

# Linear Regression with Many Controls of Limited Explanatory Power\*

Chenchuan (Mark) Li and Ulrich K. Müller

Princeton University

Department of Economics

Princeton, NJ, 08544

March 2020

## Abstract

We consider inference about a scalar coefficient in a linear regression model. One previously considered approach to dealing with many controls imposes sparsity, that is, it is assumed known that nearly all control coefficients are zero, or at least very nearly so. We instead impose a bound on the quadratic mean of the controls' effect on the dependent variable. We develop a simple inference procedure that exploits this additional information in general heteroskedastic models. We study its asymptotic efficiency properties and compare it to a sparsity-based approach in a Monte Carlo study. The method is illustrated in three empirical applications.

**Keywords:** high dimensional linear regression, limit of experiments, L2 bound, invariance to linear reparameterizations

---

\*We thank Michal Kolesar and participants at various workshops for useful comments and advice, and Guido Imbens for providing us with the data of Section 6.2. Müller gratefully acknowledges financial support from the National Science Foundation through grant SES-1627660.

# 1 Introduction

A classic issue that arises frequently in applied econometrics is how to deal with a potentially large number of control variables in a linear regression. In observational studies, the plausibility of an unconfoundedness assumption often hinges on having correctly controlled for the value of predetermined variables, which might require including higher order interactions, leading to many control variables. As is well understood, excluding controls that have non-zero coefficients in general yields estimators with omitted variable bias, and corresponding confidence intervals with less than nominal coverage. In empirical practice, this issue is often addressed by reporting results from several specifications that vary in the number and identity of included control variables.

A seemingly more systematic approach is to use a pre-test to identify which controls have non-zero coefficients, such as testing down procedures, or information criteria, and then proceed with standard inference using only the selected controls. As stressed by Leeb and Pötscher (2005) (also see Leeb and Pötscher (2008a, 2008b) and the references therein), however, this does not yield uniformly valid inference: If a control coefficient is of order  $O(n^{-1/2})$  in a sample of size  $n$ , then it is not selected with probability one, yet it induces an omitted variable bias that is still large enough to yield oversized confidence intervals. This speaks to a broader theoretical result that in the regression model with Gaussian errors, a hypothesis test either overrejects for some value of the control coefficients, or its power is uniformly dominated by the “long regression” that simply includes all potential controls. Hence, an assumption on the control coefficients is necessary to make progress.

In that context, the empirical practice of reporting several specifications amounts to two extremes: A specification that does not include a set of potential control variables is justified under the assumption that all coefficients are zero, while the specification with the control variables leaves them entirely unconstrained. A potentially more attractive middle ground is an assumption that the control coefficients are, in some sense, of limited magnitude.

One formalization of this idea that has spawned a burgeoning literature is the assumption of sparsity (Tibshirani (1996), Fan and Li (2001), etc.): Most of the control coefficients are known to be zero (or very close to zero), but it is not known which ones. A standard LASSO implementation does not lead to valid inference about the coefficient of interest. But by combining a sparsity assumption on the control coefficients with a sparsity assumption on the correlations between the regressor of interest and the control variables, recent work by Belloni, Chernozhukov, and Hansen (2014) shows how a novel LASSO based “double selection procedure” does yield uniformly valid large-sample inference.

While this work is important progress, a sparsity assumption might not always be a compelling starting point: In social science applications, it is usually not obvious why the large majority of control coefficients should be very nearly zero. In addition, the sparsity restriction does not remain invariant to linear reparameterizations of the controls. For instance, in the context of technical controls that are functions of an underlying continuous variable, sparsity drives a distinction between specifying the controls as powers or Chebyshev polynomials, and when including a set of fixed effects, in general it matters which one is dropped to avoid perfect multi-collinearity. Finally, in a LASSO implementation, the imposed degree of sparsity is implicitly controlled by a penalty parameter, which makes the small sample interpretation of the resulting inference less than straightforward.

This paper develops an alternative approach that considers *a priori* upper bounds on the weighted average of squared control coefficients, rather than on the number of non-zero control coefficients. To be precise, consider constructing a confidence interval for the scalar parameter  $\beta$  from observing  $\{y_i, x_i, \mathbf{q}_i, \mathbf{z}_i\}_{i=1}^n$ , where

$$y_i = \beta x_i + \mathbf{q}_i' \boldsymbol{\delta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_i, \quad (1)$$

the  $m \times 1$  control variables  $\mathbf{q}_i$  are the baseline specification, the  $p \times 1$  additional variables  $\mathbf{z}_i$  are potential additional control variables and  $\varepsilon_i$  is a conditionally mean zero error term. To ease notation, assume that  $\mathbf{z}_i$  has been projected off  $\mathbf{q}_i$ , so that  $\mathbf{z}_i' \boldsymbol{\gamma}$  is the contribution of  $\mathbf{z}_i$  to the conditional mean of  $y_i$  after having controlled for the baseline controls  $\mathbf{q}_i$ . We impose the bound

$$\kappa^2 = n^{-1} \sum_{i=1}^n (\mathbf{z}_i' \boldsymbol{\gamma})^2 \leq \bar{\kappa}^2. \quad (2)$$

The parameter  $\kappa^2$  is the average of the squared mean effects  $\mathbf{z}_i' \boldsymbol{\gamma}$  on  $y_i$  induced by  $\mathbf{z}_i$ , that is,  $\kappa$  is the quadratic mean of the mean effects of  $\mathbf{z}_i$  on  $y_i$ , after controlling for the baseline controls  $\mathbf{q}_i$ . Small values of  $\bar{\kappa}$  thus embed the *a priori* assumption that the explanatory power of the controls is small.

Let  $\hat{\beta}_{\text{short}}$  and  $\hat{\beta}_{\text{long}}$  be the coefficients on  $x_i$  from a linear regression of  $y_i$  on  $(x_i, \mathbf{q}_i)$ , and from a linear regression of  $y_i$  on  $(x_i, \mathbf{q}_i, \mathbf{z}_i)$ , respectively. We combine the information in  $(\hat{\beta}_{\text{short}}, \hat{\beta}_{\text{long}})$  and the bound (2) to develop a likelihood ratio (LR) procedure that is more informative than the usual confidence interval centered at  $\hat{\beta}_{\text{long}}$ . Since  $(\hat{\beta}_{\text{short}}, \hat{\beta}_{\text{long}})$  and the bound (2) are invariant to linear transformations of the additional controls, so is the new confidence interval. The new interval essentially reduces to the usual intervals centered at  $\hat{\beta}_{\text{short}}$  and  $\hat{\beta}_{\text{long}}$  for  $\bar{\kappa} = 0$  and  $\bar{\kappa} \rightarrow \infty$ , respectively.<sup>1</sup> The intervals thus provide a continuous

---

<sup>1</sup>See Section 3 for details.

bridge between omitting the additional controls and including them with unconstrained coefficients. For any given  $\bar{\kappa}$ , the center of the 95% confidence interval  $\hat{\beta}_{\text{LR}}(\bar{\kappa})$  serves as an estimator of  $\beta$  under the constraint (2).

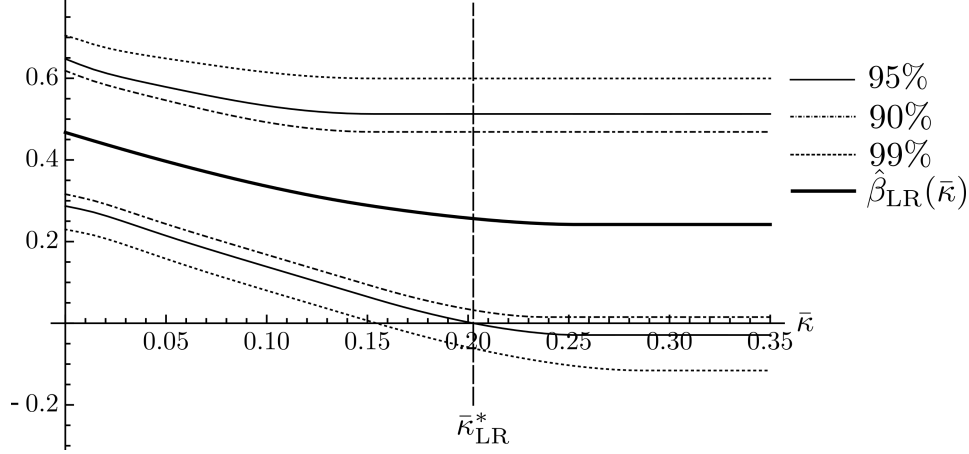
As an illustration of these ideas, consider the study by Macchiavello and Morjaria (2015), who use data on African rose exports to identify reputational effects in markets without contract enforcement. The authors construct a model with the feature that binding incentive constraints yield observable proxies for the buyer-seller relationship value during periods of maximum temptation for sellers and buyers to undercut each other. They find, empirically, that this value proxy is correlated with relationship age but not outside prices, evidence that reputation constrains trade in the absence of enforcement. We apply our approach to determine the extent to which the correlation between relationship value and age is sensitive to Macchiavello and Morjaria’s choice of control variables.

We treat the panel regression from Table 5, Column 8 of Macchiavello and Morjaria (2015) as the short regression, where relationships are the unit of observation and the time dimension corresponds to four growing seasons (years), for a total of  $n = 372$  observations. This regression has the log of the relationship value as the dependent variable, the regressor of interest is the log of relationship age, and the baseline controls are the maximum of the previous observed log auction value as well as relationship and season fixed-effects. This is a difference-in-differences model in which the main effect is identified by variation in sales across seasons for buyer-seller relationships of different ages. Macchiavello and Morjaria (2015) find that  $\beta$ , the coefficient on relationship age, is statistically significant at standard confidence levels.

We investigate the sensitivity of these results to  $p = 123$  additional buyer $\times$ season fixed effects. This specification is an extension of the baseline season controls that allows flexibility over buyers. To the extent that purchase patterns over seasons differed between buyers with relationships of various lengths for reasons unrelated to learning about seller quality, omitting these additional fixed effects could lead to bias in  $\beta$ . However, absent constraints on the coefficient of these additional controls, only variation from seller differences across seasons can identify the main effect  $\beta$ , so including these additional fixed effects in an unconstrained fashion leads to much less informative inference.

Figure 1 plots 90%, 95% and 99% confidence intervals for  $\beta$  from our new procedure as a function of  $\bar{\kappa}$ , along with the point estimates  $\hat{\beta}_{\text{LR}}(\bar{\kappa})$ . We see that the short regression strongly rejects, but the long regression does not. For any significance level, there is a unique  $\bar{\kappa}_{\text{LR}}^*$  such that for all  $\bar{\kappa} < \bar{\kappa}_{\text{LR}}^*$ , the confidence interval excludes zero, and for all  $\bar{\kappa} > \bar{\kappa}_{\text{LR}}^*$ ,

Figure 1: LR Confidence Intervals for  $\beta$  in Macchiavello and Morjaria (2015)

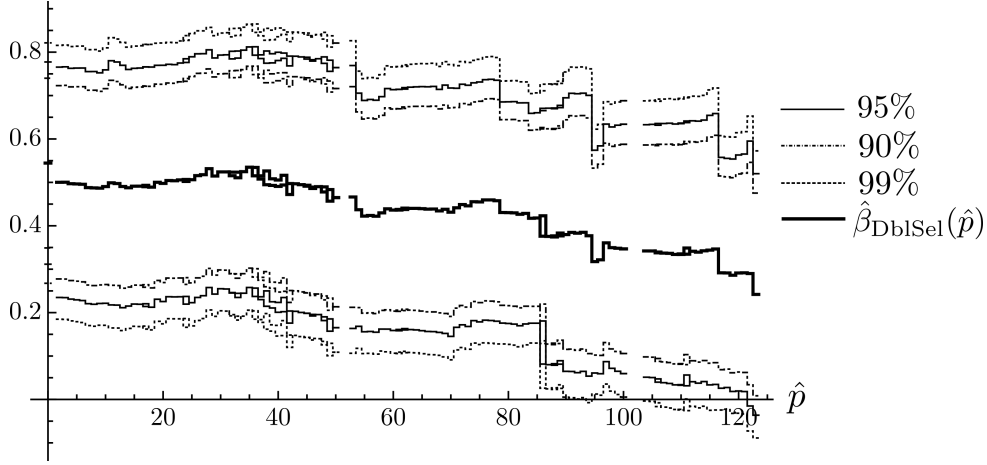


it includes zero. This is a generic property of our procedure. For instance, for the 95% level,  $\bar{\kappa}_{\text{LR}}^* = 0.202$ . Thus, since the outcome is measured in logs, as long as one believes that season-specific idiosyncratic buyer preferences not already captured by Macchiavello and Morjaria's (2015) baseline controls induce on average changes in the relationship value of no more than 20.2%, the conclusion of a statistically significant effect of the age of the relationship is upheld. It might also be useful to consider the ratio of  $n(\bar{\kappa}_{\text{LR}}^*)^2$  and the sum of squared residuals of a regression of  $y_i$  on  $\mathbf{q}_i$ , which equals 15.2% in this example; this  $R^2$ -type ratio is the fraction of the variability of  $y_i$  that is explained by the effect of  $\mathbf{z}_i$  on  $y_i$  under the null hypothesis of  $\beta = 0$ , after controlling for the baseline controls  $\mathbf{q}_i$ . In other words, significance of  $\beta$  at the 5% level prevails when the direct effect of season-specific idiosyncratic buyer preferences is responsible for less than 15.2% of residual variation in the log-relationship value. Figure 1 thus provides a comprehensive picture of the sensitivity of Macchiavello and Morjaria's (2015) results to the inclusion of buyer-season fixed effects under an interpretable range of assumptions about their potential explanatory power, and the threshold value  $\bar{\kappa}_{\text{LR}}^*$  is a useful quantitative summary of this sensitivity.

It might be useful to contrast these results to what is obtained from an analysis imposing sparsity. Figure 2 provides post-double lasso point estimates and confidence intervals for  $\beta$ , where the penalty terms suggested by Belloni, Chernozhukov, and Hansen (2014) are multiplied by a common factor that induces the sparsity index  $0 \leq \hat{p} \leq p$  in the post-double lasso regression.<sup>2</sup> The confidence interval on the very left and very right are again standard short

<sup>2</sup>The gaps in the figure arise because it seems numerically impossible to induce all values of  $0 \leq \hat{p} \leq p$  by

Figure 2: Sparsity based Confidence Intervals for  $\beta$  in Macchiavello and Morjaria (2015)



and long regression inference. But in between, the bounds on the confidence intervals are not a monotone function of the sparsity index, complicating the interpretation of a higher index as a “weaker” assumption on the control coefficients. What is more, there is no justification for the confidence intervals reported in Figure 2: The asymptotic justification for double selection lasso inference is not for a specific sparsity index, but rather, it is shown to be valid under certain asymptotic sequences of penalty terms if the model satisfies some asymptotic sparsity constraint. The default penalty choice suggested by Belloni, Chernozhukov, and Hansen (2014) applied to the example selects  $\hat{p} = 2$  controls and leads to a significant  $\beta$  at all conventional levels.

Our suggested confidence interval is based on the Likelihood Ratio (LR) test statistic obtained from the large sample normality of  $(\hat{\beta}_{\text{short}}, \hat{\beta}_{\text{long}})$  and the bound on the omitted variable bias of  $\hat{\beta}_{\text{short}}$  implied by (2), for which we tabulate appropriate critical values. From an econometric theory perspective, it is interesting to investigate whether this simple “bivariate” approach comes close to efficiently exploiting the information contained in (2). To this end, we consider the Gaussian homoskedastic version of the regression model (1) and consider asymptotics where the number of additional controls  $p$  is of the same order of magnitude as the sample size  $n$ . Our main theoretical finding is that in this model, tests that depend on the data only through  $(\hat{\beta}_{\text{short}}, \hat{\beta}_{\text{long}})$  are asymptotically efficient in a well

---

varying the penalty term factor. When there is more than one post selection regression at a given sparsity level  $\hat{p}$  (which is possible, since Lasso is based on an  $L_1$  penalty, rather than directly penalizing sparsity), we report the lower and upper envelopes.

defined sense as long as  $\kappa = o(n^{-1/4})$ . This rate corresponds to a ratio of  $n\kappa^2$  to the sum of squared residuals of a regression of  $y_i$  on  $\mathbf{q}_i$  of order  $o(n^{-1/2})$ . While converging to zero, this rate allows for finitely many non-zero coefficients of order  $o(n^{-1/4})$ , which would lead to corresponding individual t-statistics that diverge at the rate  $o(n^{1/4})$ . It also allows for a fraction  $o(n^{1/4})$  of control coefficients of the already problematic order  $O(n^{-1/2})$ . Since we expect that our procedure is most valuable in cases where the additional control coefficients are not obviously relevant *a priori*, this limited efficiency result is thus still useful. The validity of the suggested inference does not depend on any assumptions about  $\kappa$  or  $\bar{\kappa}$ .

$L_2$  penalties of the form (2) play a key role in ridge regression (Hoerl and Kennard (1970)), but our set-up uses (2) as a constraint on the nuisance parameters  $\gamma_j$  only. Furthermore, our focus is on hypothesis testing and confidence intervals, and ridge regression estimators do not automatically lead to shorter confidence intervals (see, for instance, Obenchain (1977)). Armstrong and Kolesár (2018) derive small sample minimax optimal confidence intervals in a class of Gaussian regression models with the regression function an element of a known convex set. As they point out in the Section 4.1.2. of the corresponding working paper Armstrong and Kolesár (2016), their generic results could be applied to (1) under the bound (2), and we provide some comparison with the LR confidence interval in our Section 2.2 below. Our approach of exploiting an *a priori* bound on the value of a nuisance parameter is also related in spirit to the analysis of Conley, Hansen, and Rossi (2012), who consider instrumental variable estimation with an imperfect instrument that has a direct effect on the outcome of bounded magnitude.

The rest of the paper is organized as follows. Section 2 contains the analysis of the Gaussian linear regression model (1). In this model, bivariate LR inference is exact, and we analyze and compare its properties. Section 2.3 derives the asymptotic efficiency result for bivariate inference. Section 3 discusses the implementation of feasible inference for non-normal, possibly heteroskedastic and clustered linear regressions. Section 4 contains two extensions: First, we discuss instrumental variable regression with a scalar instrument and a scalar endogenous variable, and second how to further sharpen inference under an additional bound on the explanatory power in the population regression of  $x_i$  on the potential controls  $\mathbf{z}_i$ . Section 5 provides a small sample Monte Carlo analysis of our procedure and compares it to the double selection lasso procedure proposed by Belloni, Chernozhukov, and Hansen (2014). Section 6 provides two additional empirical illustrations. Section 7 concludes. All proofs are collected in an appendix.

## 2 Gaussian Linear Model

### 2.1 Set-up

Write model (1) in vector form as

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{Q}\boldsymbol{\delta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad (3)$$

in obvious notation. Without loss of generality, assume that  $\mathbf{x}$  and the additional controls  $\mathbf{Z}$  have been projected off the baseline controls  $\mathbf{Q}$  (so that  $\mathbf{Q}'\mathbf{x} = \mathbf{0}$  and  $\mathbf{Z}'\mathbf{Q} = \mathbf{0}$ ). Our efficiency results focus on the simplest model where the regressors  $(\mathbf{x}, \mathbf{Q}, \mathbf{Z})$  are non-stochastic and  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ . In Section 3 below, we discuss the implementation in models where, after potentially conditioning on random  $(\mathbf{x}, \mathbf{Q}, \mathbf{Z})$ , the errors are non-Gaussian and heteroskedastic.

We assume throughout that  $(\mathbf{x}, \mathbf{Q}, \mathbf{Z})$  is of full column rank. The  $(1 + m + p)$  vector of OLS estimators

$$\begin{pmatrix} \hat{\beta}_{\text{long}} \\ \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'\mathbf{x} & \mathbf{0} & \mathbf{x}'\mathbf{Z} \\ \mathbf{0} & \mathbf{Q}'\mathbf{Q} & \mathbf{0} \\ \mathbf{Z}'\mathbf{x} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}'\mathbf{y} \\ \mathbf{Q}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \beta \\ \boldsymbol{\delta} \\ \boldsymbol{\gamma} \end{pmatrix}, \begin{pmatrix} \mathbf{x}'\mathbf{x} & \mathbf{0} & \mathbf{x}'\mathbf{Z} \\ \mathbf{0} & \mathbf{Q}'\mathbf{Q} & \mathbf{0} \\ \mathbf{Z}'\mathbf{x} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}^{-1} \right) \quad (4)$$

form a sufficient statistic. Inference about  $\beta$  thus becomes inference about one element of the mean of a  $p + m + 1$  dimensional multivariate normal with known covariance matrix.

Let  $\mathbf{Y} = (\mathbf{y}, \mathbf{x}, \mathbf{Q}, \mathbf{Z}) \in \mathbb{R}^{(2+m+p)n}$  be the observed data, and let  $\varphi_{\beta_0}(\mathbf{Y})$  be non-randomized level  $\alpha$  tests of the null hypothesis  $H_0 : \beta = \beta_0$ , where  $\varphi_{\beta_0}(\mathbf{Y}) = 1$  indicates rejection. A confidence set of level  $1 - \alpha$  is obtained by “inverting” the family of tests  $\varphi_{\beta_0}$ , that is by collecting the values of  $\beta_0$  for which the test does not reject,  $\text{CI}(\mathbf{Y}) = \{\beta_0 : \varphi_{\beta_0}(\mathbf{y}) = 0\}$ . By Proposition 15.2 of van der Vaart (1998), for one-sided hypothesis tests about  $\beta$ , the uniformly most powerful test is simply based on the statistic  $\hat{\beta}_{\text{long}}$ , and the uniformly most powerful unbiased test is based on the statistic  $|\hat{\beta}_{\text{long}}|$ . By Pratt (1961), the inversion of these uniformly most powerful tests yield confidence intervals of minimal expected length: Let  $(\infty, U(\mathbf{Y}))$  be a confidence interval obtained from inverting one-sided tests of the form  $H_0 : \beta \geq \beta_0$  against  $H_a : \beta < \beta_0$ . For a given realization  $\mathbf{Y}$  under  $\beta$ , the excess length of this interval is  $\max(U(\mathbf{Y}) - \beta, 0) = \int_{\beta}^{\infty} (1 - \varphi_{\beta_0}(\mathbf{Y})) d\beta_0$ . By Tonelli’s Theorem,  $E_{\beta} \left[ \int_{\beta}^{\infty} (1 - \varphi_{\beta_0}(\mathbf{Y})) d\beta_0 \right] = \int_{\beta}^{\infty} E_{\beta}[1 - \varphi_{\beta_0}(\mathbf{Y})] d\beta_0$ , and the integrand on the right hand side is minimized by a family of uniformly most powerful tests, indexed by  $\beta_0$ . Similarly, for a two-sided test, the length of the resulting confidence interval can be written as

$\int (1 - \varphi_{\beta_0}(\mathbf{y})) d\beta_0$ , so we obtain

$$E_{\beta} \left[ \int (1 - \varphi_{\beta_0}(\mathbf{y})) d\beta_0 \right] = \int (1 - E_{\beta}[\varphi_{\beta_0}(\mathbf{y})]) d\beta_0$$

and the inversion of uniformly most powerful unbiased tests thus yield the confidence interval of shortest expected length among all unbiased confidence intervals. In the Gaussian model, no procedure whatsoever can therefore do better than simply running the “long regression” that includes all controls in a well defined sense.

## 2.2 Bivariate Inference Problem

In order to exploit the bound (2) for more informative inference, consider the coefficient estimator  $\hat{\beta}_{\text{short}}$  from the regression of  $\mathbf{y}$  on  $(\mathbf{x}, \mathbf{Q})$  that excludes the additional controls  $\mathbf{Z}$ . Since  $\mathbf{Q}'\mathbf{x} = 0$ ,  $\hat{\beta}_{\text{short}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ . Let  $\rho^2 = \mathbf{x}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}/(\mathbf{x}'\mathbf{x})$ , the  $R^2$  of a regression of  $\mathbf{x}$  on  $\mathbf{Z}$ . To avoid trivial complications in notation, assume  $0 < \rho$  in the following. Straightforward algebra yields

$$\begin{pmatrix} \hat{\beta}_{\text{long}} \\ \hat{\beta}_{\text{short}} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \beta \\ \beta + \Delta \end{pmatrix}, (\mathbf{x}'\mathbf{x})^{-1} \begin{pmatrix} \frac{1}{1-\rho^2} & 1 \\ 1 & 1 \end{pmatrix} \right) \quad (5)$$

where  $\Delta = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}\gamma$  is the omitted variable bias. Equation (5) is intuitive: the long regression provides an unbiased signal  $\hat{\beta}_{\text{long}}$  about  $\beta$ , but with a variance that is larger than the (typically biased) signal  $\hat{\beta}_{\text{short}}$  from the short regression. If  $\rho \rightarrow 0$ , then  $\mathbf{Z}$  is orthogonal to  $\mathbf{x}$ , there is no bias from the short regression, and the two signals are identical,  $\hat{\beta}_{\text{long}} = \hat{\beta}_{\text{short}}$ .

Notice that  $n\kappa^2 = \gamma'\mathbf{Z}'\mathbf{Z}\gamma$  in (2) may be rewritten as

$$\begin{aligned} n\kappa^2 &= \gamma'\mathbf{Z}'\mathbf{x}(\mathbf{x}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}\gamma + \gamma'\mathbf{Z}'\mathbf{M}_{\rho}\mathbf{Z}\gamma \\ &= \rho^{-2}(\mathbf{x}'\mathbf{x})\Delta^2 + \gamma'\mathbf{Z}'\mathbf{M}_{\rho}\mathbf{Z}\gamma \end{aligned} \quad (6)$$

where  $\mathbf{M}_{\rho} = \mathbf{I}_p - \mathbf{x}(\mathbf{x}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x})^{-1}\mathbf{x}' = \mathbf{I}_p - \rho^{-2}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ . The bound  $\kappa^2 \leq \bar{\kappa}^2$  in (2) thus implies an upper bound on the omitted variable bias,

$$|\Delta| \leq \rho\bar{\kappa}/\sqrt{\mathbf{x}'\mathbf{x}/n} \quad (7)$$

and this bound is sharp. This limit on the magnitude of the omitted variable bias in (5) makes  $\hat{\beta}_{\text{short}}$  potentially valuable for inference about  $\beta$ , especially if  $\rho$  is close to one (so that  $\hat{\beta}_{\text{short}}$  is much less variable than  $\hat{\beta}_{\text{long}}$ ).

We focus in the following on tests of  $H_0 : \beta = 0$ , since the general case  $H_0 : \beta = \beta_0$  may be reduced to this case by subtracting  $\beta_0$  from  $\hat{\beta}_{\text{long}}$  and  $\hat{\beta}_{\text{short}}$ . In terms of the localized parameters  $b = \beta\sqrt{\mathbf{x}'\mathbf{x}}$ ,  $d = \rho^{-1}\Delta\sqrt{\mathbf{x}'\mathbf{x}}$  and  $\bar{k} = \sqrt{n}\bar{k}$ , the inference problem then becomes testing  $H_0 : b = 0$  from observing the bivariate normal vector  $\hat{\mathbf{b}} = (\hat{b}_{\text{long}}, \hat{b}_{\text{short}})' \sim (\sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}_{\text{long}}, \sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}_{\text{short}})'$ ,

$$\begin{pmatrix} \hat{b}_{\text{long}} \\ \hat{b}_{\text{short}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} b \\ b + \rho d \end{pmatrix}, \Sigma(\rho)\right), \Sigma(\rho) = \begin{pmatrix} \frac{1}{1-\rho^2} & 1 \\ 1 & 1 \end{pmatrix}, |d| \leq \bar{k}. \quad (8)$$

The inference problem (8) is a fairly transparent small sample problem indexed by two known parameters  $(\rho, \bar{k}) \in [0, 1) \times [0, \infty)$ , and involves a one-dimensional unknown nuisance parameter  $d \in \mathbb{R}$ . The second observation  $\hat{b}_{\text{short}}$  augments the usual Gaussian shift experiment, and there are a variety of potential approaches to exploiting this additional information. We found that a simple but effective test of  $H_0 : b = 0$  is generated by the generalized likelihood ratio statistic

$$\begin{aligned} \text{LR}(\bar{k}) &= \min_{|d| \leq \bar{k}} \begin{pmatrix} \hat{b}_{\text{long}} \\ \hat{b}_{\text{short}} - \rho d \end{pmatrix}' \Sigma(\rho)^{-1} \begin{pmatrix} \hat{b}_{\text{long}} \\ \hat{b}_{\text{short}} - \rho d \end{pmatrix} \\ &\quad - \min_{b, |d| \leq \bar{k}} \begin{pmatrix} \hat{b}_{\text{long}} - b \\ \hat{b}_{\text{short}} - b - \rho d \end{pmatrix}' \Sigma(\rho)^{-1} \begin{pmatrix} \hat{b}_{\text{long}} - b \\ \hat{b}_{\text{short}} - b - \rho d \end{pmatrix}. \end{aligned} \quad (9)$$

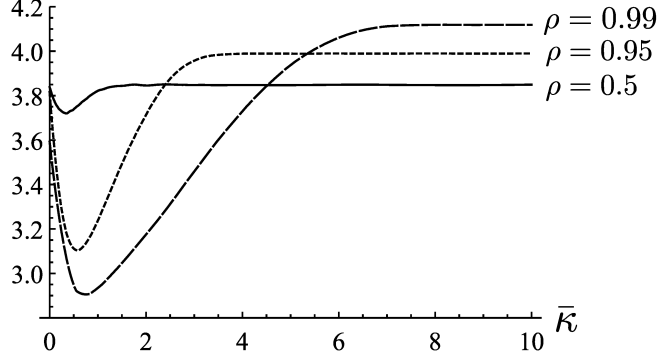
Under larger and larger bounds  $\bar{k} \rightarrow \infty$ , the limit of the inference problem (8) depends on the relative magnitude of  $d$ . In particular, suppose that as  $\bar{k} \rightarrow \infty$ , for some given sequence  $s$ ,  $d - s = a$ . Then after recentering  $\hat{b}_{\text{short}}^o$  by  $\rho s$  via  $\hat{\mathbf{b}}^o = (\hat{b}_{\text{long}}, \hat{b}_{\text{short}}^o)' \sim (\hat{b}_{\text{long}}, \hat{b}_{\text{short}} - \rho s)'$ , we obtain

$$\hat{\mathbf{b}}^o \sim \mathcal{N}\left(\begin{pmatrix} b \\ b + \rho a \end{pmatrix}, \Sigma(\rho)\right) \quad (10)$$

and the constraint on  $a$  resulting from  $|d| \leq \bar{k}$  depends on the relationship between  $s$  and  $\bar{k}$ . In particular, with  $s = \bar{k}$ ,  $a \in A_1 = (-\infty, 0]$ , and this corresponds to the case where the bound  $\bar{k}$  is very large, but  $d$  is positive and close to the bound. Similarly, with  $s = -\bar{k}$ ,  $d$  is negative and close to  $-\bar{k}$ , and the corresponding constraint in (10) becomes  $a \in A_{-1} = [0, \infty)$ . Finally, if  $\bar{k} \rightarrow \infty$  and  $\bar{k} - |s| \rightarrow \infty$ , so that the bound  $\bar{k}$  is much larger than  $|d|$ , then  $a \in A_0 = \mathbb{R}$  is unrestricted in (10). For each of these three cases  $i \in \{1, -1, 0\}$ , the  $\text{LR}(\bar{k})$  statistic converges to

$$\text{LR}_i^o = \min_{a \in A_i} \begin{pmatrix} \hat{b}_{\text{long}} \\ \hat{b}_{\text{short}}^o - \rho a \end{pmatrix}' \Sigma(\rho)^{-1} \begin{pmatrix} \hat{b}_{\text{long}} \\ \hat{b}_{\text{short}}^o - \rho a \end{pmatrix}$$

Figure 3: Five percent critical value of  $\text{LR}(\bar{k})$  as a function of  $\bar{k}$



$$= \min_{b, a \in A_i} \begin{pmatrix} \hat{b}_{\text{long}} - b \\ \hat{b}_{\text{short}}^o - \rho a - b \end{pmatrix}' \Sigma(\rho)^{-1} \begin{pmatrix} \hat{b}_{\text{long}} - b \\ \hat{b}_{\text{short}}^o - \rho a - b \end{pmatrix}.$$

Figure 3 plots the 5% level critical value  $\text{cv}_\rho(\bar{k})$  of  $\text{LR}(\bar{k})$  that is just large enough to ensure size control for all values of  $|d| \leq \bar{k}$  for  $\rho \in \{0.5, 0.95, 0.99\}$ , and Figure 4 plots the rejection region of the resulting 5% level test for  $\rho = 0.95$  and  $\bar{k} \in \{0, 1, 3, 10\}$ . For  $\bar{k} = 0$ , the LR test reduces to rejecting for large values of  $(\hat{b}_{\text{short}})^2 > \text{cv}_\rho(0) = 1.96^2$ , that is it reduces to the usual t-test based on the short regression. More generally, whenever  $|\hat{b}_{\text{short}}| \gg \bar{k}$ , that is the short regression coefficient value is much larger than  $\bar{k}$  in absolute value, then the LR test rejects. On the other hand, for  $|\hat{b}_{\text{short}}| \ll \bar{k}$  and  $\bar{k}$  large, the LR test rejects when  $(1 - \rho^2)(\hat{b}_{\text{long}})^2 > \text{cv}_\rho(\bar{k})$ , that is whenever the long regression coefficient is too large in absolute value, with a critical value that is slightly larger than what one would employ in a standard chi-squared test with one degree of freedom. Once  $\bar{k}$  is moderately large (say, larger than 8), the critical value  $\text{cv}_\rho(\bar{k})$  stabilizes at  $\text{cv}_\rho(\infty)$ , and further increases of  $\bar{k}$  simply amount to an additional outward shift of the acceptance region. Indeed, after a recentering by  $s \in \{\bar{k}, -\bar{k}\}$ , this acceptance region is equal to the acceptance region of a test based on  $\text{LR}_i^o$ ,  $i \in \{-1, 1\}$ , whose smallest valid critical value also equals  $\text{cv}_\rho(\infty)$ . The  $\text{LR}_i^o$  test is hence simply the smooth extension of the fixed  $\bar{k}$  test based on (8) to the  $\bar{k} \rightarrow \infty$  problems (10). In the unconstrained test of observing (10) with  $a \in A_0 = \mathbb{R}$ , the  $\text{LR}_0^o$  based test rejects for large values of  $(1 - \rho^2)(\hat{b}_{\text{long}})^2$ , and relying on  $\text{cv}_\rho(\infty)$  instead of  $1.96^2$  induces a slight power loss.

Note that the inversion of the LR statistic for general null hypotheses  $H_0 : b = b_0$  yields a confidence interval for  $b$  that is translation equivariant, that is, the interval obtained from

Table 1: Properties of 95% Bivariate LR Inference

Panel A: Weighted Expected Length of CI for  $b$  under  $d \sim U[-\bar{k}, \bar{k}]$

$\rho \backslash \bar{k}$	LR( $\bar{k}$ ) Interval					Lower Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	3.9	4.2	4.4	4.5	4.5	3.9	4.2	4.4	4.5	4.5
0.90	3.9	5.0	7.1	8.5	8.9	3.9	5.0	6.9	8.2	8.7
0.99	3.9	5.3	9.1	18.7	25.3	3.9	5.2	8.9	17.4	23.9

Panel B: Expected Length of CI for  $b$ , Maximized over  $|d| \leq \bar{k}$

$\rho \backslash \bar{k}$	LR( $\bar{k}$ ) Interval					Lower Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	3.9	4.3	4.5	4.5	4.6	3.9	4.2	4.4	4.5	4.5
0.90	3.9	5.0	7.4	9.1	9.1	3.9	5.0	7.1	8.4	8.9
0.99	3.9	5.3	9.1	20.0	28.7	3.9	5.2	8.9	18.4	25.1

Panel C: Ratio of Expected Length of LR CI for  $b$  Relative to Long Regression Interval

$\rho \backslash \bar{k}$	Minimized over $ d  \leq \bar{k}$					Maximized over $ d  \leq \bar{k}$				
	0	1	3	10	30	0	1	3	10	30
0.50	0.87	0.92	0.93	0.92	0.94	0.87	0.94	1.00	1.01	1.03
0.90	0.43	0.55	0.72	0.73	0.73	0.43	0.56	0.82	1.01	1.02
0.99	0.14	0.19	0.33	0.59	0.61	0.14	0.19	0.33	0.72	1.03

Panel D: Median of  $\bar{k}_\phi^*$  under  $b = 0$ ,  $P(d = d_0) = P(d = -d_0) = 1/2$

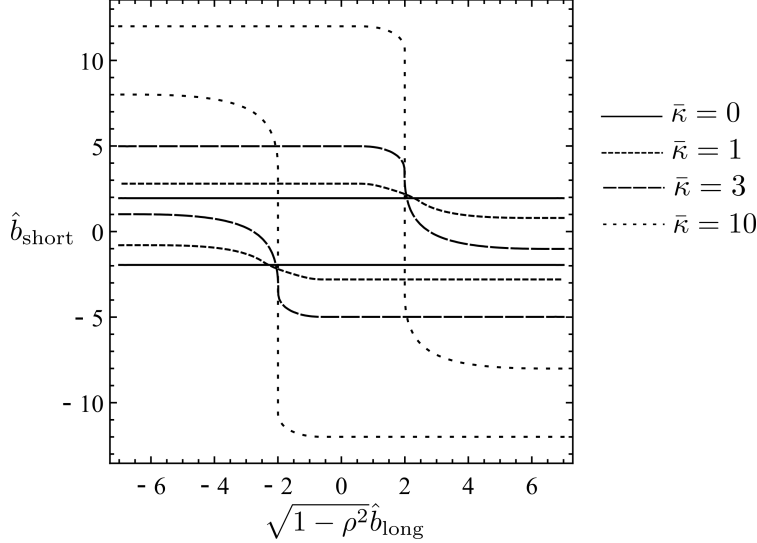
$\rho \backslash d_0$	$\bar{k}_{\text{LR}}^*$					Upper Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	0.0	0.0	0.0	3.1	13.2	0.0	0.0	0.7	4.2	14.3
0.90	0.0	0.0	0.9	7.1	25.3	0.0	0.0	1.2	7.6	25.8
0.99	0.0	0.0	1.3	8.0	28.0	0.0	0.0	1.4	8.4	28.4

Panel E: Weighted Average MSE of Equivariant Estimators of  $b$  under  $d \sim U[-\bar{k}, \bar{k}]$

$\rho \backslash \bar{k}$	$\hat{b}_{\text{LR}}$					Lower Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	1.00	1.11	1.25	1.31	1.32	1.00	1.07	1.22	1.30	1.32
0.90	1.00	1.29	2.62	4.42	4.98	1.00	1.25	2.53	4.38	4.97
0.99	1.00	1.33	3.79	21.1	39.9	1.00	1.32	3.77	20.4	39.7

Notes: Lower Bounds in Panels A, B, D and E are numerically determined using the algorithm in Elliott, Müller, and Watson (2015) and Müller and Wang (2015), and impose translation equivariance in Panels A, B and E (cf. Müller and Norets (2012)). Based on 500,000 importance sampling draws.

Figure 4: Acceptance region of  $\text{LR}(\bar{\kappa})$  for  $\rho = 0.95$



Notes: The lines are the boundaries of the acceptance region. For all values of  $\bar{\kappa}$ ,  $(0, 0)$  is in the acceptance region.

the observation  $(\hat{b}_{\text{long}} + c, \hat{b}_{\text{short}} + c)$  simply shifts the interval from  $(\hat{b}_{\text{long}}, \hat{b}_{\text{short}})$  by  $c$ , for any  $c \in \mathbb{R}$ . We consider the LR approach attractive for a number of reasons. First, it is easy to implement (we discuss implementation issues in more detail in Section 3 below). Second, it yields confidence intervals that are close to minimal weighted expected length under a weighting function where  $d$  is uniform between  $[-\bar{k}, \bar{k}]$ , among all translation equivariant confidence intervals. This is shown in panel A of Table 1, which reports an upper bound on this weighted expected length for selected values of  $(\rho, \bar{k})$ , along with the weighted expected length of the LR interval. Given the tight link between the power of tests and their expected length discussed in Section 2.1 above, this implies that the LR tests are also close to maximizing the corresponding weighted average power. Third, as shown in panel B, it is reasonably close to being maximin in terms of expected length among all equivariant confidence intervals. Fourth, its expected length is nearly uniformly shorter over all  $|d| \leq \bar{k}$  than the standard long regression interval; panel C provides corresponding numerical evidence. And finally, the LR approach has the potentially attractive feature that if  $|\hat{b}_{\text{short}}| > 1.96$  and  $\sqrt{1 - \rho^2}|\hat{b}_{\text{long}}| < 1.96$  (that is, the short regression rejects, but the long regression doesn't), then there is a unique threshold value  $\bar{k}^* > 0$  such that the LR test rejects only when  $\bar{k} < \bar{k}^*$ , so in this sense, imposing a smaller value of  $\bar{k}$  always leads to more informative inference.

The LR approach yields a continuum of confidence intervals for  $\beta = b/\sqrt{\mathbf{x}'\mathbf{x}}$ , indexed

by  $\bar{\kappa} = \bar{k}/\sqrt{n}$ , as plotted in Figure 1 of the introduction. The properties discussed so far concern the length of the interval for  $b$  for a given  $\bar{k}$ , the vertical distance between the confidence interval bounds. But in the context of the threshold value  $\bar{k}^* = \sqrt{n}\bar{\kappa}^*$ , it is also interesting to consider the horizontal distance between 0 and  $\bar{k}^*$  for  $b = 0$ , that is the ability of tests to indicate that  $H_0 : b = 0$  is empirically incompatible with small values of  $\bar{k}$ . Let  $\phi(\bar{k}, \hat{\mathbf{b}}) \in \{0, 1\}$  be a family of level  $\alpha$  tests of  $H_0 : b = 0$  in (8), indexed by  $\bar{k}$ . Then  $\bar{k}_\phi^* : \mathbb{R}^2 \mapsto [0, \infty) \cup \{+\infty\}$  is defined as

$$\bar{k}_\phi^*(\hat{\mathbf{b}}) = \inf_{\bar{k}} \{\bar{k} : \phi(\bar{k}, \hat{\mathbf{b}}) = 0\}$$

that is,  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  is the smallest value of  $\bar{k}$  for which the test  $\phi(\bar{k}, \hat{\mathbf{b}})$  does not reject. From this alternative perspective, one might prefer tests  $\phi$  that generate large  $\bar{k}_\phi^*(\hat{\mathbf{b}})$ . As  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  can be equal to  $+\infty$ , it is not sensible to maximize the expectation of  $\bar{k}_\phi^*(\hat{\mathbf{b}})$ . Instead, consider a quantile of  $\bar{k}_\phi^*(\hat{\mathbf{b}})$ , such as its median. Since  $\phi(\bar{k}, \hat{\mathbf{b}})$  is a level  $\alpha$  test of  $H_0 : b = 0$ ,  $|d| \leq \bar{k}$ , the  $1 - \alpha$  quantile of  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  must be smaller than  $|d|$  under  $b = 0$ , and  $[\bar{k}_\phi^*(\hat{\mathbf{b}}), \infty)$  is thus a  $1 - \alpha$  confidence interval for  $|d|$  under  $b = 0$ . This constrains the possibility of making the median of  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  arbitrarily large. In panel D of Table 1 we report the median of  $\bar{k}_{\text{LR}}^*$  of the LR tests under  $b = 0$  and  $P(d = d_0) = P(d = -d_0) = 1/2$  for various  $d_0$ , along with an upper bound that holds for all  $\bar{k}_\phi^*(\hat{\mathbf{b}})$ .<sup>3</sup> We find that unless  $|d|$  is very small, the median of  $\bar{k}_{\text{LR}}^*$  is only slightly smaller than the upper bound, and unreported results show this to hold also for other quantiles and assumptions about the distribution of  $b$ . The LR approach thus also performs well in the sense that it is nearly as informative as possible about values of  $\bar{k}$  that are empirically incompatible with  $b = 0$ ,  $|d| \leq \bar{k}$ .

Finally, consider the problem of improving the estimation of  $b$  under the constraint (2). Our suggested estimator is  $\hat{b}_{\text{LR}}(\bar{k})$ , the center of the 95% confidence interval constructed by inverting the  $\text{LR}(\bar{k})$  statistic. As shown in panel E of Table 1, this estimator comes close to minimizing the weighted average mean square error among all equivariant estimators of  $b$ , with a weighting function on  $d$  that is uniform on  $[-\bar{k}, \bar{k}]$ . (Unreported results show that the center of the 95% interval does particularly well compared to other levels). One might think that the maximum likelihood estimator of  $b$  in (8) under  $|d| \leq \bar{k}$  is a more natural

---

<sup>3</sup>Note that existence of an estimator  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  with median larger than  $M$  under some distribution  $F$  for  $(b, d)$  is equivalent to the existence of a test  $\phi_M(\hat{\mathbf{b}}) \in \{0, 1\}$  such that  $E[\phi_M(\hat{\mathbf{b}})] \leq \alpha$  for all  $b = 0$ ,  $|d| \leq M$  and  $E[\phi_M(\hat{\mathbf{b}})] \geq 1/2$  with  $(b, d) \sim F$ , since we can always set  $\phi_M(\hat{\mathbf{b}}) = \mathbf{1}[\bar{k}_\phi^*(\hat{\mathbf{b}}) \geq M]$  or  $\bar{k}_\phi^*(\hat{\mathbf{b}}) = M\phi_M(\hat{\mathbf{b}})$ , respectively. The upper bound can therefore be obtained from the upper bound of the power of tests in Elliott, Müller, and Watson (2015).

estimator; but unreported results show that the MLE is much more variable, resulting in a substantially larger mean squared error compared to  $\hat{b}_{\text{LR}}(\bar{k})$ .

As mentioned in the introduction, the problem (8) falls into the general class considered by Armstrong and Kolesár (2016). They construct fixed-length confidence intervals for  $b$  that are minimax among all fixed length confidence intervals centered at a linear estimator of  $b$ . Table 3 in the appendix is the analogue of Table 1 for their confidence interval, and implied estimators  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  and midpoint  $\hat{b}_\phi(\bar{k})$ . Comparing the tables reveals that the LR approach never does substantially worse, but in some dimensions does substantially better: The LR approach can yield much shorter intervals, it leads to much larger  $\bar{k}_\phi^*$ , and it has lower weighted average MSE, especially for large  $\bar{k}$ .

## 2.3 Asymptotic Efficiency of Bivariate Inference

Regardless how exactly they are constructed, confidence intervals about  $\beta$  obtained from the bivariate observation  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})'$  will in general be shorter than those based on  $\hat{\beta}_{\text{long}}$  alone. But this does not mean that they necessarily fully exploit the information in the bound (2). After all, the reduction to the bivariate problem (5) was not based on any sufficiency argument. So the question arises whether one can do systematically better than what can be achieved in the bivariate problem (8).

In general, the distribution of tests and confidence intervals about  $\beta$  that are a function of the entire set of observations  $\mathbf{Y}$  not only depends on  $\beta$ , the bias  $\Delta = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Z}\boldsymbol{\gamma}$  of the short regression and the slackness in the inequality (7)  $\tau^2 = \kappa^2 - \rho^{-2}(\mathbf{x}'\mathbf{x})\Delta^2/n = \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_\rho\mathbf{Z}\boldsymbol{\gamma}/n$ , but also of the direction of  $\boldsymbol{\gamma}$  that leads to identical values of  $\Delta$  and  $\tau$ . Consider the linear reparameterization  $\mathbf{Z}\boldsymbol{\gamma} = \mathbf{Z}^r\boldsymbol{\gamma}^r$  with  $\boldsymbol{\gamma}^r = (\mathbf{Z}'\mathbf{Z})^{1/2}\boldsymbol{\gamma}$  and  $\mathbf{Z}^r = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2}$ , and corresponding estimator  $\hat{\boldsymbol{\gamma}}^r = (\mathbf{Z}'\mathbf{Z})^{1/2}\hat{\boldsymbol{\gamma}}$ . Further, let  $\mathbf{P}_{xZ^r}$  be a  $p \times (p-1)$  matrix such that  $\mathbf{P}_{xZ^r}'\mathbf{Z}^{r'}\mathbf{x} = 0$  and  $\mathbf{P}_{xZ^r}'\mathbf{P}_{xZ^r} = \mathbf{I}_{p-1}$ . Then with  $\hat{\boldsymbol{\phi}} = \mathbf{P}_{xZ^r}'\hat{\boldsymbol{\gamma}}^r$ , it follows from (4) that

$$\hat{\boldsymbol{\xi}} = \begin{pmatrix} \hat{\beta}_{\text{long}} \\ \hat{\beta}_{\text{short}} \\ \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \beta \\ \beta + \Delta \\ \boldsymbol{\delta} \\ \sqrt{n}\tau\boldsymbol{\omega} \end{pmatrix}, (\mathbf{x}'\mathbf{x})^{-1} \begin{pmatrix} \frac{1}{1-\rho^2} & 1 & \mathbf{0} & \mathbf{0} \\ 1 & 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\mathbf{x}'\mathbf{x})(\mathbf{Q}'\mathbf{Q})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (\mathbf{x}'\mathbf{x})\mathbf{I}_{p-1} \end{pmatrix} \right) \quad (11)$$

where  $\boldsymbol{\omega} = \mathbf{P}_{xZ^r}'\boldsymbol{\gamma}^r / \|\mathbf{P}_{xZ^r}'\boldsymbol{\gamma}^r\| = n^{-1/2}\mathbf{P}_{xZ^r}'\boldsymbol{\gamma}^r/\tau$ . The parameter  $\boldsymbol{\omega}$  is an element of the surface of the  $p-1$  dimensional unit hypersphere and indicates the direction of  $\boldsymbol{\gamma}^r$  in the  $p-1$  dimensional subspace orthogonal to  $\mathbf{Z}^{r'}\mathbf{x}$ .

As noted before, by sufficiency, it suffices to consider functions of  $\hat{\xi}$ . Given the decomposition of  $\kappa^2$  in (6), it would clearly be beneficial to know the value of  $\tau^2 = \gamma' \mathbf{Z}' \mathbf{M}_\rho \mathbf{Z} \gamma / n$  for inference about  $\beta$ , as it would allow the strengthening of the bound on  $\Delta$  under (2) to

$$|\Delta| \leq \rho \sqrt{\bar{\kappa}^2 - \tau^2} / \sqrt{\mathbf{x}' \mathbf{x} / n}. \quad (12)$$

Since  $\hat{\phi}$  contains information about  $\tau$ , it thus seems that one can do better than restricting attention to tests that are a solely a function of  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})'$ .

A first observation concerns maximin properties, and follows from the logic in Donoho (1994): Consider the risk of a set estimator for  $\beta$  constructed from  $\hat{\xi}$ . Surely, if we were told the value of  $(\delta, \tau\omega)$ , the maximal risk can only decrease. In other words, the maximal risk of procedures that treat  $(\delta, \tau\omega)$  as known is a lower bound for the maximal risk in the original problem, for any  $(\delta, \tau\omega)$ . Furthermore, with  $(\delta, \tau\omega)$  known, it is evident from (11) that we may ignore  $(\hat{\delta}, \hat{\phi})$ , that is, for any loss function, the maximin procedure in the resulting problem can be written as a function of  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})$  alone. In particular, this holds for  $(\delta, \tau\omega) = \mathbf{0}$ . Thus, the maximal risk of the bivariate procedures that is optimal for  $(\delta, \tau\omega) = \mathbf{0}$  known is a lower bound on the overall risk. But it is also an upper bound on overall maximal risk, since for  $(\delta, \tau\omega) \neq \mathbf{0}$ , the risk of the bivariate procedure can only decrease, since it can only lead to a lower bound (12) if  $\tau > 0$ . Thus, all maximin procedures can be written as a function of  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})$ . To the extent that the bivariate LR confidence set is numerically close to being maximin in terms of expected length (cf. Panel B of Table 1), it is therefore also approximately maximin among all confidence sets that are functions of  $\mathbf{Y}$ .

This is a noteworthy result, but there is the usual concern that the maximin criterion is inherently too pessimistic: One might well be willing to give up a little bit of worst-case expected length in return for much smaller expected lengths in other parts of the parameter space. We now establish a further result that limits the possibility of such “adaption”, although that additional result only holds for small values of  $\tau$ .

We focus on tests  $\varphi(\bar{\kappa}, \mathbf{Y}) \in [0, 1]$  of  $H_0 : \beta = 0$ , where values between zero and one indicate the probability of rejection, so that a non-randomized test has range  $\{0, 1\}$ . In the parameterization  $\xi = (\beta, \Delta, \delta, \tau, \omega)$ , the rejection probability of  $\varphi$  is  $E_\xi[\varphi(\bar{\kappa}, \mathbf{Y})]$ . In absence of any information about the controls  $(\mathbf{Q}, \mathbf{Z})$  beyond (2), it seems natural to consider tests whose rejection probability does not depend on the baseline coefficients  $\delta$ , or the direction  $\omega$ . Otherwise, the ability of the test to reject would necessarily be higher for some values of  $(\delta, \omega)$  compared to others, which only makes substantive sense in the presence of some a priori information about  $(\delta, \omega)$ .

The following lemma shows that for any such test, there exists another test with the same rejection probability that is a function of the three dimensional statistic  $\mathbf{T} = (\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, \hat{\tau})'$ , where

$$\hat{\tau}^2 = \hat{\gamma}' \mathbf{Z}' \mathbf{M}_p \mathbf{Z} \hat{\gamma} / n = \hat{\phi}' \hat{\phi} / n.$$

Note that the distribution of  $\mathbf{T}$  only depends on  $(\beta, \Delta, \tau)$ . Thus, to the extent that one is willing to restrict attention to tests whose power function is symmetric in this sense, one might focus on tests that are functions of  $\mathbf{T}$ , with an effective parameter space equal to  $\theta = (\beta, \Delta, \tau) \in \mathbb{R}^2 \times [0, \infty)$ .

**Lemma 1** *For given  $\bar{\kappa}$  and any  $n > p + m$ , if  $E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})]$  does not vary in  $(\delta, \omega)$  for all  $(\beta, \Delta, \tau)$ , then there exists a test  $\tilde{\varphi} : \mathbb{R}^3 \mapsto [0, 1]$  such that  $E_{\xi}[\tilde{\varphi}(\mathbf{T})] = E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})]$  for all  $\xi$ .*

For the observation  $\hat{\tau}^2$  to be useful to obtain a sharper bound (12), the estimation error in  $\hat{\tau}^2$  must not be too large relative to  $\bar{\kappa}^2$ . The following Lemma shows that in large samples,  $\hat{\tau}$  does not contain useful information about  $\tau$  as long as  $\tau$  is not too large. From now on, we use subscripts to denote the value of quantities and functions that depend on the sample size  $n$ .

**Lemma 2** *Let  $L_n(\tau)$  be the likelihood of  $\tau$  based on the observation  $\hat{\tau}_n$  in the regression model (1) with  $n$  observations and  $\varepsilon_i \sim \text{iid}\mathcal{N}(0, 1)$ . If  $p_n/n \rightarrow c \in (0, 1)$ ,  $\tau_n = o(n^{-1/4})$  and  $t_n = o(n^{-1/4})$ , then  $L_n(t_n)/L_n(0) \xrightarrow{p} 1$ .*

The lemma shows that even the likelihood ratio statistic for the observation  $\hat{\tau}_n$  does not drive an asymptotic wedge between the values  $\tau_n = 0$  and  $\tau_n = o(n^{-1/4})$ , suggesting that  $\mathbf{T}_n$  does not help to determine the value of  $\tau_n$  of order  $o(n^{-1/4})$ . Combining the observations in Lemmas 1 and 2 with limit of experiments arguments leads to the following result.

**Theorem 1** *Consider a sequence of observations from the linear regression model with  $\varepsilon_i \sim \text{iid}\mathcal{N}(0, 1)$  where  $p_n/n \rightarrow c \in (0, 1)$  and  $\rho_n^2 \rightarrow \rho^2 \in [0, 1)$ , and let  $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$  be a sequence of tests that, for all sufficiently large  $n$ , satisfy the assumption of Lemma 1. If for some sequence  $s_n$  and all  $(b, a) \in \mathbb{R}^2$ ,  $E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)]$  converges along a sequence  $\theta_n$  with  $(\sqrt{\mathbf{x}'_n \mathbf{x}_n} \beta_n, \sqrt{\mathbf{x}'_n \mathbf{x}_n} \Delta_n - s_n) = (b, a)$  and  $\tau_n = o(n^{-1/4})$  (where  $\tau_n$  may depend on  $(b, a)$ ), then  $\lim_{n \rightarrow \infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,a}[\phi(\hat{\mathbf{b}}^o)]$ , with  $\hat{\mathbf{b}}^o$  distributed as in (10). Furthermore,  $\lim_{n \rightarrow \infty} E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] = E_{b,a}[\phi(\hat{\mathbf{b}}^o)]$  then holds under all sequences  $\theta_n$  with  $(\sqrt{\mathbf{x}'_n \mathbf{x}_n} \beta_n, \sqrt{\mathbf{x}'_n \mathbf{x}_n} \Delta_n - s_n) = (b, a)$  and  $\tau_n = o(n^{-1/4})$ .*

The theorem allows for  $\sqrt{n}\bar{\kappa}_n \rightarrow k_0$  and  $s_n = 0$ , so the localized problem (8) initially considered in Section 2.2 is covered as a special case with  $a = d$ , and the case with  $|s_n| \rightarrow \infty$  and  $\sqrt{n}\bar{\kappa}_n - |s_n| \rightarrow 0$  or  $\sqrt{n}\bar{\kappa}_n - |s_n| \rightarrow \infty$  correspond to the  $|\bar{k}| \rightarrow \infty$  cases discussed below (9). The theorem thus demonstrates that under  $\tau_n = o(n^{-1/4})$  the asymptotic power function of any test satisfying the condition of Lemma 1 can always be matched by the power function of a bivariate test that depends on the data only through the short and long regression coefficient estimators. For instance, with  $\tau_n = t_0/\sqrt{n}$  and  $\bar{\kappa}_n = \bar{k}_0/\sqrt{n}$  for fixed  $(t_0, \bar{k})$ , knowledge of  $\tau_n$  is helpful, as it reduces the bound on  $d$  in the corresponding localized problem (8) from  $|d| < \bar{k}$  to  $|d| \leq \sqrt{\bar{k}^2 - t_0^2}$ . Theorem 1 thus doesn't provide conditions under which knowledge of  $\tau_n$  is irrelevant; rather, it shows that even though if  $\tau_n$  is potentially helpful, the data is not informative about its value, so the asymptotically best thing to do is to make inference a function of the short and long regression coefficient estimators. The determination of inference procedures with attractive asymptotic power properties is hence reduced to the problem of identifying good bivariate inference, as discussed in the last subsection.

The implementation of asymptotically valid bivariate tests is straightforward in the Gaussian homoskedastic model. In particular, the test

$$\varphi_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n) = \mathbf{1}[\text{LR}_n(\sqrt{n}\bar{\kappa}_n) > \text{cv}_{\rho_n}(\sqrt{n}\bar{\kappa}_n)] \quad (13)$$

with  $\text{LR}_n(\bar{k})$  equal to (9) and  $\hat{\mathbf{b}} = (\sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}_{\text{long}}, \sqrt{\mathbf{x}'\mathbf{x}}\hat{\beta}_{\text{short}})'$  and  $\text{cv}_{\rho}(\bar{k})$  as defined in Section 2.2 has asymptotic rejection probability equal to the small sample rejection probability of the LR test discussed there. The attractive properties of the LR approach among all bivariate tests thus translate via Theorem 1 into attractive asymptotic properties in the Gaussian homoskedastic model among the larger class of tests that are only required to satisfy Lemma 1.

A bound of the order  $\bar{\kappa} = o(n^{-1/4})$  implies  $\tau_n = o(n^{-1/4})$  under (2). As discussed in the introduction, this suggests that the efficiency implication of Theorem 1 is particularly pertinent in small samples for values of  $\bar{\kappa}_n$  that imply that the explanatory power of the controls is small compared to the variation in  $y_i$ .

Note, however, that the theorem does not require  $\tau_n$  to be smaller than  $\bar{\kappa}_n$ ; for instance  $\sqrt{n}\bar{\kappa}_n$  may converge, while  $\sqrt{n}\tau_n$  diverges. This corresponds to a situation where the bound  $\bar{\kappa}_n$  is much smaller than the actual value  $\kappa_n$ . This is most easily interpreted along the lines discussed at the end of Section 2.2 above: The limited information in the data may make it impossible correctly exclude such values along the  $\beta = 0$  line in Figure 1, but one would still prefer a procedure that excludes as many as possible. Specifically, assuming that

$\varphi_n(\bar{\kappa}, \mathbf{Y}_n)$  is not randomized, one would prefer the threshold value  $\bar{\kappa}^*(\mathbf{Y}_n) \in [0, \infty) \cup \{+\infty\}$  defined via

$$\bar{\kappa}^*(\mathbf{Y}_n) = \inf_{\bar{\kappa}} \{\bar{\kappa} : \varphi_n(\bar{\kappa}, \mathbf{Y}_n) = 0\} \quad (14)$$

to be as large as possible, as discussed in Section 2.2, under the constraint that  $[0, \bar{\kappa}^*(\mathbf{Y}_n)]$  forms a  $1 - \alpha$  confidence set for  $|\Delta_n|$  under  $\beta = 0$ . We formulate a corresponding result in terms of a generic scalar estimator  $\psi_n(\mathbf{Y}_n)$ , which allows for potential recentering,  $\psi_n(\mathbf{Y}_n) = \bar{\kappa}^*(\mathbf{Y}_n) - c_n$ .

**Corollary 1** *In addition to the assumptions of Theorem 1, suppose  $\psi_n(\mathbf{Y}_n)$  has a distribution that depends on  $\boldsymbol{\xi}$  only through  $(\beta, \Delta, \tau)$  for all large enough  $n$ . If for some sequence  $s_n$ , and all  $(b, a) \in \mathbb{R}^2$ ,  $\psi_n(\mathbf{Y}_n)$  converges in distribution along a sequence  $\theta_n$  with  $(\sqrt{\mathbf{x}'_n \mathbf{x}_n} \beta_n, \sqrt{\mathbf{x}'_n \mathbf{x}_n} \Delta_n - s_n) = (b, a)$  and  $\tau_n = o(n^{-1/4})$  (where  $\tau_n$  may depend on  $(b, a)$ ), then the limit distribution is of the form  $\psi^o(\hat{\mathbf{b}}^o, U)$  for some function  $\psi^o : \mathbb{R}^3 \mapsto \mathbb{R} \cup \{+\infty\}$ , with  $\hat{\mathbf{b}}^o$  distributed as in (10), and  $U$  a uniform random variable on  $[0, 1]$  independent of  $\hat{\mathbf{b}}^o$ . Furthermore,  $\psi_n(\mathbf{Y}_n)$  then converges in distribution to  $\psi^o(\hat{\mathbf{b}}^o, U)$  under all sequences  $\theta_n$  with  $(\sqrt{\mathbf{x}'_n \mathbf{x}_n} \beta_n, \sqrt{\mathbf{x}'_n \mathbf{x}_n} \Delta_n - s_n) = (b, a)$  and  $\tau_n = o(n^{-1/4})$ .*

The corollary shows that the problem of constructing asymptotically attractive threshold estimators  $\bar{\kappa}^*(\mathbf{Y}_n)$  is effectively reduced to considering functions of  $\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}$ , and potentially an independent randomization device, as long as one restricts attention to  $\bar{\kappa}_n^*(\mathbf{Y}_n)$  whose distribution does not depend on the nuisance parameters  $(\boldsymbol{\delta}, \boldsymbol{\omega})$ . The local asymptotic properties of the threshold estimator (14) implied by the LR test (13) corresponds to the small sample properties of  $\bar{k}_\phi^*(\hat{\mathbf{b}})$  discussed at the end of Section 2.2, so similar to Theorem 1, their attractive features again extend to this larger class. And by setting  $\psi_n(\mathbf{Y}_n)$  in Corollary 1 equal to a potentially recentered and rescaled estimator of  $\beta$  that exploits the bound (2), the same holds for our suggested midpoint estimator  $\hat{\beta}_{\text{LR}}(\bar{\kappa})$ .

In summary, under  $p/n \rightarrow c \in (0, 1)$  asymptotics, as long as  $\tau_n = o(n^{-1/4})$ , the quality of asymptotic inference in the Gaussian homoskedastic model is limited from above by the performance of bivariate procedures. The attractive small sample features of the LR approach discussed in Section 2.2 thus translate into attractive large sample inference.

### 3 Implementation in Non-Gaussian and Potentially Heteroskedastic Models

In the Gaussian linear regression model, the bivariate tests introduced in Section 2.2 have exact small sample properties. But for applied use, it is important to have a valid implementation in non-Gaussian and potentially heteroskedastic models. With the regressors non-stochastic (or after conditioning on the regressors with a conditionally mean zero error term), the general model is still of the form (1), where now  $\varepsilon_i \sim (0, \sigma_i^2)$  independent across  $i$ . Under weak technical conditions on the tails of the distribution of  $\varepsilon_i$ , on the sequence  $\{\sigma_i^2\}_{i=1}^n$  and on the regressors  $\{x_i, \mathbf{q}_i, \mathbf{z}_i\}_{i=1}^n$ , a central limit theorem yields

$$\mathbf{\Omega}_n^{-1/2} \begin{pmatrix} \hat{\beta}_{\text{long},n} - \beta_n \\ \hat{\beta}_{\text{short},n} - \beta_n - \Delta_n \end{pmatrix} \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2) \quad (15)$$

for some suitably defined  $\mathbf{\Omega}_n$ , since  $(\hat{\beta}_{\text{long},n} - \beta_n, \hat{\beta}_{\text{short},n} - \beta_n - \Delta_n)$  are linear combinations of the heterogeneous but mean zero and independent random variables  $\{\varepsilon_i\}_{i=1}^n$ . We provide a corresponding result in Appendix A.6 that allows for dependence among the  $\varepsilon_i$  due to clustering.

Suppose  $\hat{\mathbf{\Omega}}_n$  is a consistent estimator of  $\mathbf{\Omega}_n$  in the sense that  $\mathbf{\Omega}_n^{-1} \hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$ . The natural LR statistic of  $H_0 : \beta_n = 0$  under the bound (2) then becomes

$$\begin{aligned} \widehat{\text{LR}}_n(\bar{\kappa}_n) &= \min_{|d| \leq \rho_n \bar{\kappa}_n / \sqrt{\mathbf{x}'_n \mathbf{x}_n / n}} \begin{pmatrix} \hat{\beta}_{\text{long},n} \\ \hat{\beta}_{\text{short},n} - d \end{pmatrix}' \hat{\mathbf{\Omega}}_n^{-1} \begin{pmatrix} \hat{\beta}_{\text{long},n} \\ \hat{\beta}_{\text{short},n} - d \end{pmatrix} \\ &\quad - \min_{b, |d| \leq \rho_n \bar{\kappa}_n / \sqrt{\mathbf{x}'_n \mathbf{x}_n / n}} \begin{pmatrix} \hat{\beta}_{\text{long},n} - b \\ \hat{\beta}_{\text{short},n} - b - d \end{pmatrix}' \hat{\mathbf{\Omega}}_n^{-1} \begin{pmatrix} \hat{\beta}_{\text{long},n} - b \\ \hat{\beta}_{\text{short},n} - b - d \end{pmatrix}. \end{aligned}$$

Exploiting the invariance of the LR statistic to reparameterizations, the distribution of  $\widehat{\text{LR}}_n(\bar{\kappa}_n)$  under the approximations (15) and  $\hat{\mathbf{\Omega}}_n = \mathbf{\Omega}_n$  is effectively indexed by two scalar parameters

$$\chi_1 = \frac{\Omega_{11} - \Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22} - \Omega_{12}^2}} \quad (16)$$

$$\chi_2 = \frac{\sqrt{\Omega_{11}}}{\sqrt{\Omega_{11}\Omega_{22} - \Omega_{12}^2}} \frac{\rho_n \bar{\kappa}_n}{\sqrt{\mathbf{x}'_n \mathbf{x}_n / n}} \quad (17)$$

where  $\Omega_{ij}$  is the  $i, j$ th element of  $\mathbf{\Omega}_n$ , and  $\widehat{\text{LR}}_n(\bar{\kappa}_n)$  under the null hypothesis of  $\beta_n = 0$  has

the same asymptotic distribution as

$$\begin{aligned} \min_{|g| \leq \chi_2} \begin{pmatrix} Z_1 \\ Z_2 + g_0 - g \end{pmatrix}' \begin{pmatrix} Z_1 \\ Z_2 + g_0 - g \end{pmatrix} \\ - \min_{h, |g| \leq \chi_2} \begin{pmatrix} Z_1 - h \\ Z_2 + g_0 - \chi_1 h - g \end{pmatrix}' \begin{pmatrix} Z_1 - h \\ Z_2 + g_0 - \chi_1 h - g \end{pmatrix} \end{aligned} \quad (18)$$

where  $(Z_1, Z_2)' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$  and  $g_0 = \sqrt{\Omega_{11}}\Delta_n / \sqrt{\Omega_{11}\Omega_{22} - \Omega_{12}^2}$ . Let  $\text{cv}_\Omega(\chi)$ ,  $\chi = (\chi_1, \chi_2)$  be the corresponding critical value which ensures size control of the statistic in (18) for all values of  $|g_0| \leq \chi_2$ . In the replication files, we provide an look-up table for all possible values of  $\chi$ . A subsequence argument then yields asymptotic validity of this feasible LR test  $\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n) = \mathbf{1}[\widehat{\text{LR}}_n(\bar{\kappa}_n) > \text{cv}_\Omega(\hat{\chi}_n)]$ , where  $\hat{\chi}_n = (\hat{\chi}_{n,1}, \hat{\chi}_{n,2})$  are as in (16) and (17), with the elements of  $\mathbf{\Omega}_n$  replaced by those of  $\hat{\mathbf{\Omega}}_n$ .

**Lemma 3** (a) If  $\mathbf{\Omega}_n^{-1}\hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$  and (15) holds, then  $\limsup_{n \rightarrow \infty} E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] \leq \alpha$  for all sequences  $\theta_n$  with  $\beta_n = 0$  and  $|\Delta_n| \leq \rho_n \bar{\kappa}_n / \sqrt{\mathbf{x}'_n \mathbf{x}_n / n}$ .

(b) Under the assumptions of Theorem 1,  $E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] - E_{\theta_n}[\varphi_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] \rightarrow 0$ .

Note that the asymptotic validity in part (a) holds without any assumptions about the sequences  $p_n$  or  $\bar{\kappa}_n$ . In particular, it is not required that  $p_n/n \rightarrow c \in (0, 1)$  or  $\bar{\kappa}_n = o(n^{-1/4})$ . In the Gaussian homoskedastic model,  $\mathbf{\Omega}_n$  is equal to  $(\mathbf{x}'_n \mathbf{x}_n)^{-1} \mathbf{\Sigma}(\rho_n)$ , and in large samples,  $\hat{\varphi}_{\text{LR},n}$  reduces to the bivariate LR test introduced in Section 2.2. Formally, part (b) of the Lemma shows that the large sample power properties of  $\hat{\varphi}_{\text{LR},n}$  in the Gaussian homoskedastic model are equal to the small sample power properties of the LR test as introduced in Section 2.2. Thus,  $\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)$  has the same asymptotic efficiency properties as  $\varphi_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)$  discussed below Theorem 1, even among tests that depend on the data beyond the short and long regression coefficient.

Given Lemma 3, the only obstacle to a straightforward implementation of the LR test in a more general model is the estimation of the asymptotic variance  $\mathbf{\Omega}_n$ . If the number of controls  $p$  is fixed, or only slowly increasing with  $n$ , the usual heteroskedasticity robust White (1980) estimator for  $\mathbf{\Omega}_n$  is consistent under reasonably weak assumptions. However, under asymptotics where  $p_n/n \rightarrow c \in (0, 1)$ , as employed for the asymptotic efficiency argument in Theorem 1, Cattaneo, Jansson, and Newey (2018a) show that the White (1980) estimator is no longer consistent, and Cattaneo, Jansson, and Newey (2018b) provide an alternative estimator that remains consistent. Alternatively, if the explanatory power of the additional controls is limited in the sense that  $\kappa_n = o(1)$ , one may also consistently estimate  $\hat{\mathbf{\Omega}}_n$  from

the usual White formula based on the residuals from the short regression that only includes the baseline controls. This has the advantage of being readily implementable also with clustering. We provide a corresponding result in Appendix A.6.

Figure 1 in the introduction was obtained by inverting  $\hat{\varphi}_{\text{LR}}(\bar{\kappa}, \mathbf{Y})$  to obtain a confidence interval for  $\beta$ , given any value of  $\bar{\kappa}$  (in the following, we drop  $n$  subscripts again to ease notation). For  $\bar{\kappa} = 0$ , and under homoskedasticity, this yields the same interval as obtained from standard short regression inference using the 2,2 element of  $\hat{\mathbf{\Omega}}$  as the variance estimator. In small samples, when  $\hat{\mathbf{\Omega}}$  does not impose homoskedasticity, the confidence interval for  $\bar{\kappa} = 0$  is centered at a slightly different value, since under heteroskedasticity, it is in general more efficient to estimate  $\beta$  by a linear combination of  $\hat{\beta}_{\text{long}}$  and  $\hat{\beta}_{\text{short}}$  that puts non-zero weight on  $\hat{\beta}_{\text{long}}$ . For  $\bar{\kappa} \rightarrow \infty$ , the interval is exactly centered at  $\hat{\beta}_{\text{long}}$ , but the LR test uses a slightly larger critical value, as discussed in Section 2.2 above.

## 4 Extensions

### 4.1 Instrumental Variable Regression

Suppose the scalar regressor  $x_i$  of interest in the linear regression (1) is endogenous, but we have access to a scalar instrument  $w_i$  ( $w_i$  could be a linear combination of a vector of instruments, such as in two stage least squares). As in the baseline model, we treat  $\{w_i, \mathbf{q}_i, \mathbf{z}_i\}_{i=1}^n$  as non-stochastic, or, equivalently, we condition on their realization in the following. To simplify notation, let  $w_i$  be orthogonal to the baseline controls  $\mathbf{q}_i$ . Assume that the data is generated via

$$x_i = \eta w_i + \mathbf{q}_i' \boldsymbol{\delta}_x + \mathbf{z}_i' \boldsymbol{\gamma}_x + \varepsilon_{xi} \quad (19)$$

$$y_i = \beta x_i + \mathbf{q}_i' \boldsymbol{\delta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_i \quad (20)$$

where  $(\varepsilon_{xi}, \varepsilon_i)$  is mean-zero independent across  $i$ , but potentially heteroskedastic, and if  $\varepsilon_{xi}$  is correlated with  $\varepsilon_i$ , the regressor  $x_i$  in (20) is endogenous.

Let  $(\hat{\beta}_{\text{long}}^{IV}, \hat{\beta}_{\text{short}}^{IV})$  be the IV estimators of  $\beta$  that include or exclude the additional controls  $\mathbf{z}_i$ . These estimators involve the term  $\sum_{i=1}^n w_i x_i$ , which under (19) is stochastic and depends on the realization of  $\varepsilon_{xi}$ , complicating the description of their bias. In order to avoid these difficulties, we focus on their moment condition instead, as in Anderson and Rubin (1949). Let  $\hat{w}_i^z$  be the residuals of a regression of  $w_i$  on  $\mathbf{z}_i$ . The estimators  $\hat{\beta}_{\text{long}}^{IV}$  and  $\hat{\beta}_{\text{short}}^{IV}$  are identified from the two moment conditions  $E[n^{-1} \sum_{i=1}^n \hat{w}_i^z \varepsilon_i] = 0$  and  $E[n^{-1} \sum_{i=1}^n w_i \varepsilon_i] = 0$ ,

respectively. Consider testing  $H_0 : \beta = 0$  (non-zero values can be reduced to this case by subtracting  $\beta_0 x_i$  from  $y_i$ ). Similar to (15), the empirical moment condition of these estimators then satisfies under the null hypothesis

$$(\mathbf{\Omega}^{IV})^{-1/2} \begin{pmatrix} n^{-1} \sum_{i=1}^n \hat{w}_i^z y_i \\ n^{-1} \sum_{i=1}^n w_i y_i - \Delta^{IV} \end{pmatrix} \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2) \quad (21)$$

for some suitably defined  $\mathbf{\Omega}^{IV}$ , which can be consistently estimated by  $\hat{\mathbf{\Omega}}^{IV}$ . The “bias”  $\Delta^{IV}$  in (21) is given by  $\Delta^{IV} = n^{-1} \sum_{i=1}^n w_i \mathbf{z}_i' \boldsymbol{\gamma}$ , and a straightforward calculation shows that under (2), we have the sharp bound

$$|\Delta^{IV}| \leq \bar{\kappa} \sqrt{\mathbf{w}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{w} / n}$$

with  $\mathbf{w} = (w_1, \dots, w_n)'$ . Thus, under the null hypothesis of  $H_0 : \beta = 0$ , the observations (21) have the identical structure as (15) of the previous section, and one can apply the LR test in entirely analogous fashion to obtain a valid large sample test that exploits the bound (2) to sharpen inference in instrumental variable regression. Our focus on the moments (21), rather than the estimators  $(\hat{\beta}_{\text{long}}^{IV}, \hat{\beta}_{\text{short}}^{IV})$ , has the additional appeal that no assumptions about the strength of the instrument are required.

## 4.2 Double Bounds

In Section 2 and 3, we have treated the regressors  $\{x_i, \mathbf{q}_i, \mathbf{z}_i\}_{i=1}^n$  as either non-stochastic, or the analysis conditioned on their value. In the simple Gaussian model of Section 2, it is easy to see that with the regressors random, the Gram matrix forms an ancillary statistic. It is textbook advice to condition inference on ancillary statistics in general and on the Gram matrix in particular (see, for instance, Chapter 2.2 in Cox and Hinkley (1974)), providing a rationale for our analysis. Furthermore, our approach does not require or depend on a model for the potentially stochastic properties of the regressors. This is attractive in so far as it relieves applied researchers from having to defend a particular data generating mechanism, and avoids a source of potential misspecification.

We now discuss how one could exploit additional assumptions on the generation of the regressor of interest  $x_i$  to potentially further sharpen inference about  $\beta$ . In particular, assume that  $\tilde{x}_i$  is generated by the linear model

$$\tilde{x}_i = \mathbf{q}_i' \boldsymbol{\delta}_x + \mathbf{z}_i' \boldsymbol{\gamma}_x + \varepsilon_{xi} \quad (22)$$

where  $\varepsilon_{xi}$  is conditionally mean zero given  $\{\mathbf{q}_i, \mathbf{z}_i\}_{i=1}^n$ . The regressor  $x_i$  is simply defined as the residuals of a least squares regression of  $\tilde{x}_i$  on  $\mathbf{q}_i$ , so that consistent with our notation above  $\mathbf{Q}'\mathbf{x} = \mathbf{0}$ . We maintain, as in Sections 2 and 3, that  $\varepsilon_i$  in (1) is conditionally mean zero (so  $\tilde{x}_i$  is not endogenous, and no instrument is required). Assume further that we are willing to assume that in addition to (2), also

$$\kappa_x^2 = n^{-1} \sum_{i=1}^n (\mathbf{z}_i' \boldsymbol{\gamma}_x)^2 \leq \bar{\kappa}_x^2, \quad (23)$$

so that  $\bar{\kappa}_x$  has the interpretation of an upper bound on the quadratic mean of the effect of  $\mathbf{z}_i$  on  $x_i$ , after controlling for  $\mathbf{q}_i$ . This “double bounds” structure of limiting the population coefficients in both the regression of interest (1), and the auxiliary regression (22), parallels the assumptions validating the double lasso procedure by Belloni, Chernozhukov, and Hansen (2014).

As in the previous subsection, it is convenient to focus on the moment conditions defining the OLS estimators  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})$ : Under weak regularity conditions, (2), (22) and (23) imply that under  $H_0 : \beta = 0$

$$(\boldsymbol{\Omega}^{\text{Dbl}})^{-1/2} \begin{pmatrix} n^{-1} \sum_{i=1}^n \hat{x}_i^z y_i \\ n^{-1} \sum_{i=1}^n x_i y_i - \Delta^{\text{Dbl}} \end{pmatrix} \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2) \quad (24)$$

where  $\hat{x}_i^z$  are the residuals of a regression of  $x_i$  on  $\mathbf{q}_i$ , and  $\Delta^{\text{Dbl}}$  satisfies the sharp bound  $|\Delta^{\text{Dbl}}| \leq \bar{\kappa} \cdot \bar{\kappa}_x$  (see Appendix A.7 for details). With an appropriate estimator  $\hat{\boldsymbol{\Omega}}^{\text{Dbl}}$ , this again has the same structure as the problem discussed in Section 3, so the LR test defined there can be used to exploit the additional information from (22) and (23).

## 5 Small Sample Simulations

In this section, we use Monte Carlo simulations to evaluate the finite-sample properties of confidence intervals based on  $\hat{\varphi}_{\text{LR}}$ , and we compare it to the performance of the LASSO-based post-double-selection technique of Belloni, Chernozhukov, and Hansen (2014) (abbreviated BCH in the following two sections).

As in BCH’s Monte Carlo, we set the total number of observations to  $n = 500$ , let  $p = 200$ , and generate data from a model where the baseline control is simply a constant,

$$y_i = \tilde{\delta}_1 + \tilde{x}_i \beta + \tilde{\mathbf{z}}_i' \boldsymbol{\gamma} + \varepsilon_i, \quad i = 1, \dots, n \quad (25)$$

with  $\varepsilon_i \sim iid\mathcal{N}(0, 1)$  independent of  $\{\tilde{x}_i, \tilde{\mathbf{z}}_i\}$ , and  $\tilde{\mathbf{z}}_i$  is generated by the linear model

$$\tilde{x}_i = \tilde{\mathbf{z}}_i' \boldsymbol{\mu} + \varepsilon_i^x \quad (26)$$

with  $\varepsilon_i^x \sim iid\mathcal{N}(0, 1)$  independent of  $\{\tilde{\mathbf{z}}_i\}$ . To be consistent with our previous notation, we orthogonalize the regressors in (25) off the baseline control, that is  $x_i = \tilde{x}_i - n^{-1} \sum_{l=1}^n \tilde{x}_l$  and  $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})'$  with  $z_{ij} = \tilde{z}_{ij} - n^{-1} \sum_{l=1}^n \tilde{z}_{lj}$ , so that (25) implies the linear model (1) with an appropriate definition of  $\delta_1$ . We set  $\beta = 0$  throughout. Our designs vary according to the value of four parameters: the previously introduced  $\rho^2 \in \{0.7, 0.95\}$  and  $\kappa \in \{0.2, 0.5\}$ ; the scalar  $\eta \in \{0.1, 0.3\}$  determines the degree of sparsity of  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)'$  and  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ ; and  $\nu \in \{0, 0.5, 1\}$  determines the overlap between the non-zero indices of  $\boldsymbol{\gamma}$  and  $\boldsymbol{\mu}$ . Specifically,  $\gamma_j = c_\gamma \mathbf{1}[j \leq \lfloor \eta p \rfloor]$ , where the scalar  $c_\gamma$  is chosen such that the implied value of  $\kappa^2$  is equal to the specified value, and  $\mu_j = c_\mu \mathbf{1}[\lfloor \eta(1-\nu)p \rfloor + 1 \leq j \leq \lfloor \eta \nu p \rfloor]$ ,  $j = 1, \dots, p$ , where  $c_\mu \in \mathbb{R}$  is chosen such that the sample  $R^2$  of a regression of  $x_i$  on  $\mathbf{z}_i$  is equal to  $\rho^2$ .

The parameter  $\eta$  plays a crucial role for the BCH method, since the method requires that the number of non-zero values in  $\boldsymbol{\gamma}$  and  $\boldsymbol{\mu}$  is not too large. In contrast, the test  $\hat{\varphi}_{\text{LR}}$  remains numerically invariant to any linear reparameterizations of the regressors. Finally, the parameter  $\nu$  determines the omitted variable bias in the short regression coefficient  $\hat{\beta}_{\text{short}}$  (which is the coefficient on  $x_i$  in the regression of  $y_i$  on  $(1, x_i)$ ). Under  $\nu = 0$ , there is no overlap, and the variables  $z_{ij}$  with non-zero coefficient  $\gamma_j$  are uncorrelated with the regressor of interest  $x_i$ , so there is no omitted variable bias, at least over repeated samples with random regressors. In the other extreme, with  $\nu = 1$ , every variable  $z_{ij}$  with non-zero coefficient  $\gamma_j$  is correlated with  $x_i$ , leading to a large omitted variable bias.

We consider four types of confidence intervals for  $\beta$ . First, the usual confidence interval based on  $\hat{\beta}_{\text{short}}$ . Second, the usual confidence interval based on  $\hat{\beta}_{\text{long}}$ . Third, the confidence interval obtained by inverting the feasible test  $\hat{\varphi}_{\text{LR}}$  introduced in Section 3, where we set  $\bar{\kappa}$  equal to the actual value of  $\kappa$ . Fourth, the LASSO-based post-double-selection method “LPDS” from BCH, as specified in their Monte Carlo Section 4.2. For the first three types of methods, we estimate standard errors of  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})'$  with the heteroskedasticity-robust estimator of Cattaneo, Jansson, and Newey (2018b). In addition, we report quartiles of the threshold value  $\bar{\kappa}_{\text{LR}}^* \in [0, \infty) \cup \{+\infty\}$  computed from the family of tests  $\hat{\varphi}_{\text{LR}}$  for each draw, defined to be zero if  $\bar{\kappa} = 0$  does not lead to rejection, and  $+\infty$  if none of the  $\bar{\kappa}$  values lead to rejection.

Table 2 contains the results. The confidence interval based on  $\hat{\beta}_{\text{short}}$  has coverage substantially below the nominal level whenever the overlap parameter  $\nu$  is positive. This shows

Table 2: Small Sample Properties for  $n = 500$  and  $p = 200$ 

			$\hat{\beta}_{\text{short}}$		$\hat{\beta}_{\text{long}}$		LPDS		$\hat{\varphi}_{\text{LR}}(\kappa, \mathbf{Y})$		$\bar{\kappa}_{\text{LR}}^*$		
			Cov	Lgth	Cov	Lgth	Cov	Lgth	Cov	Lgth	Q1	Q2	Q3
$\eta$	$\rho^2$	$\nu$	$\kappa = 0.20$										
0.10	0.70	0.00	0.94	0.14	0.94	0.22	0.95	0.23	0.94	0.22	0.00	0.00	0.00
0.10	0.70	0.50	0.74	0.14	0.95	0.22	0.92	0.23	0.95	0.22	0.00	0.00	0.00
0.10	0.70	1.00	0.27	0.14	0.94	0.22	0.82	0.23	0.95	0.20	0.00	0.04	0.08
0.10	0.95	0.00	0.94	0.05	0.94	0.22	0.95	0.26	0.98	0.14	0.00	0.00	0.00
0.10	0.95	0.50	0.42	0.05	0.95	0.22	0.95	0.26	0.98	0.14	0.00	0.02	0.05
0.10	0.95	1.00	0.01	0.05	0.95	0.22	0.95	0.25	0.95	0.13	0.08	0.11	0.15
0.30	0.70	0.00	0.94	0.14	0.95	0.22	0.95	0.21	0.95	0.22	0.00	0.00	0.00
0.30	0.70	0.50	0.74	0.14	0.94	0.22	0.87	0.21	0.94	0.22	0.00	0.00	0.00
0.30	0.70	1.00	0.27	0.14	0.94	0.22	0.62	0.21	0.95	0.20	0.00	0.04	0.08
0.30	0.95	0.00	0.94	0.05	0.95	0.22	0.95	0.18	0.98	0.14	0.00	0.00	0.00
0.30	0.95	0.50	0.43	0.05	0.95	0.22	0.91	0.17	0.98	0.14	0.00	0.02	0.05
0.30	0.95	1.00	0.01	0.05	0.95	0.22	0.81	0.17	0.95	0.13	0.08	0.11	0.15
$\eta$	$\rho^2$	$\nu$	$\kappa = 0.50$										
0.10	0.70	0.00	0.92	0.14	0.94	0.22	0.95	0.24	0.94	0.22	0.00	0.00	0.00
0.10	0.70	0.50	0.13	0.14	0.94	0.22	0.77	0.24	0.94	0.22	0.03	0.08	0.12
0.10	0.70	1.00	0.00	0.14	0.94	0.22	0.38	0.24	0.95	0.22	0.22	0.26	0.31
0.10	0.95	0.00	0.91	0.05	0.95	0.22	0.95	0.27	0.95	0.22	0.00	0.00	0.00
0.10	0.95	0.50	0.00	0.05	0.94	0.22	0.95	0.26	0.96	0.19	0.13	0.16	0.20
0.10	0.95	1.00	0.00	0.05	0.94	0.22	0.95	0.25	0.95	0.15	0.37	0.41	0.44
0.30	0.70	0.00	0.91	0.14	0.94	0.22	0.95	0.22	0.94	0.22	0.00	0.00	0.00
0.30	0.70	0.50	0.12	0.14	0.94	0.22	0.50	0.22	0.94	0.22	0.04	0.08	0.12
0.30	0.70	1.00	0.00	0.14	0.94	0.22	0.03	0.22	0.95	0.22	0.22	0.27	0.31
0.30	0.95	0.00	0.92	0.05	0.94	0.22	0.95	0.18	0.95	0.22	0.00	0.00	0.00
0.30	0.95	0.50	0.00	0.05	0.94	0.22	0.74	0.18	0.96	0.19	0.13	0.16	0.20
0.30	0.95	1.00	0.00	0.05	0.94	0.22	0.42	0.17	0.95	0.15	0.37	0.41	0.44

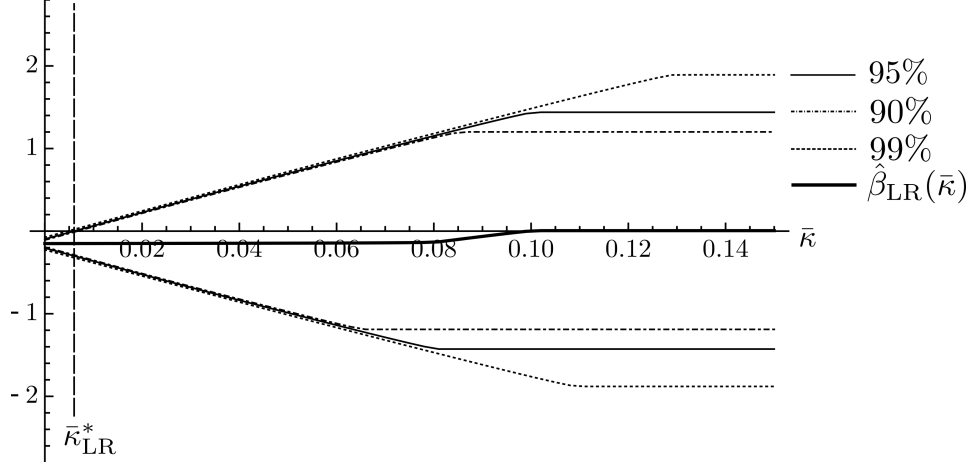
Notes: Entries are coverage and average length of 95% confidence intervals for  $\beta$ , and the quartiles of the distribution of  $\bar{\kappa}_{\text{LR}}^*$ . Rows correspond to different DGPs, with  $\eta$  measuring the sparsity of the design,  $\nu$  the overlap between the non-zero indices on  $\mathbf{z}_i$  in the regressions of  $y_i$  on  $\mathbf{z}_i$  and of  $x_i$  on  $\mathbf{z}_i$ , and  $\rho^2$  is  $R^2$  of a regression of  $x_i$  on  $\mathbf{z}_i$ . The columns are different confidence intervals, with  $\hat{\beta}_{\text{short}}$  and  $\hat{\beta}_{\text{long}}$  the confidence interval based on short and long regression coefficients,  $\hat{\varphi}_{\text{LR}}(\bar{\kappa}, \mathbf{Y})$  the LR based confidence interval developed in this paper that imposes the bound  $\kappa \leq \bar{\kappa}$ , and LPDS is BCH's LASSO-based post-double-selection procedure. Based on 20,000 Monte Carlo simulations.

that the considered values of  $\kappa$  are large enough to severely distort inference that simply sets the control coefficients to zero. In contrast, the interval associated with  $\hat{\beta}_{\text{long}}$  has size very close to the nominal level throughout, but at the cost of being fairly long. The LPDS method is sometimes substantially undercovers even in the relatively sparse design with  $\eta = 0.1$ . Apparently, the relatively small values of  $\gamma_j$  make it difficult for the method to correctly pick up the  $\eta \cdot p = 20$  non-zero coefficients, leading to a remaining omitted variable bias that is large enough to induce non-negligible overrejections. When  $\kappa = 0.5$  and  $\rho^2 = 0.95$ , so that both  $\gamma_j$  and  $\mu_j$  are relatively larger, the LPDS method reliably controls size in the  $\eta = 0.1$  sparse design, but yields somewhat longer intervals compared to  $\hat{\beta}_{\text{long}}$ . The new tests  $\hat{\varphi}_{\text{LR}}(\kappa, \mathbf{Y})$  control size throughout as they should, given that  $\bar{\kappa} = \kappa$  trivially implies  $\kappa \leq \bar{\kappa}$ , and yield intervals that are shorter than the  $\hat{\beta}_{\text{long}}$ -interval when  $\nu > 0$ , with larger gains for larger values of  $\rho$  and  $\nu$ .

Of course, with  $\kappa$  unknown in practice, one cannot apply  $\hat{\varphi}_{\text{LR}}(\bar{\kappa}, \mathbf{Y})$  with  $\kappa = \bar{\kappa}$ . So imagine one computed, for each draw of the data, a figure corresponding to Figure 1 in the introduction. Since the figures consider all values of  $\bar{\kappa}$ , there necessarily is a point on the x-axis where  $\bar{\kappa} = \kappa$ , and the length of  $\hat{\varphi}_{\text{LR}}(\kappa, \mathbf{Y})$  in Table 2 indicates that at that point, the interval exploiting the bound (2) is often considerably more informative than the interval based on  $\hat{\beta}_{\text{long}}$ . A further summary statistic of these figures is  $\bar{\kappa}_{\text{LR}}^*$ , and the last three columns in Table 3 report the quartiles of its distribution. For  $\nu = 1$ , and for  $(\kappa, \nu) = (0.5, 0.5)$ , the median of  $\bar{\kappa}_{\text{LR}}^*$  is always positive. Thus, the majority of these figures correctly indicate that small upper bounds for  $\bar{\kappa}$  are empirically incompatible with  $\beta = 0$ , so they are also informative in this sense. Unreported results show that if the true value of  $\beta$  is nonzero, these medians become larger. Thus, the figures constructed from the LR approach help sharpen inference about  $\beta$  in a meaningful way.

Still, looking over the table, it is tempting to conclude that one should use  $\hat{\beta}_{\text{short}}$  whenever there is no overlap,  $\nu = 0$ , as this leads to the shortest intervals by far, and only slight size distortions. Similarly, if  $\nu < 1$  the quartiles of  $\bar{\kappa}_{\text{LR}}^*$  are much smaller than  $\kappa$ . However, it is not possible to consistently determine the value of  $\nu$  from the observations. This is the result of the asymptotic efficiency derivations in Section 2.3: For small values of  $\bar{\kappa}$ , it is impossible to do better than to construct inference based on the bivariate statistics  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})'$  at least in large samples, and as demonstrated there, the LR approach comes close to exploiting the information contained in this pair of statistics.

Figure 5: LR Confidence Intervals for  $\beta$  in Donohue and Levitt (2001)



## 6 Further Empirical Applications

### 6.1 Abortion and Crime

In an influential paper, Donohue and Levitt (2001) found significant effect of lagged abortion rates on crime, using a panel data of U.S. states from 1985 to 1997, but these results were disputed in follow-up studies (see, for instance, Foote and Goetz (2008), and Joyce (2004, 2009)). BCH also consider this example as an illustration of their methodology.

We apply the same specification as in BCH, and focus on violent crimes (results are similar for property and murder crime rates). The regression models panel data from 48 states over 12 years, with all variables expressed in first differences to account for state fixed-effects, for a total of  $n = 576$  observations. The explanatory variable is the violent crime rate, the regressor of interest with coefficient  $\beta$  is a measure of lagged abortion rates, the “short” specification includes a set of 20 controls (including 12 time dummies) present in Donohue and Levitt’s original specification, and the potential additional controls are a set of  $p = 284$  regressors proposed by BCH, including higher-order terms, initial conditions, and interactions of variables with state-specific observables. Using standard errors clustered at the state level, estimated using short regression residuals (see Appendix A.6), the t-tests based on  $\hat{\beta}_{\text{short}}$  rejects at the 5% level, but the t-test using the estimator  $\hat{\beta}_{\text{long}}$  from the long regression does not.

Figure 5 plots the LR confidence intervals for  $\beta$  as a function of the bound  $\bar{\kappa}$ . It is apparent

from the figure that trying to control for the 284 additional controls is very ambitious, as it leads to a dramatically increased standard error compared to the short regression. Correspondingly, the cut-off value  $\bar{\kappa}_{\text{LR}}^*$  is rather small at 0.6%: As soon as one allows for a quadratic mean effect of the additional controls on the crime rate to be larger than 0.6%, one loses significance of lagged abortion rates on violent crime rates. This conclusion accords with the qualitative result of BCH, who find that post double lasso inference about  $\beta$  is not significant.

## 6.2 Earnings, Lottery Winnings and Treatment Heterogeneity

It is well understood that inference on average treatment effects is sensitive to the accommodation of treatment heterogeneity (Imbens and Wooldridge, 2009). In particular, procedures that ignore heterogeneity of treatment effects may misappropriate explanatory power from the treatment to the confounding factors in a way that biases the average treatment effect. However, allowing too much heterogeneity can lead to noisy inference. Our approach can be used to illuminate how assumptions on treatment heterogeneity shape inference on the average treatment effect. We illustrate this in this section using data from a study by Imbens, Rubin, and Sacerdote (2001).<sup>4</sup>

Using data on a cross-section of  $n = 496$  individuals participating in the Massachusetts lottery from 1984 to 1988, these authors study the effects of unearned income on the marginal propensity to earn (MPE). In their main empirical exercise, the authors regress post-lottery earnings on lottery winnings, and they interpret the coefficient on lottery winnings as the effect of income on MPE. Although winning the lottery is plausibly exogenous conditional on purchasing a lottery ticket, the frequency of lottery ticket purchases may be correlated with factors that also affect labor and wages. Hence, the authors include observable individual characteristics as control variables. In particular, their Table 4 displays the coefficient on lottery winnings from regressions with different sets of controls. For illustration, we focus on the specification from Row 1, Column 2 of their Table 4

$$y_i = \beta x_i + \sum_{j=1}^7 q_{ij} \delta_j + \varepsilon_i. \quad (27)$$

In this specification, the outcome  $y_i$  denotes the average of social security earnings in the six

---

<sup>4</sup>Also see Imbens and Rubin (2015) for further exploration of this data set from a perspective of causal inference.

years after the lottery measured in multiples of \$1000, the main regressor  $x_i$  denotes lottery winnings in multiples of \$1000, and the baseline covariates  $q_{ij}$  include years of education, age, an intercept term, and dummies for gender, some college, age greater than 55, and age greater than 65. In Row 1, Column 2 of their Table 4, the authors estimate  $\beta$ , the coefficient on winnings, to be -0.052 and statistically significant at standard levels.

Following their baseline specification, Imbens, Rubin, and Sacerdote (2001) explore heterogeneity in the main effect. In particular, in their Table 5, the authors explore how  $\beta$  differs by gender, prior earnings, age, education, and years since winning. Despite this heterogeneity, the coefficients in (27) may still be unbiased for the conditional average effect in the sample of 496 individuals, if one assumes that potential heterogeneity in  $\beta$  is uncorrelated with the observed regressors. However, without this assumption, inference based on the short regression (27) may be invalid. We apply our approach to assess the extent to which allowing heterogeneity in  $\beta$  over different subgroups of the data changes inference on the conditional average effect.

In particular, we consider potential controls  $z_{ik}$  in the form of subgroup dummies and interactions with the regressor of interest in order to allow for heterogeneity in the coefficient of interest. Formally, we identify subgroups of the data generated by the cross-products of dummies for gender, full college, age greater than 45, age greater than 55, and age greater than 65. This results in 16 potential subgroups  $G_1, \dots, G_{16}$  that partition the full set of observables  $i \in \{1, \dots, 496\}$ , with each individual  $i$  belonging to one subgroup  $G_j$ . A linear model allowing for arbitrary heterogeneity of the treatment effect across these groups is given by

$$y_i = \sum_{j=1}^{16} \mathbf{1}[i \in G_j] x_i \beta_j + \sum_{j=1}^7 q_{ij} \delta_j + \sum_{j=1}^{16} \mathbf{1}[i \in G_j] \gamma_j + \varepsilon_i$$

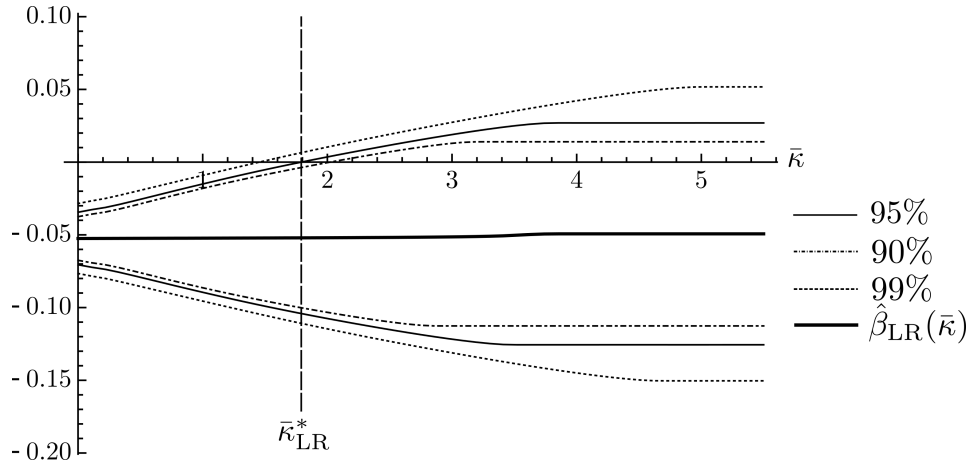
with  $\beta_j$  the conditional treatment effect in subgroup  $j$ .

Suppose that under treatment heterogeneity, the parameter of interest is the average  $\beta = n^{-1} \sum_{i=1}^n \sum_{j=1}^{16} \mathbf{1}[i \in G_j] \beta_j$ . Following Section 5.2 of Imbens and Wooldridge (2009), inference about  $\beta$  can conveniently be performed by augmenting the short regression (27) by controls  $z_{ik}$  consisting of dummies  $\mathbf{1}[i \in G_j]$  and interaction terms  $(\mathbf{1}[i \in G_j] - n^{-1} \sum_{l=1}^n \mathbf{1}[l \in G_j]) x_i$ . Dropping collinear terms, this results in the “long regression” with 25 additional controls  $z_{ik}$

$$y_i = \beta x_i + \sum_{j=1}^7 q_{ij} \delta_j + \sum_{k=1}^{25} z_{ik} \gamma_k + \varepsilon_i.$$

We use our approach to study the sensitivity of inference about  $\beta$  to varying assumptions

Figure 6: LR Confidence Intervals for  $\beta$  in Imbens, Rubin and Sacerdote (2001)



about population group heterogeneity, that is the coefficients  $\gamma_k$ . Figure 6 plots confidence intervals for  $\beta$  as a function of the bound  $\bar{\kappa}$ , using Cattaneo, Jansson, and Newey (2018b) standard errors. We find that for the 95% level,  $\bar{\kappa}_{LR}^* = \$1.79\text{k}$ , so under the assumption that the quadratic mean of heterogeneity across groups is smaller than \$1.79k, we still reject the null hypothesis that the conditional average treatment effect of lottery winnings on post-lottery earnings is zero at the 5% significance level. This translates into a ratio of 1.7% of  $n(\bar{\kappa}_{LR}^*)^2$  to the sum of squared residuals in a regression of  $y_i$  on  $\mathbf{q}_i$ . As such, we conclude that the significance of the homogeneous baseline (27) are somewhat sensitive, but not extremely sensitive to the assumption of treatment homogeneity.

## 7 Conclusion

Improving inference over including all potential controls in a “long regression” requires some *a priori* knowledge about the control coefficients. In this paper, we develop a simple inference procedure that exploits a bound on the weighted sum of squared control coefficients, which corresponds to a limit of the explanatory power of the controls. This yields a continuous bridge between excluding the controls and including them with unconstrained coefficients, as a function of the bound. The approach enables applied researchers to make interpretable statements about the robustness of an empirical result relative to set of potential controls, beyond the dichotomous conclusion that significance is, or isn’t lost with their inclusion.

# A Appendix

Table 3: Properties of 95% Armstrong and Kolsar Inference

Panel A: Weighted Expected Length of CI for  $b$  under  $d \sim U[-\bar{k}, \bar{k}]$

$\rho \backslash \bar{k}$	AK( $\bar{k}$ ) Interval					Lower Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	3.9	4.2	4.5	4.5	4.5	3.9	4.2	4.4	4.5	4.5
0.90	3.9	5.0	7.4	9.0	9.0	3.9	5.0	6.9	8.2	8.7
0.99	3.9	5.2	9.1	21.5	27.8	3.9	5.2	8.9	17.4	23.9

Panel B: Expected Length of CI for  $b$ , Maximized over  $|d| \leq \bar{k}$

$\rho \backslash \bar{k}$	AK( $\bar{k}$ ) Interval					Lower Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	3.9	4.2	4.5	4.5	4.5	3.9	4.2	4.4	4.5	4.5
0.90	3.9	5.0	7.4	9.0	9.0	3.9	5.0	7.1	8.4	8.9
0.99	3.9	5.2	9.1	21.5	27.8	3.9	5.2	8.9	18.4	25.1

Panel C: Ratio of Expected Length of AK CI for  $b$  Relative to Long Regression Interval

$\rho \backslash \bar{k}$	Minimized over $ d  \leq \bar{k}$					Maximized over $ d  \leq \bar{k}$				
	0	1	3	10	30	0	1	3	10	30
0.50	0.87	0.92	0.98	1.00	1.00	0.87	0.92	0.98	1.00	1.00
0.90	0.44	0.55	0.83	1.00	1.00	0.44	0.55	0.83	1.00	1.00
0.99	0.14	0.19	0.33	0.77	1.00	0.14	0.19	0.33	0.77	1.00

Panel D: Median of  $\bar{k}_\phi^*$  under  $b = 0$ ,  $P(d = d_0) = P(d = -d_0) = 1/2$

$\rho \backslash d_0$	$\bar{k}_{AK}^*$					Upper Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	0.0	0.0	0.0	0.7	1.5	0.0	0.0	0.7	4.2	14.3
0.90	0.0	0.0	0.8	2.9	4.7	0.0	0.0	1.2	7.6	25.8
0.99	0.0	0.0	1.3	7.6	11.7	0.0	0.0	1.4	8.4	28.4

Panel E: Weighted Average MSE of Equivariant Estimators of  $b$  under  $d \sim U[-\bar{k}, \bar{k}]$

$\rho \backslash \bar{k}$	$\hat{b}_{AK}$					Lower Bound				
	0	1	3	10	30	0	1	3	10	30
0.50	1.00	1.09	1.26	1.33	1.33	1.00	1.07	1.22	1.30	1.32
0.90	1.00	1.27	2.78	5.26	5.26	1.00	1.25	2.53	4.38	4.97
0.99	1.00	1.33	3.79	22.8	50.2	1.00	1.32	3.77	20.4	39.7

Notes: See Table 1.

## A.1 Proof of Lemma 1

Let  $\varphi_0(\hat{\xi}) = E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})|\hat{\xi}]$ , so that by sufficiency of  $\hat{\xi}$ , and the law of iterated expectations,  $E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})] = E_{\xi}[\varphi_0(\hat{\xi})]$ . Since by assumption,  $E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})]$  does not depend on  $\delta$ ,  $E_{\xi}[\varphi_0(\hat{\xi})] = E_{\xi_0}[\varphi_0(\hat{\xi})]$ , where  $\xi_0 = (\beta, \Delta, \mathbf{0}, \tau, \omega)$ . Define  $\varphi_S(\hat{\vartheta}) = E_{\xi_0}[\varphi_0(\hat{\xi})|\hat{\vartheta}]$  with  $\hat{\vartheta} = (\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, n^{-1/2}\hat{\phi}')$ . Then by the law of iterated expectations,  $E_{\xi_0}[\varphi_S(\hat{\vartheta})] = E_{\xi_0}[\varphi_0(\hat{\xi})] = E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})]$  for all  $\xi$ .

Furthermore, with  $\mathbf{O}$  a  $(p-1) \times (p-1)$  rotation matrix

$$\begin{aligned} E_{\beta, \Delta, \tau, \omega}[\varphi_S(\hat{\vartheta})] &= E_{\beta, \Delta, \tau}[\varphi_S(\hat{\vartheta})] \\ &= E_{\beta, \Delta, \tau}[\varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\tau\omega + \mathbf{e})')))] \\ &= E_{\beta, \Delta, \tau}[\varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\tau\omega + \mathbf{O}\mathbf{e})')))] \\ &= E_{\beta, \Delta, \tau}[\varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\tau\mathbf{O}\omega + \mathbf{O}\mathbf{e})')))] \end{aligned}$$

where  $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, n^{-1}\mathbf{I}_{p-1})$  is independent of  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}})$ , the first and last equality follow from assumption about the rejection probability of  $\varphi(\bar{\kappa}, \mathbf{Y})$ , and the before last equality follows from the spherical symmetry of the distribution of  $\mathbf{e}$ . Since  $\mathbf{O}$  was arbitrary, we also have

$$E_{\xi}[\varphi(\bar{\kappa}, \mathbf{Y})] = \int E_{\beta, \Delta, \tau}[\varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\mathbf{O}\tau\omega + \mathbf{O}\mathbf{e})'))]dH_{p-1}(\mathbf{O})$$

where  $H_{p-1}$  is the Haar measure on the  $p-1$  rotation matrices. Now set  $\tilde{\varphi}(\mathbf{T}) = \int \varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, \tau\iota'\mathbf{O}))dH_{p-1}(\mathbf{O}) = \int \varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, n^{-1/2}\hat{\phi}'\mathbf{O}))dH_{p-1}(\mathbf{O})$  with  $\iota = (1, 0, \dots, 0)' \in \mathbb{R}^{p-1}$ . Then

$$\begin{aligned} E_{\beta, \Delta, \tau}[\tilde{\varphi}(\mathbf{T})] &= E_{\beta, \Delta, \tau} \left[ \int \varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, n^{-1/2}\hat{\phi}'\mathbf{O}'))dH_{p-1}(\mathbf{O}) \right] \\ &= \int E_{\beta, \Delta, \tau}[\varphi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\tau\mathbf{O}\omega + \mathbf{O}\mathbf{e})'))]dH_{p-1}(\mathbf{O}) \end{aligned}$$

and the result follows.

## A.2 Proof of Lemma 2

We will make use of the following Lemma.

**Lemma 4** *Let  $\mathcal{I}_m$  denote the modified Bessel function of the first kind of degree  $m > 0$ . Then for any positive sequence  $s_m = o(m^{1/2})$ ,*

$$\lim_{m \rightarrow \infty} \mathcal{I}_m(s_m) \frac{\Gamma(m+1)}{(\frac{1}{2}s_m)^m} = 1$$

where  $\Gamma$  is the Gamma function.

**Proof.** From the definition of  $\mathcal{I}_m$ , for any  $s > 0$

$$\begin{aligned}\mathcal{I}_m(s) &= \left(\frac{1}{2}s\right)^m \sum_{j=0}^{\infty} \frac{(\frac{1}{4}s^2)^j}{j! \Gamma(m+j+1)} \\ &= \frac{(\frac{1}{2}s)^m}{\Gamma(m+1)} \left(1 + \sum_{j=1}^{\infty} \frac{(\frac{1}{4}s^2)^j}{j!} \frac{\Gamma(m+1)}{\Gamma(m+j+1)}\right).\end{aligned}$$

Now

$$\begin{aligned}\sum_{j=1}^{\infty} \frac{(\frac{1}{4}s^2)^j}{j!} \frac{\Gamma(m+1)}{\Gamma(m+j+1)} &\leq \sum_{j=1}^{\infty} \frac{s^{2j}}{j!} \frac{\Gamma(m+1)}{\Gamma(m+j+1)} \\ &\leq \sum_{j=1}^{\infty} \frac{(s^2/m)^j}{j!} = \exp[s^2/m] - 1\end{aligned}$$

where the second inequality uses the elementary inequality  $\Gamma(m+1)m^j/\Gamma(m+j+1) \leq 1$  obtained from repeatedly applying  $\Gamma(m+i+1) = (m+i)\Gamma(m+i) \leq m\Gamma(m+i)$  for all  $i \geq 0$  and  $m > 0$ . The result now follows from  $s_m^2/m \rightarrow 0$  under  $s_m = o(m^{1/2})$ . ■

For ease of notation, we omit the dependence on  $n$  (and  $p = p_n$ ), except for  $t_n$ . From (11), it follows that  $n\hat{\tau}^2 = \hat{\phi}'\hat{\phi}$  with  $\hat{\phi} \sim \mathcal{N}(\sqrt{n}\tau\boldsymbol{\omega}, \mathbf{I}_{p-1})$  is distributed non-central  $\chi^2$  with  $p-1$  degrees of freedom and non-centrality parameter  $n\tau^2$ . Without loss of generality, assume  $\boldsymbol{\omega} = \boldsymbol{\iota} = (1, 0, \dots, 0)'$ . Then, with  $\hat{\boldsymbol{\omega}} = \hat{\phi}/\|\hat{\phi}\|$ , from the density of  $\hat{\phi}$ ,

$$L_n(t_n) = C \int \exp[-\frac{1}{2}n\|\hat{\tau}\mathbf{O}\hat{\boldsymbol{\omega}} - \boldsymbol{\iota}t_n\|^2] dH_{p-1}(\mathbf{O})$$

for some constant  $C$  that does not depend on  $t_n$  (and note that  $L_n(t_n)$  does not depend on the realization of  $\hat{\boldsymbol{\omega}}$ ). Thus

$$L_n(t_n)/L_n(0) = \int \exp[nt_n\hat{\tau}\hat{\boldsymbol{\omega}}'\mathbf{O}\boldsymbol{\iota} - \frac{1}{2}nt_n^2] dH_{p-1}(\mathbf{O}).$$

We initially show the convergence under  $\tau = 0$ . It then suffices to show that  $E[(L_n(t_n)/L_n(0) - 1)^2] \rightarrow 0$  under  $\hat{\phi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p-1})$  and an arbitrary sequence  $t_n = o(n^{-1/4})$ . Observe that

$$\begin{aligned}& E \left[ (L_n(t_n)/L_n(0) - 1)^2 \right] \\ &= E \left[ \left( \int \exp[n^{1/2}t_n\hat{\phi}'\mathbf{O}\boldsymbol{\iota} - \frac{1}{2}nt_n^2] dH_{p-1}(\mathbf{O}) - 1 \right)^2 \right] \\ &= E \left[ \left( \int \exp[n^{1/2}t_n\hat{\phi}'\mathbf{O}\boldsymbol{\iota} - \frac{1}{2}nt_n^2] dH_{p-1}(\mathbf{O}) - 1 \right) \left( \int \exp[n^{1/2}t_n\hat{\phi}'\tilde{\mathbf{O}}\boldsymbol{\iota} - \frac{1}{2}nt_n^2] dH_{p-1}(\tilde{\mathbf{O}}) - 1 \right) \right]\end{aligned}$$

$$\begin{aligned}
&= E \left[ \left( \int \exp[n^{1/2}t_n \hat{\phi}' \mathbf{O}\boldsymbol{\iota} - \tfrac{1}{2}nt_n^2] dH_{p-1}(\mathbf{O}) \right) \left( \int \exp[n^{1/2}t_n \hat{\phi}' \tilde{\mathbf{O}}\boldsymbol{\iota} - \tfrac{1}{2}nt_n^2] dH_{p-1}(\tilde{\mathbf{O}}) \right) \right] \\
&\quad - 2 \cdot E \left[ \int \exp[n^{1/2}t_n \hat{\phi}' \mathbf{O}\boldsymbol{\iota} - \tfrac{1}{2}nt_n^2] dH_{p-1}(\mathbf{O}) \right] + 1 \\
&= h_n(t_n) - 2\tilde{h}_n(t_n) + 1.
\end{aligned}$$

In what follows, we show that  $h_n(t_n) \rightarrow 1$ . The convergence  $\tilde{h}_n(t_n) \rightarrow 1$  follows from the same arguments and is omitted for brevity.

Tonelli's Theorem and  $\hat{\phi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p-1})$  imply

$$\begin{aligned}
h_n(t_n) &= E \left[ \int \int \exp[n^{1/2}t_n \hat{\phi}'(\mathbf{O}\boldsymbol{\iota} + \tilde{\mathbf{O}}\boldsymbol{\iota}) - nt_n^2] dH_{p-1}(\mathbf{O}) dH_{p-1}(\tilde{\mathbf{O}}) \right] \\
&= \int \int E \left[ \exp[n^{1/2}t_n \hat{\phi}'(\mathbf{O}\boldsymbol{\iota} + \tilde{\mathbf{O}}\boldsymbol{\iota}) - nt_n^2] \right] dH_{p-1}(\mathbf{O}) dH_{p-1}(\tilde{\mathbf{O}}) \\
&= \int \int \exp[\tfrac{1}{2}nt_n^2 \|\mathbf{O}\boldsymbol{\iota} + \tilde{\mathbf{O}}\boldsymbol{\iota}\|^2 - nt_n^2] dH_{p-1}(\mathbf{O}) dH_{p-1}(\tilde{\mathbf{O}}) \\
&= \int \int \exp[nt_n^2 (\mathbf{O}\boldsymbol{\iota})' \tilde{\mathbf{O}}\boldsymbol{\iota}] dH_{p-1}(\mathbf{O}) dH_{p-1}(\tilde{\mathbf{O}}) \\
&= \int \exp[nt_n^2 \boldsymbol{\iota}' \mathbf{O}\boldsymbol{\iota}] dH_{p-1}(\mathbf{O}).
\end{aligned}$$

Using the notation of Lemma 4, the formula for the normalizing constant of the von Mises-Fisher distribution (see, for instance, equation (9.3.4) of Mardia and Jupp (2000)) implies  $h_n(t_n) = 2^{\tilde{p}/2-1} \cdot \mathcal{I}_{\tilde{p}/2-1}(nt_n^2) \cdot \Gamma(\tilde{p}/2)/(nt_n^2)^{\tilde{p}/2-1}$  where  $\tilde{p} = p - 1$ . Application of Lemma 4 with  $s_n = nt_n^2$  now yields  $h_n(t_n) \rightarrow 1$ , since under  $t_n = o(n^{-1/4})$  and  $p/n \rightarrow c \in (0, 1)$ ,  $s_n^2 = n^2 t_n^4 = o(p/2 - 1)$ .

This concludes the proof under  $\tau_n = 0$ . Now apply this very result to another sequence  $t_n$ ,  $t_n = t'_n$ . Then  $L_n(t'_n)/L_n(0) \xrightarrow{p} 1$  implies via LeCam's first lemma (see, for instance, Lemma 6.4 in van der Vaart (1998)) that in the experiment of observing  $\hat{\tau}_n^2$ , the sequence  $\tau_n = t'_n$  is contiguous to  $\tau_n = 0$ . Thus,  $L_n(t_n)/L_n(0) \xrightarrow{p} 1$  also holds under  $\tau_n = t'_n = o(n^{-1/4})$  by definition of contiguity, which was to be shown.

### A.3 Proof of Theorem 1

Let  $\ell_n(\mathbf{T}_n)$  be the log-likelihood ratio statistic based on  $\mathbf{T}_n$  of testing  $H_0 : (b, a, \tau_n) = (b_0, a_0, \tau_{n,0})$  against  $H_1 : (b, a, \tau_n) = (b_1, a_1, \tau_{n,1})$ . Let  $h_{j,n} = (b_j, b_j + \rho_n a_j)'$  and  $h_j = (b_j, b_j + \rho a_j)'$ ,  $j = 0, 1$ . From (11),

$$\begin{aligned}
\ell_n(\mathbf{T}_n) &= \sqrt{\mathbf{x}'_n \mathbf{x}_n} \left( \begin{array}{c} \hat{\beta}_{\text{long},n} \\ \hat{\beta}_{\text{short},n} - s_n \end{array} \right)' \boldsymbol{\Sigma}(\rho_n)^{-1} (h_{1,n} - h_{0,n}) - \tfrac{1}{2} h'_{1,n} \boldsymbol{\Sigma}(\rho_n)^{-1} h_{1,n} \\
&\quad + \tfrac{1}{2} h'_{0,n} \boldsymbol{\Sigma}(\rho_n)^{-1} h_{0,n} + \log \left( \frac{L_n(\tau_{n,1})}{L_n(\tau_{n,0})} \right)
\end{aligned}$$

and with  $\ell_0(\hat{\mathbf{b}}^o)$  the log-likelihood ratio statistic based on  $\hat{\mathbf{b}}^o$  of testing  $H_0 : (b, a) = (a_0, b_0)$  against  $H_1 : (b, a) = (b_1, a_1)$ ,

$$\ell_0(\hat{\mathbf{b}}^o) = \begin{pmatrix} \hat{b}_{\text{long}} \\ \hat{b}_{\text{short}}^o \end{pmatrix}' \Sigma(\rho)^{-1} (h_1 - h_0) - \frac{1}{2} h_1' \Sigma(\rho)^{-1} h_1 + \frac{1}{2} h_0' \Sigma(\rho)^{-1} h_0.$$

By Lemma 2,

$$\frac{L_n(\tau_{n,1})}{L_n(\tau_{n,0})} = \frac{L_n(\tau_{n,1})}{L_n(0)} \frac{L_n(0)}{L_n(\tau_{n,0})} \xrightarrow{p} 1$$

and from  $\rho_n \rightarrow \rho$ ,  $\sqrt{\mathbf{x}_n' \mathbf{x}_n} (\hat{\beta}_{\text{long},n}, \hat{\beta}_{\text{short},n})' \Rightarrow \hat{\mathbf{b}}^o$  and  $h_{j,n} \rightarrow h_j$  for  $j = 0, 1$ . Thus, under  $H_0$ ,  $\ell_n(\mathbf{T}_n) \Rightarrow \ell_0(\hat{\mathbf{b}}^o)$ . This straightforwardly extends more generally to  $\{\ell_n(\mathbf{T}_n)\}_{(b,a) \in H} \Rightarrow \{\ell_0(\hat{\mathbf{b}}^o)\}_{(b,a) \in H}$  for any finite  $H \subset \mathbb{R}^2$ . Thus, by Definition 9.1 in van der Vaart (1991), under the assumptions of the Lemma, the sequence of experiments of observing  $\mathbf{T}_n$  with local parameter space  $(b, a) \in \mathbb{R}^2$  converges to the experiment of observing  $\hat{\mathbf{b}}^o$ . The first claim now follows from Theorem 15.1 in van der Vaart (1991).

For the second claim, for given  $(b, a)$ , suppose  $E_{\theta_n}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)] \rightarrow E_{b,a}[\phi(\hat{\mathbf{b}}^o)]$  along  $\theta_n = \theta_{n,1}$  with  $\tau_n = \tau_{n,1} = o(n^{-1/4})$ . Let  $\tau_{n,2} = o(n^{-1/4})$  be another sequence, and denote  $\theta_{n,2}$  the corresponding sequence of  $\theta$ . Suppose  $E_{\theta_{n,2}}[\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)]$  does not converge to  $E_{b,a}[\phi(\hat{\mathbf{b}}^o)]$ . By Prohorov's Theorem (see, for instance, Theorem 2.4 in van der Vaart (1998)) and  $0 \leq \varphi_n(\bar{\kappa}_n, \mathbf{Y}_n) \leq 1$ , there exists a subsequence of  $n$  such that  $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$  converges in distribution along that subsequence. Furthermore, by Lemma 2, the likelihood ratio statistic between the corresponding sequences  $\theta_{n,1}$  and  $\theta_{n,2}$  with identical values of  $(b, a)$  converges in probability to one, and this convergence automatically holds jointly with  $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$  along the subsequence. Thus, a trivial application of LeCam's Third Lemma (see, for instance, Theorem 6.6 in van der Vaart (1998)) yields that under  $\theta_{n,2}$ ,  $\varphi_n(\bar{\kappa}_n, \mathbf{Y}_n)$  converges to the same weak limit as under  $\theta_{n,1}$  under the subsequence. But convergence in distribution implies convergence of expectations given that  $0 \leq \varphi_n \leq 1$ , and the desired contradiction follows.

## A.4 Proof of Corollary 1

We use the same notation as the proof of Lemma 1, and momentarily drop the index  $n$  to ease notation. By assumption, the distribution of  $\psi(\mathbf{Y})$  only depends on  $\boldsymbol{\xi}$  through  $(\beta, \Delta, \tau)$ , so  $\psi(\mathbf{Y})$  has the same distribution under  $\boldsymbol{\xi}$  and  $\boldsymbol{\xi}_0$ . By sufficiency, the conditional distribution of  $\psi(\mathbf{Y})$  given  $\hat{\boldsymbol{\vartheta}}$  under  $\boldsymbol{\xi}_0$  does not depend on  $\boldsymbol{\xi}_0$ , so by inverting the probability integral transform conditional on  $\hat{\boldsymbol{\vartheta}}$ , we can write  $\psi(\mathbf{Y}) \sim \psi_S(\hat{\boldsymbol{\vartheta}}, U_S)$  for some function  $\psi_S : \mathbb{R}^{p+2} \mapsto \mathbb{R}$  with  $U_S \sim [0, 1]$  independent of  $\hat{\boldsymbol{\vartheta}}$ .

Let  $\hat{\mathbf{O}}$  be a random rotation matrix drawn from the Haar measure  $H_{p-1}$ , independent of  $(\hat{\boldsymbol{\vartheta}}, U_S)$ . Since by assumption, the distribution of  $\psi(\mathbf{Y})$  does not depend on  $\boldsymbol{\omega}$ , we have

$$\psi_S(\hat{\boldsymbol{\vartheta}}, U_S) \sim \psi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\tau \hat{\mathbf{O}} \boldsymbol{\iota} + \mathbf{e})'), U_S)$$

$$\begin{aligned}
&\sim \psi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, (\tau\boldsymbol{\iota} + \mathbf{e})'\hat{\mathbf{O}}'), U_S) \\
&\sim \psi_S((\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, \hat{\tau}^2\boldsymbol{\iota}'\hat{\mathbf{O}}'), U_S)
\end{aligned}$$

where the second equality follows from the spherical symmetry of the distribution of  $\mathbf{e}$ . Since  $(\hat{\beta}_{\text{long}}, \hat{\beta}_{\text{short}}, \hat{\tau}^2\boldsymbol{\iota}'\hat{\mathbf{O}}')$  is a one-to-one function of  $\mathbf{T}$ , so we can hence write  $\psi(\mathbf{Y}) \sim \tilde{\psi}(\mathbf{T}, U_S)$  for some function  $\tilde{\psi} : \mathbb{R}^4 \mapsto \mathbb{R}$ .

Now reintroducing  $n$  subscripts, consider the sequence of experiments of observing  $(\mathbf{T}'_n, U_S)$ . Recalling that  $U_S$  is independent of  $\mathbf{T}_n$ , these experiments converge to the limit experiment of observing  $(\hat{\mathbf{b}}^o, U_S) \in \mathbb{R}^3$  under the assumptions of the corollary, by the same arguments employed in the proof of Theorem 1. Thus, by the asymptotic representation theorem (Theorem 9.3 in van der Vaart (1998)), there exists a function  $\tilde{\psi}^o : \mathbb{R}^4 \mapsto \mathbb{R} \cup \{+\infty\}$  and uniform random variable  $U$  independent of  $(\hat{\mathbf{b}}^o, U_S)$  such that the limit distribution of  $\tilde{\psi}_n(\mathbf{T}_n, U_S)$  can be written as  $\tilde{\psi}^o(\hat{\mathbf{b}}^o, U_S, U)$ , for all  $(a, b)$ . Since the distribution of  $\tilde{\psi}^o(\hat{\mathbf{b}}^o, U_S, U)$  conditional on  $\hat{\mathbf{b}}^o$  does not depend  $(a, b)$ , there exists a function  $\psi^o : \mathbb{R}^3 \mapsto \mathbb{R} \cup \{+\infty\}$  so that  $\tilde{\psi}^o(\hat{\mathbf{b}}^o, U_S, U) \sim \psi^o(\hat{\mathbf{b}}^o, U)$  for all  $(a, b)$ , as was to be shown.

## A.5 Proof of Lemma 3

We will make use of the following Lemma.

**Lemma 5** *Let  $\mathbf{H}_n$  and  $\hat{\mathbf{H}}_n$  be the Choleski decompositions of  $\boldsymbol{\Omega}_n = \mathbf{H}_n\mathbf{H}'_n$  and  $\hat{\boldsymbol{\Omega}}_n = \hat{\mathbf{H}}_n\hat{\mathbf{H}}'_n$ , respectively, and let  $w_n = (\hat{\beta}_{\text{long},n} - \beta_n, \hat{\beta}_{\text{short},n} - \beta_n - \Delta_n)'$ . Then under (15) and  $\boldsymbol{\Omega}_n^{-1}\hat{\boldsymbol{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$*

- (a)  $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n \xrightarrow{p} \mathbf{I}_2$
- (b)  $\hat{\mathbf{H}}_n^{-1}w_n \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ .

**Proof.** (a) Note that  $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}'_n\hat{\mathbf{H}}_n'^{-1}$ , by similarity, has the same eigenvalues as  $\hat{\mathbf{H}}_n'^{-1}\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}'_n = \hat{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{\Omega}_n \xrightarrow{p} \mathbf{I}_2$ , so they both converge to one in probability. But  $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}'_n\hat{\mathbf{H}}_n'^{-1}$  is symmetric, and all symmetric matrices with eigenvalues converging to one converge to the identity matrix. Thus  $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n\mathbf{H}'_n\hat{\mathbf{H}}_n'^{-1} \xrightarrow{p} \mathbf{I}_2$ , and since  $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n$  is lower triangular, this further implies  $\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n \xrightarrow{p} \mathbf{I}_2$ .

(b) Note that  $\mathbf{H}_n$  is related to  $\boldsymbol{\Omega}_n^{1/2}$  via  $\mathbf{H}_n = \boldsymbol{\Omega}_n^{1/2}\mathbf{O}_n$  for some rotation matrix  $\mathbf{O}_n$ . Thus, also  $\mathbf{H}_n^{-1}w_n = \mathbf{O}_n'\boldsymbol{\Omega}_n^{-1/2}w_n \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ . (Suppose otherwise. Then, by the Cramér-Wold device, for some  $2 \times 1$  vector  $\mathbf{v}$  and  $c \in \mathbb{R}$ ,  $\liminf_{n \rightarrow \infty} |P(\mathbf{v}'\mathbf{O}_n'\boldsymbol{\Omega}_n^{-1/2}w_n > c) - P(\mathcal{N}(\mathbf{0}, \mathbf{v}'\mathbf{v}) > c)| > 0$ . Pick a subsequence along which the liminf is attained, and  $\mathbf{O}_n$  converges. Then we have a contradiction, because the continuous mapping theorem implies the convergence  $P(\mathbf{v}'\mathbf{O}_n'\boldsymbol{\Omega}_n^{-1/2}w_n > c) - P(\mathcal{N}(\mathbf{0}, \mathbf{v}'\mathbf{v}) > c) \rightarrow 0$  along that subsequence.) Invoking Lemma 5 (a), also  $\hat{\mathbf{H}}_n^{-1}w_n = (\hat{\mathbf{H}}_n^{-1}\mathbf{H}_n)\mathbf{H}_n^{-1}w_n \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$  by the continuous mapping theorem. ■

(a) Write  $L(Z_1, Z_2 + g_0, \chi_1, \chi_2)$  for the expression in equation (18). Reparametrize  $\chi$  and  $g_0$  in (18) in terms of  $(r, \phi, u) \in [0, \infty) \times [0, \pi/2) \times [0, 1]$  via  $\chi_1 = r \cos(\phi)$ ,  $\chi_2 = r \sin(\phi)$  and  $u = g_0/\chi_2$  (with  $u = 0$  if  $\chi_2 = 0$ ). By a direct calculation, the limit of  $L(Z_1, Z_2 + ur \cos(\phi), r \cos(\phi), r \sin(\phi))$  as  $r \rightarrow \infty$  exists for almost all  $Z_1, Z_2$  and all  $(u, \phi) \in \mathbb{R} \times [0, \pi/2)$  and is equal to

$$L^\infty(Z_1, Z_2, u, \phi) = \begin{cases} (Z_1 - (1+u) \tan \phi)^2 & \text{if } Z_1 > (1+u) \tan \phi \\ (Z_1 + (1-u) \tan \phi)^2 & \text{if } Z_1 < -(1-u) \tan \phi \\ 0 & \text{otherwise.} \end{cases}$$

Correspondingly,  $\lim_{r \rightarrow \infty} \text{cv}_\Omega((r \cos(\phi), r \sin(\phi))) = \text{cv}_\Omega^\infty(\phi)$  exists, too, and satisfies  $\sup_{0 \leq u \leq 1} P(L^\infty(Z_1, Z_2, u, \phi) \geq \text{cv}_\Omega^\infty(\phi)) \leq \alpha$ . (In general, this inequality is not sharp, since the definition of  $\text{cv}_\Omega^\infty(\phi)$  also requires  $P(L(Z_1, ur \cos(\phi) + Z_2, r \cos(\phi), r \sin(\phi)) \geq \text{cv}_\Omega^\infty(\phi)) \leq \alpha$  for all finite  $r$ ). If  $r \rightarrow \infty$  and  $\phi \rightarrow \pi/2$ , then the limit still exists and is equal to  $L^\infty(Z_1, Z_2, u, \pi/2) = 0$ .

Suppose the assertion of the Lemma is false. Then there exists a subsequence of  $n$  such that along that subsequence,  $\lim_{n \rightarrow \infty} E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] = \limsup_{n \rightarrow \infty} E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] > \alpha$ . Pick a sub-subsequence, such that with  $(r_n, \phi_n, u_n)$  the parameters computed from  $\Omega = \Omega_n$  and  $g_{0,n} = \sqrt{\Omega_{n,11}}\Delta_n/\sqrt{\Omega_{n,11}\Omega_{n,22} - \Omega_{n,12}^2}$ ,  $(r_n, \phi_n, u_n)$  converge along that sub-subsequence to some value  $(r_0, \phi_0, u_0)$  in  $(\mathbb{R} \cup \{\infty\}) \times [0, \pi/2) \times [0, 1]$ . Correspondingly, let  $\text{cv}_0$  be the limit of  $\text{cv}_\Omega((r_n \cos(\phi_n), r_n \sin(\phi_n)))$  along that sub-subsequence (which exists by the above observations also when  $r_n \rightarrow \infty$ , even when  $\phi_0 = \pi/2$ ).

By Lemma 5 (a),  $(\hat{Z}_{n,1}, \hat{Z}_{n,2})' = \hat{\mathbf{H}}_n^{-1} w_n \Rightarrow (Z_1, Z_2)' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$  by the continuous mapping theorem. Since

$$\hat{\mathbf{H}}_n^{-1} = \begin{pmatrix} 1/\sqrt{\hat{\Omega}_{n,11}} & 0 \\ -\frac{\hat{\Omega}_{n,12}}{\sqrt{\hat{\Omega}_{n,11}}\sqrt{\hat{\Omega}_{n,11}\hat{\Omega}_{n,22} - \hat{\Omega}_{n,12}^2}} & \frac{\sqrt{\hat{\Omega}_{n,11}}}{\sqrt{\hat{\Omega}_{n,11}\hat{\Omega}_{n,22} - \hat{\Omega}_{n,12}^2}} \end{pmatrix}$$

the definitions of  $\hat{\chi}_{1,n} = \hat{\chi}_1$  and  $\hat{\chi}_{2,n} = \hat{\chi}_2$  yield

$$\widehat{\text{LR}}_n(\bar{\kappa}_n) = \min_{|g| \leq \hat{\chi}_{2,n}} \left\| \begin{pmatrix} \hat{Z}_{n,1} \\ \hat{Z}_{n,2} + \hat{g}_{0,n} - g \end{pmatrix} \right\|^2 - \min_{h, |g| \leq \hat{\chi}_{2,n}} \left\| \begin{pmatrix} \hat{Z}_{n,1} - h \\ \hat{Z}_{n,2} + \hat{g}_{0,n} - g - \hat{\chi}_{1,n}h \end{pmatrix} \right\|^2$$

where  $\hat{g}_{0,n} = \sqrt{\hat{\Omega}_{n,11}}\Delta_n/\sqrt{\hat{\Omega}_{n,11}\hat{\Omega}_{n,22} - \hat{\Omega}_{n,12}^2}$ . Let  $(\hat{r}_n, \hat{\phi}_n, \hat{u}_n) \in [0, \infty) \times [0, \pi/2) \times [0, 1]$  be such that  $\hat{\chi}_{1,n} = \hat{r}_n \cos(\hat{\phi}_n)$ ,  $\hat{\chi}_{2,n} = \hat{r}_n \sin(\hat{\phi}_n)$  and  $\hat{u}_n = \hat{g}_{0,n}/\hat{\chi}_{2,n}$  (with  $\hat{u}_n = 0$  if  $\hat{\chi}_{2,n} = 0$ ). Write  $\hat{\mathbf{H}}_n^{-1} \mathbf{H}_n \xrightarrow{P} \mathbf{I}_2$  from Lemma 5 (b) element-by-element to conclude that  $\hat{\Omega}_{11,n}/\Omega_{11,n} \xrightarrow{P} 1$ ,  $(\hat{\Omega}_{11,n}\hat{\Omega}_{22,n} - \hat{\Omega}_{12,n}^2)/(\Omega_{11,n}\Omega_{22,n} - \Omega_{12,n}^2) \xrightarrow{P} 1$  and  $(\hat{\Omega}_{12,n} - \Omega_{12,n})/(\Omega_{11,n}\Omega_{22,n} - \Omega_{12,n}^2) \xrightarrow{P} 0$ . Therefore, also  $(\hat{r}_n - r_n)/\max(r_n, 1) \xrightarrow{P} 0$ ,  $\hat{u}_n - u_n \xrightarrow{P} 0$  and  $\hat{\phi}_n - \phi_n \xrightarrow{P} 0$ . Thus, along the sub-subsequence defined above, by the continuous mapping theorem

$$\widehat{\text{LR}}_n(\bar{\kappa}_n) \Rightarrow L_0 = \begin{cases} L(Z_1, u_0 r_0 \cos(\phi_0) + Z_2, r_0 \cos(\phi_0), r_0 \sin(\phi_0)) & \text{if } r_0 < \infty \\ L^\infty(Z_1, Z_2, u_0, \phi_0) & \text{otherwise} \end{cases}$$

and  $E_{\theta_n}[\hat{\varphi}_{\text{LR},n}(\bar{\kappa}_n, \mathbf{Y}_n)] \rightarrow P(L_0 > \text{cv}_0)$ . But by the definition of  $\text{cv}_0$ ,  $P(L_0 > \text{cv}_0) \leq \alpha$ , yielding the desired contradiction.

## A.6 Additional Results in Heteroskedastic Models

Consider the linear model (1) with non-stochastic regressors, so that in vector form

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\beta + \mathbf{Q}\boldsymbol{\delta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ &= \mathbf{R}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \end{aligned}$$

where  $\mathbf{R} = (\mathbf{x}, \mathbf{Q})$ ,  $\boldsymbol{\alpha} = (\beta, \boldsymbol{\delta}')'$ , and as in the main text,  $\mathbf{Q}'\mathbf{x} = \mathbf{0}$  and  $\mathbf{Q}'\mathbf{Z} = \mathbf{0}$ .

We consider a set-up with  $M \rightarrow \infty$  clusters of not necessarily equal size. Write

$$\mathbf{y}_j = \mathbf{R}_j\boldsymbol{\alpha} + \mathbf{Z}_j\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_j$$

for the observations in the  $j$ th cluster (so that the sum of the lengths of the  $\mathbf{y}_j$  vectors over  $j = 1, \dots, M$  equals  $n$ , and  $n$  is implicitly a function of  $M$ ). We allow the sequence of regressors  $\mathbf{R}$  and  $\mathbf{Z}$ , the coefficients  $\boldsymbol{\alpha}$  and  $\boldsymbol{\gamma}$ , the number of observations per cluster, and the distribution of  $\boldsymbol{\varepsilon}_j$  to depend on  $M$  in a double array fashion. In particular, this allows for the number of regressors  $p$  and/or  $m$  to be proportional to the sample size. To ease notation, we do not make this dependence on  $M$  explicit.

Define the  $n \times 2$  matrix  $\mathbf{v} = (\mathbf{v}'_1, \dots, \mathbf{v}'_M)'$ . Let  $\|\cdot\|$  be the spectral norm.

**Condition 1** (a)  $\boldsymbol{\varepsilon}_j$ ,  $j = 1, \dots, M$  are independent with  $E[\boldsymbol{\varepsilon}_j] = \mathbf{0}$  and  $E[\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}'_j] = \boldsymbol{\Sigma}_j$ .

(b)  $\|(M^{-1} \sum_{j=1}^M \mathbf{v}'_j \boldsymbol{\Sigma}_j \mathbf{v}_j)^{-1}\| = O(1)$ ,  $\max_j \|\mathbf{v}_j\|^4 \cdot \sum_{j=1}^M E[\|\boldsymbol{\varepsilon}_j\|^4] = o(M^2)$ .

(c)  $\|M^{-1} \sum_{j=1}^M \mathbf{R}_j \mathbf{R}'_j\| = O(1)$ ,  $\|(M^{-1} \sum_{j=1}^M \mathbf{R}_j \mathbf{R}'_j)^{-1}\| = O(1)$ ,  $\max_j \|\boldsymbol{\Sigma}_j\| = o(M)$ ,  $\max_j \|\boldsymbol{\Sigma}_j\| \cdot \max_j \|\mathbf{v}_j\|^4 = O(M)$  and  $\max_j \|\mathbf{v}_j\|^2 = O(M)$ .

(d)  $\max_j \|\mathbf{v}_j\|^2 \cdot \kappa^2 = o(M/n)$  and  $\max_j \|\boldsymbol{\Sigma}_j\| \cdot \max_j \|\mathbf{v}_j\|^4 \cdot \kappa^2 = o(M^2/n)$ , where  $\kappa^2 = \boldsymbol{\gamma}' \mathbf{Z}' \mathbf{Z} \boldsymbol{\gamma} / n$ .

**Theorem 2** (a) Under Condition 1 (a) and (b), as  $M \rightarrow \infty$ ,

$$\boldsymbol{\Omega}_n^{-1/2} M^{-1} \sum_{j=1}^M \mathbf{v}'_j \mathbf{y}_j \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$$

where  $\boldsymbol{\Omega}_n = M^{-2} \sum_{j=1}^M \mathbf{v}'_j \boldsymbol{\Sigma}_j \mathbf{v}_j$ ;

(b) Under Condition 1 (a)-(d),  $\boldsymbol{\Omega}_n^{-1} \hat{\boldsymbol{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$ , where

$$\hat{\boldsymbol{\Omega}}_n = M^{-2} \sum_{j=1}^M \mathbf{v}'_j \hat{\boldsymbol{\varepsilon}}_j \hat{\boldsymbol{\varepsilon}}'_j \mathbf{v}_j \text{ and } \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{y}. \quad (28)$$

This result immediately implies the following.

**Corollary 2** (a) Let  $\mathbf{v}$  be such that  $M^{-1}\mathbf{v}'\mathbf{y} = M^{-1}\sum_{j=1}^M \mathbf{v}'_j \mathbf{y}_j = (\hat{\beta}_{long}, \hat{\beta}_{short})'$ , and assume that Condition 1 holds. Then (15) holds, and  $\mathbf{\Omega}_n^{-1}\hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$  with  $\hat{\mathbf{\Omega}}_n$  defined in (28).

(b) Let the  $j$ th row of  $\mathbf{v}$  be equal to  $(\hat{w}_j^z, w_j)$  as defined in Section 4, and assume that Condition 1 holds. Then under  $\beta = 0$ , (21) holds, and  $\mathbf{\Omega}_n^{-1}\hat{\mathbf{\Omega}}_n \xrightarrow{p} \mathbf{I}_2$  with  $\hat{\mathbf{\Omega}}_n$  defined in (28).

**Remark 3** Since  $(\hat{\beta}_{long}, \hat{\beta}_{short})$  and the IV score in (21) are  $\mathbf{v}_j$ -weighted averages of  $\boldsymbol{\varepsilon}_j$ , some bound on the relative magnitude of  $\|\mathbf{v}_j\|$  is necessary to obtain asymptotic normality. The bounds in Condition 1 are relatively weak, allowing for  $\max_j \|\mathbf{v}_j\| = o(M^{1/4})$  (if  $\sum_{j=1}^M E[\|\boldsymbol{\varepsilon}_j\|^4] = O(M)$ ,  $\max_j \|\boldsymbol{\Sigma}_j\| = O(1)$  and  $\kappa^2 = o(M^{1/2})$ ). At the same time, one could also imagine that  $\max_j \|\mathbf{v}_j\| = O(1)$ , which would then allow for  $E[\|\boldsymbol{\varepsilon}_j\|^4] = o(M)$ , either because of increasingly fat tails, or because the number of observations per cluster is growing.

The result in part (a) makes no assumptions on  $\boldsymbol{\gamma}$ , so no restrictions are put on the asymptotic behavior of  $\kappa_n$  or  $\tau_n$ .

The definition of  $\hat{\mathbf{\Omega}}_n$  in part (b) for  $M^{-1}\mathbf{v}'\mathbf{y} = (\hat{\beta}_{long}, \hat{\beta}_{short})'$  is the standard clustered variance estimator, except that the regression residuals are computed from the short regression. Under  $\max_j \|\mathbf{v}_j\| = O(1)$  and  $\max_j \|\boldsymbol{\Sigma}_j\| = O(1)$ ,  $\kappa^2 = o(M/n)$  is enough to obtain consistency of  $\hat{\mathbf{\Omega}}_n$ . The important special case of independent but heteroskedastic disturbances  $\varepsilon_i$  (so that  $\hat{\mathbf{\Omega}}_n$  reduces to the White (1980) standard errors based on short regression residuals), is obtained for  $M = n$ .

**Proof.** (a) By the Cramér-Wold device, it suffices to show that  $M^{-1}\mathbf{v}'\mathbf{v}'\boldsymbol{\varepsilon}/\sqrt{\mathbf{v}'\mathbf{\Omega}_n\mathbf{v}} \Rightarrow \mathcal{N}(0, 1)$  for all  $2 \times 1$  vectors  $\mathbf{v}$  with  $\mathbf{v}'\mathbf{v} = 1$ . This follows from the (triangular array version of the) Lyapunov central limit theorem applied to the  $M$  independent variables  $\mathbf{v}'\mathbf{v}'_j\boldsymbol{\varepsilon}_j \sim (0, \mathbf{v}'\mathbf{v}'_j\boldsymbol{\Sigma}_j\mathbf{v}_j\mathbf{v})$  and Condition 1 (b), since

$$\frac{\sum_{j=1}^M E[(\mathbf{v}'\mathbf{v}'_j\boldsymbol{\varepsilon}_j)^4]}{(\sum_{j=1}^M \mathbf{v}'\mathbf{v}'_j\boldsymbol{\Sigma}_j\mathbf{v}_j\mathbf{v})^2} \leq \max_j \|\mathbf{v}_j\|^4 \cdot M^{-2} \sum_{j=1}^M E[\|\boldsymbol{\varepsilon}_j\|^4] \cdot \|(M^{-1} \sum_{j=1}^M \mathbf{v}'_j\boldsymbol{\Sigma}_j\mathbf{v}_j)^{-1}\|^2 \rightarrow 0$$

and  $\text{Var}[M^{-1}\mathbf{v}'\mathbf{v}'\boldsymbol{\varepsilon}/\sqrt{\mathbf{v}'\mathbf{\Omega}_n\mathbf{v}}] = 1$ .

(b) We show convergence of  $\mathbf{v}'\hat{\mathbf{\Omega}}_n\mathbf{v}/(\mathbf{v}'\mathbf{\Omega}_n\mathbf{v}) \xrightarrow{p} 1$  for all  $2 \times 1$  vectors  $\mathbf{v}$  with  $\mathbf{v}'\mathbf{v} = 1$ . Note that  $\mathbf{v}'\hat{\mathbf{\Omega}}_n\mathbf{v} = M^{-2}\sum_{j=1}^M \hat{\boldsymbol{\varepsilon}}'_j \mathbf{V}_j \hat{\boldsymbol{\varepsilon}}_j = M^{-2}\hat{\boldsymbol{\varepsilon}}'\mathbf{V}\hat{\boldsymbol{\varepsilon}}$  with  $\mathbf{V}_j = \mathbf{v}_j\mathbf{v}'\mathbf{v}'_j$  and  $\mathbf{V} = \text{diag}(\mathbf{V}_1, \dots, \mathbf{V}_M)$ , and

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} &= \mathbf{M}_R\boldsymbol{\varepsilon} + \mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} \\ &= \boldsymbol{\varepsilon} - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\varepsilon} + \mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} \end{aligned} \tag{29}$$

with  $\mathbf{M}_R = \mathbf{I}_n - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$ , so that

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}'\mathbf{V}\hat{\boldsymbol{\varepsilon}} &= \boldsymbol{\varepsilon}'\mathbf{V}\boldsymbol{\varepsilon} + \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} + 2\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\boldsymbol{\varepsilon} - 2\boldsymbol{\varepsilon}'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\varepsilon} \\ &\quad + \boldsymbol{\varepsilon}'\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{V}\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\boldsymbol{\varepsilon}.\end{aligned}$$

Now

$$\begin{aligned}\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} &\leq \|\mathbf{V}\| \cdot \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} \\ &\leq \max_j \|\mathbf{v}_j\|^2 \cdot \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\gamma} = \max_j \|\mathbf{v}_j\|^2 \cdot n\kappa^2\end{aligned}$$

and, with  $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_M)$ ,

$$\begin{aligned}\text{Var}[\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\boldsymbol{\varepsilon}] &= \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\boldsymbol{\Sigma}\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} \\ &\leq \max_j \|\boldsymbol{\Sigma}_j\| \cdot \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{V}\mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} \\ &\leq \max_j \|\boldsymbol{\Sigma}_j\| \cdot \boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}_R\mathbf{V}^2\mathbf{M}_R\mathbf{Z}\boldsymbol{\gamma} \\ &\leq \max_j \|\boldsymbol{\Sigma}_j\| \cdot \max_j \|\mathbf{v}_j\|^4 \cdot n\kappa^2\end{aligned}$$

and

$$\begin{aligned}\|\text{Var}[\mathbf{R}'\boldsymbol{\varepsilon}]\| &= \|\mathbf{R}'\boldsymbol{\Sigma}\mathbf{R}\| \leq \max_j \|\boldsymbol{\Sigma}_j\| \cdot \sum_{j=1}^M \|\mathbf{R}_j\|^2 \\ \|\text{Var}[\mathbf{R}'\mathbf{V}\boldsymbol{\varepsilon}]\| &= \|\mathbf{R}'\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}\mathbf{R}\| \leq \max_j \|\boldsymbol{\Sigma}_j\| \cdot \max_j \|\mathbf{v}_j\|^4 \cdot \sum_{j=1}^M \|\mathbf{R}_j\|^2 \\ \|\mathbf{R}'\mathbf{V}\mathbf{R}\| &\leq \max_j \|\mathbf{v}_j\|^2 \cdot \sum_{j=1}^M \|\mathbf{R}_j\|^2.\end{aligned}$$

Furthermore,  $(\mathbf{v}'\boldsymbol{\Omega}_n\mathbf{v})^{-1} = (M^{-2} \sum_{j=1}^M \mathbf{v}'\mathbf{v}'_j\boldsymbol{\Sigma}_j\mathbf{v}_j\mathbf{v})^{-1} \leq \|(M^{-2} \sum_{j=1}^M \mathbf{v}'_j\boldsymbol{\Sigma}_j\mathbf{v}_j)^{-1}\| = O(M^{-1})$ , so that under Condition 1 (b)-(d),  $M^{-2}(\hat{\boldsymbol{\varepsilon}}'\mathbf{V}\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}'\mathbf{V}\boldsymbol{\varepsilon})/(\mathbf{v}'\boldsymbol{\Omega}_n\mathbf{v}) \xrightarrow{p} 0$ .

Finally, rewrite  $\boldsymbol{\varepsilon}'\mathbf{V}\boldsymbol{\varepsilon} = \sum_{j=1}^M \mathbf{v}'\mathbf{v}'_j\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}'_j\mathbf{v}_j\mathbf{v}$ . Then  $E[M^{-1}\boldsymbol{\varepsilon}'\mathbf{V}\boldsymbol{\varepsilon} - M\mathbf{v}'\boldsymbol{\Omega}_n\mathbf{v}] = 0$ , and

$$\begin{aligned}\text{Var}[M^{-1}\boldsymbol{\varepsilon}'\mathbf{V}\boldsymbol{\varepsilon} - M\mathbf{v}'\boldsymbol{\Omega}_n\mathbf{v}] &= M^{-2} \sum_{j=1}^M \text{Var}[\mathbf{v}'\mathbf{v}'_j(\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}'_j - \boldsymbol{\Sigma}_j)\mathbf{v}_j\mathbf{v}] \\ &\leq M^{-2} \max_j \|\mathbf{v}_j\|^4 \cdot \sum_{j=1}^M E[\|\boldsymbol{\varepsilon}_j\|^4]\end{aligned}$$

and the result follows from  $(\mathbf{v}'\boldsymbol{\Omega}_n\mathbf{v})^{-1} = O(M^{-1})$  and Condition 1 (a). ■

## A.7 Asymptotics under Double Bounds

Let  $\mathbf{S} = (\mathbf{Q}, \mathbf{Z})$ , and the following treats  $\mathbf{S}$  as non-stochastic (or conditions on its realization). Straightforward algebra yields that under  $\beta = 0$ ,

$$\sum_{i=1}^n \hat{x}_i^z y_i = \hat{\mathbf{x}}^z \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_x' \mathbf{M}_S \boldsymbol{\varepsilon} \quad (30)$$

$$\sum_{i=1}^n x_i y_i = \boldsymbol{\gamma}' \mathbf{Z}' \mathbf{Z} \boldsymbol{\gamma}_x + \boldsymbol{\varepsilon}_x' \mathbf{Z} \boldsymbol{\gamma} + \mathbf{x}' \boldsymbol{\varepsilon} \quad (31)$$

$$= \boldsymbol{\gamma}' \mathbf{Z}' \mathbf{Z} \boldsymbol{\gamma}_x + \boldsymbol{\varepsilon}_x' \mathbf{Z} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_x' \mathbf{M}_Q \boldsymbol{\varepsilon} + \boldsymbol{\gamma}_x' \mathbf{Z}' \boldsymbol{\varepsilon} \quad (32)$$

where  $\mathbf{M}_S$  and  $\mathbf{M}_Q$  are the  $n \times n$  projection matrices associated with  $\mathbf{Q}$  and  $\mathbf{S}$  with elements  $M_{Q,ij}$  and  $M_{S,ij}$ , respectively. With  $\Delta^{\text{Dbl}} = n^{-1} \boldsymbol{\gamma}' \mathbf{Z}' \mathbf{Z} \boldsymbol{\gamma}_x$ , the bound  $|\Delta^{\text{Dbl}}| \leq \bar{\kappa} \cdot \bar{\kappa}_x$  follows from the Cauchy-Schwarz inequality.

For fixed and finite  $p$ , standard arguments yield a CLT (24) and an associate asymptotic covariance estimator. For diverging  $p$ , more careful arguments are required, as discussed in Cattaneo et al. (2018a, 2018b). In particular, by the Cramér-Wold device, and arguments very similar to the ones employed in the proof of Lemma A.2 of Chao, Swanson, Hausman, Newey, and Woutersen (2012), one obtains the following result.

**Lemma 6** *Suppose that  $(\varepsilon_{x,i}, \varepsilon_i)$  are mean-zero independent across  $i$ ,  $E[\varepsilon_{x,i} \varepsilon_i] = 0$ , and for some  $C$  that does not depend on  $n$ ,  $E[\varepsilon_{x,i}^4] < C$ ,  $E[\varepsilon_i^4] < C$  and  $E[\varepsilon_{x,i}^4 \varepsilon_i^4] < C$  almost surely. If  $p \rightarrow \infty$ , then (24) holds with*

$$\boldsymbol{\Omega}^{\text{Dbl}} = n^{-2} \begin{pmatrix} \sum_{i,j} M_{S,ij}^2 E[\varepsilon_{x,i}^2 \varepsilon_j^2] & \sum_{i,j} M_{S,ij} M_{Q,ij} E[\varepsilon_{x,i}^2 \varepsilon_j^2] \\ \sum_{i,j} M_{S,ij} M_{Q,ij} E[\varepsilon_{x,i}^2 \varepsilon_j^2] & \sum_{i,j} M_{Q,ij}^2 E[\varepsilon_{x,i}^2 \varepsilon_j^2] + \sum_{i=1}^n E[(z_i' \boldsymbol{\gamma})^2 \varepsilon_{x,i}^2 + (z_i' \boldsymbol{\gamma}_x)^2 \varepsilon_i^2] \end{pmatrix}.$$

In the high-dimensional case with  $p/n \rightarrow c \in (0, 1)$ , it is not obvious how one would obtain a consistent estimator of  $n\boldsymbol{\Omega}^{\text{Dbl}}$  in general, because it is difficult to estimate  $\boldsymbol{\gamma}$  and  $\boldsymbol{\gamma}_x$  with sufficient precision. We leave this question for future research.

In order to make further progress, suppose that  $\mathbf{S}$  is such that  $\|\boldsymbol{\Omega}^{\text{Dbl}}\| = O(n)$  and  $\|(\boldsymbol{\Omega}^{\text{Dbl}})^{-1}\| = O(n^{-1})$ , where  $\|\cdot\|$  is the spectral norm. Assume further that  $\kappa = o(1)$ . Then under the assumptions of Lemma 6, or other weak dependence assumptions,  $\text{Var}[\boldsymbol{\varepsilon}_x' \mathbf{Z} \boldsymbol{\gamma}] = o(n)$ . The term  $\boldsymbol{\varepsilon}_x' \mathbf{Z} \boldsymbol{\gamma}$  in (31) thus no longer makes a contribution to the asymptotic distribution. Under these assumptions, one can therefore proceed as in Section A.6 with  $\mathbf{v} = (\hat{\mathbf{x}}^z, \mathbf{x})$  in Condition 1 to obtain both an alternative CLT (24) under clustering, and an appropriate estimator  $\hat{\boldsymbol{\Omega}}^{\text{Dbl}}$  conditional on  $\mathbf{x}$ .

## References

- ANDERSON, T. W., AND H. RUBIN (1949): “Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations,” *The Annals of Mathematical Statistics*, 21, 570–582.
- ARMSTRONG, T. B., AND M. KOLESÁR (2016): “Optimal inference in a class of regression models,” *Working Paper, Princeton University*.
- (2018): “Optimal inference in a class of regression models,” *Econometrica*, 86.
- BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): “Inference on treatment effects after selection among high-dimensional controls,” *The Review of Economic Studies*, 81(2), 608–650.
- CATTANEO, M. D., M. JANSSON, AND W. K. NEWEY (2018a): “Alternative asymptotics and the partially linear model with many regressors,” *Econometric Theory*, 34(2), 277–301.
- (2018b): “Inference in linear regression models with many covariates and heteroscedasticity,” *Journal of the American Statistical Association*, 113(523), 1350–1361.
- CHAO, J. C., N. R. SWANSON, J. A. HAUSMAN, W. K. NEWEY, AND T. WOUTERSEN (2012): “Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments,” *Econometric Theory*, 28(1), 42–86.
- CONLEY, T. G., C. B. HANSEN, AND P. E. ROSSI (2012): “Plausibly exogenous,” *Review of Economics and Statistics*, 94(1), 260–272.
- COX, D. R., AND D. V. HINKLEY (1974): *Theoretical statistics*. Chapman&Hall/CRC, New York.
- DONOHU, D. L. (1994): “Statistical estimation and optimal recovery,” *The Annals of Statistics*, 22(1), 238–270.
- DONOHUE, J. J., AND S. D. LEVITT (2001): “The Impact of Legalized Abortion on Crime,” *Quarterly Journal of Economics*, CXVI, 379–420.
- ELLIOTT, G., U. K. MÜLLER, AND M. W. WATSON (2015): “Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis,” *Econometrica*, 83, 771–811.

- FAN, J., AND R. LI (2001): “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96(456), 1348–1360.
- FOOTE, C. L., AND C. F. GOETZ (2008): “The impact of legalized abortion on crime: Comment,” *The Quarterly Journal of Economics*, 123, 407–423.
- HOERL, A. E., AND R. W. KENNARD (1970): “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, 12(1), 55–67.
- IMBENS, G. W., AND D. B. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York, An introduction.
- IMBENS, G. W., D. B. RUBIN, AND B. I. SACERDOTE (2001): “Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players,” *American Economic Review*, 91, 778–794.
- IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86.
- JOYCE, T. (2004): “Did legalized abortion lower crime?,” *Journal of Human Resources*, 39(1), 1–28.
- (2009): “A simple test of abortion and crime,” *The Review of Economics and Statistics*, 91(1), 112–123.
- LEEB, H., AND B. M. PÖTSCHER (2005): “Model Selection and Inference: Facts and Fiction,” *Econometric Theory*, 21, 21–59.
- LEEB, H., AND B. M. PÖTSCHER (2008a): “Can one estimate the unconditional distribution of post-model-selection estimators?,” *Econometric Theory*, 24(2), 338–376.
- (2008b): “Recent Developments in Model Selection and Related Areas,” *Econometric Theory*, 24(2), 319–322.
- MACCHIAVELLO, R., AND A. MORJARIA (2015): “The Value of Relationships: Evidence from a Supply Shock to Kenyan Rose Exports,” *American Economic Review*, 105, 2911–2945.
- MARDIA, K. V., AND P. E. JUPP (2000): *Directional statistics*, Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester.

- MÜLLER, U. K., AND A. NORETS (2012): “Credibility of Confidence Sets in Nonstandard Econometric Problems,” *Working Paper, Princeton University*.
- MÜLLER, U. K., AND Y. WANG (2015): “Nearly Weighted Risk Minimal Unbiased Estimation,” *Working Paper, Princeton University*.
- OBENCHAIN, R. L. (1977): “Classical F-Tests and Confidence Regions for Ridge Regression,” *Technometrics*, 19, 429–439.
- PRATT, J. W. (1961): “Length of Confidence Intervals,” *Journal of the American Statistical Association*, 56, 549–567.
- TIBSHIRANI, R. (1996): “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.
- VAN DER VAART, A. W. (1991): “An asymptotic representation theorem,” *International Statistical Review*, 259, 97–121.
- (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.
- WHITE, H. (1980): “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48, 817–830.