2) \to (\bar{x}_1 + m, \bar{x}_2 + m, s_1, s_2)$ and $(\bar{x}_1, \bar{x}_2, s_1, s_2) \to (c\bar{x}_1, c\bar{x}_2, cs_1, cs_2)$ for $m \in \mathbb{R}$ and $c > 0$ further reduces the problem to the two dimensional maximal invariant Y

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

Note that Y_β is the usual two-sample t-statistic which converges to $\mathcal{N}(0,1)$ under the null hypothesis as $n_1, n_2 \to \infty$. 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

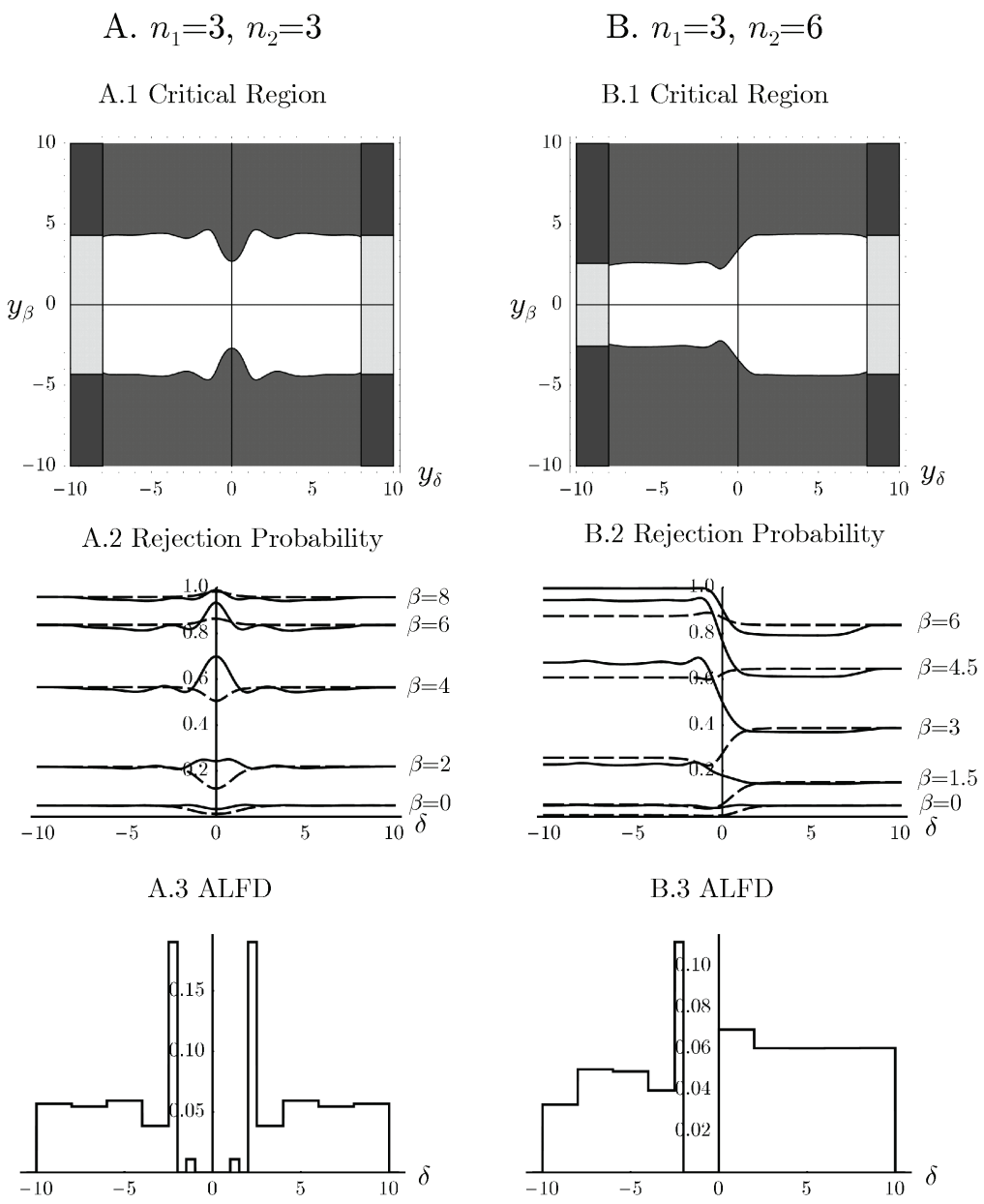
$$H_0 : \beta = 0,\ \delta \in \mathbb{R} \quad \text{against} \quad H_1 : \beta \neq 0,\ \delta \in \mathbb{R}. \tag{12}$$

While the well-known two-sided test of Welch (1947) with "data dependent degrees of freedom" approximately controls size for moderate sample sizes (Wang (1971) and Lee and Gurland (1975)), it is substantially over-sized when n_1 and n_2 are small; moreover, its efficiency properties are unknown. Thus, we employ the algorithm described above to compute nearly optimal tests for $(n_1, n_2) = (3, 3)$ and $(n_1, n_2) = (3, 6)$.

To implement the algorithm, we choose F as uniform on $\delta \in [-10, 10]$ and $\beta = \{-4, 4\}$ when $(n_1, n_2) = (3, 3)$, and $\beta = \{-3, 3\}$ when $(n_1, n_2) = (3, 6)$, where these values of β yield WAP of approximately 50% in the two cases. For extreme values of Y_δ, we switch to the test that treats one of the groups as having effectively zero variance for the sample mean: $\chi(y) = \mathbf{1}[|\log(s_1/s_2)| > 8]$, and $\varphi_S(y) = \mathbf{1}[y_\delta > 0]\mathbf{1}[|t| > T_{n_1-1}(0.975)] + \mathbf{1}[y_\delta < 0]\mathbf{1}[|t| > T_{n_2-1}(0.975)]$, where $T_n(\alpha)$ is the α^{th} quantile of a Student-t distribution with n degrees of freedom. We compare the power of the resulting $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ test to the "conservative" test obtained by using the $1 - \alpha/2$ quantile of a student-t distribution with degrees of freedom equal to $\text{df} = n_1 - 1$, which is known be of level α (cf. Mickey and Brown (1966)).

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 non-standard part ($|y_\delta| < 8$) the critical region is much like the critical region of the standard test ($|y_\beta| > T_2(0.975)$) for values of $|y_\delta| > 2$, but includes smaller values of $|y_\beta|$ when y_δ is close to zero. Evidently, small values of $|y_\delta|$ suggest that the values of σ_1 and σ_2 are close, essentially yielding more degrees of freedom for the null distribution of y_β. This feature of the critical region translates in the greater power for $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ than the conservative test when δ 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 become "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 δ, where the test has more power when the larger group has smaller variance. Finally, the conservative test has a null rejection rejection frequency substantially less than 5% when $\delta < 0$ and weighted average power substantially below the nearly optimal test.

Figure 2: Behrens-Fisher Problem



Notes: Darker shades for $|y_\delta| \geq 8$ 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 $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$, and dashed lines are for the usual test t-test with critical value computed from the student-t distribution with $n_1 - 1$ degrees of freedom.

5.2 Inference about the Break Date in a Time Series Model

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

Under the alternative embedding for moderately sized breaks $T^{1/2}d \to \delta \in \mathbb{R}$, the parameter δ 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 τ by the amount d,

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

and the aim is to conduct inference about the break date τ. 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 ($\{y_t\} \to \{y_t + m\}$ for all m) requires that tests are a function of the demeaned data $y_t - \bar{y}$. Partial summing the observations yields

$$T^{-1/2}\sum_{t=1}^{\lfloor sT \rfloor}(y_t - \bar{y}) \sim G(s) = W(s) - sW(1) - \delta(\min(\beta, s) - \beta s) \tag{13}$$

for $s = j/T$ and integer $1 \leq j \leq T$. 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 β of the kink in its mean,

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

for some $\beta_0 \in B$. Elliott and Müller (2009) formally show that this is indeed the relevant asymptotic experiment for a moderate structural break in a well behaved parametric time series model.

By Girsanov's Theorem, the Radon-Nikodym derivative of the measure of G in (13) relative to the measure ν of the standard Brownian Bridge, evaluated at G, is given by

$$f_\theta(G) = \exp[-\delta G(\beta) - \tfrac{1}{2}\delta^2\beta(1-\beta)]. \tag{14}$$

To construct the AFLD test we choose F so that β is uniform on B, and $\delta \sim \mathcal{N}(0, 100)$. This places substantial weight on large values of δ and eliminates the need to switch to

a standard test for large $|\delta|$.[5] 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 β for five values of δ; 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 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

$$G(s) \to G(s) + c(\min(\beta_0, s) - \beta_0 s) \qquad \text{for all } c. \tag{15}$$

This invariance requirement eliminates the nuisance parameter δ under the null, and thus leads to a similar test. But the transformation (15) 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 near optimal test $\varphi^{\varepsilon}_{\Lambda^*}$ has indeed substantially larger power for most alternatives.

5.3 Predictive Regression with a Local-To-Unity Regressor

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

$$\begin{aligned} y_t &= a + bx_{t-1} + \varepsilon_{y,t} \\ x_t &= rx_{t-1} + \varepsilon_{x,t}, \;\; x_0 = 0 \end{aligned} \tag{16}$$

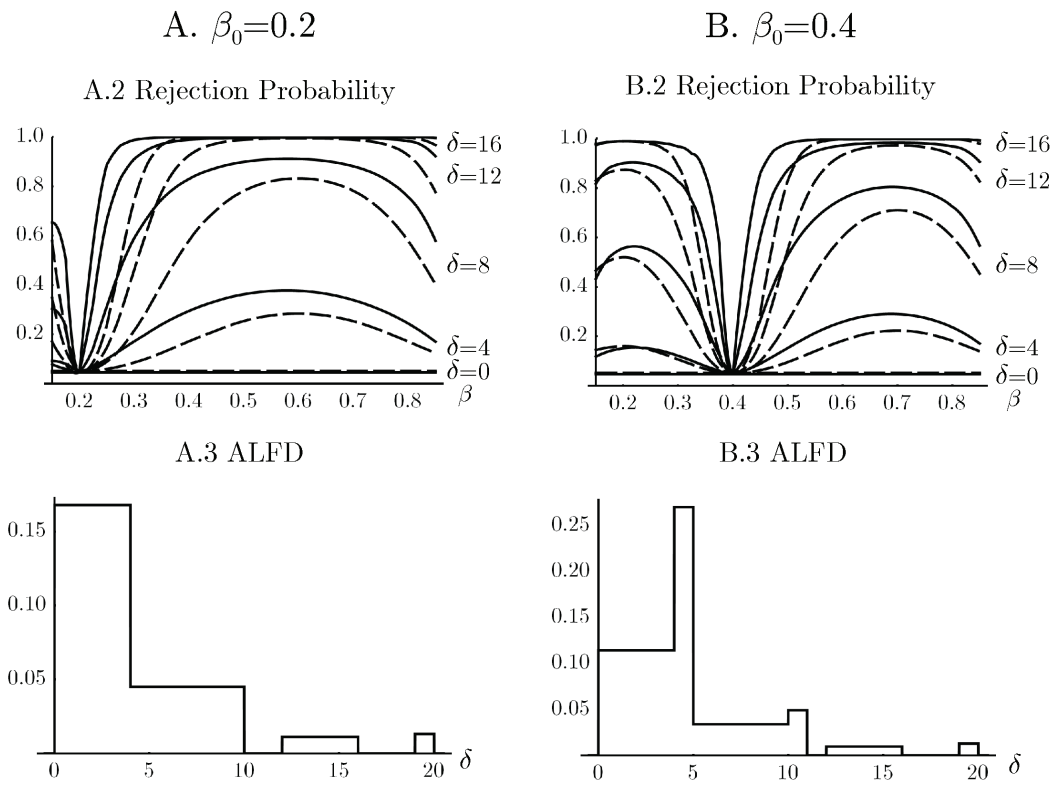
where $E(\varepsilon_{y,t}|\{\varepsilon_{x,t-j}\}_{j=1}^{t-1}) = 0$, so that the first equation is a predictive regression. The persistence in x_t is often modelled as a local-to-unity process (in the sense of Bobkoski (1983), Cavanagh (1985), Chan and Wei (1987) and Phillips (1987)) with $r = r_T = 1 - \delta/T$. Interest focuses on a particular value of b given by $H_0 : b = b_0$ (where typically $b_0 = 0$). When the long-run covariance between ε_y and ε_x is non-zero, the usual t-test on b is known to severely overreject unless δ is very large.

After imposing invariance to translations of y_t, $\{y_t\} \to \{y_t + m\}$, 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_θ of a bivariate Gaussian continuous time process G,

$$f_\theta(G) = K(G) \exp[\beta Y_1 - \delta Y_2 - \tfrac{1}{2}\left(\beta + \frac{\rho}{\sqrt{1-\rho^2}}\delta\right)^2 Y_3 - \tfrac{1}{2}\delta^2 Y_4] \tag{17}$$

[5] For $|\delta| > 20$, the discretization of the break date β becomes an important factor, even with 1,000 step approximations to Wiener processes. Since these discretizations errors are likely to dominate the analysis with typical sample sizes for even larger δ, we restrict attention to $\delta \in \Delta = [-20, 20]$.

Figure 3: Break Date



Notes: In panels A.1 and B.1, solid lines are the rejection probability of the nearly optimal test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$, and dashed lines are for Elliott and Müller's (2007) test that imposes an additional invariance.

where β is proportional to $T(b-b_0)$, $\rho \in (-1,1)$ is the known (long-run) correlation between $\varepsilon_{x,t}$ and $\varepsilon_{y,t}$, and $\theta = (\beta,\delta)' \in \mathbb{R}^2$ is unknown.

With an upper bound on the root r_T of x_t, $\delta \geq \underline{\delta}$, the one-sided asymptotic inference problem is

$$H_0 : \beta = 0, \delta \geq \underline{\delta} \text{ against } \quad H_1 : \beta > 0,\ \delta \geq \underline{\delta} \tag{18}$$

and the four dimensional sufficient statistic $Y = (Y_1, Y_2, Y_3, Y_4)$ has distribution

$$\begin{aligned} Y_1 &= \int_0^1 W^{\mu}_{x,\delta}(s) dW_y(s) + \left(\beta + \frac{\rho}{\sqrt{1-\rho^2}}\delta\right) \int_0^1 W^{\mu}_{x,\delta}(s)^2 ds \\ Y_2 &= \int_0^1 W_{x,\delta}(s) dW_{x,\delta}(s) - \frac{\rho}{\sqrt{1-\rho^2}} Y_1 \\ Y_3 &= \int_0^1 W^{\mu}_{x,\delta}(s)^2 ds, \qquad Y_4 = \int_0^1 W_{x,\delta}(s)^2 ds \end{aligned}$$

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

While several methods have been developed that control size in (18) (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 determine a weighed average power maximizing test within a parametric class of functions $\mathbb{R}^4 \mapsto \{0,1\}$, and Jansson and Moreira (2006) derive the best conditionally unbiased tests of (18), conditional on the specific ancillary (Y_3, Y_4). However, Jansson and Moreira (2006) report that Cambell and Yogo's (2006) test has higher power for most alternatives. We therefore compare the one-sided ALFD test to this more powerful benchmark. We set $\underline{\delta} = -5$, so that very explosive roots are ruled out, although in contrast to Campbell and Yogo (2006), we do not impose an upper bound on δ.

The maximum likelihood estimators $(\hat{\beta}, \hat{\delta})$ derived from (17) are

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

For δ large (where $Y_3 \approx Y_4 \approx T^{-1}(1-r_T^2)^{-1} \approx (2\delta)^{-1}$ and $\hat{\delta} \approx \mathcal{N}(\delta, 2\delta)$), $\hat{\beta}$ is approximately distributed $\mathcal{N}(\beta, 2\delta/(1-\rho^2))$. Alternatives with fixed β thus become more difficult to distinguish from the null hypothesis as δ increases. For moderate values of δ, alternatives that are roughly equally difficult to distinguish from the null hypothesis are those where

$$\beta = \tilde{\beta}\frac{1 + 0.07\delta}{\sqrt{1-\rho^2}}$$

for fixed $\tilde{\beta}$.[6] We choose F such that δ is uniform on $[0, 40]$, and $\tilde{\beta} = 5$. We further impose switching to inference based on the maximum likelihood t-test (using the observed information),

$$\varphi_S(Y) = \mathbf{1}[\frac{\hat{\beta}}{\sqrt{Y_3^{-1} + \frac{\rho^2}{1-\rho^2} Y_4^{-1}}} > \mathrm{cv}_S] \tag{19}$$

whenever $\hat{\delta} \geq 35$, that is $\chi(Y) = \mathbf{1}[\hat{\delta} \geq 35]$. The critical value cv_S equals the usual 5% level value of 1.645 when $\rho \geq 0$, but we choose $\mathrm{cv}_S = 1.75$ when $\rho < 0$. This slight adjustment compensates for the heavier tail of the t-test statistic for moderate values of δ and negative ρ.

Figure 4 shows that the resulting nearly optimal test has close to uniformly higher power than the test developed by Campbell and Yogo (2006). Unreported results show a similar picture when $\rho = \pm 0.9$.

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 the data generating value of of μ, which is only known to satisfy

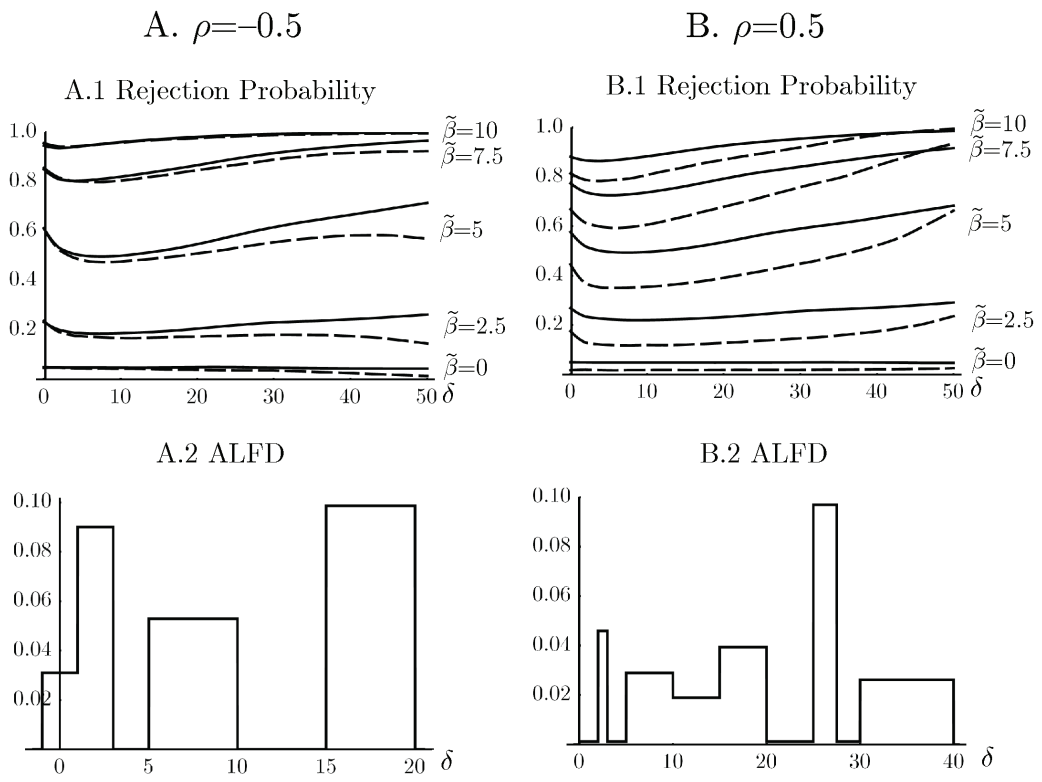
$$\mu_l \leq \mu \leq \mu_u. \tag{20}$$

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 μ_0 from Y_l and Y_u). Whilst under the null hypothesis the inequality (20) holds if and only if $\mu_l/\sigma_l \leq 0 \leq \mu_u/\sigma_u$, under the alternative the normalized means μ_l/σ_l and μ_u/σ_u may not 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 ρ without loss of generality. In the sequel, we demonstrate our approach when $\sigma_l = \sigma_u = 1$ and various value of ρ.

It is useful to reparametrize (μ_l, μ_u) in terms of $(\beta, \delta, \tau) \in \mathbb{R}^3$ as follows: Let $\delta = \mu_u - \mu_l$ be the length of the identified set $[\mu_l, \mu_u]$, let $\beta = \min(|\mu_l|, |\mu_u|)$ if $\mu_l\mu_u > 0$, and $\beta = 0$

[6] The linear relationship between β and δ is numerically convenient, since it enables analytical computation of the alternative density h under a uniform distribution for δ.

Figure 4: Predictive Regression with a Local-To-Unity Regressor



Notes: In panels A.1 and B.1, solid lines are the rejection probability of the nearly optimal test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$, and dashed lines are for Campbell and Yogo's (2006) test.

otherwise, so that β is the distance between zero and the identified set $[\mu_l, \mu_u]$, and let $\tau = -\mu_l$, so that under the null hypothesis τ is the distance between the lower bound μ_l and zero. In this parametrization, the hypothesis testing problem becomes

$$H_0 : \beta = 0,\ \delta \geq 0,\ \tau \in [0, \delta] \quad \text{against} \quad H_1 : \beta > 0,\ \delta \geq 0. \tag{21}$$

As emphasized by Imbens and Manski (2004), as $\delta \to \infty$, the natural 5% level test is $\varphi_S(y) = \mathbf{1}[y_l > 1.645 \text{ or } y_u < -1.645]$. We switch to this standard test according to $\chi(y) = \mathbf{1}[\hat{\delta} > 6]$, where $\hat{\delta} = Y_u - Y_l \sim \mathcal{N}(\delta, 2(1-\rho))$. The weighting function F is chosen to be uniform on $\delta \in [0, 8]$, with equal mass on the two points $\beta \in \{-2, 2\}$.

Note that (21) has a two-dimensional nuisance parameter under the null hypothesis, as neither the length $\delta = \mu_u - \mu_l$ nor the distance τ of μ_l from zero is specified under H_0. It is reasonable to guess, though, that the least favorable distribution only has mass at $\tau \in \{0, \delta\}$, so that one of the endpoints of the interval coincides with the hypothesized value of μ. Further, the problem is symmetric in these these two values for τ. In the computation of the ALFD, we thus impose $\tau \in \{0, \delta\}$ with equal probability, and then check that the resulting test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ does indeed control size also for $\tau \in (0, \delta)$.

Figure 5 shows results for two values of ρ. Looking first at the critical regions, when y_u is sufficiently large (say $y_u > 2$), the test rejects when $y_l > 1.645$, and similarly when y_l is sufficiently negative. The upper left-hand quadrant of the figures in panels A.1 and B.1 show 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 AFLD test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ to the test $\varphi_{ST}(y) = \mathbf{1}[y_l > 1.96 \text{ or } y_u < -1.96]$, 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 $\delta = 0$ and $\tau \in \{0, \delta\}$. Not surprisingly $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ dominates φ_{ST} when δ is large, but it also has higher power when δ is small and $\rho = 0.5$ (because when δ is small, the mean of Y_l and Y_u is more informative about μ than either Y_l or Y_u unless ρ is close to 1).

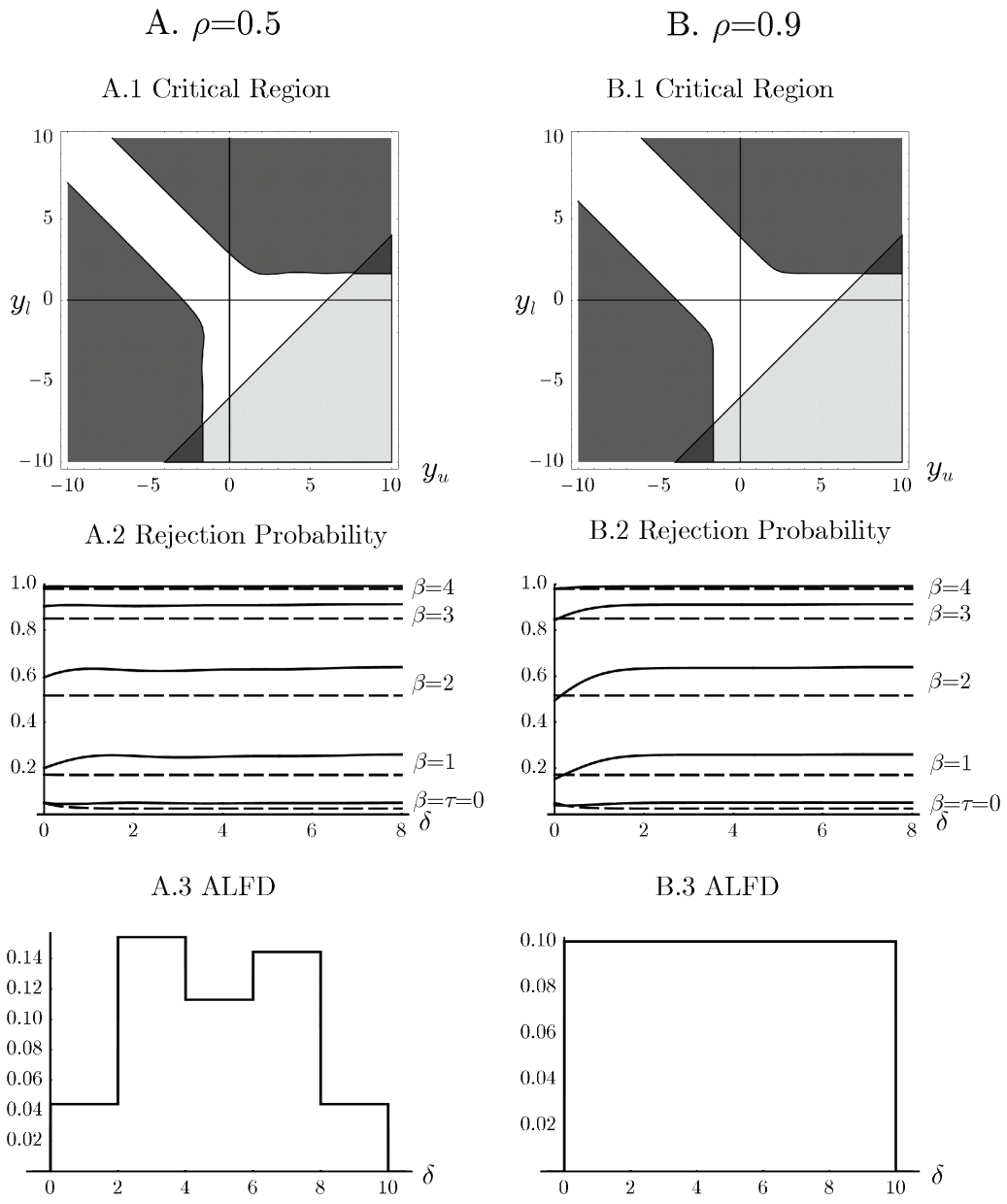
5.5 Regressor Selection

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

$$y_i = bx_i + dz_i + \varepsilon_i,\ \ i = 1, \cdots, n,\ \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{22}$$

where σ^2 is known. We are interested in testing $H_0 : b = b_0$, and d is a nuisance parameter. Suppose there is substantial uncertainty whether the additional control z_i needs to

Figure 5: Set-Identified Parameter



Notes: 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 $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$, and dashed lines are for Stoye's (2009) test $\varphi_{ST}(y) = 1[y_l > 1.96 \text{ or } y_u < -1.96]$.

be included in (22), that is $d = 0$ is deemed likely, but not certain. Let $(\hat{b}, \hat{d})$ denote the OLS estimators from the "long" regression of y_i on (x_i, z_i). Let $\beta = n^{1/2}(b - b_0)$, $\delta = n^{1/2}d$, $(Y_\beta, Y_\delta) = n^{1/2}(\hat{b} - b_0, \hat{d})$, and for notational simplicity, assume that the regressors and σ^2 have been scale normalized so that

$$Y = \begin{pmatrix} Y_\beta \\ Y_\delta \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \delta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{23}$$

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

For $\delta = 0$ known, the statistic $Y_\beta - \rho Y_\delta$ is more informative about β than is Y_β. Intuitively, $Y_\beta - \rho Y_\delta$ is the (rescaled) regression coefficient in the "short" regression of y_i on x_i, omitting z_i. 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_β. As reviewed by Leeb and Pötscher (2005), however, data-dependent model selection procedures do not perform uniformly well for all parameter values, even in large samples, so that no optimal inference is 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

$$H_0 : \beta = 0, \, \delta \in \mathbb{R} \quad \text{against} \quad H_{1,F} : \beta = \beta_1, \, \delta = 0. \tag{24}$$

Just as in footnote 4, note that rejecting for large values of Y_β 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 α test of (24) is also of level α under H_0^s, the uniformly most powerful one-sided test of (24) thus rejects for large values of Y_β. Thus, as long as one insists on uniform size control, the question of best one-sided inference about β 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 all 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 the 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

$$H_0 : \beta = 0, \, \delta \in \mathbb{R} \quad \text{against} \quad H_1 : \beta \neq 0, \, \delta = 0 \tag{25}$$

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.

Figure 6 shows the resulting critical region, power functions, and AFLDs for two values of ρ. Looking first at the critical regions, they are now discontinuous at the switching boundary. This discontinuity arises because $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ is constructed to maximize power at $\delta = 0$ and achieves this by sacrificing power for nonzero values of δ. On the other hand, the power of φ_S does not depend on δ. This is evident in panels A.2 and B.2. The $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ test dominates φ_S when $|\delta|$ is small, the roles are reversed for moderate values of $|\delta|$, and the power curves coincide for large $|\delta|$, where $\chi(Y) = 1$ with high probability.

By construction, the weighted average power at $\delta = 0$ of $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ 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 ρ, Figure 7 depicts the power bound, the power of $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ and the power of φ_S.[7] The experiment (23) becomes more informative about β as ρ increases, and correspondingly, the power bound is an increasing function of ρ.[8] 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 φ_S are never larger than 12 percentage points, and the test $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$ described in Appendix C comes very close to fully exploiting the available information.

6 Additional Remarks

From a decision theoretic perspective, the approximate least favorable distribution is related to the minimax value in the problem of distinguishing between H_0 against $H_{1,F}$: Suppose a false rejection of H_0 induces loss 1, a false rejection of $H_{1,F}$ induces loss $L_F > 0$, and a correct decision has loss 0. Then risk for a given θ and decision rule φ is given by

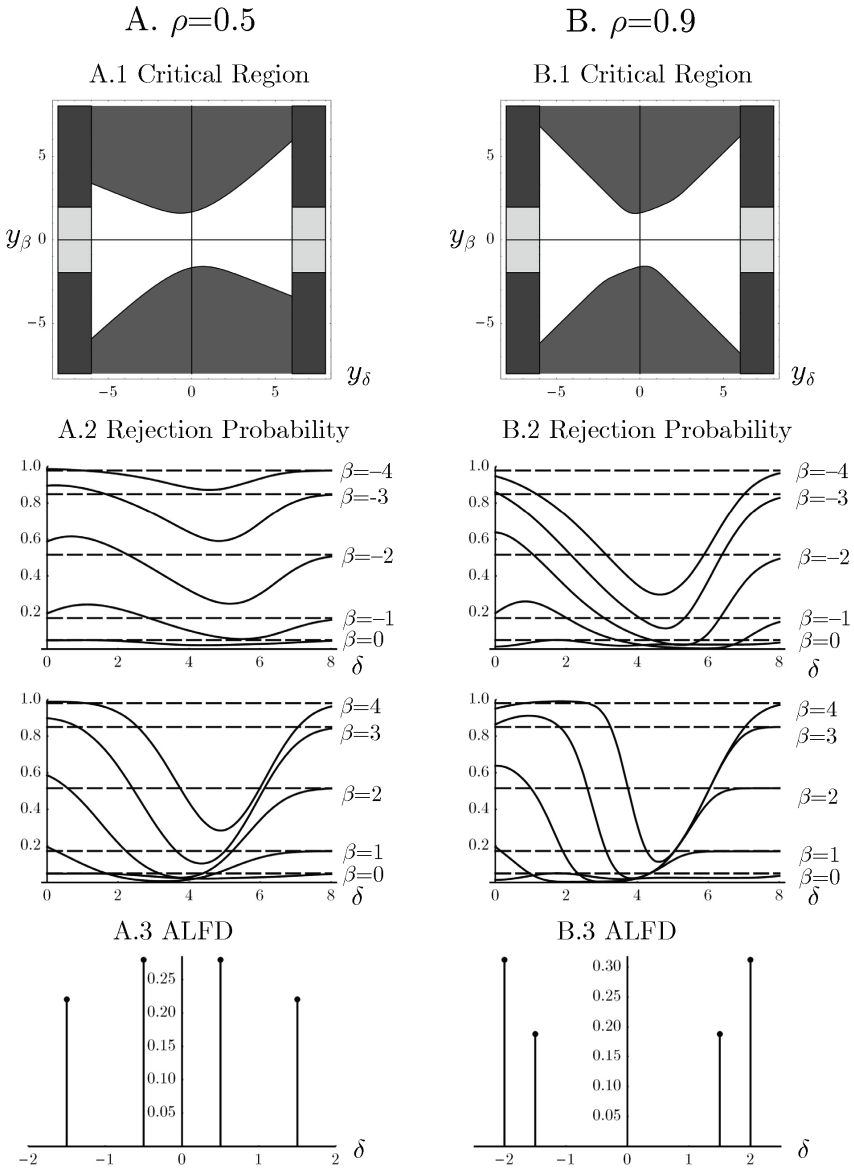
$$R(\theta, \varphi) = \mathbf{1}[\theta \in \Theta_0] \int \varphi f_\theta d\nu + L_F \mathbf{1}[\theta \in \Theta_1](1 - \int \varphi h d\nu). \tag{26}$$

As above, let Λ^{**} denote the least favorable distribution, $\varphi_{\Lambda^{**}}$ denote the test based on Λ^{**}, and let Λ^* and $\varphi^{\varepsilon}_{\Lambda^*}$ denote the ALFD and test as defined in Definition 1. Let α and

[7] The power bound is constructed from the ALFDs that underlie the family of tests described in Appendix C.

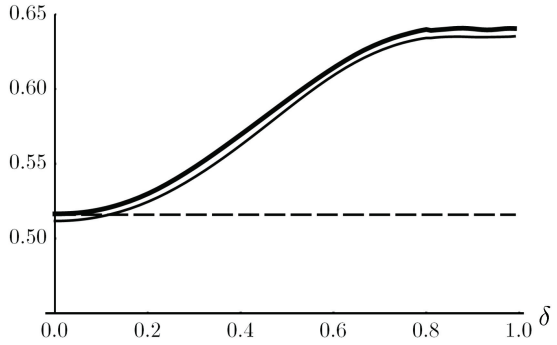
[8] Adding mean-zero Gaussian noise to Y_δ and an appropriate rescaling yields an equivalent experiment with smaller $|\rho|$.

Figure 6: Regressor Selection



Notes: 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 $\varphi^{\varepsilon}_{\Lambda^*,S,\chi}$, and dashed lines are for the usual test that ignores Y_δ, $\varphi_S(y) = 1[|y_\beta| > 1.96]$.

Figure 7: Weighted Average Power in Regressor Selection Problem as Function of ρ



Notes: 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_δ, $\varphi_S(y) = 1[|y_\beta| > 1.96]$.

π^{**} denote the size and power of $\varphi_{\Lambda^{**}}$. For arbitrary φ, let $\alpha_\varphi = \sup_{\theta\in\Theta_0} \int \varphi f_\theta d\nu$ and $\pi_\varphi = \int \varphi h d\nu$ denote size and power. Suppose that $L_F = \alpha/(1-\pi^{**})$, so that the relative loss associated with a false rejection of H_1 is equal to the ratio of rejection probabilities of the test associated with Λ^{**}. Then, for this value of L_F, $\varphi_{\Lambda^{**}}$ is the minimax decision rule and $\varphi^\varepsilon_{\Lambda^*}$ has maximum risk that is only slightly larger than the minimax value. To see why $\varphi_{\Lambda^{**}}$ is the minimax decision rule, notice that $\sup_{\theta\in\Theta} R(\theta, \varphi_{\Lambda^{**}}) = \alpha$ (by a direct calculation) and $\sup_{\theta\in\Theta} R(\theta,\varphi) \geq \alpha$ (noting that $\pi_\varphi \leq \pi^{**}$ if $\alpha_\varphi \leq \alpha$ and $\sup_{\theta\in\Theta} R(\theta,\varphi) \geq \alpha_\varphi$ if $\alpha_\varphi \geq \alpha$). Because $\varphi^\varepsilon_{\Lambda^*}$ has power within ε of π^{**} and has size α, $\sup_{\theta\in\Theta} R(\theta,\varphi^\varepsilon_{\Lambda^*}) - \sup_{\theta\in\Theta} R(\theta,\varphi_{\Lambda^{**}}) \leq \alpha\varepsilon/(1-\pi^{**})$, so that $\varphi^\varepsilon_{\Lambda^*}$ approximately achieves the minimax risk.

Minimax rules are inherently pessimistic, and they might be considered unattractive if they are rationalized by an unreasonable distribution for δ. In the context of the algorithm suggested here, this judgement can be made by inspecting the ALFD. Note that a Bayesian would decide between H_0 and H_1 by computing posterior odds, which is proportional to the nearly optimal likelihood ratio statistic for a prior of F on Θ_1, and a prior of Λ^* on Θ_0. In this context, the algorithm suggested here might be used as a prior selection device for the prior under H_0, which guarantees attractive frequentist properties of the resulting Bayes rule.

A Details on the Algorithm of Section 3

Note that the test φ_N^* with critical value cv can equivalently be written as $\varphi_N^*(y) = \mathbf{1}[\sum_{i\in J_N} p_i^* f_i(y)/h(y) < 1/\,\mathrm{cv}]$. In case of switching, replace $1/h$ by $1/\tilde{h}$, and set $1/\tilde{h}(y) = 0$ for $\tilde{h}(y) = \infty$ and $1/\tilde{h}(y) = \infty$ for $\tilde{h}(y) = 0$. We numerically approximate rejection probabilities of this test under "Y has density f" by

$$\hat{\pi}(P_N^*, \mathrm{cv}; f) = \frac{1}{m}\sum_{j=1}^{m}\left(1 + \left(\mathrm{cv}\sum_{i\in J_N} p_i^* \frac{f_i(Y_j^f)}{h(Y_j^f)}\right)^{10}\right)^{-1} \tag{27}$$

where Y_j^f, $j = 1, \cdots, m$, are i.i.d. pseudo random draws with probability density f relative to ν. For all f, the Y_j^f are suitable transformations of one set of $m = 10^5$ i.i.d. pseudo-random variables. In contrast to the standard Monte Carlo estimator based on averaging φ_N^* directly, the numerically close analogue (27) is a differentiable function of (P_N^*, cv), which facilitates numerical computations.

Step 2 of the algorithm is performed by numerically minimizing the objective function

$$\sum_{l\in J_N}(100p_l^* + \exp[8000(\hat{\pi}(P_N^*, \mathrm{cv}; f_l) - \alpha)])(\hat{\pi}(P_N^*, \mathrm{cv}; f_l) - \alpha)^2. \tag{28}$$

As a function of (P_N^*, cv), (28) is continuous with known first derivative, so that a standard quasi-Newton optimizer can be employed. Also, the $N^2 m$ numbers $f_i(Y_j^{f_l})/h(Y_j^{f_l})$ for $i = 1, \cdots, N$, $l = 1, \cdots, N$ and $j = 1, \cdots, m$ can be computed and stored once to speed up the the evaluation of (28) and its partial derivatives. After a satisfactory solution to (28) has been found, indices $i \in J_N$ where $p_i^* < 10^{-4}$ are dropped from J_N, and N reduced accordingly.

The order of the computations in Step 5 are randomized to avoid cycling behavior of the algorithm.

In Steps 5 and 6, we consider estimated null rejection probabilities of $\hat{\pi} \leq 0.0515$ as within Monte Carlo error of controlling size (with $m = 10^5$, the Monte Carlo standard error of $\hat{\pi}$ in (27) is approximately 0.0007).

B Details on Computations for the Applications

The following Lemma is useful for obtaining closed form expressions in many of the applications.

Lemma 3 *For $c > 0$, $\int_{-\infty}^{a} \exp[sd - \frac{1}{2}s^2c^2]ds = \sqrt{2\pi}c^{-1}\exp[\frac{1}{2}d^2/c^2]\Phi(ac - d/c)$, where Φ is the cdf of a standard normal.*

Proof. Follows from "completing the square". ■

In all applications, we consider first a coarse set of $M = M_c$ base distributions, and then perform Step 6 by enlarging the initial set by a set of M_f fine base distributions.

Positive Nuisance Parameter:

The coarse set contains uniform distributions on $\{[0,8],[0,0.1],[0,1],[1,2],[2,4],[4,6],[6,8],[8,10],[10,12],[12,14]\}$, and the fine set contains uniform distributions on $[(j-1)/2, j/2],\ j = 1,\cdots,25$.

Behrens Fisher:

An alternative maximal invariant is given by $(t,r) = ((\bar{x}_1 - \bar{x}_2)/s_2, s_1/s_2) = (\sqrt{\frac{e^{2Y_\delta}}{n_1} + \frac{1}{n_2}}Y_\beta, e^{Y_\delta})$. In the parameterization $\eta = \mu_1 - \mu_2$ and $\omega = \sigma_1/\sigma_2$, straightforward but tedious calculations yield for the density of (t,r)

$$\frac{(n_1-1)^{n_1/2}(n_2-1)^{n_2/2}\omega}{r^2\Gamma(\frac{1+n_1}{2})\Gamma(\frac{1+n_2}{2})}\sqrt{\frac{n_1n_2}{\pi(n_1+\omega^2n_2)}}\left(\frac{r}{\omega}\right)^{n_1}2^{(1-n_1-n_2)/2}\exp[-\tfrac{1}{2}\frac{\eta^2n_1n_2}{n_1+n_2\omega^2}]$$
$$\times\int_0^\infty s^{n_1+n_2-2}\exp[\frac{2\eta n_1n_2st - s^2((n_2-1)n_2\omega^4+n_1^2r^2-n_2\omega^2+n_1(n_2\omega^2(1+r^2+t^2)-\omega^2-r^2))/\omega^2}{2(n_1+n_2\omega^2)}]ds$$

where Γ denotes the Gamma function. The integral is recognized as being proportional to the (n_1+n_2-2)th absolute moment of a half normal. In particular, for $c>0$, $\int_0^\infty \exp[-\frac{1}{2}s^2c^2]s^n ds = 2^{\frac{n-1}{2}}\Gamma(\frac{1+n}{2})c^{-(n+1)}$, and

$$\int_0^\infty \exp[sd-\tfrac{1}{2}s^2c^2]s^n ds = \exp[\tfrac{1}{2}\frac{d^2}{c^2}]\sqrt{2\pi}\Phi(d/c^2)\frac{d^n}{c^{2n+1}}\sum_{l=0}^n\binom{n}{l}(-\frac{c}{d})^l I_l(d/c^2)$$

where $I_l(h) = \frac{1}{\Phi(h)}\int_{-\infty}^h \phi(z)z^l dz$, and ϕ and Φ are the pdf and cdf of a standard normal. The iterative relations $I_0(h)=1$, $I_1(h) = -\phi(h)/\Phi(h)$ and $I_l(h) = -h^{l-1}\phi(h)/\Phi(h) + (l-1)I_{l-2}(h)$ allow the fast numerical evaluation of $I_l(h)$, as suggested by Dhrymes (2005).

The base distributions are uniform distributions for δ. The corresponding integrals are computed via Gaussian quadrature using 10 nodes (for this purpose the integral under the alternative is split up in 8 intervals of length 2). The coarse set contains uniform distributions on $[2(j-1)-10, 2j-10]$ for $j = 1,\cdots,10$, and the fine set contains uniform distributions on $[(j-1)/2-15.5, j/2-15.5]$ for $j = 1,\cdots,61$. When $n_1 = n_2$, symmetry around zero is imposed on the ALFD.

Break Date:

Wiener processes are approximated with 1,000 steps. The coarse set of baseline distribution contains uniform distributions on $\{[0,8],[0,0.1],[0,1],[1,2],[2,4],[4,6],[6,8],[8,10],[10,12],[12,14]\}$, and the fine set contains uniform distributions on $[(j-1)/2, j/2],\ j = 1,\cdots,25$.

Predictive Regression:

Ornstein-Uhlenbeck and stochastic integrals are approximated with 1,000 steps. The coarse set of baseline distributions contains uniform distributions on $\{[-5,-3],[-3,-1],[-1,1],[0,5],[5,10],[10,15],[15,10],[20,30],[30,40],[40,50],[50,60]\}$, and the fine set contains uniform distributions on all intervals of the form $[-6+j,-5+j]$, $j = 1,\cdots,8$ and $[2.5(j-1), 2.5j]$, $j = 1,\cdots,25$.

Partially Identified Parameter:

Symmetry around zero is imposed on the ALFD. The coarse set of baseline distribution are uniform distributions on $\{[0,10],[0,2],[2,4],[4,6],[6,8],[8,10],[10,12],[12,14]\}$, and the fine set contains uniform distributions on $[(j-1)/2, j/2]$, $j = 1,\cdots,26$.

Table 1: Polynomial Coefficients for the Positive Nuisance Parameter Problem

	$0 \leq \rho \leq 0.8$					$0.8 \leq \rho \leq 0.99$				
i	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$
0	NA	NA	1.274	0.026	−0.037	NA	NA	−1.056	5.650	−3.430
1	0	8	3.002	−3.396	−2.882	0	8	9.531	−3.374	−3.382
2	0	0.1	−1.362	2.411	−0.698	0	0.1	−5.250	24.378	−16.931
3	6	8	−0.006	−0.144	1.233	6	8	6.656	−6.800	3.859
4	2	4	−0.256	0.901	0.304	2	4	9.550	−13.216	7.400
5	4	6	−0.576	1.241	0.254	4	6	7.677	−9.778	5.822
6	1	2	−0.802	−1.013	1.788	1	2	−1.052	4.539	−1.135
7	1	2	−3.095	4.627	−1.752	0.5	1	−6.574	4.250	4.367

Regressor Selection:

Symmetry around zero is imposed on the ALFD. The coarse set of baseline distribution are point masses on $j/2$, $j = 1, \cdots, 6$, and the fine set contains point masses on $j/10$, $j = 1, \cdots, 80$.

C Families of Nearly Optimal Tests

This appendix describes nearly optimal tests for the problems indexed by ρ considered in this paper, and also for the break date problem (indexed by β_0). Call the indexing parameter κ. The nearly optimal tests are of the form $\varphi(y) = \mathbf{1}[h(y) \geq \text{cv}(\kappa) \sum_{i=1}^{N} p_i(\kappa) f_i(y)]$, that is the critical value cv and the p_i's are functions of κ, but not the set of base densities f_i, $i = 1, \cdots, N$ that enter the ALFD.

Specifically, $\text{cv}(\kappa) = \exp[a_{0,0} + a_{1,0}\kappa + a_{2,0}\kappa^2]$, and $p_i(\kappa) = \tilde{p}_i(\kappa) / \sum_{l=1}^{N} \tilde{p}_l(\kappa)$, where $\tilde{p}_i(\kappa) = \exp[a_{0,i} + a_{1,i}\kappa + a_{2,i}\kappa^2]$. The ALFD is thus described by $3(N+1)$ polynomial coefficients $a_{j,i}$, $j = 0, 1, 2$, $i = 0, \cdots, N$.

In the regressor selection problem, the base distributions with density f_i have δ uniformly distributed on the two points $\{-\delta_i, \delta_i\}$. In the break date problem, f_i has δ uniformly distributed on $[\underline{\delta}_i, \overline{\delta}_i] \cup [-\overline{\delta}_i, -\underline{\delta}_i]$. In all other problems, f_i has δ uniformly distributed on $[\underline{\delta}_i, \overline{\delta}_i]$. In all cases, f_i can be computed in closed form via Lemma 3. Similarly, also h can be computed in closed form for all problems.

Tables 1-5 contain the coefficients $a_{j,i}$ for the various examples. Negative values of ρ in the running example and the regressor selection problem are not reported, as the transformation $Y_\beta \to -Y_\beta$ yield the equivalent problem with correlation $-\rho$, respectively. Also, the problem of testing the null hypothesis of a break fraction $\beta_0 > 1/2$ is transformed into the equivalent problem with break fraction $1 - \beta_0$ by reversing the time series, i.e. by the transformation $G(s) \to G(1-s)$. In the predictive regression problem, tests against the alternative $H_1 : \beta < 0$, $\delta \geq \underline{\delta}$ are obtained by transforming $\{x_t\} \to \{-x_t\}$ and $\rho \to -\rho$.

The coefficients were computed by applying the algorithm of Section 3 to a range of values for κ simultaneously, that is in Step 3, the objective function is a sum of (28) over various values of κ, and this function is numerically minimized with respect to all $a_{j,i}$.

Table 2: Polynomial Coefficients for the Break Date Problem

	$0.15 \leq \beta_0 \leq 0.3$					$0.3 \leq \beta_0 \leq 0.5$				
i	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$
0	NA	NA	0.471	0.285	0.817	NA	NA	0.378	1.117	−0.977
1	0	10	6.795	−0.926	7.253	0	10	4.541	0.448	−0.209
2	12	16	3.420	8.245	−13.577	12	16	1.916	0.041	0.180
3	0	4	8.610	−11.065	13.858	0	4	4.353	−1.439	−0.422
4	19	20	2.899	3.746	−7.533	3	4	3.329	1.191	0.412
5						12	13	1.635	−0.215	−0.085
6						19	20	1.039	−0.026	0.123

Table 3: Polynomial Coefficients for the Predictive Regression Problem

			$0 \leq \rho \leq 0.8$					$0.8 \leq \rho \leq 0.99$		
i	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$
0	NA	NA	1.413	0.087	0.078	NA	NA	1.211	0.842	−0.524
1	0	40	2.064	−0.745	−1.837	30	40	8.152	−3.016	3.176
2	30	40	0.541	2.031	−1.154	10	15	7.264	−3.221	3.425
3	10	15	−0.377	−0.525	1.970	15	20	7.124	−2.681	3.087
4	20	30	0.579	−1.027	2.305	40	50	5.339	1.750	1.197
5	40	50	−2.608	3.928	0.887	5	10	2.261	7.993	−2.995
6	1	3	−0.817	−4.889	−1.912	20	30	7.601	−2.392	3.073
7	5	10	0.094	−0.532	1.063	2.5	5	7.473	3.517	−7.360
8	15	20	0.164	0.498	−0.001	1	2	4.497	−1.950	−3.604
9	2.5	5	−1.559	1.261	−1.321					
			$0 \geq \rho \geq -0.8$					$-0.8 \geq \rho \geq -0.95$		
0	NA	NA	1.373	0.316	−0.098	NA	NA	4.961	9.590	5.801
1	0	40	11.15	4.252	−4.137	0	40	16.756	−2.845	5.646
2	5	10	9.454	3.008	1.917	−1	1	27.790	−0.073	−5.964
3	−1	1	4.733	−5.307	−0.867	−3	−1	16.863	−4.919	1.976
4	1	3	8.062	−1.426	−2.793	10	15	19.773	−1.138	2.184
5	−3	−1	−1.150	−2.353	2.836	0	5	22.098	0.650	1.597
6	10	15	5.093	−2.615	1.618	15	20	28.876	1.435	−5.106
7	15	20	8.673	−1.472	−2.749	5	10	29.923	6.889	−0.332
8	20	30	10.178	5.914	4.175	−5	−4	−1.620	0	0
			$-0.95 \geq \rho \geq -0.995$							
0	NA	NA	0.954	−0.254	−0.163					
1	−5	−3	−0.583	−0.047	0.092					
2	−1	1	4.223	−0.385	0.718					
3	−3	−1	14.229	1.833	−4.878					
4	10	15	10.736	1.136	−1.687					
5	5	10	10.450	0.956	−1.488					
6	15	20	9.787	0.821	−1.126					
7	0	5	10.474	0.798	−1.570					
8	20	30	8.761	0.177	−0.148					
9	−1	0	5.118	−0.489	0.813					
10	−4	−3	−7.494	−4.800	9.274					

Table 4: Polynomial Coefficients for the Partially Identified Parameter Problem

	$0 \leq \rho \leq 0.8$					$0.8 \leq \rho \leq 0.99$				
i	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	$\underline{\delta}_i$	$\overline{\delta}_i$	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$
0	NA	NA	1.125	0.290	0.133	NA	NA	0.251	1.948	−0.657
1	0	10	−0.955	3.516	−0.262	0	10	−1.558	1.748	3.442
2	2	4	1.176	−1.561	−0.445	4	6	0.682	−0.389	−0.663
3	4	6	1.163	−3.054	1.669	2	4	0.473	−0.220	−0.605
4	6	8	1.432	−2.758	0.909	6	6.5	0.102	−0.535	−0.854
5	1.5	2	−2.814	3.857	−1.871	1.5	2	0.518	−0.603	−1.320

Table 5: Polynomial Coefficients for the Regressor Selection Problem

	$0 \leq \rho \leq 0.8$				$0.8 \leq \rho \leq 0.99$			
i	δ_i	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	δ_i	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$
0	NA	1.260	0.019	0.055	NA	0.316	2.183	−1.174
1	2	−10.215	6.175	8.264	2	2.341	−3.038	10.425
2	0.5	6.927	−6.617	−6.360	1.6	13.802	3.038	−10.425
3	1.5	1.959	0.443	−1.904				

References

Andrews, D. W. K. (2011): "Similar-on-the-Boundary Tests for Moment Inequalities Exist, But Have Poor Power," *Cowles Foundation Discussion Paper No. 1815*.

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.

Andrews, D. W. K., and W. Ploberger (1994): "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

Bai, J. (1994): "Least Squares Estimation of a Shift in Linear Processes," *Journal of Time Series Analysis*, 15, 453–470.

——— (1997): "Estimation of a Change Point in Multiple Regressions," *Review of Economics and Statistics*, 79, 551–563.

Bobkoski, M. J. (1983): "Hypothesis Testing in Nonstationary Time Series," *unpublished Ph.D. thesis, Department of Statistics, University of Wisconsin*.

Campbell, J. Y., and M. Yogo (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27–60.

Cavanagh, C. L. (1985): "Roots Local To Unity," *Working Paper, Harvard University*.

Cavanagh, C. L., G. Elliott, and J. H. Stock (1995): "Inference in Models with Nearly Integrated Regressors," *Econometric Theory*, 11, 1131–1147.

Chamberlain, G. (2000): "Econometric Applications of Maximin Expected Utility," *Journal of Applied Econometrics*, 15, 625–644.

Chan, N. H., and C. Z. Wei (1987): "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics*, 15, 1050–1063.

Chiburis, R. C. (2009): "Approximately Most Powerful Tests for Moment Inequalities," *Working Paper, Princeton University*.

Choi, A., W. J. Hall, and A. Schick (1996): "Asymptotically Uniformly Most Powerful Tests in Parametric and Semiparametric Models," *Annals of Statistics*, 24, 841–861.

Dhrymes, P. J. (2005): "Moments of Truncated (Normal) Distributions," *Working Paper, Columbia University*.

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.

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.

——— (2009): "Pre and Post Break Parameter Inference," *Working Paper, Princeton University*.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996): "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836.

Ferguson, T. S. (1967): *Mathematical Statistics — A Decision Theoretic Approach*. Academic Press, New York and London.

Hahn, J., and G. Ridder (2011): "A Dual Approach to Confidence Intervals for Partially Identified Parameters," *Working Paper, UCLA*.

Imbens, G., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Jansson, M., and M. J. Moreira (2006): "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica*, 74, 681–714.

Kempthorne, P. J. (1987): "Numerical Specification of Discrete Least Favorable Prior Distributions," *SIAM Journal on Scientific and Statistical Computing*, 8, 171–184.

Kim, S., and A. S. Cohen (1998): "On the Behrens-Fisher Problem: A Review," *Journal of Educational and Behavioral Statistics*, 23, 356–377.

King, M. L. (1988): "Towards a Theory of Point Optimal Testing," *Econometric Reviews*, 6, 169–218.

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.

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

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypotheses*. Springer, New York.

Linnik, Y. V. (1966): "Randomized Homogeneous Tests for the Behrens-Fisher Problem," *Selected Translations in Mathematical Statistics and Probability*, 6, 207–217.

——— (1968): *Statistical Problems with Nuisance Parameters*. American Mathematical Society, New York.

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.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Müller, U. K. (2011): "Efficient Tests under a Weak Convergence Assumption," *Econometrica*, 79, 395–435.

Müller, U. K., and M. W. Watson (2009): "Low-Frequency Robust Cointegration Testing," *Working paper, Princeton University*.

Phillips, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.

Salaevskii, O. V. (1963): "On the Non-Existence of Regularly Varying Tests for the Behrens-Fisher Problem," *Soviet Mathematics, Doklady*, 4, 1043–1045.

Srianthakumar, S., and M. L. King (2006): "A New Approximate Point Optimal Test of a Composite Null Hypothesis," *Journal of Econometrics*, 130, 101–122.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H., and M. W. Watson (1996): "Confidence Sets in Regressions with Highly Serially Correlated Regressors," *Working Paper, Harvard University*.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

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

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.

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.

Welch, B. L. (1947): "The Generalization of "Student's" Problem When Several Different Population Variances are Involved," *Biometrika*, 34, 28–35.

Woutersen, T. (2006): "A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters," *Unpublished Mansucript, Johns Hopkins University*.