Inference under shape restrictions*

Joachim Freyberger† Brandon Reeves‡

June 7, 2018

Abstract

We propose a uniformly valid inference method for an unknown function or parameter vector satisfying certain shape restrictions. The method applies very generally, namely to a wide range of finite dimensional and nonparametric problems, such as regressions or instrumental variable estimation, to both kernel or series estimators, and to many different shape restrictions. One application of our inference method is to construct uniform confidence bands for an unknown function of interest. These bands are build around a shape restricted estimator and the upper and lower bound functions are consistent with the shape restrictions. Moreover, the bands are asymptotically equivalent to standard unrestricted confidence bands if the true function strictly satisfies all shape restrictions, but they can be much smaller if some of the shape restrictions are binding or close to binding. We illustrate these sizable width gains as well as the wide applicability of our method in Monte Carlo simulations and in an empirical application.

Keywords: Shape restrictions, inference, nonparametric, uniform confidence bands.

*We thank Tim Armstrong, Richard Blundell, Andrew Chesher, Ivan Canay, Bruce Hansen, Joseph Hardwick, Joel Horowitz, Philipp Ketz, Matt Masten, Francesca Molinari, Taisuke Otsu, Jack Porter, Azeem Shaikh, Xiaoxia Shi, Alex Torgovitsky, Daniel Wilhelm, as well as seminar and conference participants at UW Madison, UCL, LSE, Boston College, Northwestern, Humboldt University, Bonn University, NYU, Stanford, the CEME conference at UCLA, and the interactions conference at the University of Chicago for helpful comments and discussions. We also thank Richard Blundell, Joel Horowitz, and Matthias Parey for sharing their data.

†Department of Economics, University of Wisconsin - Madison. Email: jfreyberger@ssc.wisc.edu.

‡Department of Economics, University of Wisconsin - Madison. Email: breeves@wisc.edu.

1 Introduction

Researchers can often use either parametric or nonparametric methods to estimate the parameters of a model. Parametric estimators have favorable properties, such as good finite sample precision and fast rates of convergence, and it is usually straightforward to use them for inference. However, parametric models are often misspecified. Specifically, economic theory rarely implies a particular functional form, such as a linear or quadratic demand function, and conclusions drawn from an incorrect parametric model can be misleading. Nonparametric methods, on the other hand, do not impose strong functional form assumptions, but as a consequence, confidence intervals obtained from them are often much wider.

In this paper we explore shape restrictions to restrict the class of functions but without imposing arbitrary parametric assumptions. Shape restrictions are often reasonable assumptions, such as assuming that the return to eduction is positive, and they can be implied by economic theory. For example, demand functions are generally monotonically decreasing in prices, cost functions are monotonically increasing, homogeneous of degree 1, and concave in input prices, Engel curves of normal goods are monotonically increasing, economies of scale yield subadditive average cost functions, and utility functions of risk averse agents are concave. Additionally, statistical theory can imply shape restrictions, such as noncrossing conditional quantile curves. There is a long history of estimation under shape restrictions in econometrics and statistics and obtaining shape restricted estimators is simple in many settings. Moreover, shape restricted estimators can have much better finite sample properties, such as lower mean squared errors, compared to unrestricted estimators.

Using shape restrictions for inference is much more complicated than simply obtaining a restricted estimator. The main reason is that the distribution of the restricted estimator depends on where the shape restrictions bind, which is unknown a priori. In this paper we propose a uniformly valid inference method for an unknown function or parameter vector satisfying certain shape restrictions, which can be used to test hypotheses and to obtain confidence sets. The method applies very generally, namely to a wide range of finite dimensional and nonparametric problems, such as regressions or instrumental variable estimation, to both kernel or series estimators, and to many different shape restrictions. Our confidence sets are well suited to be reported along with shape restricted estimates, because they are build around restricted estimators and eliminate regions of the parameter space that are inconsistent with the shape restrictions.

One major application of our inference method is to construct uniform confidence bands for a function. Such a band consists of a lower bound function and an upper bound function

such that the true function is between them with at least a pre-specified probability. These bands are useful to summarize statistical uncertainty and they allow the reader to easily assess statistical accuracy and perform various hypothesis tests about the function without access to the data. Our confidence bands have desirable properties. In particular, they always include the shape restricted estimator of the function and are therefore never empty. Moreover, they are asymptotically equivalent to standard unrestricted confidence bands if the true function strictly satisfies all shape restrictions (e.g. if the true function is strictly increasing but the shape restriction is that it is weakly increasing). However, if for the true function some of the shape restrictions are binding or close to binding, our confidence bands are generally much smaller. The decrease in the width reflects the increased precision of the constrained estimator. Finally, the proposed method provides uniformly valid inference over a large class of distributions, which in particular implies that the confidence bands do not suffer from under-coverage if some of the shape restrictions are close to binding. These cases are empirically relevant. For example, demand functions are likely to be strictly decreasing, but nonparametric estimates are often not monotone, suggesting that the demand function is close to constant for some prices.[1] Our method applies very generally. For example, our paper is the first to provide such inference results for the nonparametric instrumental variables (NPIV) model under general shape constraints.

Similar to many other nonstandard inference problems, instead of trying to obtain confidence sets directly from the asymptotic distribution of the estimator, our inference procedure is based on test inversion.[2] This means that we start by testing the null hypothesis that the true parameter vector θ_0 is equal to some fixed value $\bar{\theta}$. In series estimation θ_0 represents the coefficients in the series approximation of a function and θ_0 can therefore grow in dimension as the sample size increases. The major advantage of the test inversion approach is that under the null hypothesis we know exactly which of the shape restrictions are binding or close to binding. Therefore, under the null hypothesis, we can approximate the distribution of the estimator in large samples and we can decide whether or not we reject the null hypothesis. The confidence set for θ_0 consists of all values for which the null hypothesis is not rejected.

To obtain uniform confidence bands or confidence sets for other functions of θ_0, such as average derivatives, we project onto the confidence set for θ_0 (see Section 2 for a simple illustration). We choose the test statistic in a way that our confidence sets are asymptotically equivalent to standard unrestricted confidence sets if θ_0 is sufficiently in the interior of the pa-

[1] Analogously to many other papers, closeness to the boundary is relative to the sample size.

[2] Other nonstandard inference settings include autoregressive models (e.g. Mikusheva 2007), weak identification (e.g. Andrews and Cheng 2012), and partial identification (e.g. Andrews and Soares 2010).

rameter space. Thus, in this case, the confidence sets have the right coverage asymptotically. If some of the shape restrictions are binding or close to binding, our inference procedure will generally be conservative due to the projection. However, in these cases we also obtain very sizable width gains compared to a standard unrestricted confidence set. Furthermore, due to test inversion and projections, our inference method can be computationally demanding. We provide details on the computational costs and compare them with alternative approaches in Section 6.1. We also briefly describe a method recently suggested by Kaido, Molinari, and Stoye (2016) in a computationally similar problem in the moment inequality literature, which also applies to our framework, and can reduce these costs considerably.

In Monte Carlo simulations we construct uniform confidence bands in a series regression framework and in the NPIV model under a monotonicity constraint. In the NPIV model the gains of using shape restrictions are generally much higher. For example, we show that with a fourth order polynomial approximation of the true function, the average width gains can be up to 73%, depending on the slope of the true function. We also obtain large widths gains for confidence intervals for the average derivative of the function. Finally, in an empirical application, we estimate demand functions for gasoline, subject to the functions being weakly decreasing, and we provide uniform confidence bands build around restricted estimates with monotone upper and lower bound functions. In this setting, the width gains from using these shape restrictions are between 25% and 45%.

We now explain how our paper fits into the related literature. There is a vast literature on estimation under shape restrictions going back to Hildreth (1954) and Brunk (1955) who suggest estimators under concavity and monotonicity restrictions, respectively. Other related work includes, among many others, Mukerjee (1988), Dierckx (1980), Ramsay (1988), Mammen (1991a), Mammen (1991b), Mammen and Thomas-Agnan (1999), Hall and Huang (2001), Haag, Hoderlein, and Pendakur (2009), Du, Parmeter, and Racine (2013), and Wang and Shen (2013). See also Delecroix and Thomas-Agnan (2000) and Henderson and Parmeter (2009) for additional references. Many of the early papers focus on implementation issues and subsequent papers discuss rates of convergence of shape restricted estimators. Many inference results, such as those by Mammen (1991b), Groeneboom, Jongbloed, and Wellner (2001), Dette, Neumeyer, and Pilz (2006), Birke and Dette (2007), and Pal and Woodroofe (2007) are for points of the function where the shape restrictions do not bind. It is also well known that a shape restricted estimator has a nonstandard distribution if the shape restrictions bind; see for example Wright (1981) and Geyer (1994). Freyberger and Horowitz (2015) provide inference methods in a partially identified NPIV model under shape restrictions with discrete

regressors and instruments. Empirical applications include Matzkin (1994), Lewbel (1995), Ait-Sahalia and Duarte (2003), Beresteanu (2005), and Blundell, Horowitz, and Parey (2012, 2017). There is also an interesting literature on risk bounds (e.g. Zhang (2002), Chatterjee, Guntuboyina, and Sen (2015), Chetverikov and Wilhelm (2017)) showing, among others, that a restricted estimator can have a faster rate of convergence than an unrestricted estimator when the true function is close to the boundary. In addition, there is a large, less related literature on testing shape restrictions. See also Chetverikov, Santos, and Shaikh (2018) for a recent review of the econometrics of shape restrictions.

There are several existing methods which can be used to obtain uniform confidence bands under shape restrictions. First, there are a variety of existing confidence bands, some of which are tailored to specific shape restrictions, such as the ones in Dümbgen (1998, 2003), which have the feature that they can be empty with positive probability. As a very simple example, one could take a standard unrestricted band and intersect it with all functions satisfying the shape restrictions. While one could interpret an empty band as evidence against the shape restrictions, these bands can also be arbitrarily small, and the width might therefore not adequately reflect the finite sample uncertainty. More formally, these bands do not satisfy the "reasonableness" property of Müller and Norets (2016). The method of Dümbgen (2003) only applies to a regression model with fixed regressors and normally distributed errors, but he shows that his bands adapt to the smoothness of the unknown function.[3] Our method applies much more generally and covers, among other, the NPIV model under general shape constraints, but investigating analogous adaptivity results is out of scope of the current paper. However, we briefly compare the two approaches in simulations. We then also show that our bands are on average much narrower than simple monotonized bands. The second possibility is to use the rearrangement approach of Chernozhukov, Fernandez-Val, and Galichon (2009), which works with monotonicity restrictions and is very easy to implement. However, the average width does not change by rearranging a band. Finally, in a kernel regression framework with very general constraints, one could use a two step procedure by Horowitz and Lee (2017). In the first step, they estimate the points where the shape

[3]Cai, Low, and Xia (2013) focus on confidence intervals for the function evaluated at a point in the normal regression model and they show that the intervals adapt to each individual function under monotonicity and convexity constraints. Bellec (2016) constructs polyhedron type confidence regions for the entire conditional mean vector in the normal regression model with shape restrictions, but without using any smoothness assumptions, and he shows that they adapt the dimension of the smallest face of the polyhedron. These confidence sets are not directly useful for constructing uniform confidence bands or confidence intervals for functionals, because projecting onto them would yield very conservative confidence sets.

restrictions bind. In the second step, they estimate the function under equality constraints and hence, they obtain an asymptotically normally distributed estimator, which they can use to obtain uniform confidence bands. While their approach is computationally much simpler than ours, their main result leads to bands which can suffer from under-coverage if some of the shape restrictions are close to binding. They also suggest using a bias correction term to improve the finite sample coverage probability, but they do not provide any theoretical results for this method. To the best of our knowledge, our method is the first that yields uniform confidence bands, which are uniformly valid, yield width reductions when the shape restrictions are binding or close to binding, and are never empty.

An additional closely related paper, which does not consider uniform confidence bands and therefore does not fit into any of the categories above, is Chernozhukov, Newey, and Santos (2015). They develop a general testing procedure in a conditional moments setting, which can be used to test shape restrictions and to obtain confidence regions for functionals under shape restrictions. They allow for partial identification, while we assume point identification, but we study a general setup, which includes for example maximum likelihood estimation and conditional moments models. Even though there is some overlap in the settings where both methods apply, their approach is conceptually very different to ours. Similar to many other papers in the partial identification literature, to obtain confidence regions for functionals, they use a test inversion procedure, which jointly tests certain features of the model. In particular, they invert a joint test of the null hypothesis that the shape restrictions hold and that a functional takes a particular value. Consequently, the resulting confidence regions represent both uncertainty about the value of the functional and the shape restrictions, these sets can be empty (which could be interpreted as evidence against the shape restrictions), and they can be arbitrarily small. Contrarily, we impose the shape restrictions and test a null hypothesis about the parameter vector only. Thus, we treat these restrictions and other assumptions of the model, such as moment conditions, symmetrically. Our resulting confidence sets therefore represent uncertainty about the parameter vector only. We illustrate these conceptual differences as well as the computational costs in Section 6, where we consider confidence intervals for an average derivative.

Finally, our paper builds on previous work on inference in nonstandard problems, most importantly the papers of Andrews (1999, 2001) on estimation and testing when a parameter is on the boundary of the parameter space. The main difference of our paper to Andrews' work is that we allow testing for a growing parameter vector while Andrews considers a vector of a fixed dimension. Moreover, we show that our inference method is uniformly valid

when the parameters can be either at the boundary, close to the boundary, or away from the boundary. We also use different test statistics because we invert them to obtain confidence bands. Thus, while the general approach is similar, the details of the arguments are very different. Ketz (2017) has a similar setup as Andrews but allows for certain parameter sequences that are close to the boundary under non-negativity constraints.

Outline: The remainder of the paper is organized as follows. We start by illustrating the most important features of our inference approach in a very simple example. Section 3 discusses a general setting, including high level assumptions for uniformly valid inference. Sections 4 and 5 provide low level conditions in a regression framework (for both series and kernel estimation) and the NPIV model, respectively. The remaining sections contain Monte Carlo simulations, the empirical application, and a conclusion. Proofs of the results from Sections 4 and 5, computational details, and additional simulation results are in a supplementary appendix with section numbers S.1, S.2, etc..

Notation: For any matrix A, $\|A\|$ denotes the Frobenius norm. For any square matrix A, $\|A\|_S = \sup_{\|x\|=1} \|Ax\|$ denotes the spectral norm. For a positive semi-definite matrix Ω and a vector a let $\|a\|_\Omega = \sqrt{a'\Omega a}$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and the largest eigenvalue of a symmetric square matrix A. For a sequence of random variables X_n and a class of distributions $\mathcal{P}$ we say $X_n = o_p(\varepsilon_n)$ uniformly over $P \in \mathcal{P}$ if $\sup_{P\in\mathcal{P}} P(|X_n| \geq \delta\varepsilon_n) \to 0$ for any $\delta > 0$. We say $X_n = O_p(\varepsilon_n)$ uniformly over $P \in \mathcal{P}$ if for any $\delta > 0$ there are M_δ and N_δ such that $\sup_{P\in\mathcal{P}} P(|X_n| \geq M_\delta\varepsilon_n) \leq \delta$ for all $n \geq N_\delta$.

2 Illustrative example

We now illustrate the main features of our method in a very simple example. We then explain how these ideas can easily be generalized before introducing the general setup with formal assumptions in Section 3. Suppose that $X \sim N(\theta_0, I_{2\times 2})$ and that we observe a random sample $\{X_i\}_{i=1}^n$ of X. Denote the sample average by $\bar{X}$. We are interested in estimating θ_0 under the assumption that $\theta_{0,1} \leq \theta_{0,2}$. An unrestricted estimator of θ_0, denoted by $\hat{\theta}_{ur}$, is

$$\hat{\theta}_{ur} = \underset{\theta\in\mathbb{R}^2}{\arg\min}\ (\theta_1 - \bar{X}_1)^2 + (\theta_2 - \bar{X}_2)^2.$$

Hence $\hat{\theta}_{ur} = \bar{X}$. Analogously, a restricted estimator is

$$\begin{aligned}\hat{\theta}_r &= \underset{\theta\in\mathbb{R}^2:\ \theta_1\leq\theta_2}{\arg\min}\ (\theta_1 - \bar{X}_1)^2 + (\theta_2 - \bar{X}_2)^2 \\ &= \underset{\theta\in\mathbb{R}^2:\ \theta_1\leq\theta_2}{\arg\min}\ \|\theta - \hat{\theta}_{ur}\|^2,\end{aligned}$$

which implies that $\hat{\theta}_r$ is simply the projecting of $\hat{\theta}_{ur}$ onto $\{\theta \in \mathbb{R}^2 : \ \theta_1 \leq \theta_2\}$. Adding and subtracting θ_0 and multiplying by $\sqrt{n}$ then yields

$$\hat{\theta}_r \quad = \quad \underset{\theta \in \mathbb{R}^2:\ \theta_1 - \theta_2 \leq 0}{\arg\min} \ \|\sqrt{n}(\theta - \theta_0) - \sqrt{n}(\hat{\theta}_{ur} - \theta_0)\|^2.$$

Let $\lambda = \sqrt{n}(\theta - \theta_0)$. From a change of variables it then follows that

$$\sqrt{n}(\hat{\theta}_r - \theta_0) = \underset{\lambda \in \mathbb{R}^2:\ \lambda_1 - \lambda_2 \leq \sqrt{n}(\theta_{0,2} - \theta_{0,1})}{\arg\min} \ \|\lambda - \sqrt{n}(\hat{\theta}_{ur} - \theta_0)\|^2.$$

Let $Z \sim N(0, I_{2\times 2})$. Since $\sqrt{n}(\hat{\theta}_{ur} - \theta_0) \sim N(0, I_{2\times 2})$ we get

$$\sqrt{n}(\hat{\theta}_r - \theta_0) \quad \stackrel{d}{=} \underset{\lambda \in \mathbb{R}^2:\ \lambda_1 - \lambda_2 \leq \sqrt{n}(\theta_{0,2} - \theta_{0,1})}{\arg\min} \ \|\lambda - Z\|^2,$$

where $\stackrel{d}{=}$ means that the random variables on the left and right side have the same distribution. Notice that while the distribution of $\sqrt{n}(\hat{\theta}_{ur} - \theta_0)$ does not depend on θ_0 and n, the distribution of $\sqrt{n}(\hat{\theta}_r - \theta_0)$ depends on $\sqrt{n}(\theta_{0,2} - \theta_{0,1})$, which measures how close θ_0 is to the boundary of the parameter space relative to n. We denote a random variable which has the same distribution as $\sqrt{n}(\hat{\theta}_r - \theta_0)$ by $Z_n(\theta_0)$. As an example, suppose that $\theta_{0,1} = \theta_{0,2}$. Then $Z_n(\theta_0)$ is the projection of Z onto the set $\{z \in \mathbb{R}^2 : z_1 \leq z_2\}$.

A 95% confidence region for θ_0 using the unrestricted estimator can be constructed by finding the constant c_{ur} such that

$$P(\max\{|Z_1|, |Z_2|\} \leq c_{ur}) = 0.95.$$

It then follows immediately that

$$P\left(\hat{\theta}_{ur,1} - \frac{c_{ur}}{\sqrt{n}} \leq \theta_{0,1} \leq \hat{\theta}_{ur,1} + \frac{c_{ur}}{\sqrt{n}} \ \text{ and } \ \hat{\theta}_{ur,2} - \frac{c_{ur}}{\sqrt{n}} \leq \theta_{0,2} \leq \hat{\theta}_{ur,2} + \frac{c_{ur}}{\sqrt{n}}\right) = 0.95.$$

Thus

$$CI_{ur} = \left\{\theta \in \mathbb{R}^2 : \ \hat{\theta}_{ur,1} - \frac{c_{ur}}{\sqrt{n}} \leq \theta_1 \leq \hat{\theta}_{ur,1} + \frac{c_{ur}}{\sqrt{n}} \ \text{ and } \ \hat{\theta}_{ur,2} - \frac{c_{ur}}{\sqrt{n}} \leq \theta_2 \leq \hat{\theta}_{ur,2} + \frac{c_{ur}}{\sqrt{n}}\right\}$$

is a 95% confidence set for θ_0. While there are many different 95% confidence regions for θ_0, rectangular regions are particularly easy to report (especially in larger dimensions), because one only has to report the extreme points of each coordinate.

Similarly, now looking at the restricted estimator, for each $\theta \in \mathbb{R}^2$ let $c_{r,n}(\theta)$ be such that

$$P(\max\{|Z_{n,1}(\theta)|, |Z_{n,2}(\theta)|\} \leq c_{r,n}(\theta)) = 0.95$$

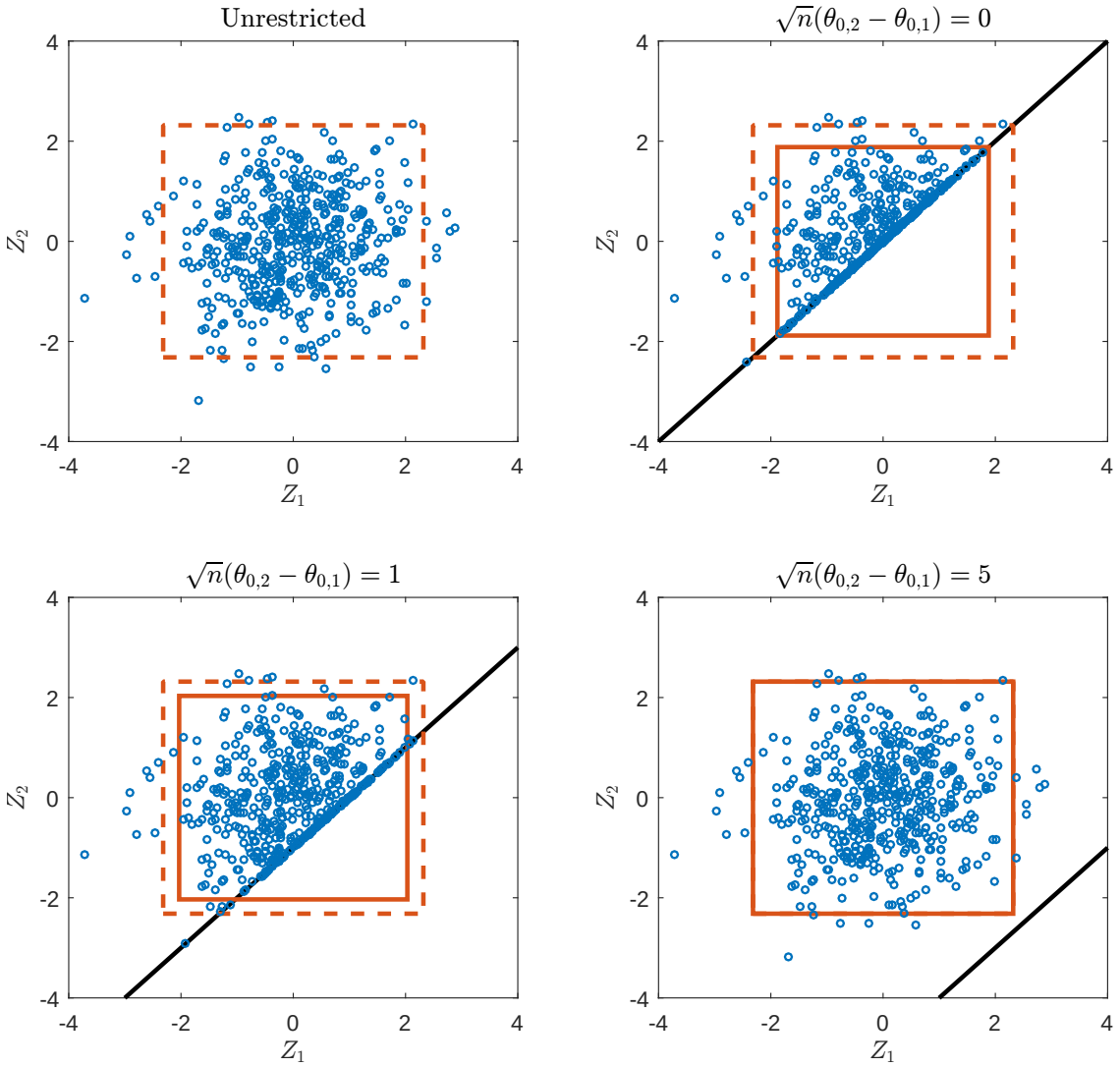
and define CI_r as

$$\left\{\theta \in \mathbb{R}^2 : \ \theta_1 \leq \theta_2, \ \hat{\theta}_{r,1} - \frac{c_{r,n}(\theta)}{\sqrt{n}} \leq \theta_1 \leq \hat{\theta}_{r,1} + \frac{c_{r,n}(\theta)}{\sqrt{n}}, \ \hat{\theta}_{r,2} - \frac{c_{r,n}(\theta)}{\sqrt{n}} \leq \theta_2 \leq \hat{\theta}_{r,2} + \frac{c_{r,n}(\theta)}{\sqrt{n}}\right\}.$$

Again, by construction $P(\theta_0 \in CI_r) = 0.95$.

Figure 1 illustrates the relation between c_{ur} and $c_{r,n}(\theta)$. The first panel shows a random sample of Z. The dashed square contains all $z \in \mathbb{R}^2$ such that $\max\{|z_1|, |z_2|\} \leq c_{ur}$. The second panel displays the corresponding random sample of $Z_n(\theta_0)$ when $\sqrt{n}(\theta_{0,2} - \theta_{0,1}) = 0$, which is simply the projection of Z onto the set $\{z \in \mathbb{R}^2 : z_1 \leq z_2\}$. In particular, for each realization z we have $z_n(\theta_0) = z$ if $z_1 \leq z_2$ and $z_n(\theta_0) = 0.5(z_1 + z_2, z_1 + z_2)'$ if $z_1 > z_2$. Therefore, if $\max\{|z_1|, |z_2|\} \leq c_{ur}$, then also $\max\{|z_{n,1}(\theta_0)|, |z_{n,2}(\theta_0)|\} \leq c_{ur}$, which immediately implies that $c_{r,n}(\theta_0) \leq c_{ur}$. The solid square contains all $z \in \mathbb{R}^2$ such that

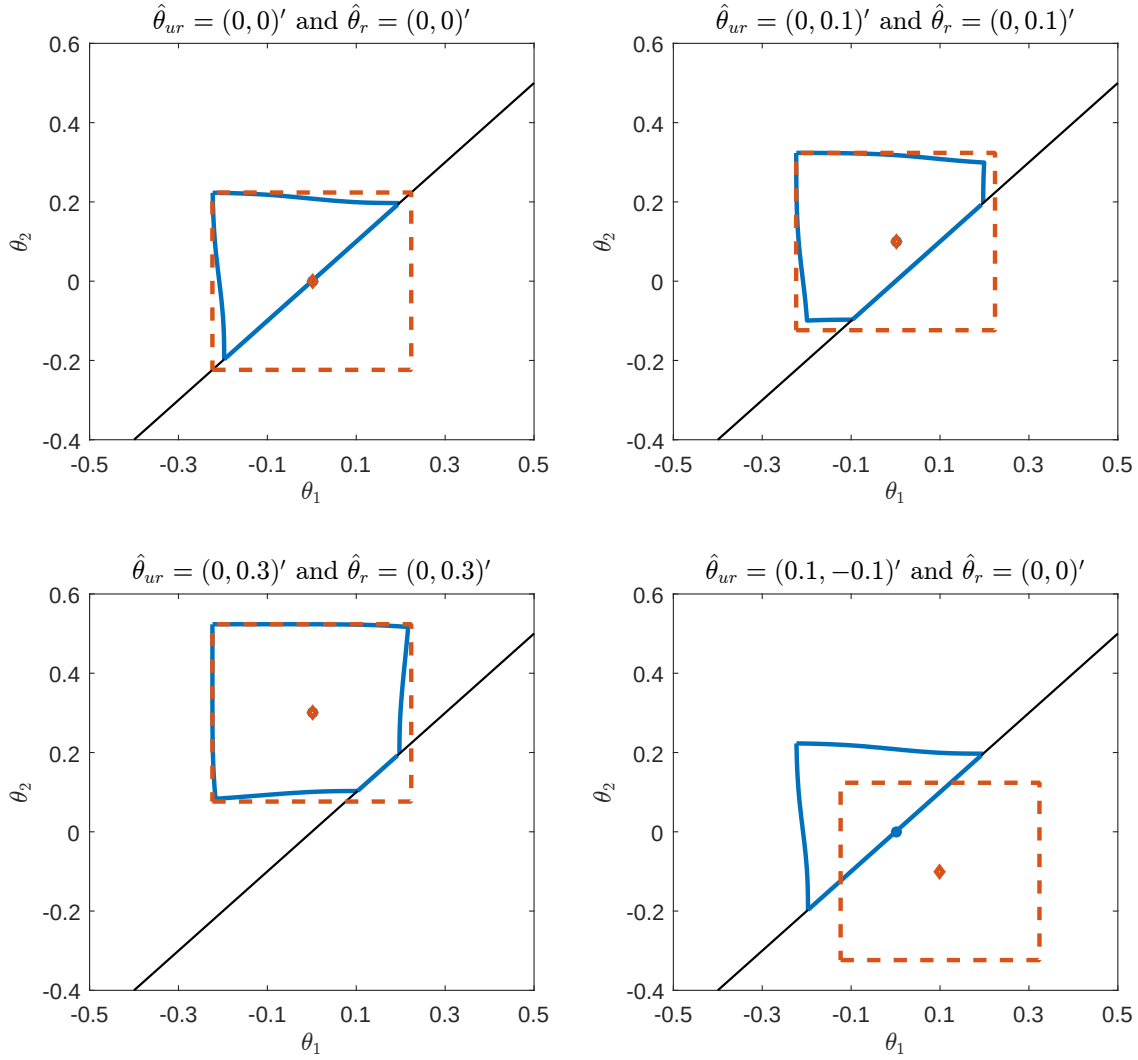
Figure 1: Scatter plots of samples illustrating relation between critical values



$\max\{|z_1|, |z_2|\} \leq c_{r,n}(\theta_0)$, which is strictly inside the dashed square. The third and fourth panel show a similar situations with $\sqrt{n}(\theta_{0,2} - \theta_{0,1}) = 1$ and $\sqrt{n}(\theta_{0,2} - \theta_{0,1}) = 5$, respectively. As $\sqrt{n}(\theta_{0,2} - \theta_{0,1})$ increases, the percentage projected onto the solid line decreases and thus $c_{r,n}(\theta_0)$ gets closer to c_{ur}. Moreover, once $\sqrt{n}(\theta_{0,2} - \theta_{0,1})$ is large enough, $c_{r,n}(\theta_0) = c_{ur}$.

Figure 2 shows the resulting confidence regions for θ_0 when $n = 100$ for specific realizations of $\hat{\theta}_{ur}$ and $\hat{\theta}_r$. The confidence sets depend on these realizations, but given $\hat{\theta}_{ur}$ and $\hat{\theta}_r$, they do not depend on θ_0. The dashed red square is CI_{ur} and the solid blue lines are the boundary of CI_r. In the first panel $\hat{\theta}_{ur} = \hat{\theta}_r = (0, 0)'$. Since $\hat{\theta}_{ur} = \hat{\theta}_r$ and $c_{r,n}(\theta) \leq c_{ur}$ for all $\theta \in \mathbb{R}^2$, it holds that $CI_r \subset CI_{ur}$. Also notice that since $c_{r,n}(\theta)$ depends on θ, CI_r is not a triangle as opposed to the set $CI_{ur} \cap \{\theta \in \mathbb{R} : \theta_1 \leq \theta_2\}$. The second and the third panel display similar situations with $\hat{\theta}_{ur} = \hat{\theta}_r = (0, 0.1)'$ and $\hat{\theta}_{ur} = \hat{\theta}_r = (0, 0.3)'$, respectively. In

Figure 2: Confidence regions



both cases, $CI_r \subset CI_{ur}$. It also follows from the previous discussion that if $\hat{\theta}_{ur} = \hat{\theta}_r$ and if $\sqrt{n}(\hat{\theta}_{ur,2} - \hat{\theta}_{ur,1})$ is large enough then $CI_{ur} = CI_r$. Consequently, for any fixed θ_0 with $\theta_{0,1} < \theta_{0,2}$, it holds that $P(CI_r = CI_{ur}) \to 1$. However, this equivalence does not hold if θ_0 is at the boundary or close to the boundary. Furthermore, it then holds with positive probability that $CI_{ur} \cap \{\theta \in \mathbb{R} : \theta_1 \leq \theta_2\} = \emptyset$, while CI_r always contains $\hat{\theta}_r$. The fourth panel illustrates that if $\hat{\theta}_{ur} \neq \hat{\theta}_r$, then CI_r is not a subset of CI_{ur}.

The set CI_r is an exact 95% confidence set for θ_0, but it cannot simply be characterized by its extreme points and it can be hard to report with more than two dimensions. Nevertheless, we can use it to construct a rectangular confidence set. To do so, for $j = 1, 2$ define

$$\hat{\theta}^L_{r,j} = \min_{\theta \in CI_r} \theta_j \quad \text{and} \quad \hat{\theta}^U_{r,j} = \max_{\theta \in CI_r} \theta_j$$

and

$$\overline{CI}_r = \left\{\theta \in \mathbb{R}^2 : \theta_1 \leq \theta_2 \text{ and } \hat{\theta}^L_{r,1} \leq \theta_1 \leq \hat{\theta}^U_{r,1} \text{ and } \hat{\theta}^L_{r,2} \leq \theta_2 \leq \hat{\theta}^U_{r,2}\right\}.$$

Then, by construction, $CI_r \subseteq \overline{CI}_r$ and thus $P(\theta_0 \in \overline{CI}_r) \geq 0.95$. Moreover, just as before, if $\hat{\theta}_{ur} = \hat{\theta}_r$, then $\overline{CI}_r \subseteq CI_{ur}$. If for example $\hat{\theta}_{ur} = \hat{\theta}_r = (0, 0)'$, then $\hat{\theta}^U_{r,2} = -\hat{\theta}^L_{r,1} = c_{ur}/\sqrt{n}$ but $\hat{\theta}^U_{r,1} = -\hat{\theta}^L_{r,2} < c_{ur}/\sqrt{n}$, which can be seen from the first panel of Figure 2. Hence, relative to the confidence set from the unrestricted estimator, we obtain width gains for the upper end of the first dimension and the lower end of the second dimension. The width gains decrease as $\hat{\theta}_{ur}$ moves away from the boundary into the interior of Θ_R. Moreover, for any $\hat{\theta}_{ur}$ and $\hat{\theta}_r$ and $j = 1, 2$ we get $\hat{\theta}^U_{r,j} - \hat{\theta}^L_{r,j} \leq 2c_{ur}/\sqrt{n}$. Thus, the sides of the square $\{\theta \in \mathbb{R}^2 : \hat{\theta}^L_{r,1} \leq \theta_1 \leq \hat{\theta}^U_{r,1} \text{ and } \hat{\theta}^L_{r,2} \leq \theta_2 \leq \hat{\theta}^U_{r,2}\}$ are never longer than the sides of the square CI_{ur}. Finally, if $\hat{\theta}_{ur}$ is sufficiently in the interior of Θ_R, then $\overline{CI}_r = CI_{ur}$, which is an important feature of our inference method. We get this equivalence in the interior of Θ_R because we invert a test based on a particular type of test statistic, namely $\max\{|Z_1|, |Z_2|\}$. If we started out with a different test statistic, such as $Z_1^2 + Z_2^2$, we would not obtain $\overline{CI}_r = CI_{ur}$ in the interior of Θ_R. We return to this result more generally in Section 3.2 and discuss possible alternative ways of constructing confidence regions in Section 8.

This method of constructing confidence sets is easy to generalize. As a first step, let Θ_R be a restricted parameter space and let $Q_n(\theta)$ be a population objective function. Suppose that the unrestricted estimator $\hat{\theta}_{ur}$ minimizes $Q_n(\theta)$. Also suppose that $Q_n(\theta)$ is a quadratic function of θ, which holds for example in the NPIV model, and which implies that $\nabla^2 Q_n(\theta)$ does not depend on θ. Then with $\hat{\Omega} = \nabla^2 Q_n(\theta)$ we get

$$Q_n(\theta) = Q_n(\hat{\theta}_{ur}) + \nabla Q_n(\hat{\theta}_{ur})'(\theta - \hat{\theta}_{ur}) + \frac{1}{2}(\theta - \hat{\theta}_{ur})'\hat{\Omega}(\theta - \hat{\theta}_{ur})$$

and since $\nabla Q_n(\hat{\theta}_{ur}) = 0$ it holds that

$$\hat{\theta}_r = \underset{\theta \in \Theta_R}{\arg\min} \ \|\theta - \hat{\theta}_{ur}\|^2_{\hat{\Omega}}.$$

Hence, $\hat{\theta}_r$ is again simply a projection of $\hat{\theta}_{ur}$ onto Θ_R. As before, we can now use a change of variables and characterize the distribution of $\sqrt{n}(\hat{\theta}_r - \theta_0)$ as a projection of $\sqrt{n}(\hat{\theta}_{ur} - \theta_0)$ onto a local parameter space that depends on θ_0 and n. Thus, when testing $H_0 : \theta_0 = \bar{\theta}$ based on a test statistic that depends on $\sqrt{n}(\hat{\theta}_r - \theta_0)$, we can use the projection of the large sample distribution of $\sqrt{n}(\hat{\theta}_{ur} - \theta_0)$ to calculate the critical values.

3 General setup

In this section we discuss a general framework and provide conditions for uniformly valid inference. We start with an informal overview of the inference method and provide the formal assumptions and results in Section 3.1. In Section 3.2 we discuss rectangular confidence regions for general functions of the parameter vector.

Let $\Theta \subseteq \mathbb{R}^{K_n}$ be the parameter space and let $\Theta_R \subseteq \Theta$ be a restricted parameter space. Inferences focuses on $\theta_0 \in \Theta_R$. In an example discussed in Section 4.2 we have

$$\theta_0 = \Big(E(Y \mid X = x_1) \quad \dots \quad E(Y \mid X = x_{K_n}) \Big)',$$

and K_n increases with the sample size. In this case, the confidence regions we obtain are analogous to the ones in the simple example above. For series estimation we take $\theta_0 \in \mathbb{R}^{K_n}$ such that $g_0(x) \approx p_{K_n}(x)'\theta_0$, where g_0 is an unknown function of interest and $p_{K_n}(x)$ is a vector of basis functions. A rectangular confidence region for certain functions of θ_0 can then be interpreted as a uniform confidence band for g_0; see Section 4.3 for details. Even though θ_0 and Θ may depend on the sample size, we omit the subscripts for brevity.

As explained in Section 2, in many applications we can obtain a restricted estimator as a projection of an unrestricted estimator onto the restricted parameter space. More generally, we assume that there exist $\hat{\theta}_{ur}$ and $\hat{\theta}_r$ such that $\hat{\theta}_r$ is approximately the projection of $\hat{\theta}_{ur}$ onto Θ_R under some norm $\|\cdot\|_{\hat{\Omega}}$ (see Assumption 1 below for a formal statement). Moreover, since the rate of convergence may be slower than $1/\sqrt{n}$, let κ_n be a sequence of numbers such that $\kappa_n \to \infty$ as $n \to \infty$. Then

$$\begin{aligned} \hat{\theta}_r &\approx \underset{\theta \in \Theta_R}{\arg\min} \|\theta - \hat{\theta}_{ur}\|^2_{\hat{\Omega}} \\ &= \underset{\theta \in \Theta_R}{\arg\min} \|\kappa_n(\theta - \theta_0) - \kappa_n(\hat{\theta}_{ur} - \theta_0)\|^2_{\hat{\Omega}}. \end{aligned}$$

Next define

$$\Lambda_n(\theta_0) = \{\lambda \in \mathbb{R}^{K_n} : \lambda = \kappa_n(\theta - \theta_0) \text{ for some } \theta \in \Theta_R\}.$$

Then

$$\kappa_n(\hat{\theta}_r - \theta_0) \approx \underset{\lambda \in \Lambda_n(\theta_0)}{\arg\min} \|\lambda - \kappa_n(\hat{\theta}_{ur} - \theta_0)\|^2_{\hat{\Omega}}.$$

We will also assume that $\kappa_n(\hat{\theta}_{ur} - \theta_0)$ is approximately $N(0, \Sigma)$ distributed (see Assumption 2 for a formal statement) and that we have a consistent estimator of Σ, denoted by $\hat{\Sigma}$.

Now let $Z \sim N(0, I_{K_n \times K_n})$ be independent of $\hat{\Sigma}$ and $\hat{\Omega}$ and define

$$Z_n(\theta, \hat{\Sigma}, \hat{\Omega}) = \underset{\lambda \in \Lambda_n(\theta)}{\arg\min} \|\lambda - \hat{\Sigma}^{1/2} Z\|^2_{\hat{\Omega}}.$$

We will use the distribution of $Z_n(\theta_0, \hat{\Sigma}, \hat{\Omega})$ to approximate the distribution of $\kappa_n(\hat{\theta}_r - \theta_0)$. This idea is analogous to Andrews (1999, 2001); see for example Theorem 2(e) in Andrews (1999). The main differences are that θ_0 can grow in dimensions as $n \to \infty$ and that our local parameter space $\Lambda_n(\theta_0)$ depends on n because we allow θ_0 to be close to the boundary.

Now for $\bar{\theta} \in \Theta_R$ consider testing

$$H_0 : \theta_0 = \bar{\theta}$$

based on a test statistic T, which depends on $\kappa_n(\hat{\theta}_r - \bar{\theta})$ and $\hat{\Sigma}$. For example

$$T(\kappa_n(\hat{\theta}_r - \bar{\theta}), \hat{\Sigma}) = \max_{k=1,\ldots,K_n} \left| \frac{\kappa_n(\hat{\theta}_{r,k} - \bar{\theta}_k)}{\sqrt{\hat{\Sigma}_{kk}}} \right|.$$

We reject H_0 if and only if

$$T(\kappa_n(\hat{\theta}_r - \bar{\theta}), \hat{\Sigma}) > c_{1-\alpha,n}(\bar{\theta}, \hat{\Sigma}, \hat{\Omega}),$$

where

$$c_{1-\alpha,n}(\bar{\theta}, \hat{\Sigma}, \hat{\Omega}) = \inf\{c \in \mathbb{R} : P(T(Z_n(\bar{\theta}, \hat{\Sigma}, \hat{\Omega}), \hat{\Sigma}) \leq c \mid \hat{\Sigma}, \hat{\Omega}) \geq 1 - \alpha\}.$$

Our $1 - \alpha$ confidence set for θ_0 is then

$$CI = \{\theta \in \Theta_R : T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega})\}.$$

To guarantee that $P(\theta_0 \in CI) \to 1 - \alpha$ uniformly over a class of distributions $\mathcal{P}$ we require

$$\sup_{P \in \mathcal{P}} \left| P\left(T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega})\right) - (1 - \alpha) \right| \to 0.$$

Notice that if $\hat{\theta}_r$ was exactly the projection of $\hat{\theta}_{ur}$ onto Θ_R, if $\kappa_n(\hat{\theta}_{ur} - \theta_0)$ was exactly $N(0, \Sigma)$ distributed, if Σ and Ω were known, and if $T(Z_n(\theta_0, \Sigma, \Omega), \Sigma)$ was continuously distributed, then by construction

$$P\left(T(\kappa_n(\hat{\theta}_r - \theta_0), \Sigma) \leq c_{1-\alpha,n}(\theta_0, \Sigma, \Omega)\right) = 1 - \alpha,$$

just as in the simple example in Section 2. Therefore, the assumptions below simply guarantee that the various approximation errors are small and that small approximation errors only have a small impact on the distribution of the test statistic.

3.1 Assumptions and main result

Let ε_n be a sequence of positive numbers with $\varepsilon_n \to 0$. We discuss the role of ε_n after stating the assumptions. Let $\mathcal{P}$ be a set of distributions satisfying the following assumptions.[4]

Assumption 1. There exists a symmetric, positive semi-definite matrix $\hat{\Omega}$ such that

$$\kappa_n(\hat{\theta}_r - \theta_0) = \underset{\lambda \in \Lambda_n(\theta_0)}{\arg\min} \|\lambda - \kappa_n(\hat{\theta}_{ur} - \theta_0)\|^2_{\hat{\Omega}} + R_n$$

and $\|R_n\| = o_p(\varepsilon_n)$ uniformly over $P \in \mathcal{P}$.

Assumption 2. There exist symmetric, positive definite matrices Ω and Σ and a sequence of random variables $Z_n \sim N(0, \Sigma)$ such that $\lambda_{\min}(\Omega)^{-1/2}\|\kappa_n(\hat{\theta}_{ur} - \theta_0) - Z_n\| = o_p(\varepsilon_n)$ uniformly over $P \in \mathcal{P}$.

Assumption 3. There exists a constant $C_\lambda > 0$ such that $1/C_\lambda \leq \lambda_{\min}(\Sigma) \leq C_\lambda$, $1/C_\lambda \leq \lambda_{\max}(\Omega) \leq C_\lambda$ and

$$\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)}\|\hat{\Sigma} - \Sigma\|^2_S = o_p(\varepsilon_n^2/K_n) \qquad \text{and} \qquad \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)^2}\|\hat{\Omega} - \Omega\|_S = o_p(\varepsilon_n^2/K_n)$$

uniformly over $P \in \mathcal{P}$.

Assumption 4. Θ_R is closed and convex and $\theta_0 \in \Theta_R$.

Assumption 5. Let Σ_1 and Σ_2 be any symmetric and positive definite matrices such that $1/B \leq \lambda_{\min}(\Sigma_1) \leq B$ and $1/B \leq \lambda_{\min}(\Sigma_2) \leq B$ for some constant $B > 0$. There exists a constant C, possibly depending on B, such that for any $z_1 \in \mathbb{R}^{K_n}$ and $z_2 \in \mathbb{R}^{K_n}$

$$|T(z_1, \Sigma_1) - T(z_2, \Sigma_1)| \leq C\|z_1 - z_2\| \quad \text{and} \quad |T(z_1, \Sigma_1) - T(z_1, \Sigma_2)| \leq C\|z_1\|\|\Sigma_1 - \Sigma_2\|_S.$$

Assumption 6. There exists $\delta \in (0, \alpha)$ such that for all $\beta \in [\alpha - \delta, \alpha + \delta]$

$$\sup_{P \in \mathcal{P}} |P\left(T(Z_n(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\beta,n}(\theta_0, \Sigma, \Omega) - \varepsilon_n\right) - (1 - \beta)| \to 0$$

and

$$\sup_{P \in \mathcal{P}} |P\left(T(Z_n(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\beta,n}(\theta_0, \Sigma, \Omega) + \varepsilon_n\right) - (1 - \beta)| \to 0.$$

[4]Even though θ_0 depends on $P \in \mathcal{P}$, we do not make the dependence explicit in the notation.

As demonstrated above, if $\hat{\theta}_{ur}$ maximizes $Q_n(\theta)$ and if $\nabla^2 Q_n(\theta)$ does not depend on θ, then Assumption 1 holds with $R_n = 0$ and $\hat{\Omega} = \nabla^2 Q_n(\theta)$. Andrews (1999) provides general sufficient conditions for a small remainder in a quadratic expansion. The assumption also holds by construction if we simply project $\hat{\theta}_{ur}$ onto Θ_R to obtain $\hat{\theta}_r$. More generally, the assumption does not necessarily require $\hat{\theta}_{ur}$ to be an unrestricted estimator of a criterion function, which may not even exist in some settings if the criterion function is not defined outside of Θ_R. Even in these cases, $\hat{\theta}_r$ is usually an approximate projection of an asymptotically normally distributed estimator onto Θ_R.[5] Assumption 2 can be verified using a coupling argument and the rate of convergence of $\hat{\theta}_{ur}$ can be slower than $1/\sqrt{n}$. Assumption 3 ensures that the estimation errors of $\hat{\Sigma}$ and $\hat{\Omega}$ are negligible. If $\lambda_{\min}(\Omega)$ is bounded away from 0 and if $\lambda_{\max}(\Sigma)$ is bounded, then the assumption simply states that $\|\hat{\Sigma} - \Sigma\|_S = o_p(\varepsilon_n/\sqrt{K_n})$ and $\|\hat{\Omega} - \Omega\|_S = o_p(\varepsilon_n^2/K_n)$, which is easy to verify in specific examples. Allowing $\lambda_{\min}(\Omega) \to 0$ is important for ill-posed inverse problems such as NPIV. We explain in Sections 4 and 5 that both $1/C_\lambda \leq \lambda_{\min}(\Sigma) \leq C_\lambda$ and $1/C_\lambda \leq \lambda_{\max}(\Omega) \leq C_\lambda$ hold under common assumptions in a variety of settings. We could adapt the assumptions to allow for $\lambda_{\min}(\Sigma) \to 0$ and $\lambda_{\max}(\Omega) \to \infty$, but this would require much more notation. Assumption 4 holds for example with linear inequality constraints of the form $\Theta_R = \{\theta \in \mathbb{R}^{K_n} : A\theta \leq b\}$. Other examples of convex shape restrictions for series estimators are monotonicity, convexity/concavity, increasing returns to scale, subadditivity, or homogeneity of a certain degree, but we rule out Slutzki restrictions, which Horowitz and Lee (2017) allow for. The assumption implies that $\Lambda_n(\theta_0)$ is closed and convex as well. The main purpose of this assumption is to ensure that the projection onto $\Lambda_n(\theta_0)$ is nonexpansive, and thus, we could replace it with a higher level assumption, which might then also allow for the Slutzki restrictions.[6] Assumption 5 imposes continuity conditions on the test statistic. We provide several examples of test statistics satisfying this assumption in Sections 4 and 5. Assumption 6 is a continuity condition on the distribution of $T(Z_n(\theta_0, \Sigma, \Omega), \Sigma)$, which requires that its distribution function does not become too steep too quickly as n increases. It is usually referred to as an anti-concentration condition and it is not uncommon in these type of testing problems; see e.g. Assumption 6.7 of Chernozhukov, Newey, and Santos (2015). If the distribution function is continuous for any fixed K_n, then the assumption is an abstract rate condition on how fast K_n can diverge relative to ε_n. As explained below, to get around this assumption we could

[5] See Ketz (2017) for the construction of such an estimator. $\hat{\theta}_{ur}$ does not even have to be a feasible estimator and we could simply replace $\kappa_n(\hat{\theta}_{ur} - \theta_0)$ by a random variable $\hat{Z}$, which is allowed for by our general formulation; specifically see Z_T in Andrews (1999).

[6] I.e. we use $\|\arg\min_{\lambda \in \Lambda_n(\theta_0)} \|\lambda - z_1\|_{\hat{\Omega}} - \arg\min_{\lambda \in \Lambda_n(\theta_0)} \|\lambda - z_2\|_{\hat{\Omega}}\|_{\hat{\Omega}} \leq C\|z_1 - z_2\|_{\hat{\Omega}}$ for some $C > 0$.

take $c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n$ instead of $c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega})$ as the critical value. Also notice that Assumptions 1 – 5 impose very little restrictions on the shape restrictions and hence, they are insufficient to guarantee that the distribution function of $T(Z_n(\theta_0, \Sigma, \Omega), \Sigma)$ is continuous.

We now get the following result.

Theorem 1. *Suppose Assumptions 1 – 5 hold. Then*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n\right) \geq 1 - \alpha.$$

If in addition Assumption 6 holds then

$$\sup_{P\in\mathcal{P}} \left| P\left(T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega})\right) - (1 - \alpha)\right| \to 0.$$

The first part of Theorem 1 implies that if we take $c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) + \varepsilon$ for any fixed $\varepsilon > 0$ as the critical value, then the rejection probability is asymptotically at most α under the null hypothesis, even if Assumption 6 does not hold. In this case, ε_n can go to 0 arbitrarily slowly. An alternative interpretation is that with $c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega})$ as the critical value and without Assumption 6, the rejection probability might be larger than α in the limit, but the resulting confidence set is arbitrarily close to the $1 - \alpha$ confidence set. The second part states that the test has the right size asymptotically if Assumptions 1 – 6 hold.

3.2 Rectangular confidence sets for functions

The previous results yield asymptotically valid confidence regions for θ_0. However, these regions might be hard to report if K_n is large and they may not be the main object of interest. For example, we might be more interested in a uniform confidence band for a function rather than a confidence region of the coefficients in the series expansion. We now discuss how we can use these regions to obtain rectangular confidence sets for functions $h : \mathbb{R}^{K_n} \to \mathbb{R}^{L_n}$ using projections, similar as in Section 2 where we used $h(\theta) = \theta$. Rectangular confidence regions are easy to report because we only have to report the extreme points of each coordinate, which is crucial when L_n is large.[7] Our method applies to general functions, such as function values or average derivatives in nonparameteric estimation. In our applications we focus on

[7]We do not impose any restrictions on L_n and in theory we could have $L_n = \infty$. For example, uniform confidence bands are projections of functionals of the form $p_{K_n}(x)'\theta$ for possibly infinitely many values of x. However, in practice, L_n is typically finite. For example, we can calculate the uniform confidence bands on an arbitrarily fine grid. See Section Sections 4 and 5 for details.

uniform confidence bands, which we can obtain using specific functions h, as explained in Sections 4 and 5. Define

$$CI = \{\theta \in \Theta_R : T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega})\}$$

and let

$$\hat{h}_l^L = \inf_{\theta \in CI} h_l(\theta) \quad \text{and} \quad \hat{h}_l^U = \sup_{\theta \in CI} h_l(\theta), \qquad l = 1, \ldots, L_n.$$

Notice that if $\theta_0 \in CI$, then $\hat{h}_l^L \leq h_l(\theta_0)$ and $\hat{h}_l^U \geq h_l(\theta_0)$ for all $l = 1, \ldots, L_n$. We therefore obtain the following corollary.[8]

Corollary 1. *Suppose Assumptions 1 – 6 hold. Then*

$$\liminf_{n\to\infty} \inf_{P \in \mathcal{P}} P\left(\hat{h}_l^L \leq h_l(\theta_0) \leq \hat{h}_l^U \text{ for all } l = 1, \ldots, L_n\right) \geq 1 - \alpha.$$

A projection for any T satisfying the assumptions above yields a rectangular confidence region with coverage probability at least $1 - \alpha$ in the limit. In the examples discussed in Sections 4 and 5 we pick T such that the resulting confidence region is nonconservative for θ_0 in the interior of Θ_R, just as the confidence sets in Figure 2. In these examples $h_l(\theta) = c_l + q_l'\theta$, where c_l is a constant and $q_l \in \mathbb{R}^{L_n}$, and possibly $L_n > K_n$. We then let

$$T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) = \sup_{l=1,\ldots,L_n} \left\{ \kappa_n \left| q_l'(\hat{\theta}_r - \theta) \right| / \sqrt{q_l' \hat{\Sigma} q_l} \right\}.$$

Now suppose that for any $\theta \in CI$, the critical value does not depend on θ, which will be the case with probability approaching 1 if θ_0 is in the interior of the parameter space. That is $c(\theta, \hat{\Sigma}, \hat{\Omega}) = \hat{c}$. Then

$$CI = \left\{ \theta \in \Theta_R : h_l(\hat{\theta}_r) - \frac{\hat{c}}{\kappa_n} \sqrt{q_l' \hat{\Sigma} q_l} \leq h_l(\theta) \leq h_l(\hat{\theta}_r) + \frac{\hat{c}}{\kappa_n} \sqrt{q_l' \hat{\Sigma} q_l} \text{ for all } l = 1, \ldots, L_n \right\}.$$

Moreover, by the definitions of the infimum and the supremum as the largest lower bound and smallest upper bound respectively, it holds that

$$\hat{h}_l^L \geq h_l(\hat{\theta}_r) - \frac{\hat{c}}{\kappa_n} \sqrt{q_l' \hat{\Sigma} q_l} \quad \text{and} \quad \hat{h}_l^U \leq h_l(\hat{\theta}_r) + \frac{\hat{c}}{\kappa_n} \sqrt{q_l' \hat{\Sigma} q_l}$$

for all $l = 1, \ldots, L_n$ and thus,

$$\hat{h}_l^L \leq h_l(\theta_0) \leq \hat{h}_l^U \text{ for all } l = 1, \ldots, L_n \quad \iff \quad \theta_0 \in CI.$$

[8] Under Assumptions 1 - 5 only, we could project onto $\{\theta \in \Theta_R : T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n\}$ to obtain the same conclusion as in Corollary 1.

Consequently

$$P\left(\hat{h}_l^L \leq h_l(\theta_0) \leq \hat{h}_l^U \text{ for all } l = 1, \ldots, L_n\right) = P\left(\theta_0 \in CI\right).$$

We state a formal result, which guarantees that the projection based confidence set does not suffer from over-coverage if θ_0 is sufficiently in the interior of the parameter space, in Corollary A1 in the appendix. The results can be extended to nonlinear functions h along the lines of Freyberger and Rai (2018).

4 Conditional mean estimation

In this section we provide sufficient conditions for Assumptions 1 – 5 when

$$Y = g_0(X) + U, \quad E(U \mid X) = 0$$

and Y, X and U are scalar random variables. We also explain how we can use the projection results to obtain uniform confidence bands for g_0. We first assume that X is discretely distributed to illustrate that the inference method can easily be applied to finite dimensional models. We then let X be continuously distributed and discuss both kernel and series estimators. Throughout, we assume that the data is a random sample $\{Y_i, X_i\}_{i=1}^n$. The proofs of all results in this and the following section are in the supplementary appendix.

4.1 Discrete regressors

Suppose that X is discretely distributed with support $\mathcal{X} = \{x_1, \ldots, x_K\}$, where K is fixed. Let

$$\theta_0 = \begin{pmatrix} E(Y \mid X = x_1) & \ldots & E(Y \mid X = x_K) \end{pmatrix}'$$

and

$$\hat{\theta}_{ur} = \begin{pmatrix} \frac{\sum_{i=1}^n Y_i \mathbf{1}(X_i = x_1)}{\sum_{i=1}^n \mathbf{1}(X_i = x_1)} & \ldots & \frac{\sum_{i=1}^n Y_i \mathbf{1}(X_i = x_K)}{\sum_{i=1}^n \mathbf{1}(X_i = x_K)} \end{pmatrix}'.$$

Define $\sigma^2(x_k) = Var(U \mid X = x_k)$ and $p(x_k) = P(X = x_k) > 0$, and let

$$\Sigma = diag\left(\frac{\sigma^2(x_1)}{p(x_1)}, \ldots, \frac{\sigma^2(x_K)}{p(x_K)}\right) \quad \text{and} \quad \hat{\Sigma} = diag\left(\frac{\hat{\sigma}^2(x_1)}{\hat{p}(x_1)}, \ldots, \frac{\hat{\sigma}^2(x_K)}{\hat{p}(x_K)}\right),$$

where $\hat{p}(x_k) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i = x_k)$ and

$$\hat{\sigma}^2(x_k) = \frac{\sum_{i=1}^n Y_i^2 \mathbf{1}(X_i = x_k)}{\sum_{i=1}^n \mathbf{1}(X_i = x_k)} - \left(\frac{\sum_{i=1}^n Y_i \mathbf{1}(X_i = x_k)}{\sum_{i=1}^n \mathbf{1}(X_i = x_k)}\right)^2.$$

Let Θ_R be a convex subset of $\mathbb{R}^K$, such as $\Theta_R = \{\theta \in \mathbb{R}^K : A\theta \leq b\}$. Now define

$$\hat{\theta}_r = \underset{\theta \in \Theta_R}{\arg\min} \, \|\theta - \hat{\theta}_{ur}\|^2_{\hat{\Sigma}^{-1}}$$

and hence $\hat{\Omega} = \hat{\Sigma}^{-1}$. Other weight functions $\hat{\Omega}$, such as the identity matrix, are possible choices as well. We discuss this issue further in Section 8. As a test statistic we use

$$T(z, \hat{\Sigma}) = \max \left\{ |z_1|/\sqrt{\hat{\Sigma}_{11}}, \ldots, |z_K|/\sqrt{\hat{\Sigma}_{KK}} \right\}$$

because the resulting confidence region of the unrestricted estimator is rectangular, analogous to the one in Section 2. We now get the following result.

Theorem 2. *Let $\mathcal{P}$ be the class of distributions satisfying the following assumptions.*

1. *$\{Y_i, X_i\}_{i=1}^n$ is an iid sample from the distribution of (Y, X) with $\sigma^2(x_k) \in [1/C, C]$, $p(x_k) \geq 1/C$, and $E(U^4 \mid X = x_k) \leq C$ for all $k = 1, \ldots, K$ and for some $C > 0$.*

2. *Θ_R is closed and convex and $\theta_0 \in \Theta_R$.*

3. *$\frac{1}{\sqrt{n}} = o(\varepsilon_n^3)$.*

Then

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{P}} P\left(T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n \right) \geq 1 - \alpha.$$

If in addition Assumption 6 holds then

$$\sup_{P \in \mathcal{P}} \left| P\left(T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \right) - (1 - \alpha) \right| \to 0.$$

Next let $h_l(\theta) = \theta_l$ for $l = 1, \ldots, K$. Then the results in Section 3.2 yield a rectangular confidence region for θ_0, which can be interpreted as a uniform confidence band for $g_0(x_1), \ldots, g_0(x_K)$. Moreover, Corollary A1 in the appendix shows that the band is nonconservative if θ_0 is sufficiently in the interior of the parameter space.

4.2 Kernel regression

We now suppose that X is continuously distributed with density f_X. We denote its support by $\mathcal{X}$ and assume that $\mathcal{X} = [\underline{x}, \overline{x}]$. Let $\{x_1, \ldots, x_{K_n}\} \subset \mathcal{X}$ and

$$\theta_0 = \Big(E(Y \mid X = x_1) \quad \ldots \quad E(Y \mid X = x_{K_n}) \Big)'.$$

Here K_n increases as the sample size increases and thus, our setup is very similar to Horowitz and Lee (2017). Let $K(\cdot)$ be a kernel function and h_n the bandwidth. The unrestricted estimator is

$$\hat{\theta}_{ur} = \begin{pmatrix} \frac{\sum_{i=1}^n Y_i K\left(\frac{x_1 - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x_1 - X_i}{h_n}\right)} & \cdots & \frac{\sum_{i=1}^n Y_i K\left(\frac{x_{K_n} - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x_{K_n} - X_i}{h_n}\right)} \end{pmatrix}'.$$

Define $B = \int_{-1}^1 K(u)^2 du$ and $\sigma^2(x) = Var(U \mid X = x)$ and let

$$\Sigma = diag\left(\frac{\sigma^2(x_1)B}{f_X(x_1)}, \ldots, \frac{\sigma^2(x_{K_n})B}{f_X(x_{K_n})}\right) \quad \text{and} \quad \hat{\Sigma} = diag\left(\frac{\hat{\sigma}^2(x_1)B}{\hat{f}_X(x_1)}, \ldots, \frac{\hat{\sigma}^2(x_{K_n})B}{\hat{f}_X(x_{K_n})}\right),$$

where $\hat{f}_X(x_k) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x_k - X_i}{h_n}\right)$ and

$$\hat{\sigma}^2(x_k) = \frac{\sum_{i=1}^n Y_i^2 K\left(\frac{x_k - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x_k - X_i}{h_n}\right)} - \left(\frac{\sum_{i=1}^n Y_i K\left(\frac{x_k - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x_k - X_i}{h_n}\right)}\right)^2.$$

Just as before, let Θ_R be convex such as $\Theta_R = \{\theta \in \mathbb{R}^{K_n} : A\theta \leq b\}$ and define

$$\hat{\theta}_r = \underset{\theta \in \Theta_R}{\arg\min} \|\theta - \hat{\theta}_{ur}\|^2_{\hat{\Sigma}^{-1}},$$

implying that $\hat{\Omega} = \hat{\Sigma}^{-1}$. Finally, as before we let

$$T(z, \hat{\Sigma}) = \max\left\{|z_1|/\sqrt{\hat{\Sigma}_{11}}, \ldots, |z_{K_n}|/\sqrt{\hat{\Sigma}_{K_n K_n}}\right\}.$$

We get the following result.

Theorem 3. *Let $\mathcal{P}$ be the class of distributions satisfying the following assumptions.*

1. *The data $\{Y_i, X_i\}_{i=1}^n$ is an iid sample where $\mathcal{X} = [\underline{x}, \overline{x}]$.*
   - *(a) $g_0(x)$ and $f_X(x)$ are twice continuously differentiable with uniformly bounded function values and derivatives.* $\inf_{x \in \mathcal{X}} f_X(x) \geq 1/C$ *for some* $C > 0$.
   - *(b) $\sigma^2(x)$ is twice continuously differentiable, the function and derivatives are uniformly bounded on $\mathcal{X}$, and* $\inf_{x \in \mathcal{X}} \sigma^2(x) \geq 1/C$ *for some* $C > 0$.
   - *(c)* $E(Y^4 \mid X = x) \leq C$ *for some* $C > 0$.
2. $x_k - x_{k-1} > 2h_n$ *for all* k *and* $x_1 > \underline{x} + h_n$ *and* $x_{K_n} < \overline{x} - h_n$.
3. $K(\cdot)$ *is a bounded and symmetric pdf with support* $[-1, 1]$.

4. Θ_R *is closed and convex and* $\theta_0 \in \Theta_R$.

5. $K_n h_n^5 n = o(\varepsilon_n^2)$ *and* $\frac{K_n^{5/2}}{\sqrt{nh_n}} = o(\varepsilon_n^3)$.

Then

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(T\left(\sqrt{nh_n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}\right) \leq c_{1-\alpha,n}\left(\theta_0, \hat{\Sigma}, \hat{\Omega}\right) + \varepsilon_n\right) \geq 1 - \alpha.$$

If in addition Assumption 6 holds then

$$\sup_{P\in\mathcal{P}} \left| P\left(T\left(\sqrt{nh_n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}\right) \leq c_{1-\alpha,n}\left(\theta_0, \hat{\Sigma}, \hat{\Omega}\right)\right) - (1-\alpha)\right| \to 0.$$

The first assumption contains standard smoothness and moment conditions. The second assumption guarantees that estimators of $g_0(x_k)$ and $g_0(x_l)$ for $k \neq l$ are independent, just as in Horowitz and Lee (2017), and it also avoids complications associated with x_k being too close to the boundary of the support. The third assumption imposes standard restrictions on the kernel function and the fourth assumption has been discussed before. The fifth assumption contains rate conditions. Notice that with a fixed K_n, these rates are the standard conditions for asymptotic normality with undersmoothing in kernel regression. The rate conditions also imply that $K_n h_n \to 0$, which is similar to Horowitz and Lee (2017).

Once again with $h_l(\theta) = \theta_l$ for $l = 1, \ldots, K_n$ the results in Section 3.2 yield a rectangular confidence region for θ_0, which is a uniform confidence band for $g_0(x_1), \ldots, g_0(x_{K_n})$.

Remark 1. While we use the Nadaraya-Watson estimator for simplicity, the general theory also applies to other estimators, such as local polynomial estimators. Another possibility is to use a bias corrected estimator and the adjusted standard errors suggested by Calonico, Cattaneo, and Farrell (2017). Finally, the general theory can also be adapted to incorporate a worst-case bias as in Armstrong and Kolesár (2016) instead of using the undersmoothing assumption; see Section S.2 for details.

4.3 Series regression

In this section we again assume that $X \in \mathcal{X}$ is continuously distributed, but we use a series estimator. One advantage of a series estimator is that it yields uniform confidence bands for the entire function g_0, rather than just a vector of function values.

Let $p_{K_n}(x) \in \mathbb{R}^{K_n}$ be a vector of basis functions and write $g_0(x) \approx p_{K_n}(x)'\theta_0$ for some $\theta_0 \in \Theta_R$. We again let Θ_R be a convex set such as $\{\theta \in \mathbb{R}^{K_n} : A\theta \leq b\}$. For example, we could impose the constraints $\nabla p_{K_n}(x_j)'\theta \geq 0$ for $j = 1, \ldots, J_n$. Notice that J_n is not

restricted, and we could even impose $\nabla p_{K_n}(x)'\theta \geq 0$ for all $x \in \mathcal{X}$ if it is computationally feasible.[9] The unrestricted and restricted estimators are

$$\hat{\theta}_{ur} = \underset{\theta \in \mathbb{R}^{K_n}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - p_{K_n}(X_i)'\theta\right)^2$$

and

$$\hat{\theta}_{r} = \underset{\theta \in \Theta_R}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - p_{K_n}(X_i)'\theta\right)^2,$$

respectively. The assumptions ensure that both minimizers are unique with probability approaching 1. Since the objective function is quadratic in θ_0 we have

$$\sqrt{n}(\hat{\theta}_r - \theta_0) = \underset{\lambda \in \Lambda_n(\theta_0)}{\arg\min} \|\lambda - \sqrt{n}(\hat{\theta}_{ur} - \theta_0)\|^2_{\hat{\Omega}},$$

where $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} p_{K_n}(X_i)p_{K_n}(X_i)'$ and $\Omega = E(\hat{\Omega})$. Define

$$\Sigma = \left(E(p_{K_n}(X_i)p_{K_n}(X_i)')\right)^{-1} E(U_i^2 p_{K_n}(X_i)p_{K_n}(X_i)') \left(E(p_{K_n}(X_i)p_{K_n}(X_i)')\right)^{-1}.$$

Also let $\hat{U}_i = Y_i - p_{K_n}(X_i)'\hat{\theta}_{ur}$ and

$$\hat{\Sigma} = \hat{\Omega}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2 p_{K_n}(X_i)p_{K_n}(X_i)'\right) \hat{\Omega}^{-1}.$$

Let $\hat{\sigma}(x) = \sqrt{p_{K_n}(x)'\hat{\Sigma}p_{K_n}(x)}$. We use the test statistic

$$T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) = \sup_{x \in \mathcal{X}} \left| \frac{p_{K_n}(x)' \left(\sqrt{n}(\hat{\theta}_r - \theta_0)\right)}{\hat{\sigma}(x)} \right|.$$

The following theorem provides conditions to ensure that confidence sets for θ_0 have the correct coverage asymptotically. We then explain how we can use these sets to construct uniform confidence bands for $g_0(x)$. To state the theorem, let $\xi(K_n) = \sup_{x \in \mathcal{X}} \|p_{K_n}(x)\|$.

Theorem 4. *Let $\mathcal{P}$ be the class of distributions satisfying the following assumptions.*

1. *The data $\{Y_i, X_i\}_{i=1}^n$ is an iid sample from the distribution of (Y, X) with $E(U^2 \mid X) \in [1/C, C]$ and $E(U^4 \mid X) \leq C$ for some $C > 0$.*

2. *The basis functions $p_k(\cdot)$ are orthonormal on $\mathcal{X}$ with respect to the L^2 norm and $f_X(x) \in [1/C, C]$ for all $x \in \mathcal{X}$ and some $C > 0$.*

[9] For example, with quadratic splines $\nabla p_{K_n}(x)'\theta \geq 0$ reduces to finitely many inequality constraints.

3. Θ_R *is closed and convex and* $\theta_0 \in \Theta_R$ *is such that for some constants* C_g *and* $\gamma > 0$

$$\sup_{x \in X} |g_0(x) - p_{K_n}(x)'\theta_0| \leq C_g K_n^{-\gamma}.$$

4. $nK_n^{-2\gamma} = o(\varepsilon_n^2)$, $\frac{\xi(K_n)^2 K_n^4}{n} = o(\varepsilon_n^6)$, *and* $\frac{\xi(K_n)^4 K_n^3}{n} = o(\varepsilon_n^2)$.

Then

$$\liminf_{n\to\infty} \inf_{P \in \mathcal{P}} P\left(T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n\right) \geq 1 - \alpha.$$

If in addition Assumption 6 holds then

$$\sup_{P \in \mathcal{P}} \left| P\left(T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega})\right) - (1 - \alpha)\right| \to 0.$$

The first assumption imposes standard moment conditions. The main role of the second assumption is to guarantee that the minimum eigenvalues of Σ and Ω are bounded and bounded away from 0. The third assumption says that g_0 can be well approximated by a function satisfying the constraints, and the fourth assumption provides rate conditions. For asymptotic normality of nonlinear functionals Newey (1997) assumes that

$$nK_n^{-2\gamma} + \frac{\xi(K_n)^4 K_n^2}{n} \to 0.$$

For orthonormal polynomials $\xi(K_n) = C_p K_n$ and for splines $\xi(K_n) = C_s\sqrt{K_n}$. Thus, our rate conditions are slightly stronger than the ones in Newey (1997), but we also obtain confidence sets for the K_n dimensional vector θ_0, which we can transform to uniform confidence bands for g_0. The last rate condition, $\frac{\xi(K_n)^4 K_n^3}{n} = o(\varepsilon_n^2)$, is not needed under the additional assumption that $var(U_i \mid X_i) = \sigma^2 > 0$.

Remark 2. In a finite dimensional regression framework with $K_n = K$, the third assumption always holds and the fourth assumption only requires that $n \to \infty$. In this case the second assumption can be replaced with the full rank condition $\lambda_{\min}(E(p_K(X)p_K(X)')) \geq 1/C$.

To obtain a uniform confidence band for $g_0(X)$, define

$$CI = \{\theta \in \Theta_R : T(\sqrt{n}(\hat{\theta}_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega})\}$$

and let

$$\hat{g}_l(x) = \min_{\theta \in CI} p_{K_n}(x)'\theta \quad \text{and} \quad \hat{g}_u(x) = \max_{\theta \in CI} p_{K_n}(x)'\theta.$$

Also notice that $\|p_{K_n}(x)\|^2$ is bounded away from 0 if the basis functions contain the constant function. We get the following result.

Corollary 2. *Suppose the assumptions of Theorem 4 and Assumption 6 hold. Further suppose that* $\inf_{x\in\mathcal{X}} \|p_{K_n}(x)\|^2 > 1/C$ *for some constant* $C > 0$*. Then*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(\hat{g}_l(x) \leq g_0(x) \leq \hat{g}_u(x) \ \ \forall x \in \mathcal{X}\right) \geq 1-\alpha.$$

Remark 3. Without any restrictions on the parameter space, inverting our test statistic results in a uniform confidence band where the width of the band is proportional to the standard deviation of the estimated function for each x. This band can also be obtained as a projecting onto the underlying confidence set for θ_0; see Freyberger and Rai (2018) for this equivalence result. If θ_0 is sufficiently in the interior of the parameter space, an application of Corollary A1 shows that the restricted band is equivalent to that band with probability approaching 1. In this case the projection based band is not conservative.

Remark 4. Similar as before, Assumption 6 is not needed if the band is obtained by projecting onto $\{\theta \in \Theta_R : T(\sqrt{n}(\hat{\theta}_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n\}$

Remark 5. The results can be extended to a partially linear model of the form $Y = g_0(X_1) + X_2'\gamma_0 + U$. The parameter vector θ_0 would then contain both γ_0 and the coefficients of the series approximation of g_0.

5 Instrumental variables estimation

As the final application of the general method we consider the NPIV model

$$Y = g_0(X) + U, \quad E(U \mid Z) = 0,$$

where X and Z are continuously distributed scalar random variables with bounded support. We assume for notational simplicity that X and Z have the same support, $\mathcal{X}$, but this assumption is without loss of generality because X and Z can always be transformed to have support on $[0,1]$. We assume that $E(U^2 \mid Z) = \sigma^2$ to focus on the complications resulting from the ill-posed inverse problem. Here, the data is a random sample $\{Y_i, X_i, Z_i\}_{i=1}^n$.

As before, let $p_{K_n}(x) \in \mathbb{R}^{K_n}$ be a vector of basis functions and write $g_0(x) \approx p_{K_n}(x)'\theta_0$ for some $\theta_0 \in \Theta_R$, where Θ_R is a convex subset of $\mathbb{R}^{K_n}$. Let P_X be the $n \times K_n$ matrix, where the ith row is $p_{K_n}(X_i)'$ and define P_Z analogously. Let Y be the $n \times 1$ vector containing Y_i. Let

$$\hat{\theta}_{ur} = \underset{\theta\in\mathbb{R}^{K_n}}{\arg\min}\ (Y - P_X\theta)'P_Z(P_Z'P_Z)^{-1}P_Z'(Y - P_X\theta)$$

and

$$\hat{\theta}_r = \underset{\theta \in \Theta_R}{\arg\min}\ (Y - P_X\theta)' P_Z (P_Z' P_Z)^{-1} P_Z' (Y - P_X\theta).$$

For simplicity we use the same basis function as well as the same number of basis functions for X_i and Z_i. Our results can be generalized to allow for different basis functions and more instruments than regressors. Since the objective function is quadratic in θ_0 we have

$$\sqrt{n}(\hat{\theta}_r - \theta_0) = \underset{\lambda \in \Lambda_n(\theta_0)}{\arg\min} \|\lambda - \sqrt{n}(\hat{\theta}_{ur} - \theta_0)\|^2_{\hat{\Omega}},$$

where $\hat{\Omega} = \frac{1}{n}(P_X' P_Z)(P_Z' P_Z)^{-1}(P_Z' P_X)$. Furthermore, let $Q_{XZ} = E(p_{K_n}(X_i) p_{K_n}(Z_i)')$. Then

$$\Sigma = \sigma^2 Q_{XZ}^{-1} E(p_{K_n}(Z_i) p_{K_n}(Z_i)')(Q_{XZ}')^{-1},$$

which we estimate by $\hat{\Sigma} = \hat{\sigma}^2 \hat{\Omega}^{-1}$ with $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2$ and $\hat{U}_i = Y_i - p_{K_n}(X_i)'\hat{\theta}_{ur}$.

As before, $\hat{\sigma}(x) = \sqrt{p_{K_n}(x)' \hat{\Sigma} p_{K_n}(x)}$ and the test statistic is

$$T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) = \sup_{x \in \mathcal{X}} \left| \frac{p_{K_n}(x)' \left(\sqrt{n}(\hat{\theta}_r - \theta_0) \right)}{\hat{\sigma}(x)} \right|.$$

The following theorem provides conditions to ensure that confidence sets for θ_0 have the correct coverage, and analogously to before we can transform these sets to uniform confidence bands for $g_0(x)$. As before, let $\xi(K_n) = \sup_{x \in \mathcal{X}} \|p_{K_n}(x)\|$.

Theorem 5. *Let $\mathcal{P}$ be the class of distributions satisfying the following assumptions.*

1. *The data $\{Y_i, X_i, Z_i\}_{i=1}^n$ is an iid sample from the distribution of (Y, X, Z) with $E(U^2 \mid Z) = \sigma^2 \in [1/C, C]$ and $E(U^4 \mid Z) \leq C$ for some $C > 0$.*

2. *The functions $p_k(\cdot)$ are orthonormal on $\mathcal{X}$ with respect to the L^2 norm and the densities of X and Z are uniformly bounded above and bounded away from 0.*

3. *Θ_R is closed and convex and for some function $b(K_n)$ and $\theta_0 \in \Theta_R$*

$$\sup_{x \in \mathcal{X}} |g_0(x) - p_{K_n}(x)'\theta_0| \leq b(K_n).$$

4. *$\lambda_{\min}(Q_{XZ} Q_{XZ}') \geq \tau_{K_n} > 0$ and $\lambda_{\max}(Q_{XZ} Q_{XZ}') \in [1/C, C]$ for some $C < \infty$.*

5. *$\frac{nb(K_n)^2}{\tau_{K_n}^2} = o(\varepsilon_n^2)$ and $\frac{\xi(K_n)^2 K_n^4}{n\tau_{K_n}^6} = o(\varepsilon_n^6)$.*

Then

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n\right) \geq 1 - \alpha.$$

If in addition Assumption 6 holds then

$$\sup_{P\in\mathcal{P}} \left| P\left(T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega})\right) - (1-\alpha)\right| \to 0.$$

Assumptions 1 – 3 of the theorem are very similar to those of Theorem 4. Assumption 4 defines a measure of ill-posedness τ_{K_n}, which affects the rate conditions. It is easy to show that $\lambda_{\max}(Q_{XZ}Q'_{XZ})$ is bounded as long as f_{XZ} is square integrable. However, $\lambda_{\max}(Q_{XZ}Q'_{XZ}) \leq C$ also allows for $X = Z$ as a special case. In fact, in this case, τ_{K_n} is bounded away from 0 and all assumptions reduce to the ones in the series regression framework with homoskedasticity. Moreover, similar to Remark 2, the assumptions also allow for K_n to be fixed in which case all conditions reduce to standard assumptions in a parametric IV framework. Finally, the results can also be extended to a partially linear model; see Remark 5.

6 Monte Carlo simulations

To investigate finite sample properties we simulate data from the model

$$Y = g_0(X) + U, \quad E(U \mid Z) = 0,$$

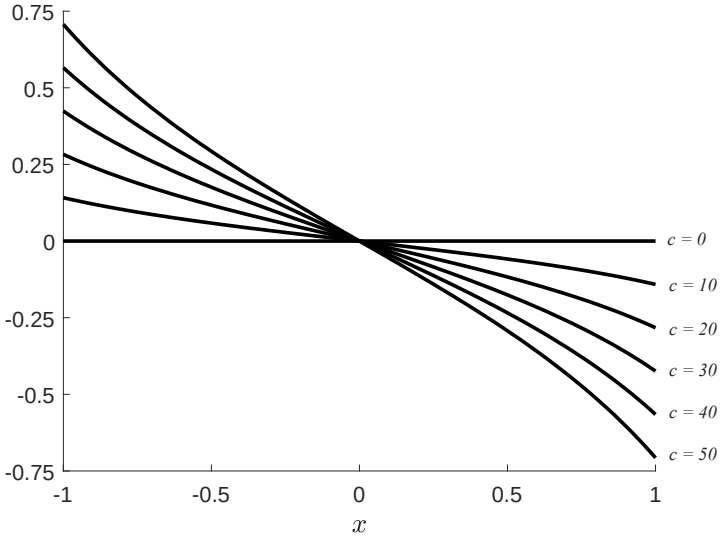
where $X \in [-1, 1]$ and

$$g_0(X) = -\frac{c}{\sqrt{n}} F^{-1}\left(\frac{1}{4}X + \frac{1}{2}\right).$$

Here, F is the cdf of a t-distribution with one degree of freedom and we vary the constant c. Figure 3 shows the function for $n = 5,000$ and $c \in \{0, 10, 20, 30, 40, 50\}$. Clearly, $c = 0$ belongs to the constant function. As c increases the slope of $g_0(x)$ increases for every x.

Let $\tilde{X}$, $\tilde{Z}$, and U be jointly normally distributed with $var(U) = 0.25$ and $var(\tilde{Z}) = var(\tilde{X}) = 1$. Let $X = 2F_{\tilde{X}}(\tilde{X}) - 1 \sim Unif[-1, 1]$ and $Z = 2F_{\tilde{Z}}(\tilde{Z}) - 1 \sim Unif[-1, 1]$. We consider two DGPs. First, we let $cov(\tilde{X}, U) = 0$. Thus, X is exogenous and we use the series estimator described in Section 4.3. Second, we let $cov(\tilde{X}, \tilde{Z}) = 0.7$ and $cov(\tilde{X}, U) = 0.5$ and use the NPIV estimator. In both cases we first focus on uniform confidence bands for g_0. In this section we report results with Legendre polynomials as basis function. In Section S.4 in the supplement we report qualitatively very similar results for quadratic splines. For the series regression setting we take $n = 1,000$ and for NPIV we use $n = 5,000$. We take

Figure 3: g_0 for different values of c



sample sizes large enough such that the unrestricted estimator has good coverage properties for a sufficiently large number of series terms, which helps in analyzing how conservative the restricted confidence bands can be. All results are based on $1,000$ Monte Carlo simulations.

We impose the restriction that g_0 is weakly decreasing and we enforce this constraint on 10 equally spaced points. We solve for the uniform confidence bands on 30 equally spaced grid point. Using finer grids has almost no impact on the results, but increases the computational costs.[10] To solve the optimization problems, we have to calculate $c_{1-\alpha}(\theta, \hat{\Sigma}, \hat{\Omega})$, which is not available in closed form. To do so, we take $2,000$ draws from a multivariate normal distribution and use them to estimate the distribution function of $T(Z_n(\theta, \hat{\Sigma}, \hat{\Omega}), \hat{\Sigma})$ using a kernel estimator and Silverman's rule of thumb bandwidth. We then take the $1-\alpha$ quantile of the estimated distribution function as the critical value. Estimating the distribution function simply as a step function yields almost identical critical values for any given θ, but our construction ensures that the estimated critical value is a smooth function of θ. The number of draws from the normal distribution is analogous to the number of bootstrap samples in other settings and using more draws has almost no impact on our results.

Tables 1 and 2 show the simulation results for the series regression model and the NPIV model, respectively. The first column is the order of the polynomial and $K_n = 2$ belongs to a linear function. We use the same number of basis functions for X and Z, but using $K_n + 3$ for the instrument matrix yields very similar results. The third and fourth columns show

[10] In the application we use a grid of 100 points for the uniform confidence bands, but we use a coarser grid for the simulations, because our reported results are based on $78,000$ estimated confidence bands in total.

the coverage rates of uniform confidence bands using the unrestricted and shape restricted method, respectively. The nominal coverage rate is 0.95. For a confidence band $[\hat{g}_l(x),\ \hat{g}_u(x)]$ define the average width as $\frac{1}{30}\sum_{j=1}^{30}(\hat{g}_u(x_j) - \hat{g}_l(x_j))$, where $\{x_j\}_{j=1}^{30}$ are the grid points. Columns 5 and 6 show the medians of the average widths of the $1,000$ simulated data sets for the unrestricted and restricted estimator, respectively. Let $width_{ur}^s$ and $width_r^s$ be the average widths in data set s. The last columns shows the median of $(width_{ur}^s - width_r^s)/width_{ur}^s$ across the $1,000$ simulated data sets. Even though the mean gains are very similar, we report the median gains to ensure that our gains are not mainly caused by extreme outcomes.

In Table 1 we can see that the unrestricted estimator has coverage rates close to 0.95 if $c = 0$. For $K_n = 2$ and $K_n = 3$, the coverage probability drops significantly below 0.95 when c is large because increasing c also increases the approximation bias. For larger values of K_n, the coverage probability of the unrestricted band is close to 0.95 for all reported values of c. Due to the projection, the coverage probability of the restricted band tends to be above the one of the unrestricted band. When c is large enough, such as $c = 10$ with $K_n = 2$, the two bands are identical with very large probability. The average width of the unrestricted band does not depend on c. On the other hand, the average width of the restricted band is much smaller when c is small. Consequently, the restricted band yields width gains of up to 26.2%. Generally, the widths gains are larger, the larger K_n and the smaller c.[11]

Table 2 shows that the results for the NPIV model are similar, but the gains from using the shape restrictions are much bigger. For example, when $K_n = 5$ and $c = 0$, the gains are 73.1%. Furthermore, the range of c values for which we achieve sizable gains is much larger for NPIV relative to the series regression framework. More generally, due to the larger variance of the estimator in the NPIV model, we observed in a variety of other simulations that the range of functions for which we obtain gains in this model is much larger than in

[11]Since U is normally distributed and independent of X, an alternative method to construct confidence bands in this setting is the one proposed by Dümbgen (2003). Since this method only applies with fixed regressors and since $X \sim Unif(0,1)$ in our simulations, we let $X \in \{-1,000/1,002, -998/1,002, \ldots, 998/1,002\}$, which is an equal spaced grid of size $1,000$. We also assume that the variance of U is known. With the monotonized bands of Dümbgen (2003) we then get coverage rates and widths of 0.948 and 0.233 when $c = 0$, 0.974 and 0.258 when $c = 2$, 0.979 and 0.281 when $c = 4$, 0.983 and 0.301 when $c = 6$, 0.986 and 0.319 when $c = 8$, and 0.988 and 0.336 when $c = 10$. This method is not conservative at the boundary, but it is conservative in the interior. The bands can be empty and arbitrarily small (with and without monotonizing), but they are empty in only 3 out of $1,000$ samples when $c = 0$. An advantage of the method is that it is smoothing parameter free, but the widths are much larger than our widths, even when $K_n = 5$, for all values of c (our method yields almost identical results as those in Table 1 when we fix the grid). Also notice that this method only applies when U is normally distributed and the regressors are fixed.

Table 1: Coverage and width comparison for regression with polynomials

K_n	c	cov_{ur}	cov_r	$width_{ur}$	$width_r$	% gains
2	0	0.957	0.948	0.107	0.090	0.175
	2	0.946	0.949	0.107	0.104	0.030
	4	0.939	0.939	0.107	0.106	0.003
	6	0.891	0.891	0.107	0.107	0.000
	8	0.858	0.858	0.107	0.107	0.000
	10	0.813	0.813	0.107	0.107	0.000
3	0	0.949	0.954	0.142	0.109	0.236
	2	0.947	0.963	0.142	0.121	0.143
	4	0.948	0.960	0.142	0.129	0.091
	6	0.925	0.939	0.142	0.134	0.050
	8	0.910	0.910	0.142	0.137	0.028
	10	0.887	0.884	0.142	0.139	0.015
4	0	0.949	0.969	0.172	0.131	0.238
	2	0.946	0.970	0.172	0.146	0.152
	4	0.945	0.969	0.172	0.155	0.097
	6	0.952	0.963	0.172	0.161	0.058
	8	0.930	0.948	0.172	0.166	0.032
	10	0.939	0.947	0.172	0.168	0.018
5	0	0.941	0.970	0.200	0.147	0.262
	2	0.943	0.971	0.200	0.162	0.187
	4	0.945	0.964	0.200	0.173	0.135
	6	0.948	0.960	0.199	0.180	0.097
	8	0.937	0.951	0.200	0.185	0.072
	10	0.948	0.960	0.200	0.189	0.051

Table 2: Coverage and width comparison for NPIV with polynomials

K_n	c	cov_{ur}	cov_r	$width_{ur}$	$width_r$	% gains
	0	0.946	0.955	0.059	0.046	0.234
	5	0.929	0.929	0.059	0.058	0.016
	10	0.879	0.879	0.060	0.060	0.000
2	20	0.608	0.608	0.059	0.059	0.000
	30	0.229	0.229	0.059	0.059	0.000
	40	0.003	0.003	0.059	0.059	0.000
	50	0.000	0.000	0.059	0.059	0.000
	0	0.933	0.963	0.107	0.061	0.426
	5	0.931	0.949	0.107	0.079	0.257
	10	0.921	0.940	0.107	0.091	0.150
3	20	0.821	0.815	0.107	0.101	0.049
	30	0.681	0.680	0.107	0.105	0.018
	40	0.426	0.426	0.107	0.106	0.002
	50	0.201	0.201	0.106	0.106	0.000
	0	0.951	0.986	0.207	0.092	0.556
	5	0.946	0.982	0.207	0.120	0.422
	10	0.944	0.967	0.208	0.143	0.310
4	20	0.942	0.947	0.208	0.171	0.176
	30	0.954	0.967	0.208	0.185	0.103
	40	0.953	0.959	0.207	0.194	0.057
	50	0.952	0.956	0.208	0.199	0.037
	0	0.959	0.989	0.456	0.122	0.731
	5	0.962	0.994	0.457	0.161	0.649
	10	0.956	0.994	0.460	0.197	0.574
5	20	0.957	0.978	0.465	0.248	0.471
	30	0.973	0.985	0.457	0.288	0.377
	40	0.966	0.978	0.462	0.322	0.310
	50	0.953	0.973	0.459	0.345	0.254

series regression for the same sample size and a similar DGP. Finally notice that when $c = 0$, the width increase as K_n increases and it appears that the width gains coverage to 1 (in fact when $K_n = 6$ and $c = 0$ we get % gains $= 0.825$). Since the gains for $c = 0$ do not depend on n, the restricted band seems to converge in probability to g_0 at a faster rate than the unrestricted band if g_0 is constant and as n and K_n both diverge. These results are in line with Chetverikov and Wilhelm (2017) who show, among others, that the restricted estimator converges at a faster rate than the unrestricted estimator if g_0 is constant.

Figure 4 shows the means of the restricted and the unrestricted bands obtained from the $1,000$ simulated data sets in the NPIV model with $K_n = 5$. Figure 5 contains four specific examples when $c = 5$. In the first example, both the restricted and the unrestricted estimator are monotone, but the restricted band is still much smaller. In the last example

Figure 4: Average confidence bands for NPIV with polynomials and $K_n = 5$

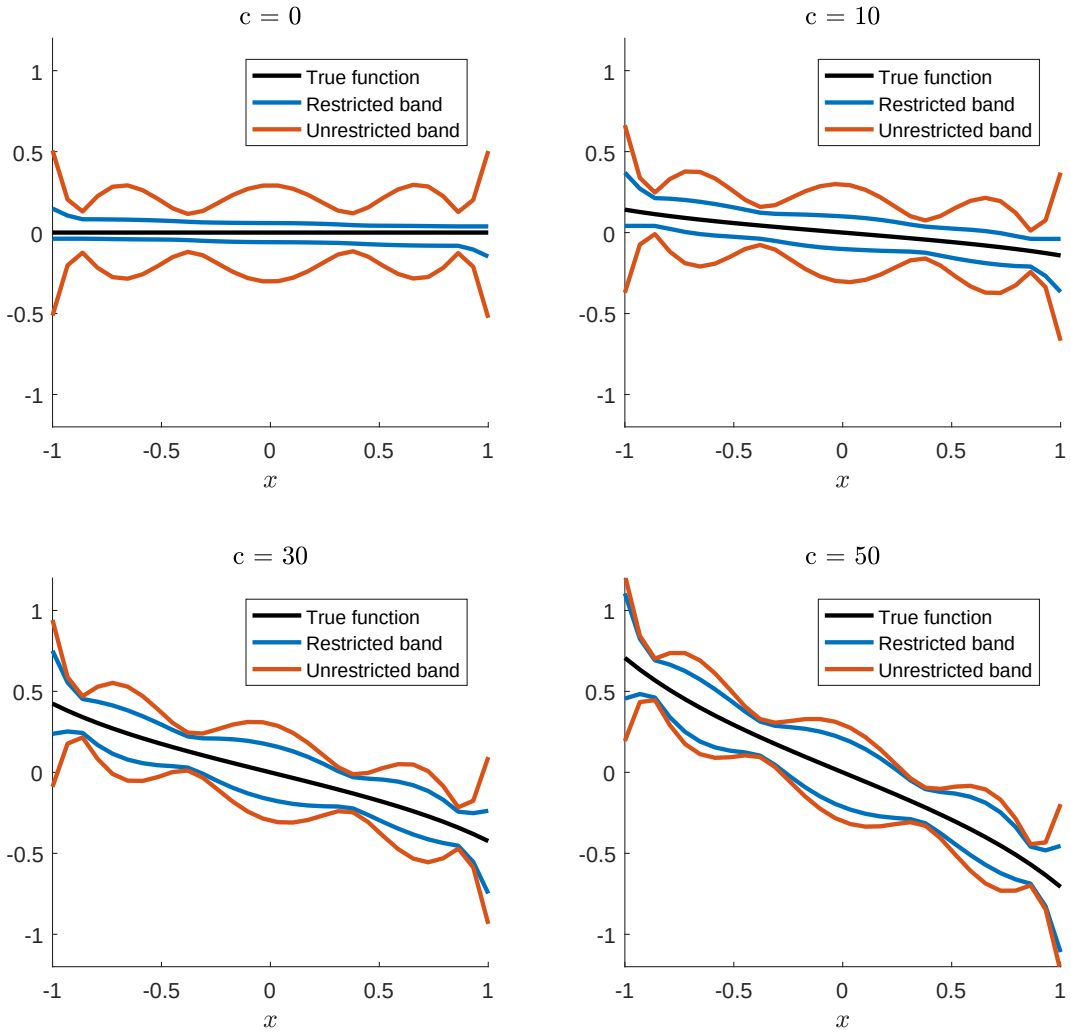
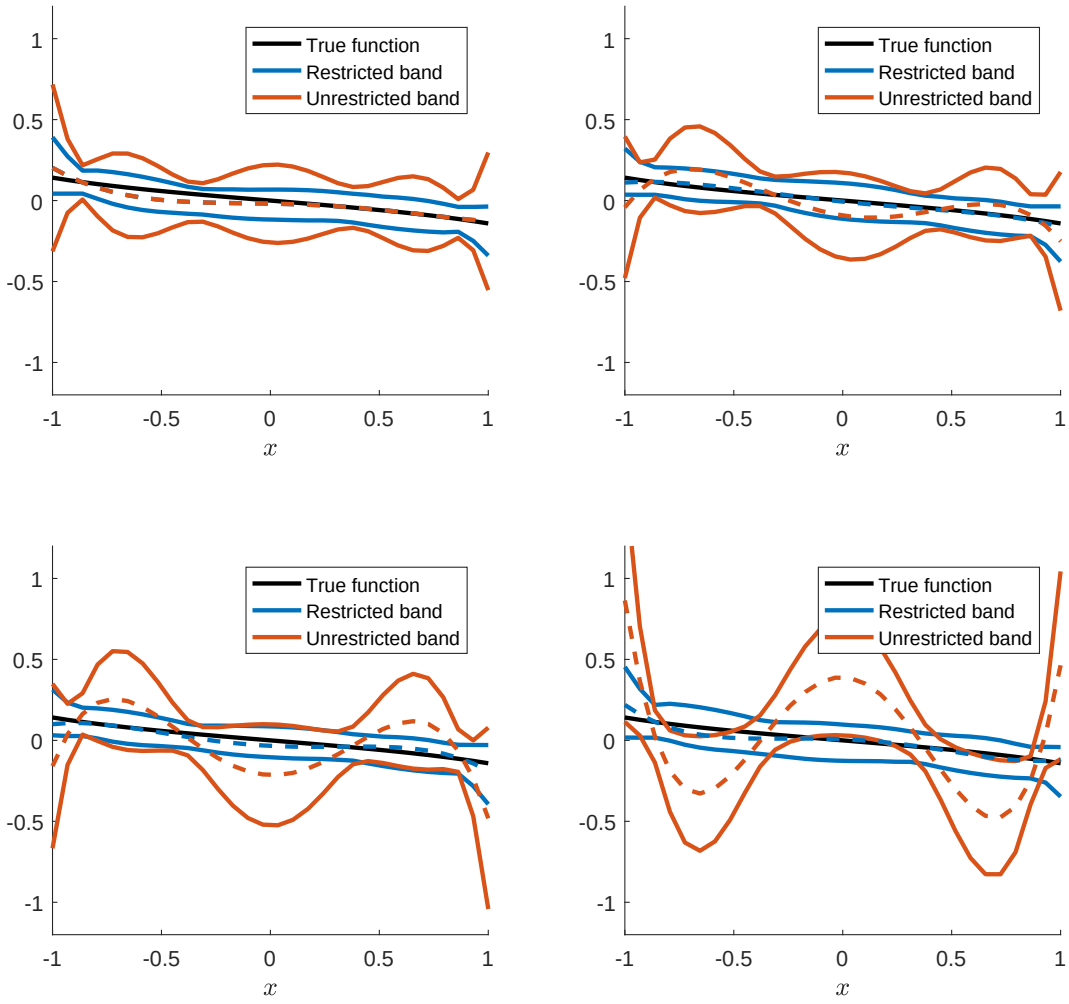


Figure 5: Example confidence bands for NPIV with polynomials and $K_n = 5$



the unrestricted band does not contain any monotone function. In contrast, the restricted bands are always centered around the restricted estimates and both the upper and lower bound functions are monotonically decreasing.

In addition to some of the information in Table 2, we report the width gains relative to monotonized bands in Table 3 for a subset of the DGPs with $K_n = 5$. To obtain these bands we simply exclude all parts of the unrestricted bands which are not consistent with a weakly decreasing function. As we can see from the reported widths, our restricted bands are considerably smaller than these bands as well, and the widths gains are up to $1 - 0.122/0.210 = 41.9\%$. Moreover, the monotonized band may be empty, which happens in 1.6% of the data sets when $c = 0$, or they can be extremely narrow.

Finally, to illustrate that our method is also applicable to functionals, Table 4 shows coverage rates and median widths of confidence intervals for the average derivative of the

Table 3: Width comparison with monotonized bands for NPIV with polynomials

K_n	c	cov_{ur}	cov_r	$widths_{ur}$	$widths_{mon}$	$widths_r$	% empty monotone
5	0	0.959	0.989	0.456	0.210	0.122	0.016
	10	0.956	0.994	0.460	0.299	0.197	0.002
	30	0.973	0.985	0.457	0.373	0.288	0.000
	50	0.953	0.973	0.459	0.411	0.345	0.000

function. Here, we compare our method to the series estimator without shape restrictions (corresponding to cov_{ur} and $widths_{ur}$ in Table 4), the unrestricted approach, but only including the non-positive part of the confidence interval (corresponding to cov_{ur} and $widths_{neg}$), and the method of Chernozhukov, Newey, and Santos (2015) (corresponding to cov_{cns} and $widths_{cns}$). Again, our method yields considerable width gains compared to the unrestricted intervals at or close to the boundary (up to $1-0.233/0.637 = 63\%$ when $c = 0$) and coverage rates above 95%. The approach of Chernozhukov, Newey, and Santos (2015) is not conservative at the boundary, their intervals are in this case narrower than ours (0.164 versus 0.233), and they are empty in 4% of the samples. As we move into the interior of the parameter space, our method performs favorably, which could be due to the choice of the user specified tuning parameters which their approach requires. We use their suggested data dependent procedures and did not explore other choices.[12] The first row of Table 4 contains results when $c = -5$ and thus, the model is misspecified and the true function is increasing. In this

Table 4: Width comparison average derivative

K_n	c	cov_{ur}	cov_r	cov_{cns}	$widths_{ur}$	$widths_{neg}$	$widths_r$	$width_{cns}$	empty neg.	empty CNS
5	-5	0.957	0.000	0.000	0.636	0.208	0.227	0.030	0.118	0.810
	0	0.962	0.957	0.939	0.637	0.325	0.233	0.164	0.026	0.040
	10	0.953	0.967	0.994	0.641	0.593	0.421	0.428	0.000	0.002
	30	0.963	0.971	0.997	0.638	0.638	0.574	0.744	0.000	0.000
	50	0.951	0.963	0.998	0.642	0.642	0.612	0.910	0.000	0.000

[12]Specifically, we use the "aggressive" data dependent choices for r_n and l_n explained in their Section 7.1. These choices might lead to a choice of r_n which is too large in the interior of the parameter space and thus, confidence intervals that are too conservative and unnecessarily wide.

case, their approach yields empty intervals in 81% of the cases, while our method can be interpreted as providing confidence sets for the projection of the true function.

6.1 Computational costs

We ran the simulations using MATLAB and the resources of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. To get accurate computational times we rerun a subset of the simulations using MATLAB R2015a on a desktop computer with an Intel Core i3 processor running at 3.5Ghz. In these simulations, the median times to solve for the uniform confidence bands in the NPIV setting were roughly 5 minutes when $K_n = 2$, 15 minutes when $K_n = 3$, 32 minutes when $K_n = 4$, and 45 minutes when $K_n = 5$. Each of these times is based on 60 simulated data sets. In Section S.3 in the supplement, we provide additional details, such as our selection of starting values. Since we use a grid of 30 points to calculate the uniform confidence bands, we solve 60 optimization problems for each band. Therefore, obtaining confidence intervals for the average derivatives is considerably faster. In particular, the median time is around 2.5 minutes, even though $K_n = 5$. The approach of Chernozhukov, Newey, and Santos (2015) is based on test inversion, where we test a particular value for the average derivative and obtain critical values using the bootstrap. With $2,000$ bootstrap samples, which is then comparable to the $2,000$ normal draws we use for our approach, each test takes around 6 seconds. Thus, the computational costs of the two approaches in this particular setting are similar if we use 25 grid points for the test inversion approach of Chernozhukov, Newey, and Santos (2015), which is considerably less than what we used to obtain the results in Table 4.

There are several possibilities to substantially reduce the computational times. First, notice that the programs for the uniform confidence bands are very easy to parallelize because the optimization problems are solved separately for each grid point. Second, in our setting we could also use an approach recently suggested by Kaido, Molinari, and Stoye (2016) in a computationally similar problem in the moment inequality literature. In our setting, their algorithm leads to essentially identical results in both the simulations and the empirical application; see Section S.3 for more details. Moreover, for the uniform confidence bands in the NPIV setting, the median times with their approach are roughly 2.5 minutes when $K_n = 2$, 8 minutes when $K_n = 3$, 13 minutes when $K_n = 4$, and 20 minutes when $K_n = 5$. Finally, we recently developed the code in Fortran, which runs approximately ten times faster than the MATLAB code in the empirical application below, where we have 16 estimated parameters, and it yields identical results.

7 Empirical application

In this section, we use the data from Blundell, Horowitz, and Parey (2012) and Chetverikov and Wilhelm (2017) to estimate US gasoline demand functions and to provide uniform confidence bands under the assumption that the demand function is weakly decreasing in the price. The data comes from the 2001 National Household Travel Survey and contains, among others, annual gasoline consumption, the gasoline price, and household income for $4,812$ households. We excludes households from Georgia because their gasoline price is much smaller than for all other regions (highest log price of 0.133 while the next largest log price observation is 0.194) and therefore $n = 4,655$. We use the model

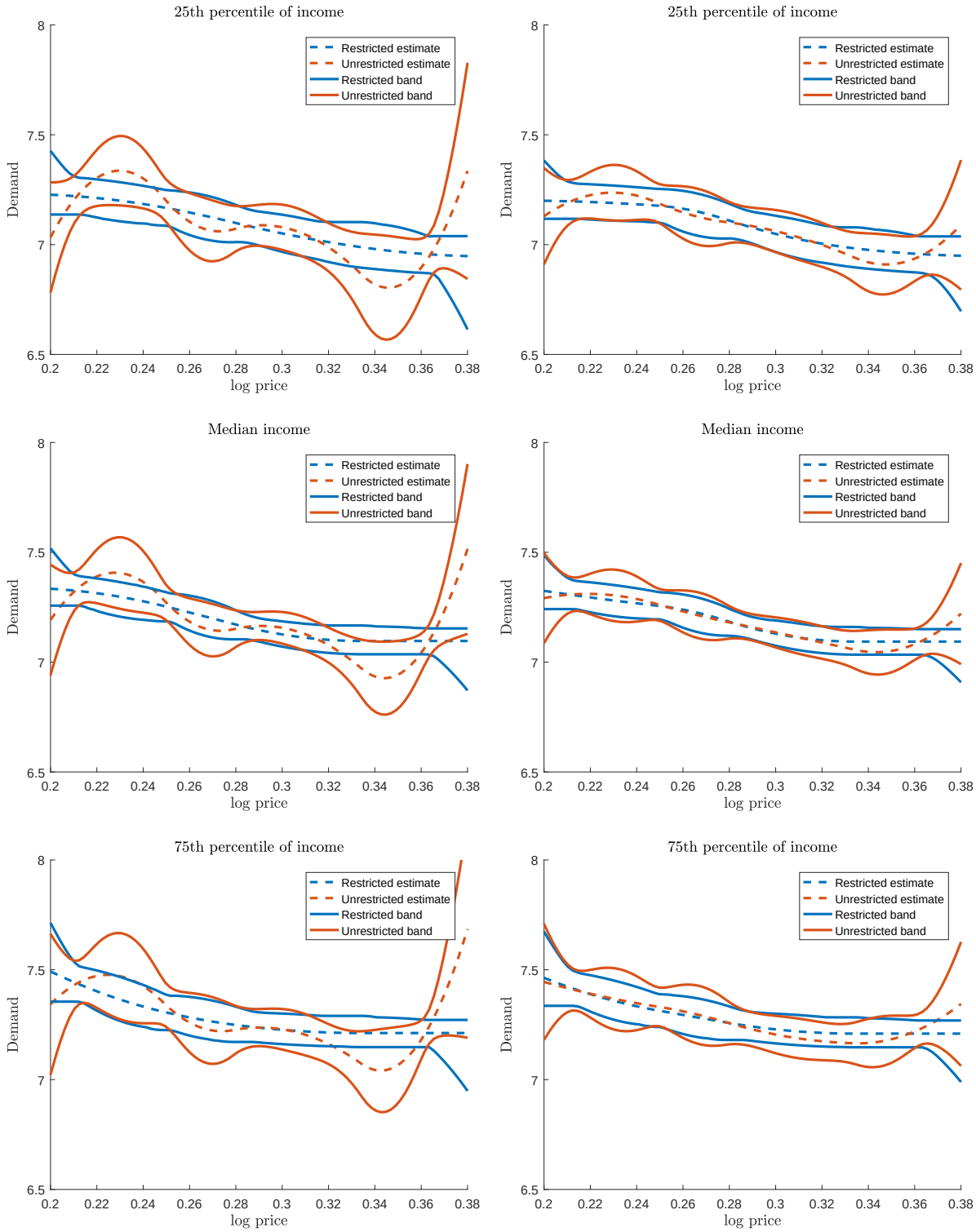
$$Y = g_0(X_1, X_2) + X_3'\gamma_0 + U, \qquad E(U \mid Z, X_2, X_3) = 0.$$

Here Y denotes annual log gasoline consumption of a household, X_1 is the log price of gasoline (the average local price), X_2 is log household income, and X_3 contains additional household characteristics, namely the log age of the household respondent, the log household size, the log number of drivers, and the number of workers in the household. Following Blundell, Horowitz, and Parey (2012) and Chetverikov and Wilhelm (2017), we use the distance to a major oil platform as an instrument, denoted by Z, for X_1. We report estimates and uniform confidence bands for $g_0(x_1, \bar{x}_2) + \bar{x}_3'\gamma_0$ as a function of x_1. We fix X_3 at their mean values and we consider three different values of $\bar{x}_2$, namely the 25th percentile, the median, and the 75th percentile of the income distribution.

The estimator is similar to the one described in Section 5 and our specification is similar to Chetverikov and Wilhelm (2017). Specifically, we use quadratic splines with three interior knots for X_1 (contained in the matrix P_{X_1}) and cubic splines with eight knots for Z (contained in the matrix P_Z). The matrix of regressors is then $(P_{X_1}, P_{X_1} \times X_2, X_3)$, where $\times$ denotes the tensor product of the columns of the matrices, and $(P_Z, P_Z \times X_2, X_3)$ is the matrix of instruments. Chetverikov and Wilhelm (2017) estimate γ in the first step and subtract $X_3'\hat{\gamma}$ from Y, while we estimate all parameters together in order to incorporate the variance of $\hat{\gamma}$ when constructing confidence bands. We also report results for a second specification using quadratic splines with six knots to construct P_Z to illustrate the sensitivity of the estimates.

Figure 6 plots unrestricted and restricted estimators for the three income levels along with 95% uniform confidence bands. The left side contains the estimates with quadratic splines and six knots for Z and right side the estimates with cubic splines and eight knots. The unrestricted estimates are generally increasing for very low and high prices, suggesting that the true demand function has a relatively small slope for these price levels. Our bands are

Figure 6: Estimated demand functions



The three figures on the left side use quadratic splines with six knots to construct P_Z. The three figures on the right side use cubic splines with eight knots.

centered around the restricted estimates with monotone upper and lower bound functions. Moreover, the average width of the restricted band is between 25% and 45% smaller than the average width of the unrestricted band. We can also see from the figures that the unrestricted estimates and bands are very sensitive to the specification, but the restricted ones are not.

8 Conclusion

We provide a general approach for conducting uniformly valid inference under shape restrictions. A main application of our method is the estimation of uniform confidence bands for an unknown function of interest, as well as confidence regions for other features of the function. Our confidence bands are well suited to be reported along with shape restricted estimates, because they are build around restricted estimators and the upper and lower bound functions are consistent with the shape restrictions. In addition, the bands are asymptotically equivalent to standard unrestricted confidence bands if the true function strictly satisfies all shape restrictions, but they can be much smaller if some of the shape restrictions are binding or close to binding. Our method is widely applicable and we provide low level conditions for our assumptions in a regression framework (for both series and kernel estimation) and the NPIV model. We demonstrate in simulations and in an empirical application that our shape restricted confidence bands can be much narrower than unrestricted bands.

There are several interesting directions for future research. First, while we prove uniform size control, we do not provide a formal power analysis. It is known that monotone nonparametric estimators can have a faster rate of convergence if the true function is close to constant (see for example Chetverikov and Wilhelm (2017)). Our simulation results suggest that our bands also converge at a faster rate than unrestricted bands in this case, but establishing this result formally is out of the scope of this paper.

Second, we assume that the restricted estimator is an approximate projection of the unrestricted estimator under a weighted Euclidean norm $\|\cdot\|_{\hat{\Omega}}$. In many settings the matrix $\hat{\Omega}$ can be chosen by the researcher (as in Section 4.2). For example, in a just identified GMM setting it is well known that the unrestricted estimator is invariant to the GMM-weight matrix. However, the restricted estimator generally depends on the GMM-weight matrix because it affects $\hat{\Omega}$. Notice that $\hat{\theta}_{ur}$ is approximately $N(\theta_0, \hat{\Sigma}/\kappa_n^2)$ distributed. To obtain the restricted estimator of θ_0 we could image maximizing the likelihood with respect to θ_0, where the data is $\hat{\theta}_{ur}$, subject to the solution being in Θ_R. It is easy to show that the restricted maximum likelihood estimator is $\arg\min_{\theta\in\Theta_R} \|\theta - \hat{\theta}_{ur}\|_{\hat{\Sigma}^{-1}}$, suggesting to use

$\hat{\Omega} = \hat{\Sigma}^{-1}$, although it is not clear that MLE has optimality properties in this setting. In a just identified GMM setting, such as our regression or IV framework, this amounts to using the standard optimal GMM-weight matrix. In simulations, we found that this weight matrix performs particularly well, but we leave optimality considerations for future research.

Finally, notice that in our setting $\hat{\theta}_r$ is a function of $\hat{\theta}_{ur}$ and hence, $\hat{\theta}_{ur}$ provides more information than $\hat{\theta}_r$. Therefore, instead of letting the test statistic depend on $\kappa_n(\hat{\theta}_r - \theta_0)$, we could let it depend on $\kappa_n(\hat{\theta}_{ur} - \theta_0)$ and incorporate the shape restrictions in the test statistic. This approach would potentially allow us to use additional test statistics. We are particularly interested in rectangular confidence sets for functions of θ_0, which are equivalent to standard rectangular confidence sets if θ_0 is in the interior of Θ_R. Such sets can be obtained using test statistics that depend on $\kappa_n(\hat{\theta}_r - \theta_0)$ and it is therefore not immediately obvious what the potential benefits of a more general formalization are.

A Non-conservative projections

We now formalize the arguments from Section 3.2. Let $h_l(\theta) = c_l + q_l'\theta$, where c_l is a constant and $q_l \in \mathbb{R}^{L_n}$. Let

$$\mathcal{Z}(\hat{\Sigma}) = \left\{ z \in \mathbb{R}^{K_n} : \sup_{l=1,\ldots,L_n} \left\{ |q_l' z| \,/ \sqrt{q_l' \hat{\Sigma} q_l} \right\} \right\},$$

where $c(\hat{\Sigma})$ is such that for $Z \sim N(0, I_{K_n \times K_n})$, $P(\hat{\Sigma}^{1/2} Z \in \mathcal{Z}(\hat{\Sigma}) \mid \hat{\Sigma}) = 1 - \alpha$. We obtain the following corollary.

Corollary A1. *Suppose that Assumptions 1 – 6 hold. Let*

$$T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) = \sup_{l=1,\ldots,L_n} \left\{ \kappa_n \left| q_l' \left(\hat{\theta}_r - \theta \right) \right| / \sqrt{q_l' \hat{\Sigma} q_l} \right\}$$

and let CI be the corresponding confidence region. Let $\Theta_R = \{\theta \in \mathbb{R}^{K_n} : A_n \theta \leq b_n\}$. Suppose that, with probability approaching 1, $A_n z < \kappa_n(b_n - A_n\theta)$ for all $\theta \in CI$ and for all $z \in \mathcal{Z}(\hat{\Sigma})$. Then for all $\theta \in CI$, $c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) = c(\hat{\Sigma})$ with probability approaching 1 and

$$\lim_{n\to\infty} P\left(\hat{h}_l^L \leq h_l(\theta_0) \leq \hat{h}_l^U \text{ for all } l = 1, \ldots, L_n \right) = 1 - \alpha.$$

Notice that $\kappa_n(b_n - A_n\theta) = \kappa_n(b_n - A_n\theta_0) + \kappa_n A_n(\theta_0 - \theta)$. If θ_0 is sufficiently in the interior of the parameter space, then each element of $\kappa_n(b_n - A_n\theta_0)$ goes to infinity. Moreover, if each element of CI converges in probability to θ_0 at rate κ_n, then each element of $\kappa_n A_n(\theta_0 - \theta)$ is bounded in probability. The condition of the corollary then holds for example if $\mathcal{Z}(\hat{\Sigma})$ is bounded with probability approaching 1, but the condition also allows the set to grow.

Proof of Corollary A1. Let $z \in \mathbb{R}^{K_n}$ and $\theta \in CI$. Let

$$z_n(\theta, \hat{\Omega}) = \underset{\lambda \in \mathbb{R}^{K_n} : A_n\lambda \leq \kappa_n(b_n - A_n\theta)}{\arg\min} \|\lambda - z\|^2_{\hat{\Omega}}.$$

Now notice that if $z \in \mathcal{Z}(\hat{\Sigma})$, then with probability approaching 1 we get $z_n(\theta, \hat{\Omega}) = z$. It therefore follows that $c(\theta, \hat{\Sigma}, \hat{\Omega}) \leq c(\hat{\Sigma})$. Now take $z_n(\theta, \hat{\Omega}) \in \mathcal{Z}(\hat{\Sigma})$. Then by assumption $A_n z_n(\theta, \hat{\Omega}) < \kappa_n(b_n - A_n\theta)$ with probability approaching 1. Since $\hat{\Omega}$ is positive definite with probability approaching 1, it follows that $z_n(\theta, \hat{\Omega}) = z$, because otherwise the projection would be on the boundary of the support. Hence $c(\theta, \hat{\Sigma}, \hat{\Omega}) \geq c(\hat{\Sigma})$ and thus $c(\theta, \hat{\Sigma}, \hat{\Omega}) = c(\hat{\Sigma})$. As shown in Section 3.2 if $c(\theta, \hat{\Sigma}, \hat{\Omega}) = c(\hat{\Sigma})$ for all $\theta \in CI$, then the projection is not conservative. □

B Useful lemmas

Lemma 1. Let Q and $\hat{Q}$ be symmetric and positive definite matrices. Then

$$\left| \min_{\|v\|=1} v'\hat{Q}v - \min_{\|v\|=1} v'Qv \right| \leq \max_{\|v\|=1} |v'(\hat{Q} - Q)v| \leq \|\hat{Q} - Q\|_S \leq \|\hat{Q} - Q\|$$

and

$$\left| \max_{\|v\|=1} v'\hat{Q}v - \max_{\|v\|=1} v'Qv \right| \leq \max_{\|v\|=1} |v'(\hat{Q} - Q)v| \leq \|\hat{Q} - Q\|_S \leq \|\hat{Q} - Q\|.$$

Proof. For both lines, the first inequality follows from basic properties of minima and maxima. The second and third inequalities follow from the Cauchy-Schwarz inequality. □

Lemma 2. Let Q and $\hat{Q}$ be symmetric and positive definite matrices. Then

$$\|Q^{1/2} - \hat{Q}^{1/2}\|_S \leq \frac{1}{\left(\lambda_{\min}(Q^{1/2}) + \lambda_{\min}(\hat{Q}^{1/2})\right)} \|Q - \hat{Q}\|_S$$

and

$$\|Q - \hat{Q}\|_S \leq \left(\lambda_{\max}(Q^{1/2}) + \lambda_{\max}(\hat{Q}^{1/2})\right) \|Q^{1/2} - \hat{Q}^{1/2}\|_S.$$

Proof. Let λ^2 be the largest eigenvalue of $(Q^{1/2} - \hat{Q}^{1/2})(Q^{1/2} - \hat{Q}^{1/2})$ with unit length eigenvector v_λ. Since $(Q^{1/2} - \hat{Q}^{1/2})$ is symmetric either λ or $-\lambda$ is an eigenvalue of $(Q^{1/2} - \hat{Q}^{1/2})$ with eigenvector v_λ. It follows that

$$\begin{aligned}
\sup_{\|v\|=1} |v'(Q - \hat{Q})v| &\geq |v_\lambda'(Q - \hat{Q})v_\lambda| \\
&= |v_\lambda' Q^{1/2}(Q^{1/2} - \hat{Q}^{1/2})v_\lambda + v_\lambda'(Q^{1/2} - \hat{Q}^{1/2})\hat{Q}^{1/2}v_\lambda| \\
&= |\lambda||v_\lambda' Q^{1/2} v_\lambda + v_\lambda' \hat{Q}^{1/2} v_\lambda| \\
&\geq |\lambda| \left(\lambda_{\min}(Q^{1/2}) + \lambda_{\min}(\hat{Q}^{1/2})\right)
\end{aligned}$$

and therefore

$$\|Q^{1/2} - \hat{Q}^{1/2}\|_S \leq \frac{1}{\left(\lambda_{\min}(Q^{1/2}) + \lambda_{\min}(\hat{Q}^{1/2})\right)} \|Q - \hat{Q}\|_S.$$

Similarly, for all v with $\|v\| = 1$ we have

$$\begin{aligned} \|(Q - \hat{Q})v\| &= \|Q^{1/2}(Q^{1/2} - \hat{Q}^{1/2})v + (Q^{1/2} - \hat{Q}^{1/2})\hat{Q}^{1/2}v\| \\ &\leq \left(\lambda_{\max}(Q^{1/2}) + \lambda_{\max}(\hat{Q}^{1/2})\right) \|Q^{1/2} - \hat{Q}^{1/2}\|_S. \end{aligned}$$

Therefore,

$$\|Q - \hat{Q}\|_S \leq \left(\lambda_{\max}(Q^{1/2}) + \lambda_{\max}(\hat{Q}^{1/2})\right) \|Q^{1/2} - \hat{Q}^{1/2}\|_S.$$

□

C Proof of Theorem 1

Proof of Theorem 1. First notice that $\lambda_{\min}(\Sigma)$ is bounded and bounded away from 0 and since $\|\Sigma - \hat{\Sigma}\|_S \xrightarrow{p} 0$ by Assumption 5 it follows from Lemma 1 that $\lambda_{\min}(\hat{\Sigma})$ is bounded and bounded way from 0 with probability approaching 1. Similarly, $\lambda_{\max}(\Omega)$ is bounded and bounded away from 0 and $\lambda_{\max}(\hat{\Omega})$ is bounded and bounded way from 0 with probability approaching 1. Hence, there exist constants $B_l > 0$ and $B_u < \infty$ such that $B_l \leq \lambda_{\min}(\Sigma), \lambda_{\max}(\Omega) \leq B_u$ and $B_l \leq \lambda_{\min}(\hat{\Sigma}), \lambda_{\max}(\hat{\Omega}) \leq B_u$ with probability approaching 1 uniformly over $P \in \mathcal{P}$.

Also notice that by Assumption 3

$$\frac{|\lambda_{\min}(\hat{\Omega}) - \lambda_{\min}(\Omega)|}{\lambda_{\min}(\Omega)} \leq \frac{\|\hat{\Omega} - \Omega\|_S}{\lambda_{\min}(\Omega)} \xrightarrow{p} 0$$

and therefore uniformly over $P \in \mathcal{P}$

$$\left| \frac{\lambda_{\min}(\hat{\Omega})}{\lambda_{\min}(\Omega)} - 1 \right| \xrightarrow{p} 0.$$

Hence $\lambda_{\min}(\hat{\Omega}) > 0$ with probability approaching 1 and, uniformly over $P \in \mathcal{P}$,

$$\left| \frac{\lambda_{\min}(\Omega)}{\lambda_{\min}(\hat{\Omega})} - 1 \right| \xrightarrow{p} 0.$$

Take Z_n as defined in Assumption 2 and $\Lambda_n(\theta_0) = \{\lambda \in \mathbb{R}^{K_n} : \lambda = \kappa_n(\theta - \theta_0)$ for some $\theta \in \Theta_R\}$ and define

$$Z_n(\theta_0, \Sigma, \Omega) = \underset{\lambda \in \Lambda_n(\theta_0)}{\arg\min} \|\lambda - Z_n\|_\Omega^2.$$

By Assumptions 5 there exists a constant C such that with probability approaching 1

$$\begin{aligned}
&\left|T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})-T(Z_n(\theta_0,\Sigma,\Omega),\Sigma)\right| \\
&\quad\le \left|T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})-T(Z_n(\theta_0,\Sigma,\Omega),\hat{\Sigma})\right|+\left|T(Z_n(\theta_0,\Sigma,\Omega),\hat{\Sigma})-T(Z_n(\theta_0,\Sigma,\Omega),\Sigma)\right| \\
&\quad\le C\left\|\kappa_n(\hat{\theta}_r-\theta_0)-Z_n(\theta_0,\Sigma,\Omega)\right\|+C\|Z_n(\theta_0,\Sigma,\Omega)\|\|\hat{\Sigma}-\Sigma\|_S \\
&\quad\le C\left\|\kappa_n(\hat{\theta}_r-\theta_0)-Z_n(\theta_0,\Sigma,\hat{\Omega})\right\|+C\left\|Z_n(\theta_0,\Sigma,\Omega)-Z_n(\theta_0,\Sigma,\hat{\Omega})\right\| \\
&\qquad\quad +C\|Z_n(\theta_0,\Sigma,\Omega)\|\|\hat{\Sigma}-\Sigma\|_S.
\end{aligned}$$

We now first prove that each term on the right hand side is $o_p(\varepsilon_n)$ uniformly over $P\in\mathcal{P}$. Since $\Lambda_n(\theta_0)$ is closed and convex it follows from Assumptions 1 and 2 that

$$\begin{aligned}
&\left\|\kappa_n(\hat{\theta}_r-\theta_0)-Z_n(\theta_0,\Sigma,\hat{\Omega})\right\| \\
&\qquad\le \left\|\arg\min_{\lambda\in\Lambda_n(\theta_0)}\|\lambda-\kappa_n(\hat{\theta}_{ur}-\theta_0)\|^2_{\hat{\Omega}}-Z_n(\theta_0,\Sigma,\hat{\Omega})\right\|+\|R_n\| \\
&\qquad\le \lambda_{\min}(\hat{\Omega})^{-1/2}\left\|\arg\min_{\lambda\in\Lambda_n(\theta_0)}\|\lambda-\kappa_n(\hat{\theta}_{ur}-\theta_0)\|^2_{\hat{\Omega}}-Z_n(\theta_0,\Sigma,\hat{\Omega})\right\|_{\hat{\Omega}}+\|R_n\| \\
&\qquad\le \lambda_{\min}(\hat{\Omega})^{-1/2}\|\kappa_n(\hat{\theta}_{ur}-\theta_0)-Z_n\|_{\hat{\Omega}}+\|R_n\| \\
&\qquad\le \sqrt{\frac{\lambda_{\max}(\hat{\Omega})}{\lambda_{\min}(\hat{\Omega})}}\|\kappa_n(\hat{\theta}_{ur}-\theta_0)-Z_n\|+\|R_n\|.
\end{aligned}$$

Also notice that $\sqrt{\lambda_{\max}(\hat{\Omega})}=O_p(1)$ and $\left|\frac{\lambda_{\min}(\Omega)}{\lambda_{\min}(\hat{\Omega})}-1\right|=o_p(1)$ uniformly over $P\in\mathcal{P}$. Combined with Assumptions 1 and 2 this implies that

$$C\left\|\kappa_n(\hat{\theta}_r-\theta_0)-Z_n(\theta_0,\Sigma,\hat{\Omega})\right\|=o_p(\varepsilon_n)$$

uniformly over $P\in\mathcal{P}$.

Next notice that the $K_n\times 1$ zero vector is in $\Lambda_n(\theta_0)$. Therefore

$$\|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|_{\Omega}\le\|Z_n\|_{\Omega}$$

and thus,

$$\sqrt{\lambda_{\min}(\Omega)}\|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|\le\sqrt{\lambda_{\max}(\Omega)}\|Z_n\|.$$

It follows that

$$\begin{aligned}
\|Z_n(\theta_0,\Sigma,\hat{\Omega})-Z_n\|^2_{\hat{\Omega}} &\le \|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|^2_{\hat{\Omega}} \\
&= \|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|^2_{\Omega}+\|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|^2_{\hat{\Omega}-\Omega} \\
&\le \|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|^2_{\Omega}+\|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|^2\|\hat{\Omega}-\Omega\|_S \\
&\le \|Z_n(\theta_0,\Sigma,\Omega)-Z_n\|^2_{\Omega}+\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)}\,\|Z_n\|^2\|\hat{\Omega}-\Omega\|_S.
\end{aligned}$$

Let

$$\hat{V}_1 = \frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} \, \|Z_n\|^2 \|\hat{\Omega} - \Omega\|_S.$$

Analogously, we get

$$\|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n\|_{\Omega}^2 \;\; \leq \;\; \|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n\|_{\hat{\Omega}}^2 + \hat{V}_2,$$

where

$$\hat{V}_2 = \frac{\lambda_{\max}(\hat{\Omega})}{\lambda_{\min}(\hat{\Omega})} \, \|Z_n\|^2 \|\hat{\Omega} - \Omega\|_S.$$

Since $\Lambda_n(\theta_0)$ is convex it follows that for any $\gamma \in (0, 1)$

$$\begin{aligned}
\|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\Omega}^2 &\leq \|\gamma Z_n(\theta_0, \Sigma, \Omega) + (1-\gamma) Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n\|_{\Omega}^2 \\
&= \gamma \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\Omega}^2 + (1-\gamma) \|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n\|_{\Omega}^2 \\
&\quad -\gamma(1-\gamma) \|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega)\|_{\Omega}^2 \\
&\leq \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\Omega}^2 + (1-\gamma)(\hat{V}_1 + \hat{V}_2) \\
&\quad -\lambda_{\min}(\Omega)\gamma(1-\gamma) \|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega)\|^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega)\|^2 &\leq \frac{1}{\lambda_{\min}(\Omega)\gamma} (\hat{V}_1 + \hat{V}_2) \\
&= \frac{1}{\lambda_{\min}(\Omega)\gamma} \left(\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\hat{\Omega})}{\lambda_{\min}(\hat{\Omega})} \right) \|Z_n\|^2 \|\hat{\Omega} - \Omega\|_S \\
&\leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)\gamma} \left(\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\hat{\Omega})}{\lambda_{\min}(\hat{\Omega})} \right) \|\Sigma^{-1/2} Z_n\|^2 \|\hat{\Omega} - \Omega\|_S.
\end{aligned}$$

Since

$$\left| \frac{\lambda_{\min}(\Omega)}{\lambda_{\min}(\hat{\Omega})} - 1 \right| \xrightarrow{p} 0,$$

$\lambda_{\max}(\hat{\Omega})$ is bounded with probability approaching 1, and $\|\Sigma^{-1/2} Z_n\|^2 = O_p(K_n)$ by Markov's inequality, it follows from Assumption 3 that

$$C^2 \|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega)\|^2 = o_p(\varepsilon_n^2)$$

uniformly over $P \in \mathcal{P}$ and thus

$$C \|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega)\| = o_p(\varepsilon_n)$$

uniformly over $P \in \mathcal{P}$.

From the arguments above and the reverse triangle inequality we have

$$\|Z_n(\theta_0,\Sigma,\Omega)\|_\Omega - \|Z_n\|_\Omega \leq \|Z_n(\theta_0,\Sigma,\Omega) - Z_n\|_\Omega \leq \|Z_n\|_\Omega$$

and therefore

$$C\|Z_n(\theta_0,\Sigma,\Omega)\|\|\hat{\Sigma}-\Sigma\|_S \leq 2C\|\hat{\Sigma}-\Sigma\|_S\sqrt{\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)}}\|Z_n\|$$

and by Assumption 3 and $\|Z_n\| = O_p(\sqrt{\lambda_{\max}(\Sigma)K_n})$

$$C\|Z_n(\theta_0,\Sigma,\Omega)\|\|\hat{\Sigma}-\Sigma\|_S = o_p(\varepsilon_n)$$

uniformly over $P \in \mathcal{P}$.

Next define

$$\begin{aligned} B_n &= C\left\|\kappa_n(\hat{\theta}_r-\theta_0) - Z_n(\theta_0,\Sigma,\hat{\Omega})\right\| + C\left\|Z_n(\theta_0,\Sigma,\Omega) - Z_n(\theta_0,\Sigma,\hat{\Omega})\right\| \\ &\quad + C\|Z_n(\theta_0,\Sigma,\Omega)\|\|\hat{\Sigma}-\Sigma\|_S, \end{aligned}$$

The previous derivations imply that

$$\sup_{P\in\mathcal{P}} P\left(B_n \geq \frac{1}{2}\varepsilon_n\right) \to 0.$$

Therefore

$$\begin{aligned} &P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right) \\ &\quad\geq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma) \leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega}) - B_n\right) - o(1) \\ &\quad\geq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma) \leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega}) - \frac{1}{2}\varepsilon_n, B_n \leq \frac{1}{2}\varepsilon_n\right) - o(1) \\ &\quad\geq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma) \leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega}) - \frac{1}{2}\varepsilon_n\right) - P\left(B_n \geq \frac{1}{2}\varepsilon_n\right) - o(1), \end{aligned}$$

where the $o(1)$ term belongs to $P(B_l \leq \lambda_{\min}(\hat{\Sigma}) \leq B_u, B_l \leq \lambda_{\max}(\hat{\Omega}) \leq B_u)$ and it converges to 0 uniformly over $P \in \mathcal{P}$. Similarly, we get

$$\begin{aligned} &P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right) \\ &\quad\leq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma) \leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})) + \frac{1}{2}\varepsilon_n\right) + P\left(B_n \geq \frac{1}{2}\varepsilon_n\right) + o(1). \end{aligned}$$

We next show that for any sufficiently small $\delta_q \in (0,\alpha)$ it holds that

$$\sup_{P\in\mathcal{P}}\left|P\left(c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega}) \geq c_{1-\alpha-\delta_q,n}(\theta_0,\Sigma,\Omega) - \frac{1}{2}\varepsilon_n\right) - 1\right| \to 0$$

and

$$\sup_{P\in\mathcal{P}}\left|P\left(c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\leq c_{1-\alpha+\delta_q,n}(\theta_0,\Sigma,\Omega)+\frac{1}{2}\varepsilon_n\right)-1\right|\to 0.$$

It then follows that

$$\begin{aligned}&\inf_{P\in\mathcal{P}}P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})+\varepsilon_n\right)\\&\quad\geq\inf_{P\in\mathcal{P}}P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma)\leq c_{1-\alpha-\delta_q,n}(\theta_0,\Sigma,\Omega)\right)-o(1)\end{aligned}$$

which implies that

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}}P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})+\varepsilon_n\right)\geq 1-\alpha-\delta_q.$$

Since δ_q was arbitrary

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}}P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})+\varepsilon_n\right)\geq 1-\alpha,$$

which is the first conclusion of Theorem 1. Similarly, for all δ_q sufficiently small

$$\begin{aligned}&P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right)\\&\quad\geq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma)\leq c_{1-\alpha-\delta_q,n}(\theta_0,\Sigma,\Omega)-\varepsilon_n\right)-o(1)\end{aligned}$$

which implies that

$$\begin{aligned}&P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right)-(1-\alpha)\\&\quad\geq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma)\leq c_{1-\alpha-\delta_q,n}(\theta_0,\Sigma,\Omega)-\varepsilon_n\right)-(1-\alpha-\delta_q)-\delta_q-o(1).\end{aligned}$$

Analogously,

$$\begin{aligned}&P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right)-(1-\alpha)\\&\quad\leq P\left(T(Z(\theta_0,\Sigma,\Omega),\Sigma)\leq c_{1-\alpha+\delta_q,n}(\theta_0,\Sigma,\Omega)+\varepsilon_n\right)-(1-\alpha+\delta_q)+\delta_q+o(1).\end{aligned}$$

Hence, if Assumption 6 holds, then for all $\delta_q\in(0,\alpha)$

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\left|P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right)-(1-\alpha)\right|\leq\delta_q$$

and since δ_q was arbitrary

$$\sup_{P\in\mathcal{P}}\left|P\left(T(\kappa_n(\hat{\theta}_r-\theta_0),\hat{\Sigma})\leq c_{1-\alpha,n}(\theta_0,\hat{\Sigma},\hat{\Omega})\right)-(1-\alpha)\right|\to 0.$$

For the final step of the proof let $\delta_q\in(0,\alpha)$ be arbitrary. Let $\delta_\varepsilon>0$, which may depend on δ_q, and define the set $\mathcal{H}_n$ as all $(\tilde{\Sigma},\tilde{\Omega})$ on the support of $(\hat{\Sigma},\hat{\Omega})$ such that $B_l\leq\lambda_{\min}(\tilde{\Sigma}),\lambda_{\max}(\tilde{\Omega})\leq B_u$,

$$\sqrt{K_n}\frac{\sqrt{\lambda_{\max}(\Sigma)}}{\sqrt{\lambda_{\min}(\Omega)}}\|\tilde{\Sigma}-\Sigma\|_S\leq\delta_\varepsilon\varepsilon_n$$

and

$$K_n \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)} \left(\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\tilde{\Omega})}{\lambda_{\min}(\tilde{\Omega})} \right) \|\tilde{\Omega} - \Omega\|_S \leq \delta_\varepsilon^2 \varepsilon_n^2.$$

Notice that Assumption 3 implies that

$$\sup_{P \in \mathcal{P}} \left| P\left((\hat{\Sigma}, \hat{\Omega}) \in \mathcal{H}_n \right) - 1 \right| \to 0.$$

Let $\tilde{Z}_n \sim N(0, I_{K_n \times K_n})$ be independent of $\hat{\Sigma}$ and $\hat{\Omega}$ and define

$$\tilde{Z}_n(\theta_0, \Sigma, \Omega) = \underset{\lambda \in \Lambda_n(\theta_0)}{\arg\min} \|\lambda - \Sigma^{1/2} \tilde{Z}_n\|_\Omega^2.$$

For any $(\Sigma^*, \Omega^*) \in \mathcal{H}_n$ we get by Assumption 5 that

$$\begin{aligned}
&\left| T(\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*), \Sigma^*) - T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma) \right| \\
&\quad \leq \left| T(\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*), \Sigma^*) - T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma^*) \right| \\
&\qquad + \left| T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma^*) - T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma) \right| \\
&\quad \leq C \left\| \tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*) - \tilde{Z}_n(\theta_0, \Sigma, \Omega^*) \right\| \\
&\qquad + C \left\| \tilde{Z}_n(\theta_0, \Sigma, \Omega) - \tilde{Z}_n(\theta_0, \Sigma, \Omega^*) \right\| + C \|\tilde{Z}_n(\theta_0, \Sigma, \Omega)\| \|\Sigma - \Sigma^*\|_S.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\|\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*) - \tilde{Z}_n(\theta_0, \Sigma, \Omega^*)\| &\leq \sqrt{\lambda_{\max}(\Omega^*)} \|(\Sigma^*)^{1/2} \tilde{Z}_n - \Sigma^{1/2} \tilde{Z}_n\| \\
&\leq \sqrt{\lambda_{\max}(\Omega^*)} \|(\Sigma^*)^{1/2} - \Sigma^{1/2}\|_S \|\tilde{Z}_n\| \\
&\leq \frac{\sqrt{\lambda_{\max}(\Omega^*)}}{\sqrt{\lambda_{\min}(\Sigma)} + \sqrt{\lambda_{\min}(\Sigma^*)}} \|\Sigma^* - \Sigma\|_S \|\tilde{Z}_n\|,
\end{aligned}$$

where the last line follows from Lemma 2. Also by the previous results

$$\begin{aligned}
&\|\tilde{Z}_n(\theta_0, \Sigma, \Omega^*) - \tilde{Z}_n(\theta_0, \Sigma, \Omega)\|^2 \\
&\quad \leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)\gamma} \left(\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\Omega^*)}{\lambda_{\min}(\Omega^*)} \right) \|\tilde{Z}_n\|^2 \|\Omega^* - \Omega\|_S.
\end{aligned}$$

and

$$\|Z_n(\theta_0, \Sigma, \Omega)\| \leq 2 \sqrt{\frac{\lambda_{\max}(\Omega)\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)}} \|\tilde{Z}_n\|.$$

Let

$$\begin{aligned}
H_n &= C \left\| \tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*) - Z_n(\theta_0, \Sigma, \Omega^*) \right\| + C \|Z_n(\theta_0, \Sigma, \Omega) - Z_n(\theta_0, \Sigma, \Omega^*)\| \\
&\quad + C \|Z_n(\theta_0, \Sigma, \Omega)\| \|\Sigma - \Sigma^*\|_S.
\end{aligned}$$

Then, for constants M_1 and M_2 that do not depend on P or δ_ε

$$\begin{aligned} H_n^2 &\leq 4C^2 \left\| \tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*) - Z_n(\theta_0, \Sigma, \Omega^*) \right\|^2 + 4C^2 \|Z_n(\theta_0, \Sigma, \Omega)\|^2 \|\Sigma - \Sigma^*\|_S^2 \\ &\quad + 4C^2 \left\| Z_n(\theta_0, \Sigma, \Omega) - Z_n(\theta_0, \Sigma, \Omega^*) \right\|^2 \\ &\leq M_1 \|\Sigma - \Sigma^*\|_S^2 \|\tilde{Z}_n\|^2 + M_1 \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)} \|\tilde{Z}_n\|^2 \|\Sigma - \Sigma^*\|_S^2 \\ &\quad + M_1 \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)} \left(\frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\Omega^*)}{\lambda_{\min}(\Omega^*)} \right) \|\tilde{Z}_n\|^2 \|\Omega^* - \Omega\|_S \\ &\leq M_2 \varepsilon_n^2 \delta_\varepsilon^2 \frac{\|\tilde{Z}_n\|^2}{K_n}. \end{aligned}$$

Since $E(\|\tilde{Z}_n\|^2) = K_n$ it follows from Markov's inequality that

$$\sup_{P \in \mathcal{P}} P\left(H_n \geq \frac{1}{2}\varepsilon_n \right) \leq 4M_2 \delta_\varepsilon^2.$$

Therefore

$$\begin{aligned} 1 - \alpha &= P\left(T(\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) \right) \\ &\leq P\left(T(\tilde{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) + H_n \right) \\ &\leq P\left(T(\tilde{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) + \frac{1}{2}\varepsilon_n \right) + 4M_2 \delta_\varepsilon^2. \end{aligned}$$

It follows that we can pick δ_ε small enough such that for any $P \in \mathcal{P}$, and any $(\Sigma^*, \Omega^*) \in \mathcal{H}_n$

$$1 - \alpha - \delta_q \leq P\left(T(\tilde{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) + \frac{1}{2}\varepsilon_n \right)$$

and thus

$$c_{1-\alpha-\delta_q,n}(\theta_0, \Sigma, \Omega) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) + \frac{1}{2}\varepsilon_n.$$

Hence

$$P\left((\hat{\Sigma}, \hat{\Omega}) \in \mathcal{H}_n \right) \leq P\left(c_{1-\alpha-\delta_q,n}(\theta_0, \Sigma, \Omega) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \frac{1}{2}\varepsilon_n \right)$$

and since

$$\sup_{P \in \mathcal{P}} \left| P\left((\hat{\Sigma}, \hat{\Omega}) \in \mathcal{H}_n \right) - 1 \right| \to 0$$

we have

$$\sup_{P \in \mathcal{P}} \left| P\left(c_{1-\alpha-\delta_q,n}(\theta_0, \Sigma, \Omega) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \frac{1}{2}\varepsilon_n \right) - 1 \right| \to 0.$$

Analogously, for any $(\Sigma^*, \Omega^*) \in \mathcal{H}_n$

$$\begin{aligned} 1 - \alpha &= P\left(T(\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) \right) \\ &\geq P\left(T(\tilde{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) - H_n \right) \\ &\geq P\left(T(\tilde{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) - \frac{1}{2}\varepsilon_n \right) - 4M_2 \delta_\varepsilon^2. \end{aligned}$$

It follows that we can pick δ_ε small enough such that for any $P \in \mathcal{P}$, and any $(\Sigma^*, \Omega^*) \in \mathcal{H}_n$

$$c_{1-\alpha+\delta_q,n}(\theta_0, \Sigma, \Omega) \geq c_{1-\alpha,n}(\theta_0, \Sigma^*, \Omega^*) - \frac{1}{2}\varepsilon_n.$$

Hence

$$P\left((\hat{\Sigma}, \hat{\Omega}) \in \mathcal{H}_n\right) \leq P\left(c_{1-\alpha+\delta_q,n}(\theta_0, \Sigma, \Omega) \geq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) - \frac{1}{2}\varepsilon_n\right)$$

and

$$\sup_{P \in \mathcal{P}} \left| P\left(c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \leq c_{1-\alpha+\delta_q,n}(\theta_0, \Sigma, \Omega) + \frac{1}{2}\varepsilon_n\right) - 1 \right| \to 0.$$

□

References

Ait-Sahalia, Y. and J. Duarte (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics 116*(1-2), 9–47.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica 67*(6), 1341–1383.

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica 69*(3), 683–734.

Andrews, D. W. K. and X. Cheng (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica 80*(5), 2153–2211.

Andrews, D. W. K. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*(1), 119–157.

Armstrong, T. and M. Kolesár (2016). Simple and honest confidence intervals in nonparametric regression. Working paper.

Bellec, P. C. (2016). Adaptive confidence sets in shape restricted regression. Working paper.

Beresteanu, A. (2005). Nonparametric Analysis of Cost Complementarities in the Telecommunications Industry. *RAND Journal of Economics 36*(4), 870–889.

Birke, M. and H. Dette (2007). Estimating a convex function in nonparametric regression. *Scandinavian Journal of Statistics 34*(2), 384–404.

Blundell, R., J. L. Horowitz, and M. Parey (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics 3*(1), 29–51.

Blundell, R., J. L. Horowitz, and M. Parey (2017). Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *The Review of Economics and Statistics 99*(2), 291–304.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics 26*(4), 607–616.

Cai, T. T., M. G. Low, and Y. Xia (2013). Adaptive confidence intervals for regression functions under shape constraints. *The Annals of Statistics 41*(2), 722–750.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2017). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association, forthcoming*.

Chatterjee, S., A. Guntuboyina, and B. Sen (2015). On risk bounds in isotonic and other shape restricted regression problems. Working paper.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika 96*(3), 559–575.

Chernozhukov, V., W. K. Newey, and A. Santos (2015). Constrained conditional moment restriction models. Working paper.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018). The econometrics of shape restrictions. *Annual Review of Economics*, forthcoming.

Chetverikov, D. and D. Wilhelm (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica 85*(4), 1303–1320.

Delecroix, M. and C. Thomas-Agnan (2000). Spline and kernel regression under shape restrictions. In M. G. Schimek (Ed.), *Smoothing and Regression: Approaches, Computation, and Application*, Chapter 5, pp. 109–133. John Wiley & Sons, Inc.

Dette, H., N. Neumeyer, and K. F. Pilz (2006, 06). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli 12*(3), 469–490.

Dierckx, P. (1980). Algorithm/algorithmus 42 an algorithm for cubic spline fitting with convexity constraints. *Computing 24*(4), 349–371.

Du, P., C. F. Parmeter, and J. S. Racine (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica 23*(3), 1347–1371.

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of Statistics 26*(1), 288–314.

Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli 9*(3), 423–449.

Freyberger, J. and J. L. Horowitz (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics 189*(1), 41–53.

Freyberger, J. and Y. Rai (2018). Uniform confidence bands: characterization and optimality. *Journal of Econometrics 204*(1), 119–130.

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics 22*(4), 1993–2010.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001). Estimation of a convex function: Characterizations and asymptotic theory. *The Annals of Statistics 29*(6), 1653–1698.

Haag, B. R., S. Hoderlein, and K. Pendakur (2009). Testing and imposing slutsky symmetry in nonparametric demand systems. *Journal of Econometrics 153*(1), 33–50.

Hall, P. and L.-S. Huang (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics 29*(3), 624–647.

Henderson, D. J. and C. F. Parmeter (2009). Imposing Economic Constraints in Nonparametric Regression: Survey, Implementation and Extension. IZA Discussion Papers 4103.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association 49*(267), 598–619.

Horowitz, J. L. and S. Lee (2017). Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics 201*(1), 108 – 126.

Kaido, H., F. Molinari, and J. Stoye (2016). Confidence intervals for projections of partially identified parameters. Working paper.

Ketz, P. (2017). Subvector inference when the true parameter vector is near the boundary. Working paper.

Lewbel, A. (1995). Consistent nonparametric hypothesis tests with an application to slutsky symmetry. *Journal of Econometrics 67*(2), 379–401.

Mammen, E. (1991a). Estimating a smooth monotone regression function. *The Annals of Statistics 19*(2), 724–740.

Mammen, E. (1991b). Nonparametric regression under qualitative smoothness assumptions. *The Annals of Statistics 19*(2), 741–759.

Mammen, E. and C. Thomas-Agnan (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics 26*(2), 239–252.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Chapter 42, pp. 2524–2558. Elsevier Science.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica 75*(5), 1411–1452.

Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics 16*(2), 741–750.

Müller, U. K. and A. Norets (2016). Credibility of confidence sets in nonstandard econometric problems. *Econometrica 84*(6), 2183–2213.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Pal, J. K. and M. Woodroofe (2007). Large sample properties of shape restricted regression estimators with smoothness adjustments. *Statistica Sinica 17*(4), 1601–1616.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science 3*(4), 425–461.

Wang, X. and J. Shen (2013). Uniform convergence and rate adaptive estimation of convex functions via constrained optimization. *SIAM Journal on Control and Optimization 51*(4), 2753–2787.

Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *The Annals of Statistics 9*(2), 443–448.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics 30*(2), 528–555.