Testing near or at the Boundary of the Parameter Space[*] (Job Market Paper)

Philipp Ketz[†]
Brown University

November 17, 2014

Statistical inference about a scalar parameter is often performed using the two-sided t-test. In extremum problems, where the estimator satisfies the restrictions on the parameter space - such as the nonnegativity of a variance parameter -, the test suffers from size distortions when the true parameter vector is near or at the boundary of the parameter space. Nevertheless, the two-sided t-test continues to be used when estimates are found to be close to the boundary. This can be attributed to a lack of inference procedures that appropriately account for boundary effects on the asymptotic distribution of the estimator. To address this issue, we propose an estimator that is asymptotically normally distributed, even when the true parameter vector is near or at the boundary, and the objective function is not defined outside the parameter space. The novel estimator allows the implementation of several existing testing procedures and a new test based on the Conditional Likelihood Ratio statistic (CLR). Compared to the existing procedures, the new test is easy to implement and has good power properties. Moreover, it offers power advantages over the two-sided t-test, when the latter controls size. We also show the test to be admissible when inference is performed with respect to a scalar parameter. We apply the test to the random coefficients logit model using data on the European car market and find more evidence of heterogeneity in consumer preferences than suggested by the two-sided t-test.

Keywords: Boundary, asymptotic normality, testing, admissibility, random coefficients.

[*] I am very grateful to my advisor Frank Kleibergen as well as Adam McCloskey, Blaise Melly, and Eric Renault for their guidance, their constant support and encouragement, and many helpful comments and discussions. I also thank Andrew Elzinga, Joachim Freyberger, Bruno Gasperini, Hyojin Han, Pepe Montiel Olea, Daniela Scida, and seminar participants at Brown University for helpful comments and suggestions.

[†] Department of Economics, Brown University, 64 Waterman Street, Providence, RI 02912. Email: philipp_ketz@brown.edu.

1 Introduction

Statistical inference about a scalar parameter is often performed using the two-sided t-test, which relies on the asymptotic normality of the underlying estimator. In extremum problems, however, the commonly employed estimator is not asymptotically normally distributed when the true parameter vector is near or at the boundary of the parameter space (Andrews, 1999). In that case, the two-sided t-test suffers from size distortions. While the test can suffer from overrejection (under some conditions), underrejection constitutes the more prevalent problem. Due to the associated loss in power, a researcher is, for example, more likely to falsely conclude that a parameter is equal to zero, which has significant consequences in the context of model selection exercises. The random coefficients logit model (Berry, Levinsohn, and Pakes, 1995), where variance parameters are restricted to be nonnegative, constitutes a prominent example of models in which estimates are frequently found to be close to the boundary, indicating that the assumption that the true parameter vector lies in the interior is violated.[1] Nevertheless, the two-sided t-test continues to be used in practice (e.g., Nevo, 2001; Goeree, 2008). This can be attributed to a lack of inference procedures that appropriately account for boundary effects on the asymptotic distribution of the estimator, which result from the restrictions on the parameter space to which the estimator is confined.[2]

In this paper, we address this gap in the literature by introducing a modified extremum estimator that is asymptotically normally distributed even when the true parameter vector is near or at the boundary of the parameter space. The novel estimator does not require the original extremum objective function to be defined outside the parameter space and is, therefore, available in a wide range of nonlinear models. The estimator is given by the unconstrained minimizer of a quadratic approximation to the original objective function. It is easy to implement and obtained by a single step of the Newton-Raphson algorithm starting at the constrained extremum estimator.

Although the two-sided t-test based on the novel estimator constitutes a valid and size-correct test, the nonstandard nature of the testing problem can be utilized to construct tests that are more powerful. The testing problem is nonstandard in that it is characterized by the presence of nuisance parameters that satisfy certain restrictions. We introduce two tests that take these restrictions into account. The tests are based on the generalized Likelihood Ratio statistic (gLR) and the Conditional Likelihood Ratio statistic (CLR), which previously have

[1] Other models with similar restrictions on the parameter space are given by random coefficients regression models (Andrews, 1999), censored panel data models with slope heterogeneity (Abrevaya and Shen, 2014), multinomial discrete response models with random coefficients (Hausman and Wise, 1978) or random effects (McFadden, 1989).

[2] Andrews and Guggenberger (2010) show that bootstrap procedures are also inconsistent.

not been considered for the testing problem at hand. The tests are easy to implement and, in many cases, lead to tighter confidence intervals than the "benchmark", i.e., the two-sided t-test based on the constrained extremum estimator.

The null distribution of the gLR depends on the true value of the nuisance parameter. In order to construct a valid test based on the gLR we consider the least favorable configuration approach. When the dimension of the nuisance parameter is small, the resulting test displays good power properties. In particular, it offers power advantages over the "benchmark" for a wide range of alternatives. However, the conservativeness of the test, resulting from the use of the least favorable configuration, leads to low power in large parts of the parameter space when the number of nuisance parameters is large.[3]

Conditional on a sufficient statistic for the nuisance parameter, the (conditional) null distribution of the CLR is independent of the true value of that parameter. As a result, the test based on the CLR displays good power properties regardless of the dimension of the nuisance parameter. Furthermore, we show the test to be admissible when inference is performed with respect a scalar parameter. Another appealing feature is that the test reduces to the two-sided t-test when the true parameter vector lies in the interior of the parameter space. Consequently, confidence intervals obtained by inverting the test can be interpreted as a natural extension of standard confidence intervals to the nonstandard setting where the true parameter vector might be near or at the boundary.

Based on our novel estimator, other tests recently proposed in the literature become available. Elliott, Müller, and Watson (2013) propose tests that maximize weighted average power (WAP), while Montiel Olea (2013) proposes tests that maximize WAP subject to a similarity constraint. WAP maximizing tests are attractive when the researcher has a particular weight function in mind, where the weight function specifies the alternatives towards which the test directs power. The tests proposed in this paper are "ad hoc" in that they are not designed to maximize WAP (or minimize a general loss function). However, from the results in Montiel Olea (2013) it follows that there exist weights with respect to which the test based on the CLR maximizes WAP subject to a similarity constraint. For the testing problem with a scalar nuisance parameter, we conduct a power comparison of the tests based on the gLR, the CLR, and the two tests proposed by Elliott, Müller, and Watson (2013) and Montiel Olea (2013) for certain choices of the weight function.[4] All tests are found to have comparable power properties, displaying advantages over some parts of the parameters

[3] Alternative methods that do not rely on the use of the least favorable configuration can also be implemented, see e.g., McCloskey (2012).

[4] The weight functions are taken from the two papers. Elliott, Müller, and Watson (2013) use "uniform" weights that assign equal weights to all alternatives. Montiel Olea (2013) uses weights that yield a simple closed form solution for his test statistic.

space, while lacking in power over others.

Regardless of whether the researcher has a particular weight function in mind, the choice of a test is also guided by computational feasibility. The tests proposed by Elliott, Müller, and Watson (2013) and Montiel Olea (2013) become computationally expensive as the number of nuisance parameters increases.[5] The computational costs associated with the tests introduced in this paper, on the other hand, are invariant to the number of nuisance parameters, making them more attractive in practice.

In order to illustrate their usefulness, we apply the proposed testing procedures to the random coefficients logit model (Berry, Levinsohn, and Pakes, 1995), which is widely used in the industrial organization and marketing literatures to model demand for differentiated products. The random coefficients in this model are typically parameterized by a vector of means and variances and allow for heterogeneity in consumer preferences with respect to different product characteristics. In many applications, it is a priori unknown which of the product characteristics interact with a random coefficient, i.e., which variance parameters are non-zero. As a result, the empirical analysis often starts with a baseline model that allows for a random coefficient on all product characteristics. Then, a powerful test such as the one based on the CLR can prove useful in determining a good model specification. In an application of the random coefficients model to the European car market using data from Reynaert and Verboven (2014), we find evidence of consumer heterogeneity with respect to price (divided by income), horse power (divided by weight), and height of the car when using the CLR.[6] The two-sided t-test, on the other hand, which represents common practice, only suggests the presence of consumer heterogeneity with respect to horse power.

The plan of this paper is as follows. Section 2 introduces the testing problem, the two test statistics, the gLR and the CLR, and the tests based on them. This section also shows that the test based on the CLR is admissible for testing hypotheses about a scalar parameter. Section 3 provides asymptotic theory for general extremum estimators and introduces a novel estimator which is shown to be asymptotically normal. This section also contains details on how to implement the proposed testing procedures when using an asymptotically normal estimator. Section 4 contains a power comparison of our testing procedures, along with other tests recently proposed in the literature, in the context of several leading examples. Section 5 contains an application to the random coefficients logit model. In this section, we perform a small Monte Carlo study to illustrate the finite sample behavior of our testing procedures

[5] For certain choices of the weight function, the test statistic proposed by Montiel Olea (2013) can be obtained in closed form. Then the computational cost is equivalent to that of the test based on the CLR and the gLR. However, for a general weight function, the test statistic needs to be evaluated numerically, leading to an increase in computational cost.

[6] The gLR is not the preferred choice in this setting due to the large number of nuisance parameters.

and present the empirical findings. Section 6 concludes.

Throughout this paper, (a, b) denotes the vector $(a', b')'$, where a and b are column vectors. $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ denote $(-\infty, \infty)$, $[0, \infty)$, and $(0, \infty)$, respectively. Furthermore, if A denotes an interval in $\mathbb{R}$, then A^N denotes $A \times \cdots \times A$ with $N \in \mathbb{N}$ copies.

2 Testing

In this section, we first describe the testing problem and motivate its relevance. Then, we introduce two testing procedures that have previously not been considered for the testing problem at hand. They are based on the generalized Likelihood Ratio statistic (gLR) and the Conditional Likelihood Ratio statistic (CLR). The section concludes by showing that the test based on the CLR is admissible for the testing problem at hand.

2.1 Testing problem

This paper studies the problem of testing hypotheses about a scalar parameter, β. We consider the case where β is an element of a $(K+1) \times 1$ dimensional unknown parameter vector, $\theta = (\beta, \delta)$, that enters a general extremum objective function, $Q_n(\theta)$, whose dependence on the data $\{W_i : i \leq n\}$ is suppressed for notational convenience. Section 3 contains detailed information on what kind of objective functions are permitted in our framework. The $K \times 1$ dimensional vector δ is not specified under the null hypothesis and, therefore, constitutes a nuisance parameter in the testing problem. The parameter space for θ is assumed to be restricted, and we are interested in testing hypotheses about β when the true parameter vector is near or at the boundary of the parameter space. This is modeled by means of drifting sequences of true parameters, θ_n. For the purpose of illustration, assume $\theta \in \mathbb{R}_+^K$. Then, the drifting sequences of interest would be such that $\sqrt{n}\theta_n \rightarrow \mu$, where μ denotes a localization parameter with $\|\mu\| < \infty$. Such drifting sequences are essential in order to derive asymptotic theory that provides good approximations to the finite sample behavior of an estimator or a test statistic, when the true parameter vector is close to the boundary relative to the sample size. Throughout this paper, we use near and close interchangeably. Furthermore, the case where the true parameter vector is at the boundary, $\mu_k = 0$ for some $k \in \{1, \ldots, K+1\}$, constitutes a special case of the true parameter vector being near the boundary and, hereafter, is sometimes referred to implicitly.

Motivated by the recent literature, we introduce the hypothesis testing problem in the limit experiment, or limiting problem (see e.g., Van der Vaart, 2000; Elliott, Müller, and Watson, 2013; Müller, 2011). In particular, we assume that we observe a draw from a random variable, Y, whose distribution is fully determined by a $(K+1) \times 1$ dimensional unknown parameter vector μ. This μ corresponds to the localization parameter of the original testing problem and is partitioned as $\mu = (b, d)$, where b is scalar and d is a $K \times 1$ dimensional vector. We are interested in testing hypotheses about b, while treating d as a nuisance parameter. Similar testing problems are also considered in Elliott, Müller, and Watson (2013) and Montiel Olea (2013). Our hypothesis testing problem of interest is

$$H_0 : b = b_0 \in B^\infty, d \in D^\infty \text{ vs. } H_1 : b \neq b_0, b \in B^\infty, d \in D^\infty, \tag{1}$$

where B^∞ equals $(-\infty, 0], [0, \infty)$, or $(-\infty, \infty)$, and D^∞ is a Cartesian product of intervals that equal $(-\infty, 0]$ or $[0, \infty)$. Let $M^\infty = B^\infty \times D^\infty$ denote the parameter space for μ. Although the shape of the parameter space is restrictive, it arises in many models of interest. Random coefficient models, where variance parameters are restricted to be nonnegative, constitute a leading example. Another example is given by (semi-) parametric regression models, where some coefficients are known to satisfy sign restrictions. The motivation for why D^∞ does not contain intervals of the form $(-\infty, \infty)$ is given by an invariance argument. In Section 3, we allow for an additional nuisance parameter in θ, whose true value is assumed to lie in the interior of the parameter space and which can be thought of as being "partialled out". The end points of the intervals in M^∞ are equal to zero, when they are not infinite, without loss of generality. When the original parameter space for (β, δ) is given by a Cartesian product of intervals, M^∞ can always be written as above by appropriately recentering (β, δ). Similarly, the focus on two-sided testing problems is without loss of generality.

In this paper, we introduce tests for the hypothesis testing problem given in (1) based on the Gaussian shift experiment, i.e.,

$$Y = \begin{pmatrix} Y_b \\ Y_d \end{pmatrix} \sim N\left(\begin{pmatrix} b \\ d \end{pmatrix}, \Sigma \right), \tag{2}$$

where Σ is known and positive definite. Alternatively, we could derive tests based on the limiting problem that is obtained under the sequence of (constrained) extremum estimators commonly employed in practice. The corresponding Y, say Y^c, is distributed as the projection of (2) onto a linear subspace of $\mathbb{R}^{K+1}$, see Section 3.1. However, tests based on (2) are not only easier to derive, but also subsume tests based on Y^c, i.e., any test based on Y^c can also be implemented given (2), while the reverse is not true. Intuitively, a draw from a normal random variable is more informative (about μ) than a draw from a "truncated" normal. Other tests recently proposed in the literature are also based on the Gaussian shift experiment (see e.g., Elliott, Müller, and Watson, 2013; Montiel Olea, 2013). A main contribution of this paper is to propose an asymptotically normal estimator, which is available under no additional assumptions beyond those made to derive the asymptotic distribution of a constrained extremum estimator, such that tests based on the Gaussian shift experiment become available in a wide range of nonlinear models.

In standard settings, the support of $\mu = (b, d)$ is $\mathbb{R}^{K+1}$, and there are good reasons to ignore Y_d when making inference about b. For example, for the two-sided testing problem, $H_0 : b = b_0$ vs. $H_1 : b \neq b_0$, the test that rejects when $|\frac{Y_b - b_0}{\sqrt{\Sigma_{\beta\beta}}}| > \text{cv}$, where cv denotes the critical value of a standard normal distribution, is the uniformly most powerful unbiased test. In the nonstandard setting considered in this paper, Y_d contains information about b that can be exploited when testing (1).

Before introducing our testing procedures, we illustrate how tests developed for (1) based on (2) can be used under drifting sequences of true parameters when an asymptotically normal estimator is available. To that end, we consider a simple example taken from Andrews and Guggenberger (2010).

Example 1: We consider the testing problem where a scalar nuisance parameter is near the boundary. Suppose $\{X_i \in \mathbb{R}^2 : i \leq n\}$ forms a triangular array of i.i.d. random vectors with

$$X_i = \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix}, \; \theta_n \equiv \mathrm{E}(X_i) = \begin{pmatrix} \beta_n \\ \delta_n \end{pmatrix}, \text{ and } \Sigma \equiv \mathrm{Var}(X_i),$$

where Σ is positive definite. θ_n is a drifting sequence of true parameters. The parameter space is given by $\Theta = B \times D$, where $B = (-\infty, \infty)$ and $D = [0, \infty)$, i.e., we know a priori that the mean of $X_{i,2}$ is nonnegative. We assume that the drifting sequence of true parameters satisfies $\sqrt{n}\beta_n \to b$, where $|b| < \infty$, and $\sqrt{n}\delta_n \to d < \infty$. We are interested in testing $H_0 : b = b_0$ while leaving d unspecified under H_0.[7] The sample averages, $\bar{X}_1$ and $\bar{X}_2$, are estimators of β_n and δ_n, respectively, and by a central limit theorem (CLT) for triangular arrays satisfy

$$\sqrt{n} \begin{pmatrix} \bar{X}_1 - \beta_n \\ \bar{X}_2 - \delta_n \end{pmatrix} \xrightarrow{d} N(0, \Sigma)$$

or, equivalently,

$$\sqrt{n} \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} b \\ d \end{pmatrix}, \Sigma \right).$$

Therefore, under the above drifting sequences the scaled estimator, $\sqrt{n}\bar{X}$, is asymptotically distributed as Y given in equation (2). Furthermore, the parameter spaces for b and d are

[7] Testing $H_0 : b = b_0$ can be understood as testing $H_0 : \beta_n = b_0/\sqrt{n}$ at a given sample size, n.

given by $B^\infty = (-\infty, \infty)$ and $D^\infty = [0, \infty)$, respectively.[8] As a result, any test derived for the testing problem given in (1) and (2) can be applied here by replacing Y by its sample analogue, $\sqrt{n}\bar{X}$, and by replacing Σ by a consistent estimator, such as the sample variance of X_i.

2.2 Testing procedures

In the problem without a nuisance parameter, i.e., without Y_d in (2), and $B = [0, \infty)$, Feldman and Cousins (1998) propose the use of the generalized Likelihood Ratio statistic (gLR) in order to make inference. To the best of our knowledge, there has been no attempt yet to use the gLR to make inference in the general testing problem given in (1) and (2).

Let $\phi(y, \mu, \Sigma)$ denote the pdf of a $(K+1) \times 1$ dimensional normal random vector with mean, μ, and positive definite variance, Σ. Then, the gLR is defined as follows

$$\begin{aligned} \mathrm{gLR}(b_0, y_b, y_d) &= 2\log\left(\frac{\sup_{\mathrm{b}\in B^\infty, \mathrm{d}\in D^\infty} \phi((y_b, y_d), (\mathrm{b}, \mathrm{d}), \Sigma)}{\sup_{\mathrm{d}\in D^\infty} \phi((y_b, y_d), (b_0, \mathrm{d}), \Sigma)}\right) \\ &= \inf_{\mathrm{d}\in D^\infty} \begin{pmatrix} y_b - b_0 \\ y_d - \mathrm{d} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - b_0 \\ y_d - \mathrm{d} \end{pmatrix} - \inf_{\mathrm{b}\in B^\infty, \mathrm{d}\in D^\infty} \begin{pmatrix} y_b - \mathrm{b} \\ y_d - \mathrm{d} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - \mathrm{b} \\ y_d - \mathrm{d} \end{pmatrix}. \end{aligned}$$

The distribution of $\mathrm{gLR}(b_0, Y_{b_0}, Y_d)$ depends on the true value of d, which is not specified under the null. One possibility to construct a level α test based on the gLR is by means of the "least favorable configuration". The least favorable configuration, say d_{LFC}, is such that

$$\inf_{d\in D^\infty} P(\mathrm{gLR}(b_0, Y_{b_0}, Y_d) < \mathrm{cv}_{d_{\mathrm{LFC}}}) = 1 - \alpha,$$

where $\mathrm{cv}_{d_{\mathrm{LFC}}}$ denotes the $(1-\alpha)$ quantile of the $\mathrm{gLR}(b_0, Y_{b_0}, Y_{d_{\mathrm{LFC}}})$ when the true parameter equals (b_0, d_{LFC}). The corresponding test rejects whenever $\mathrm{gLR}(b_0, y_b, y_d) > \mathrm{cv}_{d_{\mathrm{LFC}}}$.

Another possible testing procedure is based on the Conditional (generalized) Likelihood Ratio statistic (CLR), which was originally suggested by Moreira (2003) for the linear Instrumental Variables regression model with weak instruments. Variates thereof have also been used in the context of weak identification in the Generalized Method of Moments (GMM) (see e.g., Kleibergen, 2005). The test utilizes the gLR without relying on the least favorable configuration. We consider the following transformation of Y

$$\begin{pmatrix} Y_b \\ Y_d - \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1} Y_b \end{pmatrix} \sim N\left(\begin{pmatrix} b \\ d - \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1} b \end{pmatrix}, \begin{bmatrix} \Sigma_{\beta\beta} & 0 \\ 0 & \Sigma_{\delta\delta} - \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1}\Sigma_{\beta\delta} \end{bmatrix}\right),$$

[8] This relies on $|b| < \infty$ and $d < \infty$.

where

$$\begin{bmatrix} \Sigma_{\beta\beta} & \Sigma_{\beta\delta} \\ \Sigma_{\delta\beta} & \Sigma_{\delta\delta} \end{bmatrix}$$

denotes a conformable partition of Σ. Let $X \equiv Y_d - \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1}Y_b$. Note that the distribution of X depends on d, while that of Y_b does not. Since X and Y_b are independent, X is a sufficient statistic for d. In fact, X is a complete sufficient statistic for d, where completeness follows from Theorem 4.3.1 in Lehmann and Romano (2005). Now define

$$\begin{aligned} \mathrm{CLR}(b_0, y_b, x) = & \inf_{\mathrm{d}\in D^\infty} \begin{pmatrix} y_b - b_0 \\ x + \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1}y_b - \mathrm{d} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - b_0 \\ x + \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1}y_b - \mathrm{d} \end{pmatrix} \\ & - \inf_{\mathrm{b}\in B^\infty, \mathrm{d}\in D^\infty} \begin{pmatrix} y_b - \mathrm{b} \\ x + \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1}y_b - \mathrm{d} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - \mathrm{b} \\ x + \Sigma_{\delta\beta}\Sigma_{\beta\beta}^{-1}y_b - \mathrm{d} \end{pmatrix}. \end{aligned} \tag{3}$$

The CLR test, $\varphi_{\mathrm{CLR}}(y_b, x)$, is given by

$$\varphi_{\mathrm{CLR}}(y_b, x) = \begin{cases} 1 & \text{if } \mathrm{CLR}(b_0, y_b, x) > \mathrm{cv}(\alpha, x) \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $\mathrm{cv}(\alpha, x)$ denotes the conditional $(1-\alpha)$ quantile of the distribution of $\mathrm{CLR}(b_0, Y_b, x)$, where $Y_b \sim N(b_0, \Sigma_{\beta\beta})$. Unlike the (unconditional) distribution of the gLR statistic, the (conditional) distribution of the CLR statistic does not depend on the true value of d due to the conditioning on $X = x$, which is a sufficient statistic for d.

The critical values, $\mathrm{cv}_{d_{\mathrm{LFC}}}$ and $\mathrm{cv}(\alpha, x)$, although not available in closed form, can easily be obtained by means of simulation.

Another test statistic of interest is the two-sided t-statistic based on the constrained maximum likelihood estimator for b, t_{ML} hereafter, where the constraint is given by $\mu \in M^\infty$. The power function of the test that compares the t_{ML} to the standard critical value of a normal random variable matches, under certain conditions, the local asymptotic power function of the two-sided t-test based on a constrained extremum estimator when the true parameter vector is near or at the boundary. The latter corresponds to common practice where potential boundary effects on the distributions of the underlying estimator are ignored. Therefore, the performance of the test based on t_{ML} is of particular interest and analyzed, in detail, below.

In what follows, gLR, CLR, and t_{ML} can refer to the statistic or the respective test based on that statistic with the understanding that the gLR uses the least favorable configuration approach, while the t_{ML} uses the critical value of a standard normal random variable. Before turning to the asymptotic theory, we show that the CLR is admissible in the above testing problem.

2.3 Admissibility of the CLR

In order to show that the CLR is admissible in the class of all tests for the testing problem at hand, we first introduce some additional notation. Define M_0^∞ and M_1^∞ such that the testing problem given in equation (1) can be written as

$$H_0 : \mu \in M_0^\infty \text{ vs. } H_1 : \mu \in M_1^\infty,$$

where by definition $M_1^\infty = M^\infty \setminus M_0^\infty$. Let C denote the class of all tests. A test is defined as a measurable function $\varphi : \mathcal{Y} \to [0,1]$.[9] Here, $\mathcal{Y} = \mathbb{R}^{K+1}$ denotes the support of Y. Similarly, let $\mathcal{X} = \mathbb{R}^K$ denote the support of X. $\varphi(y)$ is to be understood as the probability of rejecting the null hypothesis given a realization of the data, y. The type I error of φ for $\mu \in M_0^\infty$ is defined as

$$E_\mu[\varphi(Y)] = \int_{\mathcal{Y}} \varphi(y) f(y|\mu) dy,$$

where $f(y|\mu)$ denotes the pdf of Y, which in the model on hand is the pdf of a multivariate normal. Since Σ is known, we suppress the dependence of $f(y|\mu)$ on Σ. The type I error is only defined for parameter values in the null set, M_0^∞, and signifies the probability with which the null hypothesis is falsely rejected. The type II error of φ for $\mu \in M_1^\infty$ is defined as

$$1 - E_\mu[\varphi(Y)] = 1 - \int_{\mathcal{Y}} \varphi(y) f(y|\mu) dy.$$

The type II error is only defined for parameter values in the alternative set, M_1^∞, and signifies the probability of falsely failing to reject the null hypothesis. Define the risk function associated with φ as

$$R_\varphi(\mu) = \begin{cases} E_\mu[\varphi(Y)] & \text{if } \mu \in M_0^\infty \\ 1 - E_\mu[\varphi(Y)] & \text{if } \mu \in M_1^\infty \end{cases}.$$

The risk function allows the comparison of tests. In particular, a test φ' is said to *dominate* φ, if $R_{\varphi'}(\mu) \leq R_\varphi(\mu)$ for all $\mu \in M^\infty$ with strict inequality, $R_{\varphi'}(\mu) < R_\varphi(\mu)$, for some $\mu \in M^\infty$. A test φ is called *admissible* in a class of tests, $C^* \subseteq C$, if there exists no test $\varphi' \in C^*$ such that φ' dominates φ. Admissibility is a minimal optimality requirement for a test within a certain class of tests. If a test is admissible, it cannot uniformly be improved

[9] The CLR above is defined as a function of Y_b and X, but since Y is a one-to-one function of Y_b and X, the CLR could equivalently be expressed as a function of Y.

upon, i.e., the probability of making an incorrect decision cannot be lowered somewhere in the parameter space without an increase in that probability elsewhere.

Before stating the admissibility result for the CLR, we introduce the concept of similarity, which is used in its derivation. A test, φ, is said to be *conditionally similar* if $E_\mu[\varphi(Y)|X = x] = \alpha \ \forall \ x \in \mathcal{X}$ and $\forall \ \mu \in M_0^\infty$ and *similar* if $E_\mu[\varphi(Y)] = \alpha \ \forall \ \mu \in M_0^\infty$. The CLR is by construction conditionally similar. It follows, by the law of iterated expectations, that it is also (unconditionally) similar.[10]

The following Theorem asserts that the CLR introduced in Section 2.2 is admissible in the class of all tests pertaining to the testing problem given in (1) and (2).

Theorem 1. *The φ_{CLR} defined in* (4) *is admissible in the class of all tests, C, pertaining to the testing problem given in* (1) *and* (2)*.*

The proof of Theorem 1 is given in Appendix A.1. It consists of two parts. First, it is shown that any similar test with convex acceptance sections is admissible. Second, it is shown that the CLR has convex acceptance sections. A test is said to have *convex acceptance sections* if for any $x \in \mathcal{X}$ the acceptance region of the test as a function of y_b is closed and convex.[11] The acceptance region of a test is the part of the sample space, for which the test fails to reject the null hypothesis. The first part of the proof follows from Theorem 3.1 in Matthes and Truax (1967). For the problem at hand, the Theorem asserts that the class of similar tests with convex acceptance sections is complete. A class of tests, $C^* \subseteq C$, is *complete*, if for any $\varphi \notin C^*$, there exists a $\varphi^* \in C^*$ such that φ^* dominates φ. Admissibility of a similar test with convex acceptance sections can then be derived when Y_b is scalar. The result relies on X being a complete sufficient statistic for d. The second part of the proof follows from the definition of the CLR statistic given in equation (3).

The testing problem at hand satisfies assumption F0-F2 in Montiel Olea (2013). Therefore, his Theorem 1 applies, which states that any admissible and similar test is an extended Efficient Conditionally Similar (ECS) test. The appeal of an extended ECS test is that there exist weights (with full support on M^∞) with respect to which the test is arbitrarily close to the weighted average power (WAP) maximizer subject to a similarity constraint. Put differently, there exists weights such that the CLR is the essential WAP maximizing test (with respect to those weights) subject to a similarity constraint.

[10] A priori, it is not clear whether similarity implies good power properties. For example, in moment inequality models similar tests have been shown to have poor power (Andrews, 2012). The power analysis in Section 4 shows that the CLR has good power properties for the testing problem at hand.

[11] The definition of convex acceptance sections given in equation (3.1) in Matthes and Truax (1967) is slightly different, since it allows for randomized tests. Here, we can restrict ourselves to non-randomized tests, since Y is a continuous random variable.

As mentioned above, the proof of Theorem 1 depends crucially on Y_b being scalar. For testing problems, where Y_b is vector-valued, the CLR can be shown to be admissible in the conditional problem, but it appears to be an open question whether it is admissible in the unconditional problem.

3 Asymptotic Theory

In this section, we introduce a general class of extremum problems. In Section 3.1, we show that constrained extremum estimators, which by construction satisfy the restrictions on the parameter space, are not asymptotically normally distributed when the true parameter vector is near or at the boundary. Consequently, the gLR and the CLR, as well as other tests defined in the Gaussian shift experiment cannot be implemented based on such estimators. Unconstrained estimators are often unavailable, because the objective function is not defined outside the parameter space, e.g., a likelihood function may not be defined for negative values of a variance parameter. In Section 3.2, we propose a modified extremum estimator that is asymptotically normally distributed and that does not require the objective function to be defined outside the parameter space. This novel estimator considerably broadens the applicability of tests defined in the Gaussian shift experiment. Details on how to implement such tests based on an asymptotically normal estimator are provided in Section 3.3.

Throughout this section, we borrow notation from Andrews and Cheng (2012a) (AC1 hereafter) and Andrews and Cheng (2012b) (AC2 hereafter). The criterion function of the extremum problem is denoted $Q_n(\theta)$. The class of extremum problems is large and includes (quasi) Maximum Likelihood (ML), Generalized Method of Moments (GMM) and Minimum Distance (MD) problems among others.

The constrained estimator $\hat{\theta}_n$ is defined as the approximate minimizer of $Q_n(\theta)$ over Θ, i.e.,

$$Q_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o_p(1/n), \tag{5}$$

where Θ denotes the true parameter space. In AC1, Θ denotes the optimization parameter space, and it is assumed that the true parameter space is a strict subset of the optimization parameter space. This is done to assume away boundary effects on the asymptotic distribution of the estimator. Here, we make the assumption that the true parameter space and the optimization parameter space are identical, since we are interested in the behavior of the estimator near the boundary.

We assume that θ can be partitioned as follows: $\theta = (\beta, \delta, \xi)$, where β denotes the scalar parameter of interest, δ denotes a $K \times 1$ dimensional nuisance parameter, and ξ denotes an additional $L \times 1$ dimensional nuisance parameter. Our setup differs from that of AC1 as none of the parameters determine the identification strength of any other parameter. In fact, all parameters are assumed to be well identified. The difference between δ and ξ is that δ is modeled as close to the boundary, whereas ξ is modeled as in the interior of the parameter space. The objective function $Q_n(\theta)$ depends on data $\{W_i : i \leq n\}$, which may be i.i.d., independent and nonidentically distributed, or temporally dependent. In most applications the distribution of the data is not fully specified by the vector θ, but it depends on an additional, commonly infinite-dimensional, parameter, ϕ. For example in conditional maximum likelihood problems, ϕ denotes the distribution of the data on which we condition. The parameter $\gamma = (\theta, \phi)$ is assumed to fully specify the distribution of the data. The true parameter space is assumed to be of the following form

$$\Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta, \phi \in \Phi(\theta)\}, \tag{6}$$

where the true parameter space for θ, $\Theta \subset \mathbb{R}^{K+L+1}$, is compact. In particular, we assume that Θ equals a Cartesian product of intervals equal to $[-c, 0]$, $[0, c]$ or $[-c, c]$ for $c \in \mathbb{R}_+$, i.e., some of the parameters are bounded below or above by 0, where the normalization to 0 is without loss of generality.[12] The boundary we are interested in is at 0, and not at c.[13] The form of the parameter space is restrictive, but it is obtained for many models of interest, most notably in the context of random coefficients models, see e.g., Berry, Levinsohn, and Pakes (1995) or Abrevaya and Shen (2014).[14] As in AC1, we assume that $\Phi(\theta) \subset \Phi \ \forall \ \theta \in \Theta$ for some compact metric space Φ with a metric that induces weak convergence of the bivariate distribution (W_i, W_{i+m}) for all $i, m \geq 1$.

When the true parameter vector lies in the interior of the parameter space, standard asymptotic theory provides good approximations to the finite sample behavior of the estimator for large enough sample sizes. But, at any given sample size, the true parameter vector might be too close to the boundary for standard asymptotic theory to provide good approximations. Intuitively, standard asymptotic theory provides poor approximations if the estimates are within “a few” standard errors from the boundary. In that case, modeling the true parameter vector as close to the boundary relative to the sample size provides better finite sample approximations. This is achieved by means of drifting sequences of true param-

[12] The use of c as common endpoint for all intervals is merely for notational convenience. The endpoints are free to vary between all $K + L + 1$ sets in Θ.

[13] In fact, here we implicitly assume that the true parameter space is a strict subset of the optimization parameter space, since we do not allow the true parameter to be equal to c, but crucially we do not assume that for the boundary at 0.

[14] With the exception of random coefficients models which allow the random coefficients to be correlated, see e.g., Andrews (2001).

eters, which approach the boundary at a rate inversely related to the sample size, n, where the rate is chosen such that the distance to the boundary "shows up" in the asymptotic distribution.

A drifting sequence of true parameters is denoted $\gamma_n = (\theta_n, \phi_n)$. The set of all such drifting sequences is given by

$$\Gamma(\gamma^*) = \{\{\gamma_n \in \Gamma : n \geq 1\} : \gamma_n \rightarrow \gamma^* \in \Gamma\}.$$

The drifting sequences of primary interest are given by

$$\Gamma(\gamma^*, b, d) = \{\{\gamma_n \in \Gamma(\gamma^*) : n \geq 1\} : \sqrt{n}\beta_n \rightarrow b \in B^\infty \text{ and } \sqrt{n}\delta_n \rightarrow d \in D^\infty\}, \tag{7}$$

where B^∞ equals $(-\infty, 0]$, $[0, \infty)$, or $(-\infty, \infty)$ when the first coordinate of Θ equals $[-c, 0]$, $[0, c]$, or $[-c, c]$, and D^∞ equals the product space of sets equaling $(-\infty, 0]$ or $[0, \infty)$ in accordance with the coordinates of Θ, which equal $[-c, 0]$ or $[0, c]$. Thus, the set of drifting sequences given in (7) implicitly imposes $|b| < \infty$ and $\|d\| < \infty$. Throughout this paper, we use the terminology "under $\{\gamma_n\} \in \Gamma(\gamma^*)$" to mean "when the true parameters are $\gamma_n \in \Gamma(\gamma^*)$ for any $\gamma^* \in \Gamma$" and "under $\{\gamma_n\} \in \Gamma(\gamma^*, b, d)$" to mean "when the true parameters are $\gamma_n \in \Gamma(\gamma^*, b, d)$ for any $\gamma^* \in \Gamma$ with $\beta^* = 0$, $\delta^* = 0$, $b \in \mathbb{R}$, and $d \in \mathbb{R}^K$".

In order to help fix ideas, we introduce a running example:

Example 2: Consider the following random coefficients model

$$y_i = x_i \eta_i + u_i,$$

where for expository purposes x_i is assumed to be scalar. (x_i, y_i) is assumed to be i.i.d.. We assume further that $\eta_i \sim N(\mu_\eta, \sigma_\eta^2)$ and $u_i \sim N(0, \sigma_u^2)$ such that $\eta_i \perp\!\!\!\perp u_i$. Then, the model can be written as

$$y_i = x_i \mu_\eta + u_i + x_i v_i \sigma_\eta = x_i \mu_\eta + \varepsilon_i,$$

where $v_i \sim N(0, 1)$ and $\varepsilon_i = u_i + x_i v_i \sigma_\eta$. Note that $\varepsilon_i | x_i \sim N(0, \sigma_u^2 + x_i^2 \sigma_\eta^2)$. The conditional individual log likelihood function is given by

$$\mathrm{l}(\mu_\eta, \sigma_u^2, \sigma_\eta^2 | y_i, x_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_u^2 + x_i^2 \sigma_\eta^2\right) - \frac{(y_i - x_i \mu_\eta)^2}{2\left(\sigma_u^2 + x_i^2 \sigma_\eta^2\right)}.$$

The (scaled) conditional log likelihood function, which for notational consistency is denoted

$Q_n(\theta)$, is given by

$$Q_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{l}(\mu_\eta, \sigma_u^2, \sigma_\eta^2 | y_i, x_i).$$

We can "change" the order of the elements in θ to conform with the above notation according to which parameter we are interested in. The parameter space for μ_η, σ_η^2, and σ_u^2 is given by $[-c, c]$, $[0, c]$, and $[a, c]$, respectively, where $0 < a < c$. We assume that $\sigma_u^{2*} > a$, i.e., only the variance parameter σ_η^{2*} may be at the boundary.[15] A reparameterization from σ_u^2 to $\sigma_u^2 - a$ results in a parameter space of the form $[-c, c]$, which conforms with the above definitions. There are three possible orderings of the elements in θ, depending on which of the three parameters is the parameter of interest: 1) $\beta = \mu_\eta$, $\delta = \sigma_\eta^2$, and $\xi = \sigma_u^2 - a$, 2) $\beta = \sigma_\eta^2$, δ empty, and $\xi = (\mu_\eta, \sigma_u^2 - a)$, and 3) $\beta = \sigma_u^2 - a$, $\delta = \sigma_\eta^2$, and $\xi = \mu_\eta$. The parameter space for ϕ is given by

$$\Phi = \{\phi : E_\phi |x_i|^{8+\epsilon} \leq C\}, \tag{8}$$

where $\epsilon \in \mathbb{R}_{++}$ and $C \in \mathbb{R}_{++}$. Note that here, Φ does not depend on θ.

Next, we introduce the assumptions underlying the asymptotic distribution theory derived in this paper. The assumptions are stronger than those in Andrews (1999), A1 hereafter, as they to not allow for non-stationary time series or complicated shapes of the parameter space. However, the assumptions allow for drifting sequences of true parameters.

We make the high level assumption that $\hat{\theta}_n$ is consistent for θ^*.

Assumption 1. $\hat{\theta}_n = \theta^* + o_p(1) \ \forall \ \gamma^* \in \Gamma$.

Note that Assumption 1 implies that

$$\hat{\theta}_n = \theta_n + o_p(1).$$

This follows trivially as $\theta_n \to \theta^*$. Thus, $\hat{\theta}_n$ can be thought of as a consistent estimator for the drifting sequence of true parameters, θ_n. A sufficient condition for Assumption 1 is given by the following.

Assumption 1*.

(a) Under $\{\gamma_n\} \in \Gamma(\gamma^*)$, $\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta; \gamma^*)| = o_p(1)$ for some non-stochastic real-valued function $Q(\theta; \gamma^*)$.

(b) $Q(\theta; \gamma^*)$ is continuous on $\Theta \ \forall \ \gamma^* \in \Gamma$.

[15] As above the superscript * denotes the limit of the drifting sequence of true parameters.

(c) $Q(\theta;\gamma^*)$ is uniquely minimized by $\theta^* \; \forall \; \gamma^* \in \Gamma$.

(d) Θ is compact.

As in A1 and AC1, we consider a quadratic approximation to the objective function. In particular, we consider an approximation around the true value as in A1 with the difference that here the true value is given by a drifting sequence and, thus, depends on the sample size, n. We do not have to consider the approximation around the point of discontinuity as in AC1, since we assume that identification does not depend on the true parameter value. The quadratic approximation is given by

$$Q_n(\theta) = Q_n(\theta_n) + \frac{\partial}{\partial\theta}Q_n(\theta_n)'(\theta-\theta_n) + \frac{1}{2}(\theta-\theta_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\theta-\theta_n) + R_n(\theta), \qquad (9)$$

where $\frac{\partial}{\partial\theta}Q_n(\theta_n)$ and $\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)$ are defined in the following assumption, which assures that the quadratic approximation exists and that the remainder term is asymptotically negligible.

Assumption 2.

(a) $Q_n(\theta)$ has continuous left/right (l/r) partial derivatives of order two on $\Theta \; \forall \; n \geq 1$ with probability 1.

(b) Under $\{\gamma_n\} \in \Gamma(\gamma^*)$, for all constants $\epsilon_n \to 0$,

$$\sup_{\theta\in\Theta:\|\theta-\theta_n\|\leq\epsilon_n}\left\|\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right\| = o_p(1),$$

where $(\partial/\partial\theta)Q_n(\theta)$ and $(\partial^2/\partial\theta\partial\theta')Q_n(\theta)$ denote the $(K+L+1)\times 1$ vector and $(K+L+1)\times(K+L+1)$ matrix of l/r partial derivatives of $Q_n(\theta)$ of orders one and two, respectively.

In many cases, Assumption 2 (b) can be verified using a uniform LLN, see e.g., Andrews (1992).[16] The last two assumptions concern the asymptotic behavior of the first and second order partial derivatives of the objective function under drifting sequences of true parameters.

Assumption 3. Under $\{\gamma_n\} \in \Gamma(\gamma^*)$, $J_n = \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n) \xrightarrow{p} J(\gamma^*)$, where $J(\gamma^*)$ is nonsingular and symmetric.

Assumption 4. (i) Under $\{\gamma_n\} \in \Gamma(\gamma^*)$, $\sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_n) \xrightarrow{d} N(0, V(\gamma^*))$ for some symmetric $V(\gamma^*)$. (ii) $V(\gamma^*)$ is positive definite $\forall\gamma^* \in \Gamma$.

[16] Assumption 2 corresponds to Assumption 2^{2*} in A1 and Assumption Q1 in AC2. Assumption 2^{2*} in A1 is sufficient for Assumption 2^* in A1, while Assumption Q1 in AC2 is sufficient for D1 in AC1.

Assumption 3 can often be verified using a uniform LLN for triangular arrays, while Assumption 4 typically follows from a CLT for triangular arrays.[17]

Example 2 (continued): The verification of Assumptions 1-4 can be found in Appendix A.4.

We introduce some additional notation. Let

$$Z_n = -J_n^{-1}\sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_n),$$

such that $Z_n \xrightarrow{d} Z(\gamma^*)$, where

$$Z(\gamma^*) \sim N(0, J(\gamma^*)^{-1}V(\gamma^*)J(\gamma^*)^{-1}). \tag{10}$$

With J_n and Z_n thus defined, the quadratic approximation to the objective function given in equation (9) can be written as

$$Q_n(\theta) = Q_n(\theta_n) - \frac{1}{2n}Z_n'J_nZ_n + \frac{1}{2n}q_n(\sqrt{n}(\theta-\theta_n)) + R_n(\theta), \tag{11}$$

where

$$q_n(\lambda) = (\lambda - Z_n)'J_n(\lambda - Z_n).$$

Under the assumptions made above, the remainder, $R_n(\theta)$, is small enough such that the asymptotic distribution of the centered and scaled minimizer of $Q_n(\theta)$, $\sqrt{n}(\hat{\theta}_n - \theta_n)$, is asymptotically equivalent to the asymptotic distribution of the centered and scaled minimizer of $Q_n(\theta) - R_n(\theta)$. The latter function only depends on θ through the function $q_n(\cdot)$, which is quadratic in θ. The distribution of the minimizer of a quadratic function is much easier to characterize than the distribution of the minimizer of $Q_n(\theta)$ explaining the use of rewriting (9) as (11). The asymptotic distribution of the minimizer of $q_n(\lambda)$ is given by the distribution of

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda(b,d)} q(\lambda), \tag{12}$$

where

$$q(\lambda) = (\lambda - Z(\gamma^*))'J(\gamma^*)(\lambda - Z(\gamma^*))$$

and where $\Lambda(b,d)$ denotes the limit of the shifted and scaled parameter space, $\sqrt{n}(\Theta - \theta_n)$.

[17] Assumptions 3 and 4 correspond to Assumptions D2 and D3 in AC1, respectively, and to Assumption 3 in A1.

Below, we formally show that $\sqrt{n}(\hat{\theta}_n - \theta_n)$ is asymptotically distributed as $\hat{\lambda}$. The distribution of $\hat{\lambda}$ crucially depends on $\Lambda(b,d)$ and is given by the projection of a normal random variable, $Z(\gamma^*)$, onto $\Lambda(b,d)$ with respect to the norm $q(\lambda)^{\frac{1}{2}}$.

In the standard case, where the limit of drifting sequence of true parameters, θ^*, lies in the interior of the parameter space, we have that $\sqrt{n}(\Theta - \theta_n) \to \Lambda = \mathbb{R}^{K+L+1}$ such that $\hat{\lambda} = Z(\gamma^*)$, recall equation (12). This illustrates that the approach of quadratically approximating the objective function, which conceptually differs slightly from the typical linear approximation to the first order condition, constitutes another way of obtaining the standard asymptotic normality result for $\sqrt{n}(\hat{\theta}_n - \theta_n)$ when θ^* lies in the interior of the parameter space.

3.1 Asymptotic distribution of constrained extremum estimator

In this section, we present the asymptotic distribution result for $\sqrt{n}(\hat{\theta}_n - \theta_n)$ when the true parameter vector is near or at the boundary. Although the result is not new (Andrews, 1999), it is helpful in establishing uniformity results, as it is derived using drifting sequences of true parameters as in Andrews and Cheng (2012a). We also discuss conditions under which the two-sided t-test based on a constrained extremum estimator controls asymptotic size.

For all sequences $\{\gamma_n\} \in \Gamma(\gamma^*, b, d)$, we have that $\sqrt{n}(\Theta - \theta_n) \to \Lambda(b,d)$, where $\Lambda(b,d)$ denotes a cone with nonzero vertex. In what follows, we also refer to $\Lambda(b,d)$ as the local parameter space. We illustrate the shape of $\Lambda(b,d)$ in the context of our running example.

Example 2 (continued): We consider ordering number 1) $\beta = \mu_\eta$, $\delta = \sigma_\eta^2$, and $\xi = \sigma_u^2 - a$ such that $B = [-c, c]$, $D = [0, c]$, and $\Xi = [-c, c]$. Furthermore, let $\sqrt{n}\beta_n \to b$, where $|b| < \infty$, $\sqrt{n}\delta_n \to d < \infty$, and $\xi_n \to \xi^*$, where $-c < \xi^* < c$. Then, $\sqrt{n}(\Theta - \theta_n) \to \Lambda(b,d) = (-\infty, \infty) \times [-d, \infty) \times (-\infty, \infty)$, which is a cone with vertex $(0, -d, 0)$.[18] Here, $\Lambda(b,d)$ does not depend on b.

More generally, $\Lambda(b,d)$ is equal to a product set of intervals. In particular it takes the form $(-\infty, -b]$, $[-b, \infty)$ or $(-\infty, \infty)$ when B equals $[-c, 0]$, $[0, c]$ or $[-c, c]$ times a K dimensional product set, where the k^{th} set equals $(-\infty, -d_k]$ or $[-d_k, \infty)$ when D_k equals $[-c, 0]$ or $[0, c]$ for $k = 1, \ldots, K$, times a L dimensional product set, where each set equals $(-\infty, \infty)$.

[18] Note, that since $\Lambda(b,d)$ is not a proper cone the vertex is not uniquely defined with respect to the first and the third element. We choose 0 without loss of generality.

Proposition 1. *Under Assumptions 1-4 and under* $\{\gamma_n\} \in \Gamma(\gamma^*, b, d)$*,* $\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \hat{\lambda}$*, where* $\hat{\lambda}$ *is defined in* (12) *with* $\Lambda(b, d)$ *defined as in the preceding paragraph.*

The proof of Proposition 1 follows from the arguments in A1. Details can be found in Appendix A.2. The results in Section 6 of A1 concerning the asymptotic distribution of subvectors of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ apply here as well with the slight modification that here $\Lambda(b, d)$ is a cone with non zero vertex. Since the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ is not the main interest of this paper, we refrain from reproducing the results here.

As mentioned in Section 2.2, the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_n)$ is given by the distribution of the maximum likelihood estimator for b in (2) subject to $\mu \in M^\infty$ under some condition. This condition is given by $V(\gamma^*) = aJ(\gamma^*)$ for some constant $a \in \mathbb{R}_+$. Put differently, the correspondence is obtained if the matrix defining the norm $q(\lambda)^{\frac{1}{2}}$ is proportional to the variance matrix of $Z(\gamma^*)$. To gain some intuition for this, we consider a simple example. Let $K = 1$ and $L = 0$ with $B = [-c, c]$ and $D = [0, c]$. Then, letting $\overset{a}{\sim}$ denote "asymptotically distributed as", it can be shown that

$$\sqrt{n}\hat{\beta}_n \overset{a}{\sim} Y_b + J_{11}^{-1} J_{12} \min(0, Y_d) \tag{13}$$

where Y_b and Y_d are scalar and

$$J = J(\gamma^*) = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{bmatrix}.$$

Generally, the variance matrix of $Y = (Y_b, Y_d)$, Σ, is given by $J(\gamma^*)^{-1}V(\gamma^*)J(\gamma^*)^{-1}$. However, if $V(\gamma^*) = cJ(\gamma^*)$, Σ simplifies, and the expression in equation (13) reduces to

$$\sqrt{n}\hat{\beta}_n \overset{a}{\sim} Y_b - \Sigma_{\beta\delta}\Sigma_{\delta\delta}^{-1} \min(0, Y_d).$$

It is easy to see that the distribution of $Y_b - \Sigma_{\beta\delta}\Sigma_{\delta\delta}^{-1} \min(0, Y_d)$ is also the distribution of the maximum likelihood estimator for b in the corresponding Gaussian shift experiment. Since the t_{ML} controls size, as illustrated in Section 4 below, it follows that the two-sided t-test based on $\hat{\beta}_n$ controls asymptotic size. $V(\gamma^*) = aJ(\gamma^*)$ holds, for instance, when, in the context of GMM, the efficient weighting matrix is employed or when, in the context of ML, the likelihood function is correctly specified.

If $V(\gamma^*) \neq cJ(\gamma^*)$, the two-sided t-statistic based on $\hat{\beta}_n$ can suffer from overrejection. Note, however, that overrejection does not occur in the part of the sample space where the estimate is not restricted or, put differently, not at the boundary. One way to think of this is that, if the estimate is not found to be restricted, then it could have been obtained using an unrestricted estimator, which, if used in the construction of the two-sided t-test, allows for size-correct inference.

Next, we illustrate the asymptotic distribution result given in Proposition 1 in the context of our running example.

Example 2 (continued): Since our model is a well-specified likelihood model, we make use of the fact that the information equality holds. In particular, this implies that $J(\gamma^*) = V(\gamma^*)$, such that $Z(\gamma^*) \sim N(0, V(\gamma^*)^{-1})$. We consider the three different orderings of elements in θ separately. For the last two orderings the asymptotic distribution is not normal. In order to investigate how well the asymptotic distribution approximates the finite sample distribution, we provide Monte Carlo results for these two orderings. The asymptotic distribution and the finite sample distribution are both evaluated using 10.000 Monte Carlo draws. We choose $n = 400$. x_i is drawn from a $U(1,2)$ distribution. The parameter values are given by $\mu_\eta = 0$, $\sigma_u^2 = 1$, and σ_η^2 is varied as indicated below.

1) $\beta = \mu_\eta$, $\delta = \sigma_\eta^2$, and $\xi = \sigma_u^2 - a$: Since $\beta_n \to 0$ and $\delta_n \to 0$, we have that $\theta^* = (0, 0, \xi^*) = (0, 0, \sigma_u^{2*} - a)$. The Fisher Information, $V(\gamma^*)$, is given by[19]

$$E_{\phi^*}\begin{bmatrix} \frac{1}{\sigma_u^{2*}} x_i^2 & 0 & 0 \\ 0 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} x_i^4 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} x_i^2 \\ 0 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} x_i^2 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} \end{bmatrix},$$

where $E_{\phi^*}[\cdot]$ denotes the expectation with respect to the distribution of x_i under ϕ^*. Due to the information orthogonality the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_n)$ does not depend on δ, and we obtain

$$\sqrt{n}(\hat{\beta}_n - \beta_n) \xrightarrow{d} N(0, \Sigma_{\beta\beta}),$$

where $\Sigma_{\beta\beta}$ denotes $\frac{\sigma_u^{2*}}{E_{\phi^*}[x_i^2]}$.

2) $\beta = \sigma_\eta^2$, δ empty, and $\xi = (\mu_\eta, \sigma_u^2 - a)$: Since $\beta_n \to 0$, we have that $\theta^* = (0, \xi^*) = (0, \mu_\eta^*, \sigma_u^{2*} - a)$. The Fisher Information, $V(\gamma^*)$, is given by

$$E_{\phi^*}\begin{bmatrix} \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} x_i^4 & 0 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} x_i^2 \\ 0 & \frac{1}{\sigma_u^{2*}} x_i^2 & 0 \\ \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} x_i^2 & 0 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} \end{bmatrix}.$$

[19] The first and second order partial derivatives of the likelihood function can be found in Appendix A.4.

Since ξ does not impact the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_n)$ (see e.g., A1), we have that $\sqrt{n}(\hat{\beta}_n - \beta_n) \xrightarrow{d} \max(-b, N(0, \Sigma_{\beta\beta}))$ or, equivalently,

$$\sqrt{n}\hat{\beta}_n \xrightarrow{d} \max(0, N(b, \Sigma_{\beta\beta})), \tag{14}$$

where $\Sigma_{\beta\beta}$ denotes $\frac{2(\sigma_u^{2*})^2}{E_{\phi^*}[x_i^4]}$.

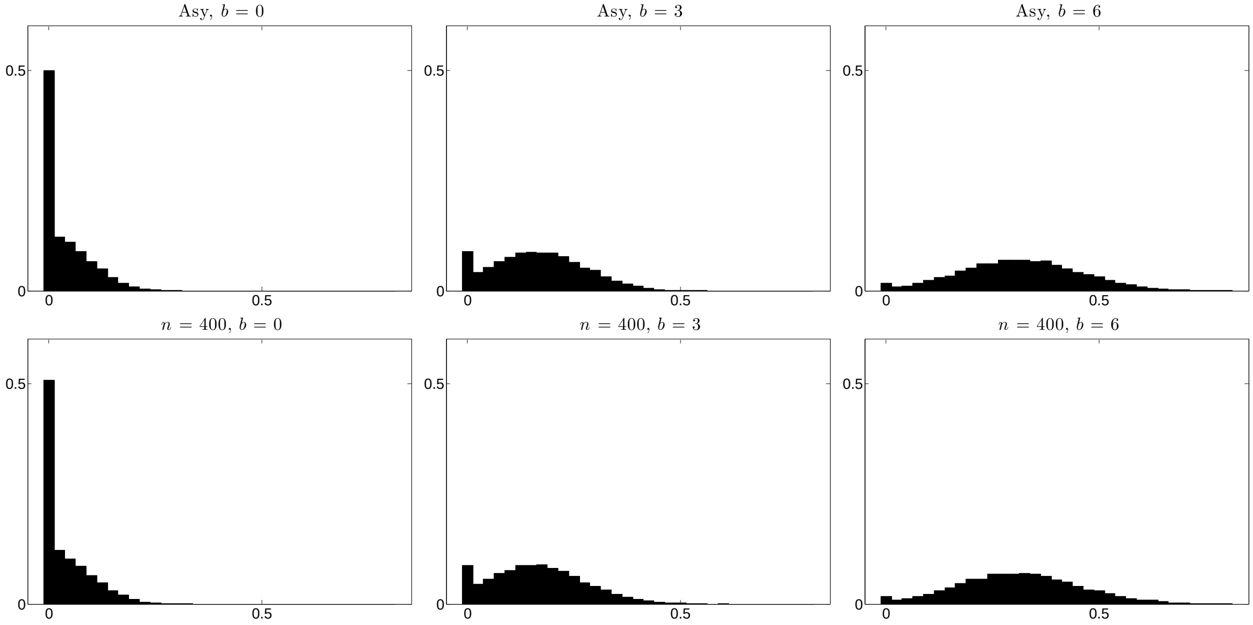

Figure 1: Asymptotic and finite sample densities of $\hat{\beta}_n = \hat{\sigma}_\eta^2$ for different values of b.

Figure 1 shows the asymptotic and finite sample densities of $\hat{\beta}_n = \hat{\sigma}_\eta^2$ for different values of $b = \sqrt{n}\beta_n$. b takes on the values 0, 3, and 6 from left to right, which corresponds to β_n taking on the values 0, 0.15, and 0.3, respectively.[20]

3) $\beta = \sigma_u^2 - a$, $\delta = \sigma_\eta^2$, and $\xi = \mu_\eta$: Since $\beta_n \to 0$ and $\delta_n \to 0$, we have that $\theta^* = (0, 0, \xi^*) = (0, 0, \mu_\eta^*)$. The Fisher Information, $V(\gamma^*)$, is given by

$$E_{\phi^*}\begin{bmatrix} \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2} & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2}x_i^2 & 0 \\ \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2}x_i^2 & \frac{1}{2}\frac{1}{(\sigma_u^{2*})^2}x_i^4 & 0 \\ 0 & 0 & \frac{1}{\sigma_u^{2*}}x_i^2 \end{bmatrix}.$$

[20] In fact, the asymptotic approximation is obtained by evaluating $V(\gamma^*)$ at β equal to 0, 0.15, and 0.3 rather than 0, 0, and 0. This provides a much better finite sample approximation and reflects common practice, where plug-in estimates of $V(\gamma^*)$ are utilized. This also holds true for Figure 2 below.

The asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_n)$ can be deduced from that of $\sqrt{n}\hat{\beta}_n$, which is given by the distribution of

$$Y_b - \Sigma_{\beta\delta}\Sigma_{\delta\delta}^{-1}\min(0, Y_d),$$

where

$$\begin{pmatrix} Y_b \\ Y_d \end{pmatrix} \sim N\left(\begin{pmatrix} b \\ d \end{pmatrix}, \Sigma\right) \text{ with } \Sigma = \begin{bmatrix} \Sigma_{\beta\beta} & \Sigma_{\beta\delta} \\ \Sigma_{\beta\delta} & \Sigma_{\delta\delta} \end{bmatrix}.$$

Here, Σ denotes the inverse of

$$E_{\phi^*}\begin{bmatrix} \frac{1}{2}\frac{1}{(\sigma_u^{2\,*})^2} & \frac{1}{2}\frac{1}{(\sigma_u^{2\,*})^2}x_i^2 \\ \frac{1}{2}\frac{1}{(\sigma_u^{2\,*})^2}x_i^2 & \frac{1}{2}\frac{1}{(\sigma_u^{2\,*})^2}x_i^4 \end{bmatrix}.$$

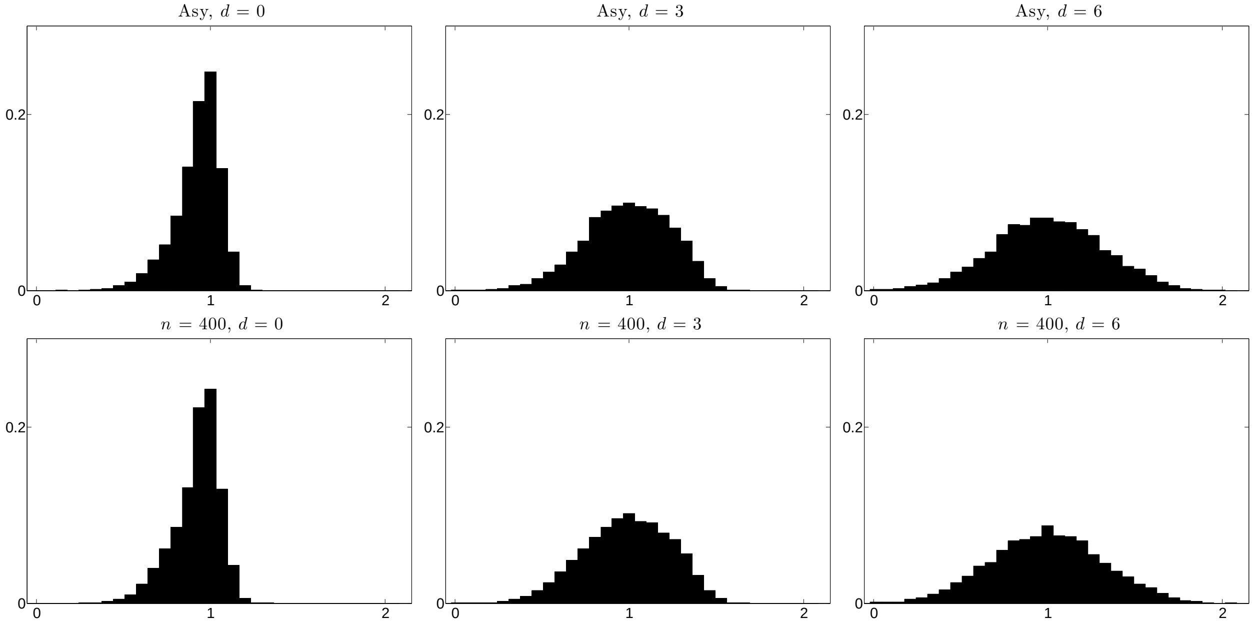

Figure 2: Asymptotic and finite sample densities of $\hat{\beta}_n = \hat{\sigma}_u^2$ for different values of d.

Figure 2 shows the asymptotic and finite sample densities of $\hat{\beta}_n = \hat{\sigma}_u^2$ for different values of $d = \sqrt{n}\delta_n$. d takes on the values 0, 3, and 6 from left to right, which corresponds to δ_n taking on the values 0, 0.15, and 0.3, respectively.

The Monte Carlo results illustrate that the finite sample distribution is well approximated by the asymptotic distribution derived above.

The reason why the asymptotic distribution of a constrained extremum estimator is not normal is that the estimator is restricted to the parameter space. If it were possible to obtain an unrestricted estimator, then that estimator would be asymptotically normal under standard regularity conditions. But as seen in Example 2, an unrestricted estimator is not always available.

3.2 Asymptotic distribution of modified extremum estimator

In the preceding section, we showed that constrained extremum estimators are not asymptotically normally distributed when the true parameter vector is near or at the boundary of the parameter space. In this section, we show that it is possible to obtain an asymptotically normal estimator, even if the objective function is not defined outside the parameter space. As mentioned above, this is useful as it broadens the applicability of testing procedures defined in the Gaussian shift experiment.

In order to give some intuition for the estimator proposed below, consider the asymptotic distribution of the estimator of σ_η^2 in Example 2 given in equation (14). It is given by a normal distribution truncated below at 0. The truncation of the asymptotic distribution results from the restriction on the parameter space, which prevents the estimator from taking on negative values. If an unrestricted estimator were available, it would be asymptotically distributed as the underlying normal random variable, $N(\beta, \omega_{\beta\beta})$. Our objective is, thus, to construct an estimator that behaves like an unrestricted estimator.

The distributional result in the previous section is obtained by showing that constrained extremum estimators behave asymptotically like minimizers of quadratic functions over a strict subset of the Euclidean space, the local parameter space. A quadratic function, in contrast to the original objective function, is always defined over the entire Euclidean space and can, therefore, also be minimized over the entire Euclidean space. The proposed estimator is given by the unrestricted minimizer of a quadratic function that approximates the original objective function.

Another way of motivating the estimator is as follows. When the estimate is at the boundary of the parameter space, we know that the estimate does not satisfy the first order condition of the optimization problem. But intuitively, there is a point outside the parameter space that would satisfy the first order condition, if that point were allowed in the sense of the objective function being well defined at that point or, more precisely, in an open set around that point. This point is approximated by the minimum of the quadratic function that approximates the objective function and that passes through the estimate at the boundary. Figure 3 illustrates the above intuition graphically, where the quadratic approximation is denoted $\hat{M}_n(\theta)$, θ is scalar and non-negative, and where the estimate is at the boundary, i.e., $\hat{\theta}_n = 0$.

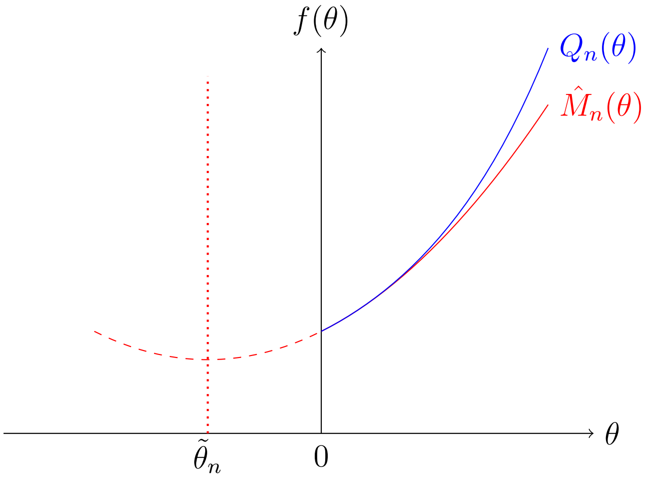

Figure 3: Illustration of proposed estimator.

We now formally define the estimator. Let

$$M_n(\theta) = Q_n(\theta_n) + \frac{\partial}{\partial \theta} Q_n(\theta_n)(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)' \frac{\partial^2}{\partial \theta \partial \theta'} Q_n(\theta_n)(\theta - \theta_n) \tag{15}$$

such that

$$Q_n(\theta) = M_n(\theta) + R_n(\theta). \tag{16}$$

Put differently, $M_n(\theta)$ constitutes the "main" (or quadratic) part of the objective function, i.e., it excludes the asymptotically negligible remainder. The minimizer of $M_n(\theta)$, say $\ddot{\theta}_n$, appropriately centered and scaled is asymptotically normally distributed. To see this, note that $\ddot{\theta}_n$ satisfies the following first order condition

$$0 = \frac{\partial}{\partial \theta} M_n(\ddot{\theta}_n) = \frac{\partial}{\partial \theta} Q_n(\theta_n) + \frac{\partial^2}{\partial \theta \partial \theta'} Q_n(\theta_n)(\ddot{\theta}_n - \theta_n).$$

Solving for $(\ddot{\theta}_n - \theta_n)$ and multiplying both sides by $\sqrt{n}$, we get

$$\sqrt{n}(\ddot{\theta}_n - \theta_n) = \left(\frac{\partial^2}{\partial \theta \partial \theta'} Q_n(\theta_n) \right)^{-1} \sqrt{n} \frac{\partial}{\partial \theta} Q_n(\theta_n).$$

Now, by Assumptions 3 and 4 (in combination with a Continuous Mapping Theorem and Slutzky), we have that

$$\sqrt{n}(\ddot{\theta}_n - \theta_n) \xrightarrow{d} Z(\gamma^*), \tag{17}$$

where $Z(\gamma^*)$ is given in equation (10). In other words, $\sqrt{n}(\ddot{\theta}_n - \theta_n)$ is asymptotically distributed like the normal random variable that underlies the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ given in Proposition 1. The estimator, thus, has the desired property described above.

However, the above estimator is infeasible, since $M_n(\theta)$ depends on the true parameter value, θ_n. A feasible estimator is given by the minimizer of the following objective function

$$\hat{M}_n(\theta) = Q_n(\hat{\theta}_n) + \frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n)'(\theta - \hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)'\frac{\partial}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n)(\theta - \hat{\theta}_n), \qquad (18)$$

which equals $M_n(\theta)$ with the unknown true parameter value, θ_n, replaced by $\hat{\theta}_n$, which by Assumption 1 is a consistent estimator of θ_n. The following Theorem formally shows that the centerd and scaled minimizer of $\hat{M}_n(\theta)$ is also asymptotically distributed as $Z(\gamma^*)$.

Theorem 2. *Let* $\tilde{\theta}_n = \arg\sup_\theta \hat{M}_n(\theta)$. *Under Assumptions 1-4 and under* $\{\gamma_n\} \in \Gamma(\gamma^*, b, d)$, $\sqrt{T}(\tilde{\theta}_n - \theta_n) \to Z(\gamma^*)$.

Remark 1. In terms of implementation, $\tilde{\theta}_n$ is easily obtained by performing a single step of the Newton-Raphson algorithm, namely by computing $\tilde{\theta}_n = \hat{\theta}_n + (\frac{\partial}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n))^{-1}\frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n)$. When $\hat{\theta}_n$ is in the interior of the parameter space, this amounts to setting $\tilde{\theta}_n = \hat{\theta}_n$, since the first order condition holds for an interior solution, i.e., $\frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n) = 0$.

The proof of Theorem 2 can be found in Appendix A.3. The proof consists of two parts. First, it is shown that the two quadratic functions given in equations (15) and (18) are "close" to each other around θ_n. This relies on $\hat{\theta}_n$ being a $\sqrt{n}$-consistent estimator of θ_n, which follows from Proposition 1. Then, the proof proceeds by showing that the centered and scaled minimizers of the two quadratic functions are "close" to each other. This then suffices to show the desired result since the centered and scaled minimizer of (15) converges in distribution to $Z(\gamma^*)$, see equation (17).

In order to illustrate the finite sample performance of the novel estimator, we return to our running example.

Example 2 (continued): For the sake of coherency, we present the finite sample densities of $\tilde{\theta}_n$ in terms of the different orderings. We only present orderings 2 and 3.

2) $\beta = \sigma^2_\eta$, δ empty, and $\xi = (\mu_\eta, \sigma^2_u - a)$:

   Figure 4 shows the finite sample distribution of $\tilde{\beta}_n = \tilde{\sigma}^2_\eta$ in the center panel and the finite sample distribution of $\hat{\beta}_n = \hat{\sigma}^2_\eta$ in the first panel, for ease of reference. The third panel shows the asymptotic density of $\tilde{\beta}_n = \tilde{\sigma}^2_\eta$.

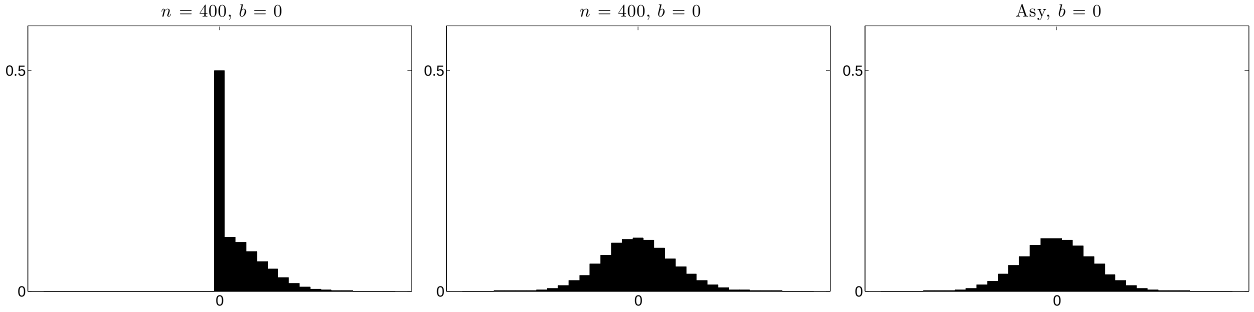

Figure 4: Finite sample (and asymptotic) densities of $\hat{\beta}_n = \hat{\sigma}^2_\eta$ and $\tilde{\beta}_n = \tilde{\sigma}^2_\eta$ for $b = 0$.

3) $\beta = \sigma^2_u - a$, $\delta = \sigma^2_\eta$, and $\xi = \mu_\eta$:

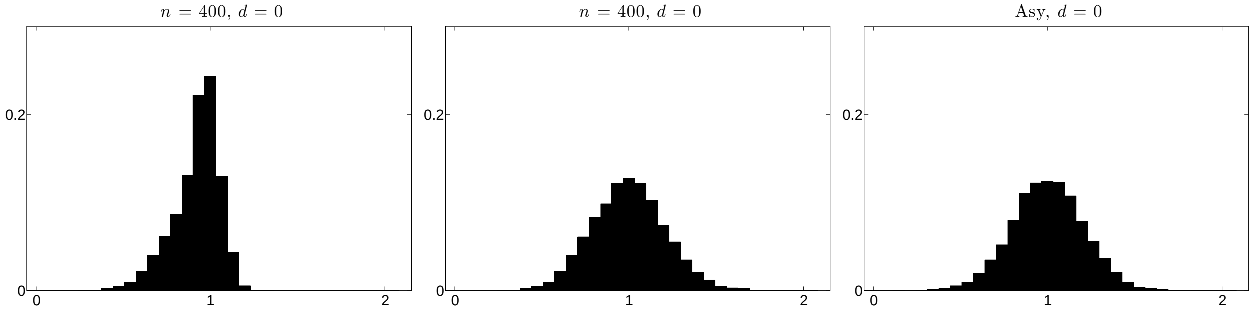

Figure 5: Finite sample (and asymptotic) densities of $\hat{\beta}_n = \hat{\sigma}^2_u$ and $\tilde{\beta}_n = \tilde{\sigma}^2_u$ for $d = 0$.

Figure 5 shows the finite sample distribution of $\tilde{\beta}_n = \tilde{\sigma}^2_u$ in the center panel and the finite sample distribution of $\hat{\beta}_n = \hat{\sigma}^2_u$ in the first panel, for ease of reference. The third panel shows the asymptotic density of $\tilde{\beta}_n = \tilde{\sigma}^2_u$.

Again, the asymptotic distribution provides a good approximation to the finite sample distribution of the estimator.

3.3 Testing in extremum problems

Theorem 2 shows that in a general class of extremum problems it is possible to obtain an asymptotically normal estimator, even when the true parameter value is near or at the boundary. This asymptotically normal estimator together with our assumptions on the parameter space allows the application of any test derived for (1) and (2) to empirically relevant testing problems. In this section, we briefly comment on the implementation details.

The local parameter space corresponding to ξ equals $\mathbb{R}^L$ and is, therefore, unrestricted. Invoking an invariance argument, we restrict our attention to the estimator of (β, δ) in order to make inference about β. This parallels the standard approach to inference, when the true parameter vector is in the interior of the parameter space. In that case, inference with respect to a scalar parameter is conducted using the two-sided t-test, which also ignores the remaining coordinates of the parameter vector. By Theorem 2, we have

$$\sqrt{n}\begin{pmatrix} \tilde{\beta}_n \\ \tilde{\delta}_n \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} b \\ d \end{pmatrix}, \Sigma\right),$$

where Σ denotes the block of $J(\gamma^*)^{-1}V(\gamma^*)J(\gamma^*)^{-1}$ corresponding to (β, δ). Furthermore, the parameter space for b and d is given by the product set of B^∞ and D^∞, which are defined in accordance to the shape of the original parameter space, e.g., if $\beta_n \in [0, c]$, then $B^\infty = [0, \infty)$. As $J(\gamma^*)^{-1}V(\gamma^*)J(\gamma^*)^{-1}$ typically depends on ϕ^* only through expectations, a consistent estimator is obtained by replacing expectations with sample averages and by replacing θ^* with $\hat{\theta}_n$, which by Assumption 1 is consistent. Let $\hat{\Sigma}$ denote the corresponding estimator of Σ. The null hypothesis $H_0 : \beta_n = \beta_0$ can then be tested using any test derived for (1) and (2) by testing $b = b_0 = \beta_0 \cdot \sqrt{n}$ with Y evaluated at its sample analogue, $\sqrt{n}\begin{pmatrix} \tilde{\beta}_n \\ \tilde{\delta}_n \end{pmatrix}$, and with $\hat{\Sigma}$ in place of Σ.

Since the tests considered in Section 2.2 are continuous in their argument, Y, it follows by the continuous mapping theorem that their evaluations at sample analogues are asymptotically valid in the sense that they asymptotically control size. Furthermore, the power curves obtained in Section 4 below, for the Gaussian shift experiment, represent the corresponding asymptotic power curves. Moreover, Theorem 1 in Müller (2011) can be invoked to show that the CLR is, in some sense, asymptotically admissible, see also Comment 3 in Müller (2011).

Since the CLR is contiguous in the localization parameters, b and d,[21] it reduces to the regular two-sided t-test, when the true parameter vector is in the interior of the parameter space. Confidence intervals obtained by inverting the CLR can thus be interpreted as a natural extension of standard confidence intervals to the nonstandard setting, where the true parameter vector is near or at the boundary. Since the (conditional) distribution of the CLR does not have a closed form expression, the inversion needs to be performed by a grid search, which is quick if β is scalar.

[21] Assume $K = 1$, i.e., d is scalar, and $D^\infty = [0, \infty)$. As $d \to \infty$, the constraint d $\in D^\infty$ (in the definition of the CLR) becomes less binding and in the limit, $d = \infty$, which corresponds to δ^* in the interior of the parameter space, never binds. The limit case is thus modeled by $K = 0$. This argument generalizes to arbitrary dimensions of the localization parameters, b and d.

4 Power comparison

In this section, we conduct a power comparison of the testing procedures introduced above and several other testing procedures that have recently been proposed in the literature. The power comparison is conducted in the Gaussian shift experiment and, thus, corresponds to a comparison of local asymptotic power when the testing procedures are based on an asymptotically normal estimator, and when the true parameter vector is close to the boundary. The goal of the power comparison is twofold. First, we want to assess the performance of the t_{ML}, which represents common practice, in comparison to alternative testing procedures. Second, we want to illustrate that our testing procedures perform well in comparison to existing ones. Note that all tests considered in this section, with the exception of the t_{ML}, require an asymptotically normal estimator and, thus, in many models of interest require the novel estimator introduced in Section 3.2.

The power comparison is conducted in the context of several leading examples that are characterized by having at most one nuisance parameter present. Power curves are numerically evaluated using 10,000 Monte Carlo draws. Throughout this section, tests are conducted with a 5% significance level. To ease the notational burden, we take Σ to be a correlation matrix, i.e., its main diagonal equals a vector of 1s. This is without loss of generality and can always be achieved through an appropriate normalization. Furthermore, $\Sigma_{\beta\delta} = \Sigma_{\delta\beta}$ is scalar throughout this section and is therefore be denoted by ρ, the correlation parameter.

4.1 No nuisance parameter

We first consider the case with one parameter of interest near the boundary and no nuisance parameter present. The Gaussian shift experiment for this case is given by

$$Y_b \sim N(b, 1).$$

The hypothesis testing problem of interest is

$$H_0 : b = b_0 \text{ vs. } H_1 : b \neq b_0, b \geq 0.$$

In the absence of nuisance parameters the gLR and the CLR are identical and given by

$$\mathrm{CLR}(b_0, y_b) = \frac{\sup_{\mathrm{b} \geq 0} \exp(-\frac{1}{2}(y_b - \mathrm{b})^2)}{\exp(-\frac{1}{2}(y_b - b_0)^2)}.$$

The rejection region of the corresponding test is given by $\{y_b : \mathrm{CLR}(b_0, y_b) > \mathrm{cv}(b_0)\}$, where $\mathrm{cv}(b_0)$ denotes the 95% quantile of $\mathrm{CLR}(b_0, Y_b)$ with $Y_b \sim N(b_0, 1)$. It is illustrative to plot the rejection region of the CLR.

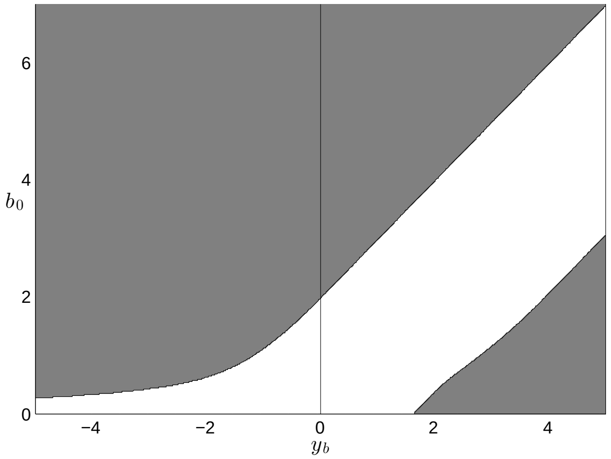

Figure 6: Rejection region of the CLR (grey) for testing $H_0 : b = b_0$ (y-axis).

Figure 6 shows the rejection region for the CLR. Two interesting features are that for testing $b_0 = 0$ the CLR equals the one-sided t-test that rejects for $y_b > 1.645$ and that for testing large values of b_0 the rejection region of the CLR approaches the rejection region of the regular two-sided t-test, which is given by $\{y_b : |y_b - b_0| > 1.96\}$.[22]

The t_{ML} is given by

$$\mathrm{t}_{\mathrm{ML}} = |\max(0, y_b) - b_0|.$$

The rejection region of the corresponding test is $\{y_b : \mathrm{t}_{\mathrm{ML}} > 1.96\}$.

Figure 7 shows the power curves of the CLR and the t_{ML} for different values of b_0. A power curve displays the rejection frequency of a test (on the y-axis) as a function of the true parameter value (on the x-axis). We find that the CLR is preferable to the t_{ML} when testing $b_0 = 0$ and $b_0 = 1$. For small alternatives, i.e., for small values of the true parameter b, the power curve of the CLR lies above the power curve of the t_{ML}, while they both asymptote to 1 for large alternatives. For testing $b_0 = 2$, the two tests are numerically indistinguishable, since they both approach the regular two-sided t-test as b_0 increases, cf. Figure 6.

Figure 8 plots the power curve of another test recently proposed by Müller and Norets (2013).[23] For ease of comparison, the figure also reproduces the power curve of the CLR. For

[22] For the sake of completeness, the power curve of the regular two-sided t-test is shown in Figure 14 in Appendix B. Overall, the CLR seems preferable, lacking only some power in comparison for small alternatives, when testing $b_0 = 1$.

[23] To be precise, Müller and Norets (2013) propose a confidence interval. The test considered here underlies the construction of their confidence interval.

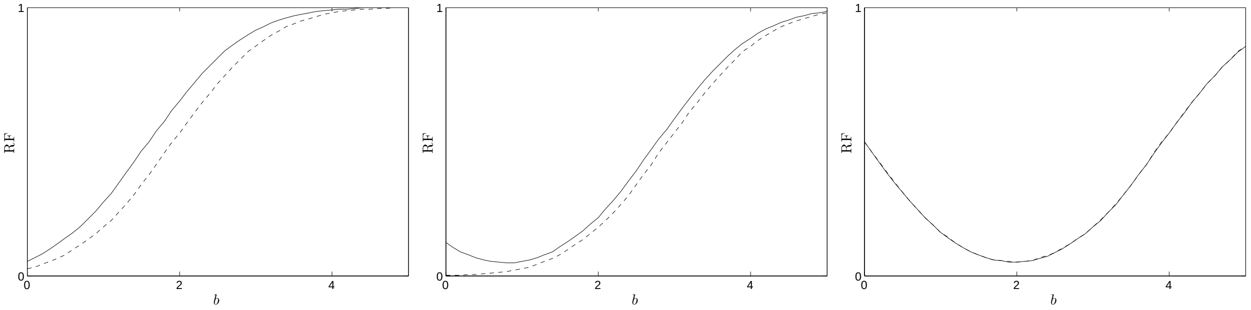

Figure 7: Power of the CLR (solid) and the t_{ML}-test (dashed) for testing $H_0 : b = 0, 1, 2$ from left to right.

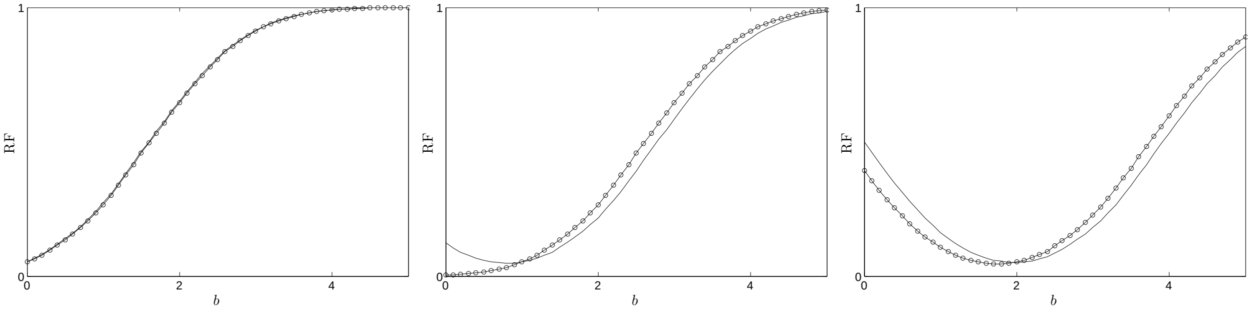

Figure 8: Power of the CLR (solid) and the MN test (solid with circles) for testing $H_0 : b = 0, 1, 2$ from left to right.

the example at hand, the test proposed by Müller and Norets (2013) (MN) is numerically indistinguishable from the (point-optimal) likelihood ratio test that is obtained by integrating out the values of b under the alternative with respect to (improper) uniform weights. For testing $b_0 = 0$, the CLR and the MN coincide, while for testing $b_0 > 0$, the power curves cross, and neither test is better than the other.

4.2 Nuisance parameter near the boundary

The testing problem becomes more interesting under the presence of nuisance parameters. A leading example is given by

$$H_0 : b = b_0, d \geq 0 \text{ vs. } H_1 : b \neq b_0, b \in \mathbb{R}, d \geq 0.$$

This testing problem is also considered in Andrews and Guggenberger (2010), Elliott, Müller, and Watson (2013) and Montiel Olea (2013). Andrews and Guggenberger (2010) discuss the

t_{ML}, which is given by

$$\mathrm{t}_{\mathrm{ML}} = |y_b - \rho \min(0, y_d) - b_0|.$$

The rejection region of the corresponding test is $\{y : \mathrm{t}_{\mathrm{ML}} > 1.96\}$. The gLR statistic for this case is given by[24]

$$\mathrm{gLR}(b_0, y_b, y_d) = \\ 2\log\left(\frac{\sup_{\mathrm{b}\in\mathbb{R},\mathrm{d}\geq 0}\exp(-\frac{1}{2(1-\rho^2)}\left[(y_b-\mathrm{b})^2+(y_d-\mathrm{d})^2-2\rho(y_b-\mathrm{b})(y_d-\mathrm{d})\right])}{\sup_{\mathrm{d}\geq 0}\exp(-\frac{1}{2(1-\rho^2)}\left[(y_b-b_0)^2+(y_d-\mathrm{d})^2-2\rho(y_b-b_0)(y_d-\mathrm{d})\right])}\right).$$

The rejection region of the gLR is given by $\{y : \mathrm{gLR}(b_0, y_b, y_d) > \mathrm{cv}_{d_{\mathrm{LFC}}}\}$, where $d_{\mathrm{LFC}} = 0$. The CLR statistic is given by the gLR statistic conditioned on the value of $x = y_d - \rho y_b$. In particular, for the example at hand,

$$\mathrm{CLR}(b_0, y_b, x) = \\ 2\log\left(\frac{\sup_{\mathrm{b}\in\mathbb{R},\mathrm{d}\geq 0}\exp(-\frac{1}{2(1-\rho^2)}\left[(y_b-\mathrm{b})^2+(x+\rho y_b-\mathrm{d})^2-2\rho(y_b-\mathrm{b})(x+\rho y_b-\mathrm{d})\right])}{\sup_{\mathrm{d}\geq 0}\exp(-\frac{1}{2(1-\rho^2)}\left[(y_b-b_0)^2+(x+\rho y_b-\mathrm{d})^2-2\rho(y_b-b_0)(x+\rho y_b-\mathrm{d})\right])}\right).$$

The rejection region is $\{(y_b, x) : \mathrm{CLR}(b_0, y_b, x) > \mathrm{cv}(\alpha, x)\}$.

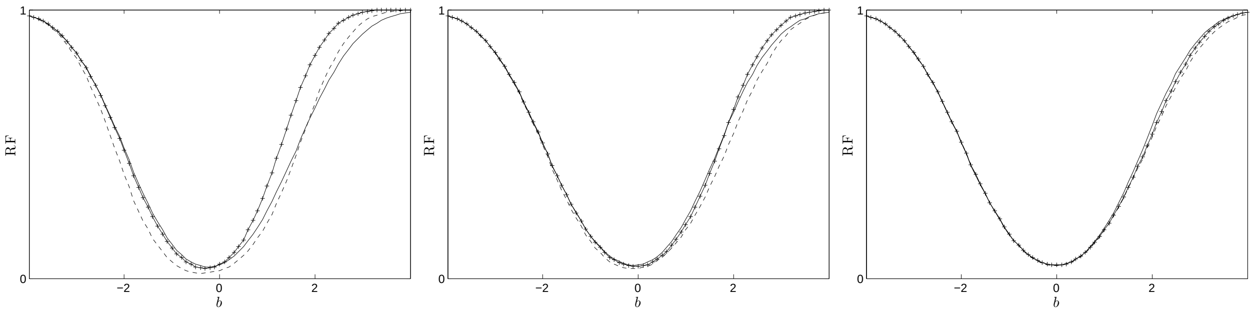

Figure 9: Power of the CLR (solid), gLR (solid with plusses), and the t_{ML}-test (dashed) for testing $H_0 : b = 0$ with $d = 0, 1, 2$ from left to right. $\rho = 0.7$.

Figure 9 shows the power curves of the three tests introduced above for different values of the nuisance parameter d and $\rho = 0.7$. The t_{ML} performs very poorly compared to the

[24] Some computations show that the closed form solution for the gLR statistic reads as follows

$$\mathrm{gLR}(b_0, y_b, y_d) = \begin{cases} (y_b - b_0)^2 & \text{if } y_d \geq 0 \text{ and } y_d - \rho(y_b - b_0) \geq 0 \\ \frac{(y_b-b_0)^2+y_d^2-2\rho(y_b-b_0)y_d}{1-\rho^2} & \text{if } y_d \geq 0 \text{ and } y_d - \rho(y_b - b_0) < 0 \\ (y_b - b_0)^2 - y_d^2 & \text{if } y_d < 0 \text{ and } y_d - \rho(y_b - b_0) \geq 0 \\ \frac{(y_b-b_0)^2+y_d^2-2\rho(y_b-b_0)y_d}{1-\rho^2} - y_d^2 & \text{if } y_d < 0 \text{ and } y_d - \rho(y_b - b_0) < 0. \end{cases}$$

gLR, displaying lower power for a wide range of alternatives. The CLR performs well. The power advantage over the gLR is limited, and arguably for this simple testing problem, the gLR seems to be preferable. But except for small values of d and large positive values of b, the CLR enjoys better power than the t_{ML}.

Recently, other tests have been proposed for the testing problem at hand. Montiel Olea (2013) suggests the use of Efficient Conditionally Similar (ECS) tests. The tests maximize weighted average power (WAP) subject to a similarity constraint. The weights are user specified. Montiel Olea (2013) suggests a certain weight function that leads to a simple expression for his test, which only depends on a scalar tuning parameter, λ. In Figure 10 below, the test is implemented for $\lambda = 0.1$.[25]

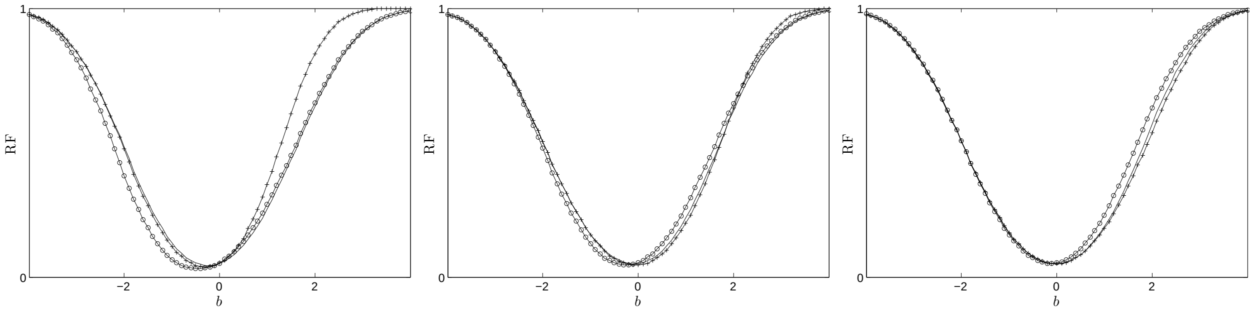

Figure 10: Power of the CLR (solid), the gLR (solid with plusses), and the ECS (solid with circles) for testing $H_0 : b = 0$ with $d = 0, 1, 2$ from left to right. $\rho = 0.7$.

Besides the power curve for the ECS, Figure 10 plots the power curves of the gLR and CLR for comparison. The ECS performs comparably to the CLR with the CLR enjoying power advantages for negative values of b, especially when d is small, whereas the ECS enjoys greater power for positive values of b at larger values of d.

Elliott, Müller, and Watson (2013) propose another testing procedure, which is based on the approximation of the least favorable distribution. The test, EMW hereafter, maximizes WAP in an otherwise unconstrained problem. For example, similarity of the test is not enforced, which distinguishes this approach from the one suggested by Montiel Olea (2013).

Figure 11 shows the power curves of the EMW, the gLR, and the CLR. The EMW is implemented using the same weights as in Elliott, Müller, and Watson (2013). The gLR performs well compared to the EMW. The gLR enjoys higher power for small values of d and compared to the EMW only lacks power marginally for larger values of d. For small values of d, the CLR has higher power for negative values of b and lower power for positive values of b, while performing similar to the EMW for larger values of d.

[25] For larger values of λ, the test becomes more skewed with very high power for positive alternatives, $b > 0$, and very low power for negative alternatives, $b < 0$.

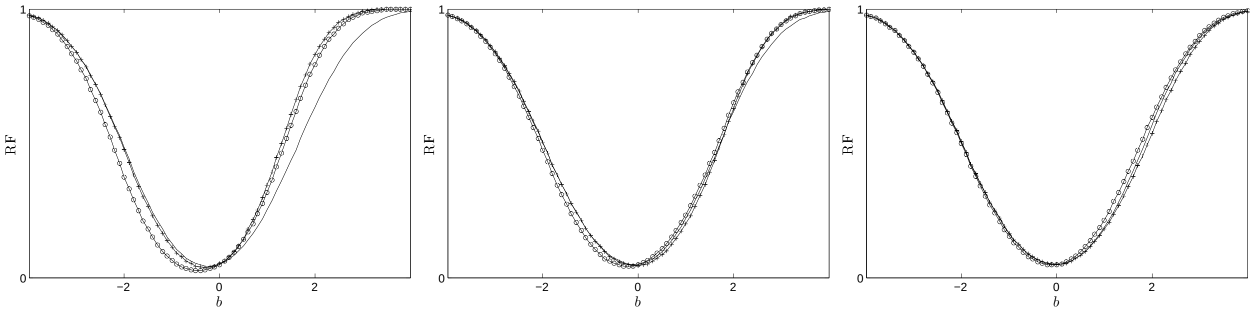

Figure 11: Power of the CLR (solid), the gLR (solid with plusses), and the EMW (solid with circles) for testing $H_0 : b = 0$ with $d = 0, 1, 2$ from left to right. $\rho = 0.7$.

4.3 Parameter of interest and nuisance parameter near the boundary

Another interesting leading example is given by

$$H_0 : b = b_0, d \geq 0 \text{ vs. } H_1 : b \neq b_0, b \geq 0, d \geq 0.$$

Here, both the parameter of interest and a scalar nuisance parameter are near or at the boundary. We refrain from implementing the tests proposed by Montiel Olea (2013) and Elliott, Müller, and Watson (2013). Note, however, that both tests would require the choice of a weight function for every b_0. We implement the gLR, the CLR, and the t_{ML}, but omit their expressions for the sake of brevity. Note that for the testing problem at hand, the sign of the correlation parameter, ρ, matters.[26] The critical configurations for the gLR are given by $d = 0$ when $\rho > 0$ and $d = \infty$ when $\rho < 0$.

Figure 12 shows the power curves of the three tests for different values of b and d with $\rho = 0.7$. Figure 15 shows the corresponding power curves for $\rho = -0.7$ and can be found in Appendix B. As before, the t_{ML} performs poorly compared to the gLR and the CLR. It performs particularly poorly for small values of d when ρ is negative. The CLR performs well and, notably, constitutes a good test for $b_0 = 0$ regardless of the sign of ρ and the value of d. This seems an appealing feature for applied work, where it is often of interest whether a parameter is significantly different from zero. Compared to the t_{ML}, the CLR enjoys greater power for alternatives greater than the null with some exceptions when d is small and ρ positive.

[26] Alternatively, we could vary $B = [0, \infty)$ and $B = (-\infty, 0]$, or $D = [0, \infty)$ and $D = (-\infty, 0]$.

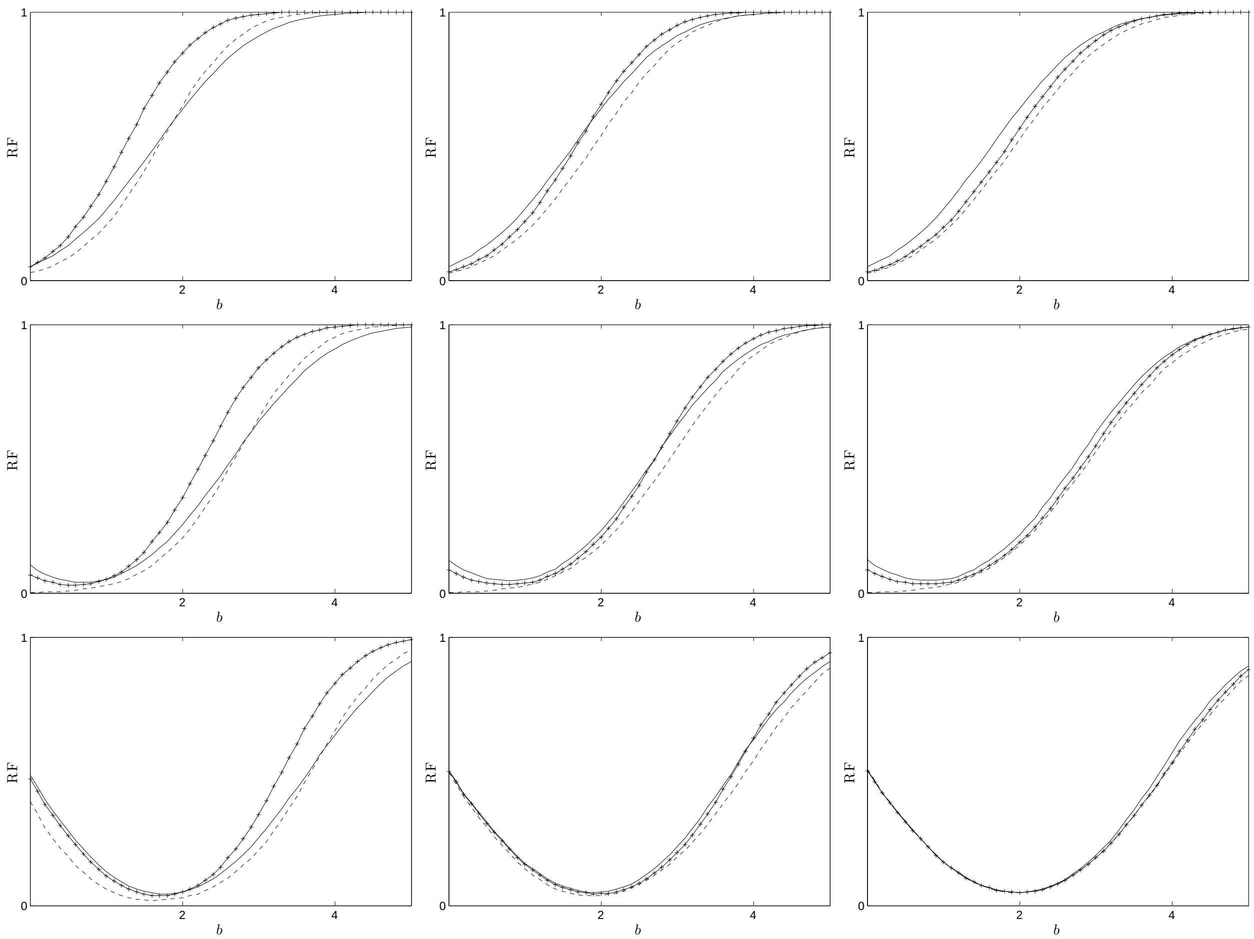

Figure 12: Power of the CLR (solid), gLR (solid with plusses), and the t_{ML}-test (dashed) for testing $b_0 = 0, 1, 2$ from top to bottom with $d = 0, 1, 2$ from left to right. $\rho = 0.7$.

4.4 Comments

The t_{ML} controls size in all three leading examples considered above. For the examples considered in Sections 4.1 and 4.2, this was already shown in Müller and Norets (2013) and Andrews and Guggenberger (2010). Size control of the t_{ML} for the example considered in Section 4.3 previously had not been shown. We conjecture that size control is obtained more generally, even for vector-valued Y_b, but we do not have a proof. The intuition is that the t_{ML} without the absolute value is asymptotically distributed as the projection of a normal random variable onto the parameter space. This amounts to a contraction or a reduction of the spread of the test-statistic compared to a normal distribution, at least when the average spread is considered, which is achieved by taking the absolute value.

For the example considered in Section 4.1, we find that the t_{ML} is outperformed by the CLR, providing a strong argument for the use of the CLR in applications. The MN constitutes another good alternative. However, its power curve crosses that of the $\mathrm{t_{ML}}$, such that no strict preference order can be established. For the example considered in Section 4.3, we find that the ECS and the EMW, and our tests offer power gains over the $\mathrm{t_{ML}}$. The EMW is associated with high computational cost, making it unattractive for empirical work. The gLR, on the other hand, which in this example performs equally well, is computationally very cheap and constitutes an attractive choice. The CLR and the ECS perform comparably, which is not surprising, since both test maximize WAP subject to a similarity constraint. In Section 4.3, we again find power advantages of the CLR and, in particular, the gLR over the $\mathrm{t_{ML}}$.

In summary, the $\mathrm{t_{ML}}$ performs poor in comparison, and from our analysis it is evident that alternative testing procedures offer valuable improvements. When the dimension of the nuisance parameter is small, the gLR performs very well, is easy to implement, and can be recommended for applied work. As the dimension of the nuisance parameter increases, the gLR looses its appeal, as illustrated in Section 4.5 below. The CLR, on the other hand, remains attractive. Furthermore, it is not associated with high computational cost as the EMW or the complexity of determining the test statistic as is the case for the ECS. Therefore, we recommend its use when the number of nuisance parameter is large.

4.5 Extensions to $K > 1$

The least favorable configuration, d_{LFC}, for the gLR in Sections 4.2 and 4.3 was determined by simulation and a numerical search over the parameter space. Further simulations studies, not reported here, indicate that the above pattern generalizes to higher dimensional problems, where d is vector-valued. In particular, when $B^{\infty} = (-\infty, \infty)$, d_{LFC} is given by the zero vector. When $B^{\infty} = [0, \infty)$ and $D_k = [0, \infty)$ for $k = 1, \ldots, K$, then the k^{th} entry of d_{LFC} equals 0 if $\Sigma_{\delta\beta,k} > 0$ and ∞ if $\Sigma_{\delta\beta,k} < 0$, where $\Sigma_{\delta\beta,k}$ denotes the k^{th} entry of $\Sigma_{\delta\beta}$.[27]

Given the power curves in Sections 4.2 and 4.3, the gLR may seem preferable to the CLR, as mentioned above. However, the power advantages of the gLR over the CLR and, in particular, the $\mathrm{t_{ML}}$ diminish and can even be reversed in parts of the parameters space as the dimension of the nuisance parameter increases. This is a direct consequence of using the least favorable configuration approach.[28] When the true value of the nuisance parameter coincides with the least favorable configuration the gLR has considerably more power than the CLR and the $\mathrm{t_{ML}}$, cf. the top-left panel of Figure 12. However, when the true value is far

[27] Since the problem is symmetric, the least favorable configurations are reversed when $B^{\infty} = (-\infty, 0]$ or $D_k = (-\infty, 0]$.

[28] It is possible to implement testing procedures based on the gLR that do not rely on the least favorable configuration. For example, the approaches based on Bonferroni bounds proposed in McCloskey (2012) offer potential gains in power.

from the least favorable configuration, the gLR lacks power in comparison, cf. the top-right panel of Figure 12. In higher dimensional cases, the gLR can even considerably lack in power compared to the t_{ML}. This is illustrated in Figure 16 shown in Appendix B. The figure plots the power curves of the gLR and the t_{ML} for $B = (-\infty, \infty)$, $D = [0, \infty)^6$, and Σ equal to the correlation matrix for one particular test considered in the application below, see Section 5.[29] The nuisance parameter is varied between 0^6, 1^6, and 2^6, where e.g., 0^6 denotes a 6×1 dimensional vector of zeros.

4.6 Ignoring Y_d

The reason why we considered the t_{ML} above is that it represents common practice, when the construction of the two-sided t-test is based on a constrained extremum estimator and the true parameter vector is near or at the boundary. However, given the availability of the asymptotically normal estimator proposed in Section 3.2, it is, of course, possible to construct a two-sided t-test based on it. More generally, it is possible to construct tests that only use the information contained in Y_b. In this section, we analyze the performance of such tests in comparison to our proposed testing procedures that utilize the information contained in Y_d.

For the example in Section 4.2, where $B^\infty = (-\infty, \infty)$, an obvious candidate is the regular two-sided t-test, $\mathrm{t} = |y_b - b_0|$, whose rejection region is $\{y : \mathrm{t} > 1.96\}$. Figure 13 below plots the power curves of the regular two sided t-test for different values of the nuisance parameter d and $\rho = 0.7$. For ease of reference, the power curves of the gLR and the CLR as seen in Figure 9 are also shown.

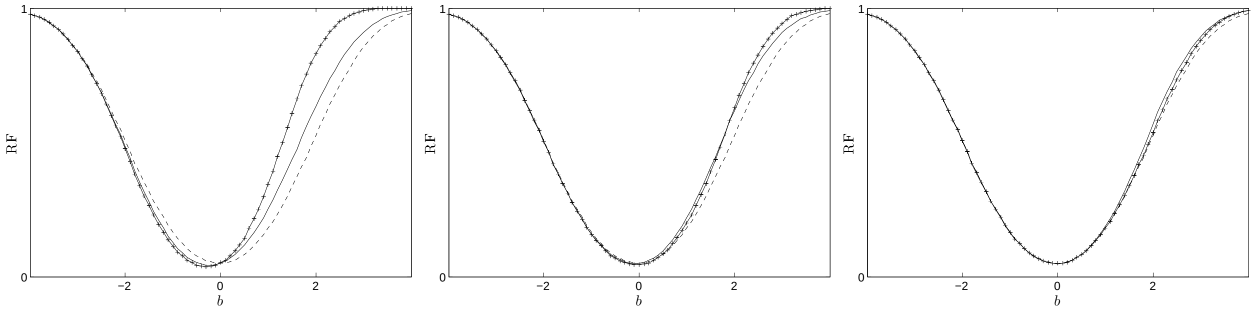

Figure 13: Power of the CLR (solid), gLR (solid with plusses), and the regular t-test (dashed) for testing $H_0 : b = 0$ with $d = 0, 1, 2$ from left to right. $\rho = 0.7$.

[29] The correlation matrix is taken from the test on the mean parameter of horse power per weight (Hp./We.). The motivation for this is to provide a possible explanation for why the confidence interval based on the gLR is wider than that based on the t_{ML}, see Table 3 in Appendix C.

Figure 13 shows that the regular two-sided t-test has greater power than the gLR and the CLR for negative values of b when d is small. But overall the power curves in Figure 13 seem favorable to the gLR and CLR.

For the example discussed in Section 4.3, where $B^{\infty} = [0, \infty)$, a partial CLR, which disregards Y_d, can be employed. Figures 17 and 18 given in Appendix B show the power curves of the partial CLR for $\rho = 0.7$ and $\rho = -0.7$ and, additionally, reproduce the power curves of the CLR as shown in Figures 12 and 15. Overall the power curves of the CLR look good in comparison. Here, it should be noted that the power gains of the CLR are more pronounced for a larger values of $|\rho|$ and, more importantly, in case of a higher dimensional nuisance parameter.

In summary, the information contained in Y_d is useful for constructing powerful tests, and we propose to exploit it by means of the gLR and the CLR.

5 Application

Berry, Levinsohn, and Pakes (1995), BLP hereafter, introduced the random coefficients logit model, which is now widely used in the industrial organization and marketing literatures to model demand for differentiated products. The random coefficients in this model interact with the characteristics of the products, allowing for heterogeneity in consumer preferences. Typically, the random coefficients are assumed to be independently normally distributed. Then, the model parameters are given by a vector of means and a vector of variances. Since variances are naturally bounded below by zero, the model readily fits our theoretical framework. There is good reason to believe that not all variance parameters are equal to zero, because in that case the model reduces to the simple multinomial logit model, which is known to suffer from the Independence of Irrelevant Alternatives (IIA). However, estimates of the variance parameters are often found to be close to the boundary (see e.g., Nevo, 2001; Goeree, 2008). Furthermore, applied researchers seldom know a priori which of the product characteristics interact with a random coefficient (positive variance) and which do not (zero variance). Therefore, it is reasonable to assume that the empirical analysis often starts with a baseline model that allows for a random coefficient on all product characteristics. In that case, powerful testing procedures as the ones proposed in this paper are deemed useful in model selection exercises.[30]

Before turning to the empirical application using data on the European car market, we

[30] For example, Nevo (2001) in his analysis of the demand for ready-to-eat cereals performs a robustness check of his main results by setting some of the variance parameters equal to zero, illustrating that the correct model specification is of concern in applied work.

present a small Monte Carlo study to investigate the finite-sample behavior of the proposed testing procedures when they are implemented using the estimator proposed in Section 3.2.

5.1 Monte Carlo

In what follows, we describe the data generating process and, at the same time, introduce the model in more detail. For ease of reference, we use the same data generating process as Reynaert and Verboven (2014), which satisfies Assumptions 1-4 (Ketz, 2014). The BLP model defines the demand for product $j \in 1, \ldots, J$ in market $t = 1, \ldots, T$ as a function of the product characteristics of all products in that market. Denote these product characteristics by $x_t = (x_{1t}, \ldots, x_{Jt})$ and $\xi_t = (\xi_{1t}, \ldots, \xi_{Jt})$. They differ in that x_t is assumed to be observed by the consumers and the econometrician, while ξ_t is assumed to be observed by the consumers only. x_{jt} and ξ_{jt} are vector-valued and scalar, respectively. Here, we consider three product characteristics in x_{jt}, $x_{jt,1}$ through $x_{jt,3}$. However, we only allow for a random coefficient on $x_{jt,3}$. The mean parameters, μ_1 through μ_3, are unrestricted, while the variance parameter, σ^2, is restricted to be nonnegative. Let $\phi(\cdot, \mu, \sigma^2)$ denote the pdf of a normal random variable with mean μ and variance σ^2.[31] Then, the market share for product j in market t is given by

$$\mathrm{s}_j(\mu_1, \mu_2, \mu_3, \sigma^2, x_t, \xi_t) = \int_{-\infty}^{\infty} \frac{e^{\mu_1 x_{jt,1} + \mu_2 x_{jt,1} + v x_{jt,3} + \xi_{jt}}}{1 + \sum_{l=1}^{J} e^{\mu_1 x_{lt,1} + \mu_2 x_{lt,1} + v x_{lt,3} + \xi_{lt}}} \phi(v, \mu_3, \sigma^2) dv. \qquad (19)$$

Equating model implied market shares given in (19) with market shares observed in the data, the model parameters are estimated by GMM relying on a zero moment condition which interacts the error term, ξ_{jt}, with a set of instruments. For more details on the estimation procedure, see e.g., Nevo (2000). We follow Reynaert and Verboven (2014) and implement an approximation to the optimal instruments. Furthermore, we employ the efficient weighting matrix in the construction of the GMM objective function.

We model $x_{jt,1}$ to be endogenous. In particular, $x_{jt,1}$ is generated as

$$x_{jt,1} = w_{jt}' \pi_1 + z_{jt}' \pi_2 + \zeta_{jt},$$

where $w_{jt} = (x_{jt,2}, x_{jt,3})$ denotes the exogenous product characteristics and z_{jt} is 3×1 dimensional vector of instruments. The instruments can be thought of as cost shifters if we think of $x_{jt,1}$ as price, which under the assumption of perfect competition equals marginal cost. The endogeneity of $x_{jt,1}$ arises, because the error terms, ξ_{jt} and ζ_{jt}, are generated

[31] $\phi(\cdot, \mu, 0)$ is defined as $\mathbb{1}(v = \mu)$, where $\mathbb{1}(\cdot)$ denotes the indicator function.

according to

$$\begin{pmatrix} \xi_{jt} \\ \zeta_{jt} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \right).$$

The two exogenous product characteristics, $x_{jt,2}$ and $x_{jt,3}$, are given by 1 and $U[1,2]$, respectively, where $U[a,b]$ denotes a uniform random variable with support on $[a,b]$. z_{jt} is generated as a vector of independent $U[0,1]$. The true parameter values are chosen as $\pi_1 = (0.7, 0.7)$ and $\pi_2 = (3,3,3)$, while we vary the true parameter values of μ_1, μ_2, μ_3, and σ^2 in order to investigate size and power for testing $H_0 : \mu_1 = -2$, $H_0 : \mu_2 = 2$, $H_0 : \mu_3 = 2$, and $H_0 : \sigma^2 = 0$. We choose $T = 25$ and $J = 10$ totaling 250 products over all markets.

Table 1 below reports the rejection frequencies of the gLR, the CLR, and the $\mathrm{t_{ML}}$ over 1000 Monte Carlo replications. Since the market share given in (19) and, thus, the GMM objective function cannot be evaluated at negative values of σ^2, the estimator introduced in Section 3.2 needs to be employed in order to construct the gLR and the CLR.

Table 1: Rejection Frequencies

Alternatives				$H_0 : \mu_1 = -2$			$H_0 : \mu_2 = 2$			$H_0 : \mu_3 = 2$			$H_0 : \sigma^2 = 0$	
μ_1	μ_2	μ_3	σ^2	$\mathrm{t_{ML}}$	gLR	CLR	$\mathrm{t_{ML}}$	gLR	CLR	$\mathrm{t_{ML}}$	gLR	CLR	$\mathrm{t_{ML}}$	CLR
-2	2	2	0	0.035	0.035	0.037	0.035	0.057	0.041	0.032	0.053	0.049	0.023	0.049
-2	2	2	0.25	0.037	0.034	0.038	0.042	0.052	0.049	0.038	0.040	0.045	0.264	0.391
-2	2	1.5	0	0.035	0.032	0.037	0.030	0.058	0.043	0.324	0.651	0.383	0.029	0.054
-2	2	2.5	0	0.036	0.033	0.038	0.046	0.059	0.049	0.393	0.418	0.442	0.020	0.045

The first row of Table 1 shows that all tests control size, which is line with expectations as the underlying estimator employs the efficient weighting matrix. The second row displays power advantages of the CLR over the $\mathrm{t_{ML}}$ for testing $H_0 : \sigma^2 = 0$. This is not surprising, since we are comparing a one-sided test to a two-sided test, recall Figure 7. The last two rows show power advantages of the gLR and the CLR over the $\mathrm{t_{ML}}$ for testing $H_0 : \mu_3 = 2$, in accordance with the local asymptotic power curves given in Figure 9.

5.2 Empirical Results

We now turn to the empirical application using data from Reynaert and Verboven (2014), RV hereafter. RV estimate the demand for cars in several European countries spanning the years from 1998 to 2010. The product characteristics are price divided by income (Price/Inc.), horse power per weight (Hp/We.), a dummy variable indicating whether the car brand is foreign (Foreign), size (Size) obtained as length times width, height (Height), and fuel efficiency (€/km) given by price in € per kilometer. See RV for more details on the dataset and its construction.

The first two columns of Table 2 below show the estimates and the corresponding standard error estimates that RV obtain for the baseline specification that allows for a random coefficient on each product characteristic (using an approximation to the optimal instruments), see the last two columns of their Table 6.[32] As the estimates of the variance parameters (in the bottom half of the Table) are all in the interior of the parameter space, the same estimates are obtained for the modified estimator proposed in Section 3.2, see Remark 1, and the construction of the gLR and the CLR can be based on them.[33] The 95% and the 90% confidence intervals based on the CLR and the t_{ML} are given in the last eight columns of Table 2.[34] Due to the high dimension of the nuisance parameter, the CLR as argued in Section 4.5 might be preferred to the gLR. For comparison, Table 3 in Appendix C replicates Table 2 with the confidence intervals constructed using the gLR rather than the CLR. Indeed, for the application at hand, the confidence intervals based on the gLR do not offer a considerable improvement over those based on the t_{ML}.

Table 2: Table 6 in RV - Optimal instruments (ii)

			95% CI				90% CI			
	Est.	Std. Err.	CLR		t_{ML}		CLR		t_{ML}	
			Mean valuations μ							
Price/Inc.	-2.322	0.498	-3.317	-1.385	-3.297	-1.347	-3.151	-1.534	-3.141	-1.504
Hp/We.	-0.918	1.192	-3.212	1.108	-3.253	1.418	-2.854	0.780	-2.878	1.043
Foreign	-0.858	0.216	-1.212	-0.422	-1.275	-0.430	-1.129	-0.494	-1.207	-0.498
Size	0.667	0.674	-0.650	2.014	-0.654	1.987	-0.445	1.788	-0.441	1.775
Height	0.183	0.052	0.088	0.288	0.080	0.285	0.101	0.270	0.097	0.269
€/km	-3.972	1.344	-6.351	-1.412	-6.606	-1.338	-5.968	-1.849	-6.183	-1.761
			Variances σ^2							
Price/Inc.	0.274	0.176	0.000	0.626	0.000	0.619	0.048	0.567	0.000	0.564
Hp/We.	10.252	4.346	2.885	18.944	1.733	18.771	3.820	17.488	3.103	17.401
Foreign	0.515	0.736	0.000	1.987	0.000	1.958	0.000	1.741	0.000	1.726
Size	0.057	0.189	0.000	0.432	0.000	0.427	0.000	0.371	0.000	0.367
Height	0.011	0.006	0.000	0.022	0.000	0.023	0.002	0.020	0.000	0.021
€/km	4.424	19.835	0.000	44.092	0.000	43.299	0.000	37.445	0.000	37.049

The bottom half of Table 2 shows that the CLR and the t_{ML} agree at the 95% significance level with respect to the significance of the variance parameters. Both indicate that only

[32] In fact, Table 6 in RV report estimates and corresponding standard error estimates for the standard deviation. In Ketz (2014), I show that this approach to inference is statistically invalid, as the Jacobian of the moment equation is of reduced rank at the boundary of the parameter space, i.e., when one or more standard deviations are equal to zero. A reparameterization in terms of variances solves the problem of a reduced rank Jacobian and is therefore employed here.

[33] The implementation of the gLR and the CLR requires an estimate of the variance matrix of the estimator, which is not reported in RV. I thank Mathias Reynaert and Frank Verboven for sharing their estimate of the variance matrix with me.

[34] Given the discussion following equation (13), the confidence intervals based on the t_{ML} reported in Table 2 are valid.

the variance on Hp/We. is significantly different from zero. At the 90% significance level, however, the CLR also rejects the null of a zero variance for Price/Inc. and Height, while the $\mathrm{t_{ML}}$ continues to reject that null only for Hp/We.. Put differently, the CLR allows us to detect the presence of (additional) heterogeneity in consumer preferences, which is not picked up by the "benchmark".

In addition, as can be seen from the top half of Table 2, the CLR can also yield tighter confidence intervals for the mean parameters. For example the 95% confidence interval for the mean parameter of HP/We. is around 8% shorter when based on the CLR. In other applications, such a difference can be enough to infer the sign of a coefficient.

6 Conclusion

In general extremum problems, there is a lack of inference procedures for testing hypotheses about subsets of parameters when the true parameter vector is near or at the boundary. The two-sided t-test based on a constrained estimator is inadequate and suffers from size distortions. Recently, alternative inference methods, such as the m-out-of-n bootstrap, have been used in an attempt to account for boundary effects on the asymptotic distribution of the estimator despite, as acknowledged by the authors, being theoretically invalid (Abrevaya and Shen, 2014). We fill the gap in the literature by introducing a novel extremum estimator that is asymptotically normally distributed, even when the objective function is not defined outside the parameter space. The estimator is available in a wide range of models, allowing the implementation of powerful testing procedures. When there is only one parameter that may be near or at the boundary, we suggest the use of the gLR. When there is multiple, we propose the use of the CLR, which we show to possess attractive optimality properties, while enjoying low computational cost compared to alternative testing procedures. As illustrated in the application of the random coefficients logit model to the European car market, conclusions concerning the model specification based on the CLR can differ from those based on the two-sided t-test.

In standard settings, where the true parameter vector is in the interior of the parameter space, point estimates and corresponding standard error estimates have the same information content as confidence intervals. In the nonstandard setting considered in this paper, the commonly reported standard error estimate looses its interpretation when a constrained estimator is employed, because it no longer measures the spread of the (asymptotic) distribution of the estimator. Confidence intervals, on the other hand, obtained by inverting the CLR can be interpreted as a natural extension of standard confidence intervals and continue to provide information that is useful and easy to interpret.

Next to allowing the implementation of tests concerning scalar parameters or, more generally, subsets of parameters, the asymptotically normal extremum estimator introduced in this paper can also be used to construct valid confidence sets for nonlinear functions of the parameter vector using the results in Woutersen and Ham (2013). For more details on this in the context of the random coefficients logit model, see Ketz (2014).

The assumptions on the parameter space we make in this paper are restrictive, and while they are satisfied in many models of interest they do not cover random coefficient models where the coefficients are allowed to be correlated (Andrews, 2001). In those models, a further complication arises as covariance parameters are not identified if one of the variance parameters is equal to zero (Andrews and Cheng, 2012a). We leave possible extensions of our results to these more general settings for future research.

References

Abrevaya, J. and S. Shen (2014). Estimation of censored panel-data models with slope heterogeneity. *Journal of Applied Econometrics 29*, 523–548.

Andrews, D. W. (1992). Generic uniform convergence. *Econometric Theory 8*, 241–257.

Andrews, D. W. (1997). Estimation When a Parameter Is on a Boundary: Theory and Applications. Cowles Foundation Discussion Paper No. 1153.

Andrews, D. W. (1999). Estimation when a parameter is on a boundary. *Econometrica 67*, 1341–1383.

Andrews, D. W. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica 69*, 683–734.

Andrews, D. W. (2012). Similar-on-the-boundary tests for moment inequalities exist, but have poor power. Working paper.

Andrews, D. W. and X. Cheng (2011). Maximum Likelihood Estimation and Uniform Inference With Sporadic Identification Failure. Cowles Foundation Discussion Paper No. 1824.

Andrews, D. W. and X. Cheng (2012a). Estimation and Inference With Weak, Semi-Strong, and Strong Identification. *Econometrica 80*, 2153–2211.

Andrews, D. W. and X. Cheng (2012b). Supplement to 'Estimation and Inference With Weak, Semi-Strong, and Strong Identification'. *Econometrica Supplemental Material 80.*

Andrews, D. W. and P. Guggenberger (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory 26*, 426.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*, 841–890.

Elliott, G., U. K. Müller, and M. W. Watson (2013). Nearly optimal tests when a nuisance parameter is present under the null hypothesis. Working paper.

Feldman, G. J. and R. D. Cousins (1998). Unified approach to the classical statistical analysis of small signals. *Physical Review D 57*, 3873–3889.

Goeree, M. S. (2008). Limited information and advertising in the US personal computer industry. *Econometrica 76*, 1017–1074.

Hausman, J. A. and D. A. Wise (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica 46*, 403–426.

Ketz, P. (2014). Asymptotic theory for the random coefficients logit model. Working Paper.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica 73*, 1103–1123.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses.* Springer.

Matthes, T. K. and D. R. Truax (1967). Tests of composite hypotheses for the multivariate exponential family. *The Annals of Mathematical Statistics 38*, 681–697.

McCloskey, A. (2012). Bonferroni-based size-correction for nonstandard testing problems. Working paper.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*, 995–1026.

Montiel Olea, J. L. (2013). Efficient conditionally similar tests: Finite-sample theory and large-sample applications. Working Paper.

Moreira, M. (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 A. Norets (2013). Credibility of confidence sets in nonstandard econometric problems. Working Paper.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy 9*, 513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica 69*, 307–342.

Reynaert, M. and F. Verboven (2014). Improving the performance of random coefficients demand models: The role of optimal instruments. *Journal of Econometrics 179*, 83–98.

Van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press.

Woutersen, T. and J. C. Ham (2013). Confidence Sets for Continuous and Discontinuous Functions of Parameters. Working Paper.

A Proofs

A.1 Proof of Theorem 1

Proof of Theorem 1. As described in the main text, the proof proceeds in two steps. 1. I show that any similar test that has convex acceptance sections is admissible in the class of all tests pertaining to the testing problem at hand. 2. I show that the CLR, which is similar by construction, has convex acceptance sections.

1. Let Φ_c denote the class of similar tests with convex acceptance sections. Then we have to show that any test in Φ_c is admissible. Let φ be any test in Φ_c. Assume it is not admissible. Then there exists a test φ' that dominates φ. Below, I show that a test that dominates φ needs to satisfy the following two equations a.e.:[35]

$$\int_{\mathcal{Y}_b} (\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b_0) dy_b = 0 \tag{20}$$

and

$$\int_{\mathcal{Y}_b} y_b(\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b_0) dy_b = 0, \tag{21}$$

[35] a.e. is short for almost everywhere and means that the two equations have to hold for all $x \in \mathcal{X}$ except for x of Lebesque measure zero.

where $f(y_b|b) = f(y_b|x, b, d)$ is the pdf of a standard normal when the true parameter value is b.[36] Equation (20) implies that φ' needs to have the same size as φ conditional on x. Equation (21) implies that φ' needs to have the same "center of gravity" as φ conditional on x, i.e., the same conditional expected value (over the rejection region). Since Φ_c is complete, which follows from Theorem 3.1 in Matthes and Truax (1967), we can assume without loss of generality that $\varphi' \in \Phi_c$, i.e., φ' has convex acceptance sections.[37] Equations (20) and (21) together with the fact that the acceptance sections of φ' are convex implies $\varphi' = \varphi$, because the size and the expected value of a test together with the convexity of the acceptance region uniquely define a test. This argument is also presented in Section 4 of Matthes and Truax (1967).

In order to show that φ' needs to satisfy equations (20) and (21), note that by the definition of dominance we have

$$\begin{aligned}&\int_{\mathcal{Y}_b \times \mathcal{X}} (\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b) f(x|b, d) dy_b dx \geq 0 \\ \Leftrightarrow &\int_{\mathcal{X}} \left[\int_{\mathcal{Y}_b} (\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b) dy_b \right] f(x|b, d) dx \geq 0. \qquad (22)\end{aligned}$$

Here, $f(x|b, d)$ denotes the pdf of a multivariate normal, whose mean is given by b and d. Equation (22) holds with equality when $b = b_0$, because φ is similar and by continuity of the power function in b cannot be dominated by a non-similar test.[38] Thus, at $b = b_0$ equation (22) reads

$$\int_{\mathcal{X}} \left[\int_{\mathcal{Y}_b} (\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b_0) dy_b \right] f(x|b_0, d) dx = 0.$$

[36] $f(y_b|b)$ equals $f(y_b|x, b, d)$, since x is a sufficient statistic for d, and y_b and x are independent.

[37] The proof of Theorem 3.1 in Matthes and Truax (1967) goes through if Ω_1 (using their notation) is replaced with a strict, convex subset of $\mathbb{R}^{K+1}$ such as M^∞.

[38] The continuity of the power function holds for all members of the exponential family, see e.g., Theorem 2.7.1 in Lehmann and Romano (2005).

By completeness, we obtain[39]

$$\int_{\mathcal{Y}_b} (\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b_0) dy_b = 0. \tag{20}$$

Given equation (22), we know that the derivative of the left hand side of equation (22) must equal zero at $b = b_0$. The derivative is given by

$$\int_{\mathcal{X}} \left[\int_{\mathcal{Y}_b} (y_b - b)(\varphi'(y_b, x) - \varphi(y_b, x)) f(y|b) dy_b \right] f(x|b, d) dx +$$
$$\int_{\mathcal{X}} \left[\int_{\mathcal{Y}_b} (\varphi'(y_b, x) - \varphi(y_b, x)) f(y_b|b) dy_b \right] f'(x|b, d) f(x|b, d) dx = 0,$$

where $f'(x|b, d)$ denotes the derivative of $f(x|b, d)$ with respect to b. Note that $(y_b - b)$ is the derivative of $f(x|b)$ with respect to b. Evaluating the above equation at $b = b_0$, we obtain

$$\int_{\mathcal{X}} \left[\int_{\mathcal{Y}_b} (y_b - b_0)(\varphi'(y_b, x) - \varphi(y_b, x)) f(y|b_0) dy_b \right] f(x|b_0, d) dx = 0,$$

because the second term is zero given equation (20). Writing the obtained equation as the sum of two terms, we see that the second term is again zero given equation (20). We obtain

$$\int_{\mathcal{X}} \left[\int_{\mathcal{Y}_b} y_b(\varphi'(y_b, x) - \varphi(y_b, x)) f(y|b_0) dy_b \right] f(x|b_0, d) dx = 0.$$

Again, by completeness

$$\int_{\mathcal{Y}_b} y_b(\varphi'(y_b, x) - \varphi(y_b, x) f(y|b_0) dy_b = 0. \tag{21}$$

2. Assume without loss of generality that Σ is a correlation matrix. In order to show that the CLR has convex acceptance sections it suffices to show that the CLR is convex as

[39] A (parametric) family of distributions $f(x|\theta)$ is said to be *complete* if

$$\int_{\mathcal{X}} h(x) f(x|\theta) dx = 0 \qquad \forall \theta \in \Theta$$

implies

$$h(x) = 0 \qquad \text{a.e.}$$

In the case on hand, $\theta = \mu$ and $\Theta = M^{\infty}$.

a function of y_b. Define

$$FP(y_b, \mathrm{d}) = \begin{pmatrix} y_b - b_0 \\ x + \Sigma_{\delta\beta} y_b - \mathrm{d} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - b_0 \\ x + \Sigma_{\delta\beta} y_b - \mathrm{d} \end{pmatrix}$$

and

$$SP(y_b, \mathrm{b}, \mathrm{d}) = \begin{pmatrix} y_b - \mathrm{b} \\ x + \Sigma_{\delta\beta} y_b - \mathrm{d} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - \mathrm{b} \\ x + \Sigma_{\delta\beta} y_b - \mathrm{d} \end{pmatrix}.$$

Let $\mathrm{d}^*(y_b) = \arg\min_{\mathrm{d} \in D^\infty} FP(y_b, \mathrm{d})$ and $(\mathrm{b}^{**}(y_b), \mathrm{d}^{**}(y_b)) = \arg\min_{\mathrm{b} \in B^\infty, \mathrm{d} \in D^\infty} SP(y_b, \mathrm{b}, \mathrm{d})$ such that $\mathrm{CLR}(b_0, y_b, x) = FP(y_b, \mathrm{d}^*(y_b)) - SP(y_b, \mathrm{b}^{**}(y_b), \mathrm{d}^{**}(y_b))$. Note that $\mathrm{d}^*(y_b)$ does not vary with y_b such that $FP(y_b, \mathrm{d}^*(y_b))$ is a quadratic function of y_b. To see this consider the first order derivative:

$$\frac{\partial FP(y_b, \mathrm{d}^*(y_b))}{\partial y_b} = 2 \begin{pmatrix} 1 \\ \Sigma_{\delta\beta} \end{pmatrix}' \Sigma^{-1} \begin{pmatrix} y_b - b_0 \\ x + \Sigma_{\delta\beta} y_b - \mathrm{d}^*(y_b) \end{pmatrix}.$$

Using

$$\begin{pmatrix} 1 \\ \Sigma_{\delta\beta} \end{pmatrix}' \Sigma^{-1} = (1\ 0 \ldots\ 0),$$

which follows from the formula for the inverse of a partitioned matrix, the first order condition reads

$$2(y_b - b_0). \tag{23}$$

Claim 1: If $y_b'' \geq y_b'$, then $\mathrm{b}^{**}(y_b'') \geq \mathrm{b}^{**}(y_b')$. Claim 2: There exists a $\tilde{y}_b$ such that $\mathrm{CLR}(b_0, \tilde{y}_b, x) = 0$. Given these two claims, we show that for any $y_b \neq \tilde{y}_b$, the difference between $FP(y_b, \mathrm{d}^*(y_b))$ and $SP(y_b, \mathrm{b}^{**}(y_b), \mathrm{d}^{**}(y_b))$ is strictly increasing in $|y_b - \tilde{y}_b|$ if it is non-zero. Then, it immediately follows that CLR is convex. Take $y_b' > \tilde{y}_b$ without loss of generality. Note that $\mathrm{CLR}(b_0, y_b', x) \geq 0$ by definition. Now, imagine fixing $\mathrm{b}^{**}(y_b')$ and $\mathrm{d}^{**}(y_b')$ and moving towards $y_b'' \geq y_b'$, i.e., consider a modification of the CLR where the second part is given by $SP(y_b, \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b'))$. Now, by an argument similar to the one leading to equation (23), the derivative of $SP(y_b, \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b'))$ is given by

$$2(y_b - \mathrm{b}^{**}(y_b')).$$

If $\mathrm{b}^{**}(y_b') > b_0$, the derivative of $SP(y_b, \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b'))$ is smaller than the derivative of $FP(y_b, \mathrm{d}^*(y_b))$. It follows that

$$FP(y_b'', \mathrm{d}^*(y_b)) - SP(y_b'', \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b')) > FP(y_b', \mathrm{d}^*(y_b)) - SP(y_b', \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b')).$$

Since, by definition $SP(y_b'', \mathrm{b}^{**}(y_b''), \mathrm{d}^{**}(y_b'')) < SP(y_b'', \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b'))$,

$$FP(y_b'', \mathrm{d}^*(y_b)) - SP(y_b'', \mathrm{b}^{**}(y_b''), \mathrm{d}^{**}(y_b'')) > FP(y_b', \mathrm{d}^*(y_b)) - SP(y_b', \mathrm{b}^{**}(y_b'), \mathrm{d}^{**}(y_b')).$$

Claim 1 follows because $SP(y_b, \mathrm{b}, \mathrm{d})$ is quadratic. And Claim 2 follows from Claim 1 and continuity of $SP(y_b, \mathrm{b}^{**}(y_b), \mathrm{d}^{**}(y_b))$.

□

A.2 Proof of Proposition 1

Before stating the proof of Proposition 1, we introduce several Lemmas, which are used in the proof of Proposition 1.

Lemma 1. *Assumption 2 implies that for all constants $\epsilon_n \to 0$,*

$$\sup_{\theta \in \Theta : \|\theta - \theta_n\| \leq \epsilon_n} \frac{|nR_n(\theta)|}{(1 + \|\sqrt{n}(\theta - \theta_n)\|)^2} = o_p(1),$$

under all $\{\gamma_n\} \in \Gamma(\gamma^*)$.

This result corresponds to Lemma 1 in Andrews (1999) (A1).

Proof of Lemma 1. By Theorem 6 in A1,

$$R_n(\theta) = \frac{1}{2}(\theta - \theta_n)' \left(\frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta_n) \right) (\theta - \theta_n),$$

where $\theta^\dagger$ lies on the line segment between θ and θ_n, when $\theta \neq \theta_n$ and $R_n(\theta) = 0$ when $\theta = \theta_n$. Thus,

$$\begin{aligned}
&\sup_{\theta \in \Theta : \|\theta - \theta_n\| \leq \epsilon_n} \frac{|nR_n(\theta)|}{(1 + \|\sqrt{n}(\theta - \theta_n)\|)^2} \\
\leq &\sup_{\theta \in \Theta : \|\theta - \theta_n\| \leq \epsilon_n} \frac{1}{2} \frac{\left| \sqrt{n}(\theta - \theta_n)' \left(\frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta_n) \right) \sqrt{n}(\theta - \theta_n) \right|}{\|\sqrt{n}(\theta - \theta_n)\|^2} \\
\leq &\sup_{\theta \in \Theta : \|\theta - \theta_n\| \leq \epsilon_n} \frac{1}{2} \left\| \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta_n) \right\| \\
&= o_p(1).
\end{aligned}$$

□

Given Lemma 1, it is easy to see that Assumption 2 also implies

$$\sup_{\theta\in\Theta:\|\sqrt{n}(\theta-\theta_n)\|\leq\epsilon} |nR_n(\theta)| = o_p(1) \tag{24}$$

under all $\{\gamma_n\} \in \Gamma(\gamma^*)$, for all $0 < \epsilon < \infty$.

Lemma 2. *Under Assumptions 1 - 4 and under $\{\gamma_n\} \in \Gamma(\gamma^*)$, $\sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1)$.*

This result is equivalent to Theorem 1 in A1, and its proof is given by a slight modification of the corresponding proof.

Proof of Lemma 2. Let $\kappa_n = J_n^{1/2}\sqrt{n}(\hat{\theta}_n - \theta_n)$. θ_n is in the closure of Θ by Assumption 1. We have

$$\begin{aligned} o_p(1) &= n\left(Q_n(\hat{\theta}_n) - \inf_{\theta\in\Theta} Q_n(\theta)\right) \\ &\geq n\left(Q_n(\hat{\theta}_n) - Q_n(\theta_n)\right) \\ &= \sqrt{n}DQ_n(\theta_n)'J_n^{-1/2}\kappa_n + \frac{1}{2}\|\kappa_n\|^2 + nR_n(\hat{\theta}_n) \\ &= O_p(\|\kappa_n\|) + \frac{1}{2}\|\kappa_n\|^2 + (1 + \|J_n^{-1/2}\kappa_n\|)^2 o_p(1) \\ &= O_p(\|\kappa_n\|) + \frac{1}{2}\|\kappa_n\|^2 + o_p(\|\kappa_n\|) + o_p(\|\kappa_n\|^2) + o_p(1). \end{aligned}$$

The first equality holds by definition of $\hat{\theta}_n$, see equation (5). The inequality then follows by the definition of the infimum. The next equality follows from equation (9). The following equality holds by Assumption 3 and 4 and Lemma 1 combined with Assumption 1. Upon rearrangement, after dropping $o_p(\|\kappa_n\|)$ and $o_p(\|\kappa_n\|^2)$, we get

$$\|\kappa_n\|^2 \leq 2\|\kappa_n\|O_p(1) + o_p(1).$$

Let ξ_n denote the $O_p(1)$ term. Then,

$$(\|\kappa_n\| - \xi_n)^2 \leq \xi_n^2 + o_p(1) = O_p(1).$$

Taking square roots gives $\|\kappa_n\| \leq O_p(1)$. Given Assumption 3, this establishes the result. □

Lemma 3. *Under Assumptions 1 - 4 and under $\{\gamma_n\} \in \Gamma(\gamma^*)$, $q_T(\sqrt{n}(\hat{\theta}_n-\theta_n)) = \inf_{\theta\in\Theta} q_T(\sqrt{n}(\theta-\theta_n)) + o_p(1)$.*

This result is equivalent to Theorem 2(e) in A1. Note that

$$\inf_{\theta\in\Theta} q_n(\sqrt{n}(\theta-\theta_n)) = \inf_{\lambda\in\sqrt{n}(\Theta-\theta_n)} q_n(\lambda),$$

where $\sqrt{n}(\Theta-\theta_n) = \{\lambda \in \mathbb{R}^{K+L+1} : \lambda = \sqrt{n}(\theta-\theta_n) \text{ for some } \theta\in\Theta\}$.

Proof of Lemma 3. Let $\hat{\theta}_n^q$ satisfy $\hat{\theta}_n^q \in cl(\Theta)$ and

$$q_n(\sqrt{n}(\hat{\theta}_n^q-\theta_n)) = \inf_{\theta\in\Theta} q_n(\sqrt{n}(\theta-\theta_n)) + o_p(1). \tag{25}$$

With this definition in place, it is sufficient to show that

$$q_n(\sqrt{n}(\hat{\theta}_n-\theta_n)) = q_n(\sqrt{n}(\hat{\theta}_n^q-\theta_n)) + o_p(1). \tag{26}$$

First, we show that

$$\sqrt{n}(\hat{\theta}_n^q-\theta_n) = O_p(1). \tag{27}$$

Let $\kappa_n^q = J_n^{1/2}\sqrt{n}(\hat{\theta}_n^q-\theta_n)$. We have

$$\left\|\kappa_n^q - J_n^{1/2}Z_n\right\|^2 = q_n(\sqrt{n}(\hat{\theta}_n^q-\theta_n)) \leq q_n(0)+o_p(1) = Z_n'J_nZ_n + o_p(1) = O_p(1),$$

where the inequality follows from equation (25) and the last equality follows from Assumptions 3 and 4. It follows that $\kappa_n^q = J_n^{1/2}Z_n + O_p(1) = O_p(1)$, where the last equality follows from Assumptions 3 and 4 again. Then, by Assumption 3, equation (27) follows.

Then, by equation (11), equation (24), Lemma 2, and equation (27), we have

$$nQ_n(\hat{\theta}_n) = nQ_n(\theta_n) + \frac{1}{2}Z_n'J_nZ_n + \frac{1}{2}q_n(\sqrt{n}(\hat{\theta}_n-\theta_n)) + o_p(1) \tag{28}$$

and

$$nQ_n(\hat{\theta}_n^q) = nQ_n(\theta_n) + \frac{1}{2}Z_n'J_nZ_n + \frac{1}{2}q_n(\sqrt{n}(\hat{\theta}_n^q-\theta_n)) + o_p(1). \tag{29}$$

The result of the Lemma then follows, since

$$\begin{aligned}
o_p(1) = n\left(Q_n(\hat{\theta}_n) - \inf_{\theta\in\Theta} Q_n(\theta)\right) &\geq n\left(Q_n(\hat{\theta}_n) - Q_n(\hat{\theta}_n^q)\right)\\
&= \frac{1}{2}q_n(\sqrt{n}(\hat{\theta}_n-\theta_n)) - \frac{1}{2}q_n(\sqrt{n}(\hat{\theta}_n^q-\theta_n)) + o_p(1)\\
&\geq \frac{1}{2}\inf_{\theta\in\Theta} q_n(\sqrt{n}(\theta-\theta_n)) - \frac{1}{2}q_n(\sqrt{n}(\hat{\theta}_n^q-\theta_n)) + o_p(1) = o_p(1),
\end{aligned}$$

where the first and last equality follow from equations (5) and (25), respectively. The inequalities follow from the definition of the infimum, while the equality in the middle follows from equations (28) and (29). □

Lemma 4. *If* $\sqrt{n}(\Theta - \theta_n) \to \Lambda$ *as* $n \to \infty$*, where* Λ *is a convex cone with possible nonzero vertex, then* $\inf_{\lambda \in \sqrt{n}(\Theta-\theta_n)} q_n(\lambda) = \inf_{\lambda \in \Lambda} q_n(\lambda) + o_p(1)$.

The proof of Lemma 4 closely resembles the proof of Lemma 2 in A1. For the sake of completeness, we give it here.

Proof of Lemma 4. For any set $\Delta \subset \mathbb{R}^{K+L+1}$ and any $z \in \mathbb{R}^{K+L+1}$, let

$$\text{dist}(z, \Gamma) \equiv \inf_{\lambda \in \Delta} \|\lambda - z\|$$

and

$$\text{dist}_n(z, \Gamma) \equiv \inf_{\lambda \in \Delta} ((\lambda - z)' J_n (\lambda - z))^{\frac{1}{2}}.$$

Note that

$$\text{dist}_n(Z_n, \Lambda) = \inf_{\lambda \in \Lambda} q_n^{\frac{1}{2}}(\lambda)$$

and

$$\text{dist}_n(Z_n, \sqrt{n}(\Theta - \theta_n)) = \inf_{\lambda \in \sqrt{n}(\Theta - \theta_n)} q_n^{\frac{1}{2}}(\lambda).$$

Let $C_n = \text{dist}_n(Z_n, \Lambda) - \text{dist}_n(Z_n, \sqrt{n}(\Theta - \theta_n))$. Then, it suffices to show that $C_n = o_p(1)$. Note that for any $Z_{\Theta,n} \in \sqrt{n}(\Theta - \theta_n)$, we have $\text{dist}(Z_{\Theta,n}, \Lambda) = o(1)$, since $\sqrt{n}(\Theta - \theta_n) \to \Lambda$.[40] This, together with $J_n \xrightarrow{p} J(\gamma^*)$, where $J(\gamma^*)$ is of full rank, implies that $\text{dist}_n(Z_{\Theta,n}, \Lambda) = o_p(1)$. Now, choose $Z_{\Theta,n} \in \sqrt{n}(\Theta - \theta_n)$ such that $\text{dist}_n(Z_n, \sqrt{n}(\Theta - \theta_n)) = \text{dist}_n(Z_n, \{Z_{\Theta,n}\}) + o_p(1)$. Such an element always exists. Then, by the triangle inequality

$$\begin{aligned} C_n &= \text{dist}_n(Z_n, \Lambda) - \text{dist}_n(Z_n, \sqrt{n}(\Theta - \theta_n)) \\ &\leq \text{dist}_n(Z_n, \{Z_{\Theta,n}\}) + \text{dist}_n(\{Z_{\Theta,n}\}, \Lambda) - \text{dist}_n(Z_n, \sqrt{n}(\Theta - \theta_n)) = o_p(1). \end{aligned}$$

Analogously, choose $Z_{\Lambda,n} \in \Lambda$ such that $\text{dist}_n(Z_n, \Lambda) = \text{dist}_n(Z_n, \{Z_{\Lambda,n}\}) + o_p(1)$. Again, we have $\text{dist}(Z_{\Lambda,n}, \sqrt{n}(\Theta - \theta_n)) = o(1)$, since $\sqrt{n}(\Theta - \theta_n) \to \Lambda$, and $\text{dist}_n(Z_{\Lambda,n}, \sqrt{n}(\Theta - \theta_n)) = o_p(1)$. By the triangle inequality

$$\begin{aligned} C_n &= \text{dist}_n(Z_n, \Lambda) - \text{dist}_n(Z_n, \sqrt{n}(\Theta - \theta_0)) \\ &\geq \text{dist}_n(Z_n, \Lambda) - \text{dist}_n(Z_n, \{Z_{\Lambda,n}\}) - \text{dist}_n(Z_{\Lambda,n}, \sqrt{n}(\Theta - \theta_n)) = o_p(1). \end{aligned}$$

[40] In particular, this holds true for $Z_{\Theta,n} = \sqrt{n}(\hat{\theta}_n - \theta_n)$, which we use in the proof of Theorem 2, see Appendix A.3.

□

Given the comment below Lemma 3, it follows from Lemma 3 and 4 that

$$q_n(\sqrt{n}(\hat{\theta}_n - \theta_n)) = \inf_{\lambda \in \Lambda} q_n(\lambda) + o_p(1). \tag{30}$$

Proof of Proposition 1. The proof of Proposition 1 follows from the proof of Theorem 3(a)-(b) in A1. Let $\hat{\lambda}_n$ be such that $\hat{\lambda}_n \in cl(\Lambda)$ and

$$q_n(\hat{\lambda}_n) = \inf_{\lambda \in \Lambda} q_n(\lambda).$$

Then, equation (30), which holds because Lemma 3 applies under $\{\gamma_n\} \in \Gamma(\gamma^*, b, d)$, reads

$$q_n(\sqrt{n}(\hat{\theta}_n - \theta_n)) = q_n(\hat{\lambda}_n) + o_p(1). \tag{31}$$

The proof proceeds by showing that $\sqrt{n}(\hat{\theta}_n - \theta_n) = \hat{\lambda}_n + o_p(1)$. The desired result then follows by the continuous mapping theorem, see the proof of Theorem 3(b) in A1. By convexity of Λ, there exists a unique $\lambda_n^* \in cl(\Lambda)$ such that[41]

$$\|\sqrt{n}(\hat{\theta}_n - \theta_n) - \lambda_n^*\| = o_p(1). \tag{32}$$

Therefore, it suffices to show that $\|\lambda_n^* - \hat{\lambda}_n\| = o_p(1)$. Define $\| \cdot \|_n$ by $\|x\|_n = (x'J_n x)^{1/2}$. Then, by Assumption 3 it suffices to show that

$$\|\lambda_n^* - \hat{\lambda}_n\|_n = o_p(1).$$

We have

$$\|\lambda_n^* - Z_n\|_n = \|\sqrt{n}(\hat{\theta}_n - \theta_n) - Z_n\|_n + o_p(1) = \|\hat{\lambda}_n - Z_n\|_n + o_p(1),$$

where the first equality follows from the triangle inequality and the fact that equation (32) also holds with $\| \cdot \|$ replaced by $\| \cdot \|_n$ by Assumption 3. The last equality holds by equation (31), which as mentioned above relies on Lemmas 3 and 4. The rest of the proof follows by the arguments presented below equation (7.19) in the proof of Theorem 3(a) in A1 with T replaced by n. The proof goes through despite Λ being a cone with possible nonzero vertex, as convexity constitutes the crucial property of Λ. □

[41] See also footnote 40.

A.3 Proof of Theorem 2

Proof of Theorem 2. The first part of the proof consists of showing that $M_n(\theta)$ and $\hat{M}_n(\theta)$, defined in equations (15) and (18), respectively, are close in the sense that for all $\breve{\theta}_n$ for which $\sqrt{T}(\breve{\theta}_n - \theta_n) = O_p(1)$, we have

$$\hat{M}_n(\breve{\theta}_n) = M_n(\breve{\theta}_n) + o_p(1/n). \tag{33}$$

By equation (24)

$$\sup_{\theta\in\Theta:\|\sqrt{n}(\theta-\theta_n)\|\leq\gamma} |R_n(\theta)| = o_p(1/n). \tag{34}$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1)$ (by Proposition 1)

$$Q_n(\hat{\theta}_n) = Q_n(\theta_n) + \frac{\partial}{\partial\theta}Q_n(\theta_n)'(\hat{\theta}_n - \theta_n) + \frac{1}{2}(\hat{\theta}_n - \theta_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\hat{\theta}_n - \theta_n) + o_p(1/n). \tag{35}$$

Furthermore,[42]

$$\frac{\partial}{\partial\theta}Q_n(\theta) = \frac{\partial}{\partial\theta}Q_n(\theta_n) + \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\theta - \theta_n) + \left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \theta_n),$$

such that

$$\frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n) = \frac{\partial}{\partial\theta}Q_n(\theta_n) + \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\hat{\theta}_n - \theta_n) + \left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\hat{\theta}_n - \theta_n).$$

It follows that

$$\begin{aligned}\frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n)'(\theta - \hat{\theta}_n) &= \frac{\partial}{\partial\theta}Q_n(\theta_n)'(\theta - \hat{\theta}_n) + (\hat{\theta}_n - \theta_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\theta - \hat{\theta}_n)\\ &\quad + (\hat{\theta}_n - \theta_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n).\end{aligned} \tag{36}$$

Lastly,

$$\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n) = \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n) + \left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right]. \tag{37}$$

[42] Here and in what follows, $\theta^\dagger$ denotes a (generic) vector on the line segment between the argument of the function (that differs from one equation to the other) and θ_n.

Plugging in (35), (36) and (37) into (18), we get

$$
\begin{aligned}
\hat{M}_n(\theta) = {} & Q_n(\theta_n) + \frac{\partial}{\partial\theta}Q_n(\theta_n)'(\hat{\theta}_n - \theta_n) + \frac{1}{2}(\hat{\theta}_n - \theta_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\hat{\theta}_n - \theta_n) + o_p(1/n) \\
& + \frac{\partial}{\partial\theta}Q_n(\theta_n)'(\theta - \hat{\theta}_n) + (\hat{\theta}_n - \theta_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\theta - \hat{\theta}_n) \\
& + (\hat{\theta}_n - \theta_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n) \\
& + \frac{1}{2}(\theta - \hat{\theta}_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\theta - \hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n).
\end{aligned}
$$

Upon rearrangement, we obtain

$$
\begin{aligned}
\hat{M}_n(\theta) = {} & Q_n(\theta_n) + \frac{\partial}{\partial\theta}Q_n(\theta_n)'(\theta - \theta_n) + \frac{1}{2}(\theta - \theta_n)'\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)(\theta - \theta_n) + o_p(1/n) \\
& + (\hat{\theta}_n - \theta_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n) \\
& + \frac{1}{2}(\theta - \hat{\theta}_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n).
\end{aligned}
$$

Next, we show that the second and third line are asymptotically negligible. Note that

$$
\begin{aligned}
& \sup_{\theta\in\Theta:\|\sqrt{n}(\theta-\theta_n)\|\le\epsilon}\left|(\hat{\theta}_n - \theta_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n)\right| \\
\le & \sup_{\theta\in\Theta:\|\sqrt{n}(\theta-\theta_n)\|\le\epsilon}\frac{1}{n}\|\sqrt{n}(\hat{\theta}_n - \theta_n)\|\left\|\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right\|\|\sqrt{n}(\theta - \hat{\theta}_n)\| \\
= & o_p(1/n),
\end{aligned}
$$

by Assumption 2, $\sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1)$, and the triangle inequality $\|\sqrt{n}(\theta-\hat{\theta}_n)\| \le \|\sqrt{n}(\theta-\theta_n)\| + \|\sqrt{n}(\hat{\theta}_n - \theta_n)\|$. By a similar argument, we have

$$
\sup_{\theta\in\Theta:\|\sqrt{n}(\theta-\theta_n)\|\le\epsilon}\left|\frac{1}{2}(\theta - \hat{\theta}_n)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right](\theta - \hat{\theta}_n)\right| = o_p(1/n).
$$

Thus, we conclude that equation (33) holds.

The rest of the proof follows from arguments similar to those underlying Theorem 2 and the second comment after Theorem 3 in A1, where $\hat{M}_n(\theta)$ takes on the role of $l_T(\theta)$ (divided by T). First, we show that $\sqrt{n}(\tilde{\theta}_n - \theta_n) = O_p(1)$. This follows from

$$
\tilde{\theta}_n = \arg\min_{\theta\in\mathbb{R}^k}\hat{M}_n(\theta) = \hat{\theta}_n + \left(\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n)\right)^{-1}\frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n).
$$

Using $\frac{\partial}{\partial\theta}Q_n(\hat{\theta}_n) = \frac{\partial}{\partial\theta}Q_n(\theta_n) + \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger)(\hat{\theta}_n - \theta_n)$ together with Assumption 2 and 3, we obtain

$$\begin{aligned}\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) = &\left(\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n)\right)^{-1}\sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_n) \\ &+ \left(\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\hat{\theta}_n)\right)^{-1}\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^\dagger)\sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1).\end{aligned}$$

Now, $\sqrt{n}(\tilde{\theta}_n - \theta_n) = \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_n) = O_p(1)$. Furthermore, using (33) and the fact that $M_n(\theta)$ can also be written as

$$M_n(\theta) = Q_n(\theta_n) - \frac{1}{2n}Z_n'J_nZ_n + \frac{1}{2n}q_n(\sqrt{n}(\theta - \theta_n))$$

we get

$$\hat{M}_n(\tilde{\theta}_n) = Q_n(\theta_n) + \frac{1}{2n}Z_n'J_nZ_n - \frac{1}{2n}q_n(\sqrt{n}(\tilde{\theta}_n - \theta_n)) + o_p(1/n). \tag{38}$$

Let $\ddot{\theta}_n = \arg\inf_{\theta\in\mathbb{R}^k} M_n(\theta)$. As argued above Theorem 2, $\sqrt{n}(\ddot{\theta}_n - \theta_n) \to Z(\gamma^*) = O_p(1)$. Similar to above, we get

$$\hat{M}_n(\ddot{\theta}_n) = Q_n(\theta_n) + \frac{1}{2n}Z_n'J_nZ_n - \frac{1}{2n}q_n(\sqrt{n}(\ddot{\theta}_n - \theta_n)) + o_p(1/n). \tag{39}$$

By definition,

$$\hat{M}_n(\tilde{\theta}_n) = \inf_{\theta\in\mathbb{R}^k}\hat{M}_n(\theta) + o_p(1/n) \tag{40}$$

and

$$M_n(\ddot{\theta}_n) = \inf_{\theta\in\mathbb{R}^k} M_n(\theta) + o_p(1/n),$$

where latter implies

$$q_n(\sqrt{n}(\ddot{\theta}_n - \theta_n)) = \inf_{\theta\in\mathbb{R}^k} q_n(\sqrt{n}(\theta - \theta_n)) + o_p(1). \tag{41}$$

Now,

$$\begin{aligned}o_p(1/n) = \hat{M}_n(\tilde{\theta}_n) - \inf_{\theta\in\mathbb{R}^k}\hat{M}_n(\theta) &\geq \hat{M}_n(\tilde{\theta}_n) - \hat{M}_n(\ddot{\theta}_n) \\ &= \frac{1}{2n}q_n(\sqrt{n}(\tilde{\theta}_n - \theta_n)) - \frac{1}{2n}q_n(\sqrt{n}(\ddot{\theta}_n - \theta_n)) + o_p(1/n) \\ &\geq \frac{1}{2n}\inf_{\theta\in\mathbb{R}^k} q_n(\sqrt{n}(\theta - \theta_n)) - \frac{1}{2n}q_n(\sqrt{n}(\ddot{\theta}_n - \theta_n)) + o_p(1/n) = o_p(1/n),\end{aligned}$$

where the first equality follows from (40), the first inequality by definition of the infimum, the second equality by (38) and (39), the second inequality by definition of the infimum, and the last equality from (41). We conclude that

$$q_n(\sqrt{n}(\tilde{\theta}_n - \theta_n)) = q_n(\sqrt{n}(\ddot{\theta}_n - \theta_n)) + o_p(1). \tag{42}$$

Plugging (41) in (42), we get

$$q_n(\sqrt{n}(\tilde{\theta}_n - \theta_n)) = \inf_{\theta \in \mathbb{R}^k} q_n(\sqrt{n}(\theta - \theta_n)) + o_p(1).$$

Trivially, $\inf_{\theta \in \mathbb{R}^k} q_n(\sqrt{n}(\theta - \theta_n)) = \inf_{\lambda \in \mathbb{R}^k} q_n(\lambda)$, such that

$$q_n(\sqrt{T}(\tilde{\theta}_n - \theta_n)) = \inf_{\lambda \in \mathbb{R}^k} q_n(\lambda) + o_p(1).$$

The argument in the second comment after Theorem 3 of A1 now applies. In particular, $\hat{\lambda}_n = Z_n$ and $\inf_{\lambda \in \mathbb{R}^k} q_n(\lambda) = 0$. Then,

$$q_n(\sqrt{n}(\tilde{\theta}_n - \theta_n)) = (\sqrt{n}(\tilde{\theta}_n - \theta_n) - Z_n)' J_n (\sqrt{n}(\tilde{\theta}_n - \theta_n) - Z_n) = o_p(1).$$

Since $J_n(\to J(\gamma^*))$ is asymptotically of full rank by Assumption 4, we conclude that $\sqrt{n}(\tilde{\theta}_n - \theta_n) = Z_n + o_p(1)$, i.e., $\sqrt{n}(\tilde{\theta}_n - \theta_n) \xrightarrow{d} Z(\gamma^*)$. □

A.4 Details for the Random Coefficient Example

A.4.1 First and second-order derivatives

The first order partial derivatives of $\mathrm{l}(\mu_\beta, \sigma_u^2, \sigma_\beta^2 | y_i, x_i)$ are given by

$$\frac{\partial \mathrm{l}(\mu_\beta, \sigma_u^2, \sigma_\beta^2 | y_i, x_i)}{\partial \mu_\beta} = \frac{y_i - x_i \mu_\beta}{\sigma_u^2 + x_i^2 \sigma_\beta^2} x_i,$$

$$\frac{\partial \mathrm{l}(\mu_\beta, \sigma_u^2, \sigma_\beta^2 | y_i, x_i)}{\partial \sigma_u^2} = -\frac{1}{2} \frac{1}{\sigma_u^2 + x_i^2 \sigma_\beta^2} + \frac{(y_i - x_i \mu_\beta)^2}{2\left(\sigma_u^2 + x_i^2 \sigma_\beta^2\right)^2}$$

and

$$\frac{\partial \mathrm{l}(\mu_\beta, \sigma_u^2, \sigma_\beta^2 | y_i, x_i)}{\partial \sigma_\beta^2} = -\frac{1}{2} \frac{1}{\sigma_u^2 + x_i^2 \sigma_\beta^2} x_i^2 + \frac{(y_i - x_i \mu_\beta)^2}{2\left(\sigma_u^2 + x_i^2 \sigma_\beta^2\right)^2} x_i^2.$$

The second order partial derivatives of $\mathrm{l}(\mu_\beta, \sigma_u^2, \sigma_\beta^2 | y_i, x_i)$ are given by

$$\frac{\partial^2 \mathrm{l}(\mu_\beta, \sigma_u^2, \sigma_\beta^2 | y_i, x_i)}{\partial^2 \mu_\beta} = -\frac{1}{\sigma_u^2 + x_i^2 \sigma_\beta^2} x_i^2,$$

$$\frac{\partial^2 \mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i)}{\partial\mu_\beta\partial\sigma_u^2} = -\frac{y_i - x_i\mu_\beta}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^2}x_i,$$

$$\frac{\partial^2 \mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i)}{\partial\mu_\beta\partial\sigma_\beta^2} = -\frac{y_i - x_i\mu_\beta}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^2}x_i^3,$$

$$\frac{\partial^2 \mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i)}{\partial^2\sigma_u^2} = \frac{1}{2}\frac{1}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^2} - \frac{\left(y_i - x_i\mu_\beta\right)^2}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^3},$$

$$\frac{\partial^2 \mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i)}{\partial\sigma_u^2\partial\sigma_\beta^2} = \frac{1}{2}\frac{1}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^2}x_i^2 - \frac{\left(y_i - x_i\mu_\beta\right)^2}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^3}x_i^2$$

and

$$\frac{\partial^2 \mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i)}{\partial^2\sigma_\beta^2} = \frac{1}{2}\frac{1}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^2}x_i^4 - \frac{\left(y_i - x_i\mu_\beta\right)^2}{\left(\sigma_u^2 + x_i^2\sigma_\beta^2\right)^3}x_i^4.$$

A.4.2 Verification of Assumptions 1-4

We verify Assumption 1*, which is sufficient for Assumption 1. Assumption 1*(a) and (b) follow from Lemma 11.3 in Andrews and Cheng (2011), AC3 hereafter, with $Q(\theta;\gamma^*)$ given by

$$-E_{\gamma^*}\mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i).$$

In what follows, let $\mathrm{l}(\theta)$ denote $-\mathrm{l}(\mu_\beta,\sigma_u^2,\sigma_\beta^2|y_i,x_i)$. The assumptions of Lemma 11.3 can be verified as follows. (i) follows from the i.i.d. assumption. (ii) follows from a mean value expansion with $s(w,\theta)$ given by $\mathrm{l}(\theta)$. (iii) is implied by (8). In particular, the expected value of $M_1(w)$ is bounded by (a constant times) x_i^4 yielding the desired integrability condition in (iii). (iv) is directly assumed. Assumption 1*(c) follows from Section 8.1 in Andrews (1997). Equation (3.18) in Andrews (1997) is satisfied as long as x_i is random. Assumption 1*(d) is satisfied by construction. Assumption 2 (a) holds by the definition of $\mathrm{l}(\theta)$. Assumption 2 (b) can be verified using Lemma 11.3 in AC3 with $s(w,\theta)$ given by $\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta)$. The conditions of the lemma can be verified as above given (8). In particular, by the triangle inequality we have

$$\left\|\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right\| \leq \left\|\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta) - E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta)\right\| + \\ \left\|E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta) - E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta_n)\right\| + \left\|E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta_n) - \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n)\right\|.$$

Now, the first and the last term on the right hand side are $o_p(1)$ by Lemma 11.3 in AC3, while the middle term is $o(1)$ when $\sup_{\theta\in\Theta:\|\theta-\theta_n\|\leq\epsilon_n}$ is applied due to the uniform continuity of $E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta)$ in θ, which also follows from Lemma 11.3 in AC3. Assumption 3 follows by a similar argument. In particular, we have

$$\begin{aligned}&\left\|\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n) - E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta^*)\right\|\\ \leq &\left\|\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n) - E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta_n)\right\| + \left\|E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta_n) - E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta^*)\right\|.\end{aligned}$$

The first term is $o_p(1)$ by Lemma 11.3 in Andrews and Cheng (2011) and the second term is $o(1)$ because of the uniform continuity of $E_{\gamma^*}\frac{\partial^2}{\partial\theta\partial\theta'}\mathrm{l}(\theta)$ in θ, which again follows from Lemma 11.3 in AC3, and $\theta_n \to \theta^*$. Non-singularity is satisfied as long as x_i is random. Assumption 4 holds by Lemma 11.5 in AC3 with $s(w, \theta)$ given by $\frac{\partial}{\partial\theta}\mathrm{l}(\theta)$. The integrability condition is implied by (8). Given the information equality, non-singularity of the variance matrix follows immediately.

B Additional figures

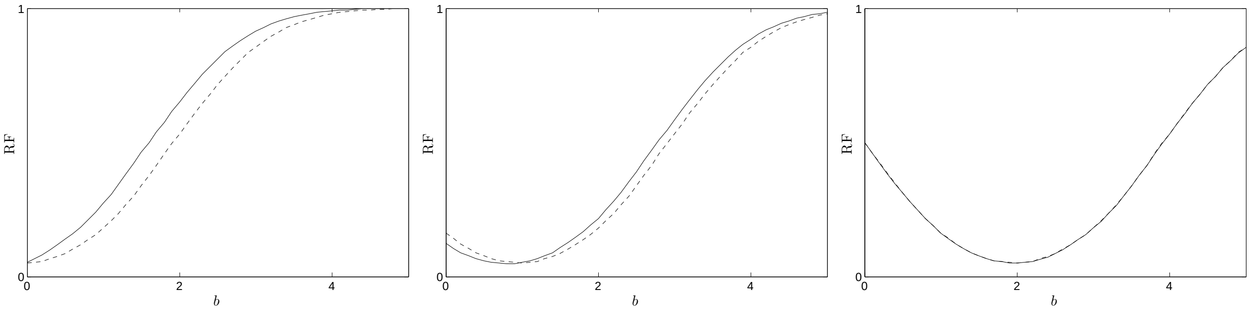

Figure 14: Power of the CLR (solid) and the regular t-test (dashed) for testing $H_0 : b = 0, 1, 2$ from left to right.

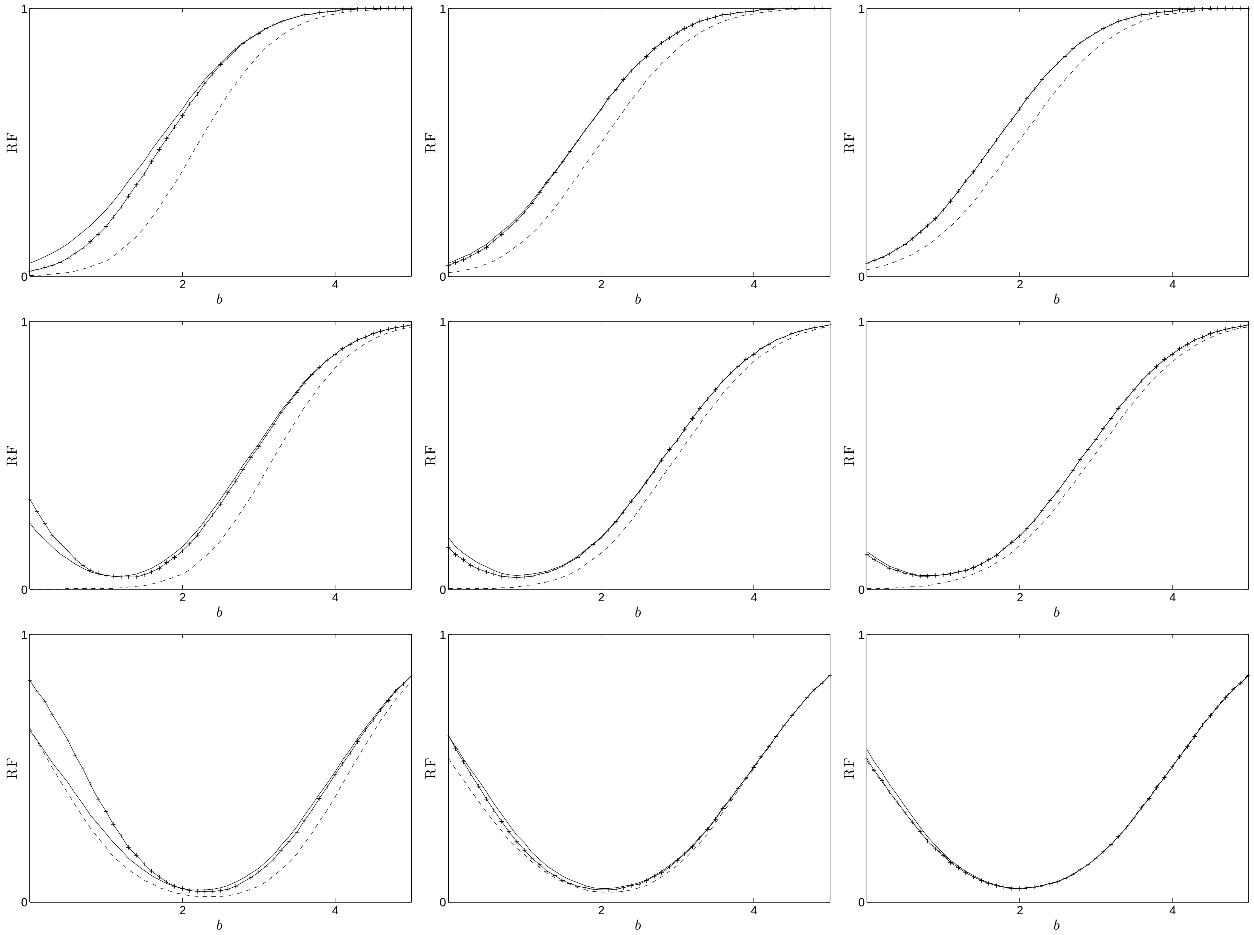

Figure 15: Power of the CLR (solid), gLR (solid with plusses), and the t_{ML}-test (dashed) for testing $b_0 = 0, 1, 2$ from top to bottom with $d = 0, 1, 2$ from left to right. $\rho = -0.7$.

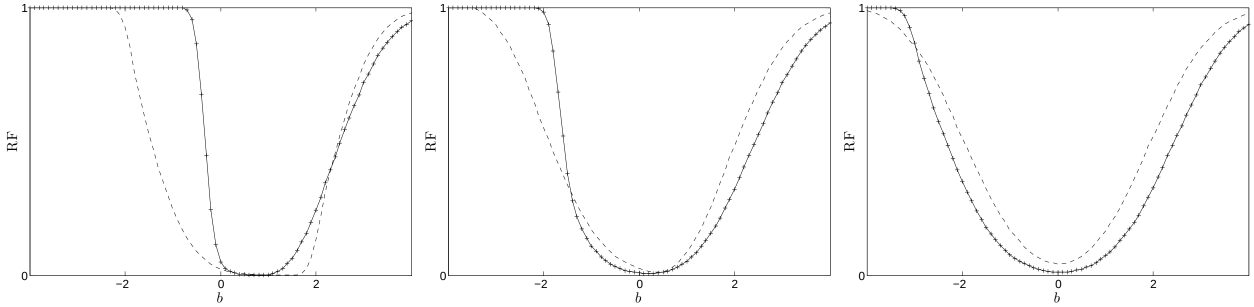

Figure 16: Power of gLR (solid with plusses) and the t_{ML}-test (dashed) for testing $H_0 : b = 0$ with $d = 0^6, 1^6, 2^6$ from left to right.

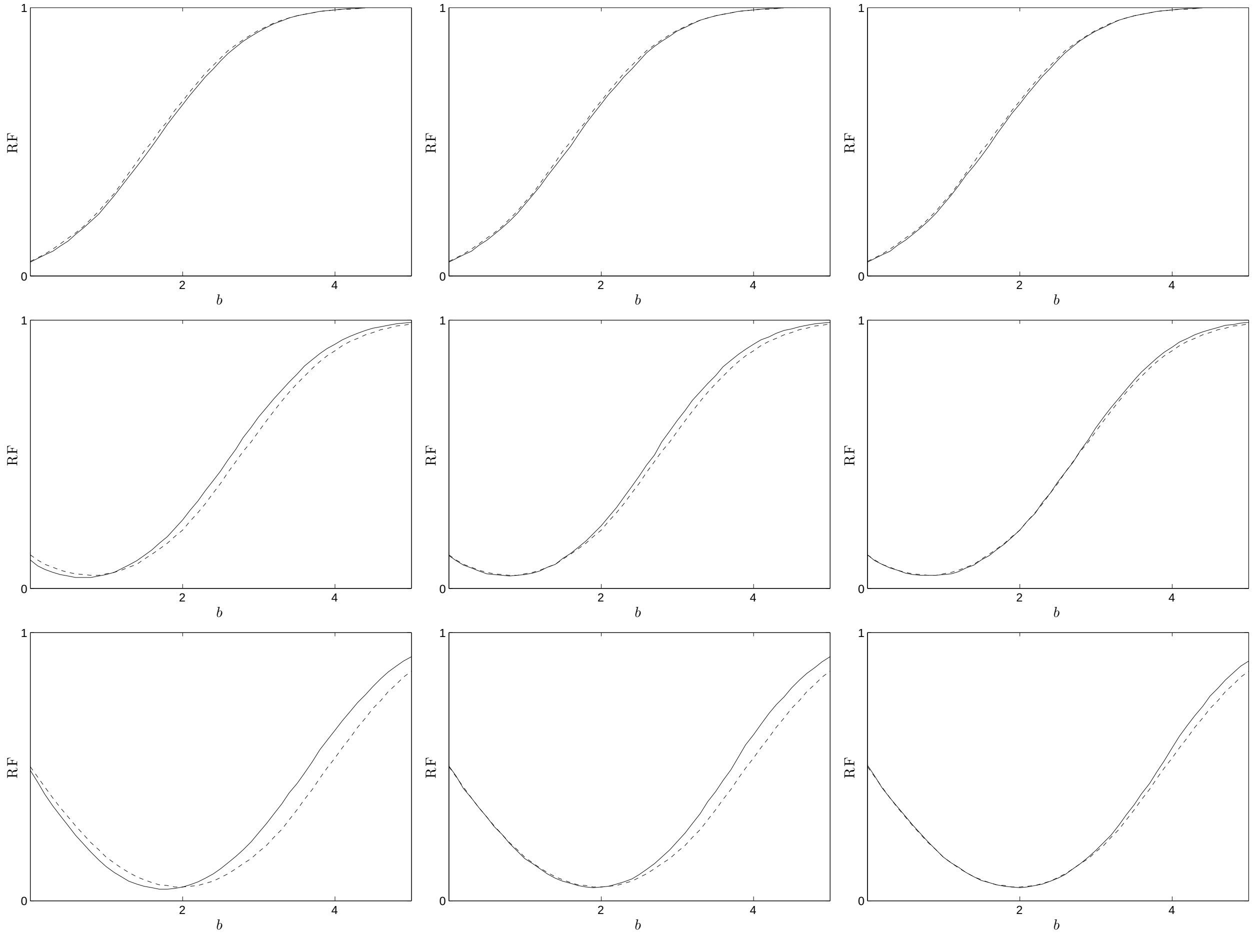

Figure 17: Power of the CLR (solid) and the partial CLR (dashed) for testing $b_0 = 0, 1, 2$ from top to bottom with $d = 0, 1, 2$ from left to right. $\rho = 0.7$.

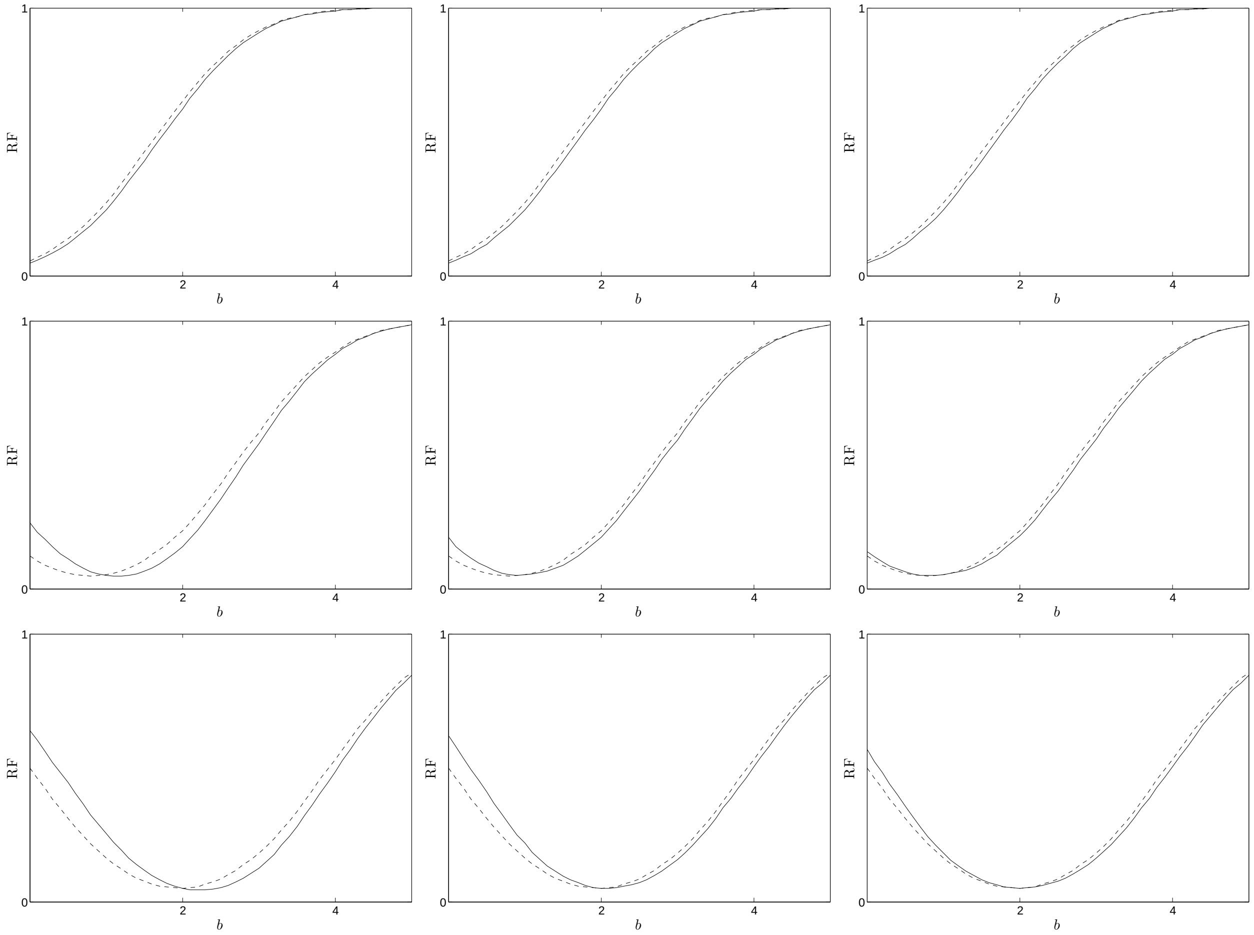

Figure 18: Power of the CLR (solid) and the partial CLR (dashed) for testing $b_0 = 0, 1, 2$ from top to bottom with $d = 0, 1, 2$ from left to right. $\rho = -0.7$.

C Additional tables

Table 3: Table 6 in RV - Optimal instruments (ii)

			95% CI				90% CI			
	Est.	Std. Err.	gLR		t_{ML}		gLR		t_{ML}	
Mean valuations μ										
Price/Inc.	-2.322	0.498	-3.188	-1.464	-3.297	-1.347	-3.051	-1.596	-3.141	-1.504
Hp/We.	-0.918	1.192	-3.790	1.788	-3.253	1.418	-3.408	1.466	-2.878	1.043
Foreign	-0.858	0.216	-1.089	-0.313	-1.275	-0.430	-1.077	-0.380	-1.207	-0.498
Size	0.667	0.674	-0.630	1.963	-0.654	1.987	-0.425	1.758	-0.441	1.775
Height	0.183	0.052	0.073	0.294	0.080	0.285	0.089	0.277	0.097	0.269
€/km	-3.972	1.344	-6.425	-1.418	-6.606	-1.338	-6.062	-1.802	-6.183	-1.761
Variances σ^2										
Price/Inc.	0.274	0.176	0.000	0.642	0.000	0.619	0.000	0.587	0.000	0.564
Hp/We.	10.252	4.346	1.885	19.097	1.733	18.771	3.103	17.749	3.103	17.401
Foreign	0.515	0.736	0.000	2.186	0.000	1.958	0.000	1.943	0.000	1.726
Size	0.057	0.189	0.000	0.430	0.000	0.427	0.000	0.372	0.000	0.367
Height	0.011	0.006	0.000	0.022	0.000	0.023	0.002	0.021	0.000	0.021
€/km	4.424	19.835	0.000	44.489	0.000	43.299	0.000	38.338	0.000	37.049