Inference under shape restrictions*

Joachim Freyberger† Brandon Reeves‡

July 31, 2017

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. A major application of our inference method is to construct uniform confidence bands for an unknown function of interest. Our confidence 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 Richard Blundell, Ivan Canay, Bruce Hansen, Joel Horowitz, Philipp Ketz, Matt Masten, Francesca Molinari, Taisuke Otsu, Jack Porter, Azeem Shaikh, Xiaoxia Shi, Alex Torgovitsky, Daniel Wilhelm, and seminar participants at UW Madison, UCL, LSE, Boston College, Northwestern University, and Humboldt University 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 in order 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 education is positive, and they can be implied by economic theory. For example, demand functions are generally monotonically decreasing in prices, cost functions are monotone increasing, homogeneous of degree 1, and concave in input prices, Engel curves of normal goods are monotonically increasing, and utility functions of risk averse agents are concave. There is a long history of estimation under shape restrictions in econometrics and statistics, going back to Hildreth (1954) and Brunk (1955), 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. One would therefore hope that the improved finite sample precision translates to smaller confidence sets.

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, which incorporates shape restrictions and can be used to test hypotheses about an unknown function or parameter vector. 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 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. Our confidence bands have desirable properties. In particular, 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, the confidence bands are generally much smaller. The decrease in the width reflects the increased precision of the constrained estimator. Moreover, the bands always include the shape restricted estimator of the function and are therefore never empty. 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.\footnote{Analogously to many other papers, closeness to the boundary is relative to the sample size.} To the best of our knowledge, even in a regression model under monotonicity, there are no existing uniform confidence bands, which are never empty, uniformly valid, and yield width reductions when the shape restrictions are binding or close to binding. Furthermore, 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 inference problems in nonstandard settings, instead of trying to obtain confidence sets directly from the asymptotic distribution of the estimator, our inference procedure is based on test inversion.\footnote{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).} This means that we start by testing the null hypothesis that the true parameter vector $\theta_0$ is equal to some fixed value $\overline{\theta}$. In series estimation $\theta_0$ represents the coefficients in the series approximation of a function and $\theta_0$ 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. We can then collect all values for which we do not reject, which form a confidence set for $\theta_0$.

To obtain uniform confidence bands, or confidence sets for other functions of $\theta_0$, we project on the confidence set for $\theta_0$ (see Section 2 for a simple illustration). We choose the test statistic in such a way that our confidence bands are asymptotically equivalent to standard unrestricted confidence bands if $\theta_0$ is sufficiently in the interior of the parameter space. Thus, in this case, the confidence bands 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 band. Furthermore, due to
test inversion and projections, our inference method can be computationally demanding.
We give a sense of the computational costs in Section 6. We also briefly describe recent
computational advances, which might help to mitigate these costs.

In Monte Carlo simulations we construct uniform confidence bands in a series regression
framework as well as 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 provide an empirical
application, where we estimate demand functions for gasoline, subject to the functions being
weakly decreasing. The width gains from using shape restrictions are between 25% and 45%
in this setting.

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 Pflitz (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) 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. Specifically, the results in Chetverikov and Wilhelm (2017) imply that a monotone estimator in the NPIV setting does not suffer from a slow rate of convergence due to the ill-posed inverse problem if the true function is close to constant. There is also a large, less related literature on testing shape restrictions.

Using existing methods, uniform confidence bands under shape restrictions can be obtained in three distinct ways. First, one could obtain a standard unrestricted confidence band and intersect it with all functions which satisfy the shape restrictions (see for example Dümbgen (1998, 2003)). A drawback of the resulting bands is that they can be empty with positive probability and hence, they do not satisfy the “reasonableness” property of Müller and Norets (2016). Furthermore, in our simulations, our bands are on average much narrower than such 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 the band. Finally, one could use a two step procedure recently suggested by Horowitz and Lee (2017) in a kernel regression framework with very general constraints. In the first step, they estimate the points where the shape 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.

Chernozhukov, Newey, and Santos (2015) develop a general testing procedure, which allows, among others, testing shape restrictions and obtaining confidence regions for functionals under shape restrictions. Even though there is some overlap in the settings where both methods apply, the technical arguments are very different. Their method is also based on test inversion and it is robust to partial identification, but it is restricted to conditional moment models and series estimators. We allow for a general setup and estimators, but we assume point identification. Since our focus is on testing a growing parameter vector, we are able to obtain uniform confidence bands next to confidence sets for functionals. However, if the main object of interest is a single functional, their approach might be computationally simpler because they test fixed values of the functional directly, rather than projecting on a confidence set for the entire parameter vector.

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 $\Omega$
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 that $X_n =
o_p(\varepsilon_n)$ uniformly over $P \in \mathcal{P}$ if $\sup_{P \in \mathcal{P}} P(\|X_n\| \leq \delta \varepsilon_n) \to 0$ for any $\delta > 0$. We say that
$X_n = O_p(\varepsilon_n)$ uniformly over $P \in \mathcal{P}$ if for any $\delta > 0$ there are $M_\delta$ and $N_\delta$ 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 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 $\theta_0$ under the assumption that
$\theta_{0,1} \leq \theta_{0,2}$. An unrestricted estimator of $\theta_0$, denoted by $\hat{\theta}_{ur}$, is

$$\hat{\theta}_{ur} = \arg \min_{\theta \in \mathbb{R}^2} (\theta_1 - X_1)^2 + (\theta_2 - X_2)^2.$$
Hence $\hat{\theta}_{ur} = \bar{X}$. Analogously, a restricted estimator is

$$
\hat{\theta}_r = \arg\min_{\theta \in \mathbb{R}^2 : \theta_1 \leq \theta_2} (\theta_1 - \bar{X}_1)^2 + (\theta_2 - \bar{X}_2)^2
= \arg\min_{\theta \in \mathbb{R}^2 : \theta_1 - \theta_2 \leq 0} \|\theta - \hat{\theta}_{ur}\|^2
= \arg\min_{\theta \in \mathbb{R}^2 : \theta_1 - \theta_2 \leq 0} \|\sqrt{n}(\theta - \theta_0) - \sqrt{n}(\hat{\theta}_{ur} - \theta_0)\|^2.
$$

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

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

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

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

where $\overset{d}{=}$ 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 $\theta_0$ 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 $\theta_0$ 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 $Z$ on the set $\{z \in \mathbb{R}^2 : z_1 \leq z_2\}$.

A 95% confidence region for $\theta_0$ 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 $\theta_0$. While there are many different 95% confidence regions for $\theta_0$, 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

$$\left\{ \theta \in \mathbb{R}^2 : \theta_1 \leq \theta_2, \hat{\theta}_{r,1} - \frac{c_{r,n}(\theta)}{\sqrt{n}} \leq \theta_1 \leq \hat{\theta}_{r,1} + \frac{c_{r,n}(\theta)}{\sqrt{n}}, \theta_2 \leq \hat{\theta}_{r,2} - \frac{c_{r,n}(\theta)}{\sqrt{n}} \leq \theta_2 \leq \hat{\theta}_{r,2} + \frac{c_{r,n}(\theta)}{\sqrt{n}} \right\}.$$ 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 $\sqrt{n}(\theta_{0,2} - \theta_{0,1}) = 0$.

Figure 1: Scatter plots of samples illustrating relation between critical values
which is simply the projection of \( Z \) on 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 \( \max\{|z_1|, |z_2|\} \leq c_{r,n}(\theta_0) \), which is strictly inside the dashed square. The third and fourth panel show a similar situation 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 on the solid line decreases and therefore \( c_{r,n}(\theta_0) \) 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 \( \theta_0 \) when \( n = 100 \), conditional on 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 \( \theta_0 \). The dashed red square is \( CI_{ur} \) and the solid

![Figure 2: Confidence regions](image-url)
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 $CI_r \subseteq CI_{ur}$. Also notice that since $c_{r,n}(\theta)$ depends on $\theta$, $CI_r$ 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 both cases, $CI_r \subseteq CI_{ur}$. 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 $\theta_0$ with $\theta_{0,1} < \theta_{0,2}$, it holds that $P(CI_r = CI_{ur}) \to 1$. However, this equivalence does not hold if $\theta_0$ 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 $\theta_0$, 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

$$\hat{\theta}_{r,j}^L = \min_{\theta \in CI_r} \theta_j \quad \text{and} \quad \hat{\theta}_{r,j}^U = \max_{\theta \in CI_r} \theta_j$$

and

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

It then holds by construction that $CI_r \subseteq \overline{CI}_r$ and therefore $P(\theta_0 \in \overline{CI}_r) \geq 0.95$. 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 $\hat{\theta}_{r,2}^U = -\hat{\theta}_{r,1}^L = c_{ur}/\sqrt{n}$ but $\hat{\theta}_{r,2}^U = -\hat{\theta}_{r,2}^L < c_{ur}/\sqrt{n}$, 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 $\Theta_R$. Moreover, for any $\hat{\theta}_{ur}$ and $\hat{\theta}_r$ and $j = 1, 2$ we get $\hat{\theta}_{r,j}^U - \hat{\theta}_{r,j}^L \leq 2c_{ur}/\sqrt{n}$. Thus, the sides of the square $\{\theta \in \mathbb{R}^2 : \hat{\theta}_{r,1}^L \leq \theta_1 \leq \hat{\theta}_{r,1}^U \text{ and } \hat{\theta}_{r,2}^L \leq \theta_2 \leq \hat{\theta}_{r,2}^U \}$ are never longer than the sides of the square $CI_{ur}$. Finally, if $\hat{\theta}_{ur}$ is sufficiently in the interior of $\Theta_R$, then $\overline{CI}_r = CI_{ur}$, which is an important feature of our inference method. We get this equivalence in the interior of $\Theta_R$ 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 $\Theta_R$. 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 $\Theta_R$ 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 in $\theta$ which implies that $\nabla^2 Q_n(\theta)$ does not depend on $\theta$. 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

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

Hence, $\hat{\theta}_r$ is simply the projection of $\hat{\theta}_{ur}$ on $\Theta_R$. Thus, just as before, we can use a change of variables and then characterize the distribution of $\sqrt{n}(\hat{\theta}_r - \theta_0)$ in terms of the distribution of $\sqrt{n}(\hat{\theta}_{ur} - \theta_0)$ and a local parameter space that depends on $\theta_0$ and $n$.

## 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

$$\theta_0 = \left( E(Y \mid X = x_1) \ldots E(Y \mid X = x_{K_n}) \right)' ,$$

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 $g_0(x) \approx p(x)' \theta_0$, where $g_0$ is an unknown function of interest and $p(x)$ is a vector of basis functions. A rectangular confidence region for certain functions of $\theta_0$ can then be interpreted as a uniform confidence band for $g_0$; see Section 4.3 for details. Even though $\theta_0$ and $\Theta$ 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 on 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 $\hat{\theta}_{ur}$ on $\Theta_R$ 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 $\kappa_n$ be a sequence of numbers such that $\kappa_n \to \infty$ as $n \to \infty$. Then

$$\hat{\theta}_r \approx \arg\min_{\theta \in \Theta_R} \|\theta - \hat{\theta}_{ur}\|_\hat{\Omega}^2 = \arg\min_{\theta \in \Theta_R} \|\kappa_n(\theta - \theta_0) - \kappa_n(\hat{\theta}_{ur} - \theta_0)\|_\hat{\Omega}^2 .$$
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
\[ \kappa_n(\hat{\theta}_r - \theta_0) \approx \arg\min_{\lambda \in \Lambda_n(\theta_0)} \| \lambda - \kappa_n(\hat{\theta}_{ur} - \theta_0) \|_\Omega^2. \]

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

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 \( \theta_0 \) can grow in dimensions as \( n \to \infty \) and that our local parameter space \( \Lambda_n(\theta_0) \) depends on \( n \) because we allow \( \theta_0 \) 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
\[ T(\kappa_n(\hat{\theta}_r - \bar{\theta}), \hat{\Sigma}) = \max_{k=1,...,K_n} \left| \frac{\kappa_n(\hat{\theta}_{r,k} - \bar{\theta})}{\sqrt{\hat{\Sigma}_{kk}}} \right|. \]

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

Our \( 1 - \alpha \) confidence set for \( \theta_0 \) 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 \( \hat{\theta}_{ur} \) on \( \Theta_R \), if \( \kappa_n(\hat{\theta}_{ur} - \theta_0) \) was exactly \( N(0, \Sigma) \) distributed, if \( \Sigma \) and \( \Omega \) 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 $\varepsilon_n$ be a sequence of positive numbers with $\varepsilon_n \to 0$. We discuss the role of $\varepsilon_n$ after stating the assumptions. Let $\mathcal{P}$ be a set of distributions satisfying the following assumptions.\(^3\)

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

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

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

**Assumption 2.** There exist symmetric, positive definite matrices $\Omega$ and $\Sigma$ 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

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

and

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

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

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

**Assumption 5.** Let $\Sigma_1$ and $\Sigma_2$ 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 $z_2 \in \mathbb{R}^{K_n}$

$$|T(z_1, \Sigma_1) - T(z_2, \Sigma_1)| \leq C \| z_1 - z_2 \|$$

and

$$|T(z_1, \Sigma_1) - T(z_1, \Sigma_2)| \leq C \| z_1 \| \| \Sigma_1 - \Sigma_2 \|_S.$$

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

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

and

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

---

\(^3\)Even though $\theta_0$ 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 \( \theta \), then Assumption 1 holds with \( R_n = 0 \) and \( \Omega = \nabla^2 Q_n(\theta) \). Andrews (1999) provides general sufficient conditions for a small remainder in a quadratic expansion. The assumption also holds by construction if we simply project \( \hat{\theta}_{ur} \) on \( \Theta_R \) 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 \( \Theta_R \). Even in these cases, \( \hat{\theta}_r \) is usually an approximate projection of an asymptotically normally distributed estimator on \( \Theta_R \).

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 \( \| \hat{\Omega} - \Omega \|_S = o_p(\varepsilon^2_n/K_n) \), 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, or homogeneity of a certain degree, but we rule out Slutski 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 on \( \Lambda_n(\theta_0) \) is nonexpansive, and thus, we could replace it with a higher level assumption. 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 \( \varepsilon_n \). As explained below, to get around this assumption we could 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

---

4See 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).

5I.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 \( C > 0 \).
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 $\alpha$ under the null hypothesis, even if Assumption 6 does not hold. In this case, $\varepsilon_n$ 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 $\alpha$ 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 $\theta_0$. 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 $h(\theta) = \theta$. 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. Our method applies to general functions, such as function values or average derivatives in nonparameteric estimation. In our applications we focus on 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

$$\hat{h}_l^L = \inf_{\theta \in CI} h_l(\theta) \quad \text{and} \quad \hat{h}_l^U = \sup_{\theta \in CI} h_l(\theta), \quad l = 1, \ldots, L_n.$$
Notice that if \( \theta_0 \in CI \), then \( \hat{h}^L_l \leq h_l(\theta_0) \) and \( \hat{h}^U_l \geq h_l(\theta_0) \) for all \( l = 1, \ldots, L_n \). We therefore obtain the following corollary.\(^6\)

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

\[
\liminf_{n \to \infty} \inf_{P \in P} P \left( \hat{h}^L_l \leq h_l(\theta_0) \leq \hat{h}^U_l \text{ for all } l = 1, \ldots, L_n \right) \geq 1 - \alpha.
\]

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 \( \theta_0 \) in the interior of \( \Theta_R \), 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

\[
T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) = \max \left\{ \left| \frac{\kappa_n q_l' (\hat{\theta}_r - \theta)}{\sqrt{q_l' \hat{\Sigma} q_l}} \right|, \ldots, \left| \frac{\kappa_n q_{L_n} (\hat{\theta}_r - \theta)}{\sqrt{q_{L_n}' \hat{\Sigma} q_{L_n}}} \right| \right\}.
\]

Now suppose that for any \( \theta \in CI \), the critical value does not depend on \( \theta \), which will be the case with probability approaching 1 if \( \theta_0 \) is in the interior of the parameter space. That is \( c(\theta, \hat{\Sigma}, \hat{\Omega}) = \hat{c} \). Then

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

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

\[
\hat{h}^L_l \geq h_l(\hat{\theta}_r) - \hat{c} \frac{\sqrt{q_l' \hat{\Sigma} q_l}}{\kappa_n} \quad \text{and} \quad \hat{h}^U_l \leq h_l(\hat{\theta}_r) + \hat{c} \frac{\sqrt{q_l' \hat{\Sigma} q_l}}{\kappa_n}
\]

for all \( l = 1, \ldots, L_n \) and thus,

\[
\hat{h}^L_l \leq h_l(\theta_0) \leq \hat{h}^U_l \text{ for all } l = 1, \ldots, L_n \iff \theta_0 \in CI.
\]

Consequently

\[
P \left( \hat{h}^L_l \leq h_l(\theta_0) \leq \hat{h}^U_l \text{ for all } l = 1, \ldots, L_n \right) = P \left( \theta_0 \in CI \right).
\]

We state a formal result, which guarantees that the projection based confidence set does not suffer from over-coverage if \( \theta_0 \) 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 (2017).

---

\(^6\)Under Assumptions 1 - 5 only, we could project on \( \{ \theta \in \Theta_R : T(\kappa_n(\theta_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n \} \) to obtain the same conclusion as in Corollary 1.
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 \( X = \{x_1, \ldots, x_K\} \), where \( K \) is fixed. Let

\[ \theta_0 = \begin{pmatrix} E(Y \mid X = x_1) \\ \vdots \\ E(Y \mid X = x_K) \end{pmatrix}, \]

and

\[ \hat{\theta}_{ur} = \begin{pmatrix} \frac{\sum_{i=1}^n Y_i 1(X_i = x_1)}{\sum_{i=1}^n 1(X_i = x_1)} \\ \vdots \\ \frac{\sum_{i=1}^n Y_i 1(X_i = x_K)}{\sum_{i=1}^n 1(X_i = x_K)} \end{pmatrix}. \]

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

\[ \Sigma = \text{diag}\left( \frac{\sigma^2(x_1)}{p(x_1)}, \ldots, \frac{\sigma^2(x_K)}{p(x_K)} \right) \]

and

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

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

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

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

\[ \hat{\theta}_r = \arg\min_{\theta \in \Theta_R} \|\theta - \hat{\theta}_{ur}\|_{\hat{\Sigma}^{-1}}^2 \]

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

\[ T(z, \hat{\Sigma}) = \max \left\{ \frac{|z_1|}{\sqrt{\hat{\Sigma}_{11}}}, \ldots, \frac{|z_K|}{\sqrt{\hat{\Sigma}_{KK}}} \right\}. \]
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 $E(U^4 | X = x_k) \leq C$ for all $k = 1, \ldots, K$ and for some $C > 0$.
2. $\Theta_R$ 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 $\theta_0$, 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 $\theta_0$ 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

$$\theta_0 = \left( E(Y | X = x_1) \ldots E(Y | X = x_{K_n}) \right)'.$$

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

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

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

$$\Sigma = \text{diag} \left( \frac{\sigma^2(x_1)B}{f_X(x_1)}, \ldots, \frac{\sigma^2(x_{K_n})B}{f_X(x_{K_n})} \right),$$
and
\[ \hat{\Sigma} = \text{diag} \left( \hat{\sigma}_1^2 B, \ldots, \hat{\sigma}_K^2 B \right), \]
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 \( \Theta_R \) be convex such as \( \Theta_R = \{ \theta \in \mathbb{R}^K : A\theta \leq b \} \) and define
\[ \hat{\theta}_r = \arg \min_{\theta \in \Theta_R} \| \theta - \hat{\theta}_u \|_{\hat{\Sigma}^{-1}}, \]
implying that \( \hat{\Omega} = \hat{\Sigma}^{-1} \). Finally, as before we let
\[ T(z, \hat{\Sigma}) = \max \left\{ \left| z_1 \right| / \sqrt{\hat{\Sigma}_{11}}, \ldots, \left| z_{Kn} \right| / \sqrt{\hat{\Sigma}_{KnKn}} \right\}. \]

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 \( \mathcal{X} = [\underline{x}, \bar{x}] \).
   (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 \( x_{Kn} < \bar{x} - h_n \).
3. \( K(\cdot) \) is a bounded and symmetric pdf with support \([-1, 1]\).
4. \( \Theta_R \) is closed and convex and \( \theta_0 \in \Theta_R \).
5. \( K_n h_n^5 n = o(\varepsilon_n^2) \) and \( K_n^{5/2} / \sqrt{nh_n} = o(\varepsilon_n^3) \).

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 $\theta_0$, 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 $\Theta_R$ 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.\(^7\) The unrestricted and restricted estimators are

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

and

$$
\hat{\theta}_r = \arg \min_{\theta \in \Theta_R} \frac{1}{n} \sum_{i=1}^{n} (Y_i - p_{K_n}(X_i)' \theta)^2,
$$

\(^7\)For example, with quadratic splines $\nabla p_{K_n}(x)' \theta \geq 0$ reduces to finitely many inequality constraints.
respectively. The assumptions ensure that both minimizers are unique with probability approaching 1. Since the objective function is quadratic in \( \theta_0 \) we have
\[
\sqrt{n}(\hat{\theta}_r - \theta_0) = \arg\min_{\lambda \in \Lambda_n(\theta_0)} \| \lambda - \sqrt{n}(\hat{\theta}_{ur} - \theta_0) \|^2, 
\]
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
\[
\hat{\Sigma} = (E(p_{K_n}(X_i)p_{K_n}(X_i)'))^{-1} E(U_i^2 p_{K_n}(X_i)p_{K_n}(X_i)') (E(p_{K_n}(X_i)p_{K_n}(X_i)'))^{-1}.
\]
Also let \( \hat{U}_i = Y_i - p_{K_n}(X_i)'\hat{\theta}_{ur} \) and
\[
\hat{\Sigma} = \hat{\Omega}^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{U}_i^2 p_{K_n}(X_i)p_{K_n}(X_i)') \hat{\Omega}^{-1}. \right.
\]
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 \( \theta_0 \) 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 \( C > 0 \).

3. \( \Theta_R \) is closed and convex and \( \theta_0 \in \Theta_R \) is such that for some constants \( C_g \) and \( \gamma > 0 \)
\[
\sup_{x \in \mathcal{X}} |g_0(x) - p_{K_n}(x)'\theta_0| \leq C_g K_n^{-\gamma}.
\]

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

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
\[
\sup_{P \in \mathcal{P}} \left| P \left( T(\sqrt{n}(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \right) - (1 - \alpha) \right| \to 0.
\]
The first assumption imposes standard moment conditions. The main role of the second assumption is to guarantee that the minimum eigenvalues of \( \Sigma \) and \( \Omega \) 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

\[
n K_n^{-2\gamma} + \frac{\xi(K_n)K_n^2}{n} \to 0.
\]

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 \( \theta_0 \), which we can transform to uniform confidence bands for \( g_0 \). The last rate condition, \( \frac{\xi(K_n)K_n^3}{n} = o(\varepsilon_n^2) \), is not needed under the additional assumptions that \( \text{var}(U_i | X_i) = \sigma^2 > 0 \).

**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 \( \inf_{x \in X} \|p_{K_n}(x)\|^2 > 1/C \) for some constant \( C > 0 \). Then

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

**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 on the underlying confidence set for \( \theta_0 \); see Freyberger and Rai (2017) for this equivalence result. If \( \theta_0 \) 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 on \( \{ \theta \in \Theta_R : T(\sqrt{n}(\hat{\theta}_r - \theta), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n \} \)

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 \( \theta_0 \) would then contain both \( \gamma_0 \) 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 \( \Theta_R \) is a convex subset of \( \mathbb{R}^{K_n} \). Let \( P_X \) be the \( n \times K_n \) matrix, where the \( i \)th row is \( p_{K_n}(X_i)' \) and define \( P_Z \) analogously. Let \( Y \) be the \( n \times 1 \) vector containing \( Y_i \). Let

\[
\hat{\theta}_u r = \arg \min_{\theta \in \mathbb{R}^{K_n}} (Y - P_X \theta)'P_Z(P_Z'P_Z)^{-1}P'_Z(Y - P_X \theta)
\]

and

\[
\hat{\theta}_r = \arg \min_{\theta \in \Theta_R}(Y - P_X \theta)'P_Z(P_Z'P_Z)^{-1}P'_Z(Y - P_X \theta).
\]

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 \( \theta_0 \) we have

\[
\sqrt{n}(\hat{\theta}_r - \theta_0) = \arg \min_{\lambda \in \Lambda_n(\theta_0)} \| \lambda - \sqrt{n}(\hat{\theta}_u r - \theta_0) \|^2_{\hat{\Omega}},
\]

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 \( \hat{U}_i = Y_i - p_{K_n}(X_i)'\hat{\theta}_u r \).

23
As before, $\hat{\sigma}(x) = \sqrt{p_{K_n}(x)'\tilde{\Sigma} p_{K_n}(x)}$ 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 $\theta_0$ 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 $E(U^2 | Z) = \sigma^2 \in [1/C, C]$ and $E(U^4 | Z) \leq C$ 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. $\Theta_R$ 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. $\lambda_{\min}(Q_{XZ}Q_{XZ}'Z_0) \geq \tau_{K_n} > 0$ and $\lambda_{\max}(Q_{XZ}Q_{XZ}') \in [1/C, C]$ for some $C < \infty$.

5. $\frac{n b(K_n)^2}{\tau_{K_n}^4} = o(\varepsilon_n^2)$ 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 $\tau_{K_n}$, 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, $\tau_{K_n}$ 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 of our inference method 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 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 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.

Figure 3: \( g_0 \) for different values of \( c \)
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.\(^8\) 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 2000 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 \( \theta \), but our construction ensures that the estimated critical value is a smooth function of \( \theta \). 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 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 \( \text{width}_{ur}^s \) and \( \text{width}_{sr}^s \) be the average widths in data set \( s \). The last columns shows the median of \( (\text{width}_{ur}^s - \text{width}_{sr}^s)/\text{width}_{ur}^s \) 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

---

\(^8\)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.
Table 1: Coverage and width comparison for regression with polynomials

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

<table>
<thead>
<tr>
<th>$K_n$</th>
<th>$c$</th>
<th>$cov_{ur}$</th>
<th>$cov_r$</th>
<th>$width_{ur}$</th>
<th>$width_r$</th>
<th>% gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.946</td>
<td>0.955</td>
<td>0.059</td>
<td>0.046</td>
<td>0.234</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.929</td>
<td>0.929</td>
<td>0.059</td>
<td>0.058</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.879</td>
<td>0.879</td>
<td>0.060</td>
<td>0.060</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0.608</td>
<td>0.608</td>
<td>0.059</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.229</td>
<td>0.229</td>
<td>0.059</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.003</td>
<td>0.003</td>
<td>0.059</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.000</td>
<td>0.000</td>
<td>0.059</td>
<td>0.059</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>0.821</td>
<td>0.815</td>
<td>0.107</td>
<td>0.101</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.681</td>
<td>0.680</td>
<td>0.107</td>
<td>0.105</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.426</td>
<td>0.426</td>
<td>0.107</td>
<td>0.106</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.201</td>
<td>0.201</td>
<td>0.106</td>
<td>0.106</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>0.942</td>
<td>0.947</td>
<td>0.208</td>
<td>0.194</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.954</td>
<td>0.967</td>
<td>0.208</td>
<td>0.185</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.953</td>
<td>0.959</td>
<td>0.207</td>
<td>0.194</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.952</td>
<td>0.956</td>
<td>0.208</td>
<td>0.199</td>
<td>0.037</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.957</td>
<td>0.978</td>
<td>0.456</td>
<td>0.345</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.973</td>
<td>0.985</td>
<td>0.457</td>
<td>0.288</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.966</td>
<td>0.978</td>
<td>0.462</td>
<td>0.322</td>
<td>0.310</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.953</td>
<td>0.973</td>
<td>0.459</td>
<td>0.345</td>
<td>0.254</td>
</tr>
</tbody>
</table>
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 $c$.

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 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

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

- $c = 0$
- $c = 10$
- $c = 30$
- $c = 50$
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 confidence 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 the unrestricted band does not contain any monotone function.

In addition to some of the information in Table 2, we report the width gains relative to monotonized bands in Table 3. To obtain these bands we simply exclude all parts of the

Figure 5: Example confidence bands for NPIV with polynomials and $K_n = 5$
unrestricted bands which are not consistent with a weakly decreasing function. As we can see from the sixth columns, our restricted bands are considerably smaller than these bands as well, and the widths gains are up to 41.5%. 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.

<table>
<thead>
<tr>
<th>( K_n )</th>
<th>( c )</th>
<th>( \text{cov}_{ur} )</th>
<th>( \text{cov}_r )</th>
<th>% gains</th>
<th>% gains</th>
<th>% empty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.959</td>
<td>0.989</td>
<td>0.731</td>
<td>0.415</td>
<td>0.016</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.962</td>
<td>0.994</td>
<td>0.649</td>
<td>0.394</td>
<td>0.006</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.956</td>
<td>0.994</td>
<td>0.574</td>
<td>0.353</td>
<td>0.002</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.957</td>
<td>0.978</td>
<td>0.471</td>
<td>0.293</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.973</td>
<td>0.985</td>
<td>0.377</td>
<td>0.240</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.966</td>
<td>0.978</td>
<td>0.310</td>
<td>0.196</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>0.953</td>
<td>0.973</td>
<td>0.254</td>
<td>0.165</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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. In our simulations, the median times to solve for the uniform confidence bands in the NPIV setting were roughly 10 minutes when \( K_n = 2 \), 40 minutes when \( K_n = 3 \), 2 hours when \( K_n = 4 \), and 3.5 hours when \( K_n = 5 \). In Section S.3 in the supplement, we provide additional details, such as our selection of starting values. Also notice that these programs are very easy to parallelize because the optimization problems are solved separately for each grid point, which can reduce to computation time to just a few minutes, even when \( K_n = 5 \). Another possibility, which reduces the computational times considerably, is to 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 result in both the simulations and the empirical application; see Section S.3 for more details. Finally, we recently developed the code in Fortran, which runs approximately ten times faster than the MATLAB code in the empirical application below and yields identical results.
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 exclude 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

\[
Y = g_0(X_1, X_2) + X_3' \gamma_0 + U, \quad E(U \mid Z, X_2, X_3) = 0.
\]

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 \( \gamma \) 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. We achieve very
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.
sizable width gains using the shape restrictions when constructing the confidence bands. 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 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. Moreover, the bands are constructed such that they always include the shape restricted estimator of the function and are therefore never empty. 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 \( N(\theta_0, \hat{\Sigma}/\kappa^2_n) \) distributed. To obtain the restricted estimator of \( \theta_0 \) we could imagine maximizing the likelihood with respect to \( \theta_0 \), subject to the solution being in \( \Theta_R \). 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 \( \theta_0 \), which are equivalent to standard rectangular confidence sets if \( \theta_0 \) is in the interior of \( \Theta_R \). 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 \( h_l(\theta) = c_l + q_l^\top \theta \), where \( c_l \) is a constant and \( q_l \in \mathbb{R}^{L_n} \). Let

\[
Z(\hat{\Sigma}) = \left\{ z \in \mathbb{R}^{K_n} : \max \left\{ \left| \frac{q_1^\top z}{\sqrt{q_1^\top \hat{\Sigma} q_1}} \right|, \ldots, \left| \frac{q_{L_n}^\top z}{\sqrt{q_{L_n}^\top \hat{\Sigma} q_{L_n}}} \right| \right\} \leq c(\hat{\Sigma}) \right\},
\]

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

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

\[
T(\kappa_n(\hat{\theta}_r - \theta), \hat{\Sigma}) = \max \left\{ \left| \frac{\kappa_n q_1^\top (\hat{\theta}_r - \theta)}{\sqrt{q_1^\top \hat{\Sigma} q_1}} \right|, \ldots, \left| \frac{\kappa_n q_{L_n}^\top (\hat{\theta}_r - \theta)}{\sqrt{q_{L_n}^\top \hat{\Sigma} q_{L_n}}} \right| \right\}
\]

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, \( A_n z < \kappa_n(b_n - A_n \theta_0) \) for all \( \theta \in CI \) and for all \( z \in Z(\hat{\Sigma}) \). Then for all \( \theta \in CI \), \( c_{1 - \alpha,n}(\theta, \hat{\Sigma}, \hat{\Omega}) = c(\hat{\Sigma}) \) with probability approaching 1 and

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

Notice that \( \kappa_n(b_n - A_n \theta_0) = \kappa_n(b_n - A_n \theta_0) + \kappa_n A_n(\theta_0 - \theta) \). If \( \theta_0 \) 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 \( \theta_0 \) at rate \( \kappa_n \), 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 \( Z(\hat{\Sigma}) \) 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
\[
z_n(\theta, \hat{\Omega}) = \arg \min_{\lambda \in \mathbb{R}^{K_n}} \min_{\lambda \leq \kappa_n(b_n - A_n \theta)} \| \lambda - z \|_{\hat{\Omega}}^2.
\]
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.

\[\square\]

B Useful lemmas

Lemma 1. Let \( Q \) and \( \hat{Q} \) be symmetric and positive definite matrices. Then
\[
\left| \min_{\|v\|=1} v' \hat{Q} v - \min_{\|v\|=1} v' Q v \right| \leq \max_{\|v\|=1} |v'(\hat{Q} - Q)v| \leq \|\hat{Q} - Q\|_S \leq \|\hat{Q} - Q\|
\]
and
\[
\left| \max_{\|v\|=1} v' \hat{Q} v - \max_{\|v\|=1} v' Q v \right| \leq \max_{\|v\|=1} |v'(\hat{Q} - Q)v| \leq \|\hat{Q} - Q\|_S \leq \|\hat{Q} - Q\|.
\]

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
\[
\|Q^{1/2} - \hat{Q}^{1/2}\|_S \leq \frac{1}{\lambda_{\min}(Q^{1/2}) + \lambda_{\min}(\hat{Q}^{1/2})} \|Q - \hat{Q}\|_S
\]
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 \( \lambda^2 \) be the largest eigenvalue of \( (Q^{1/2} - \hat{Q}^{1/2})(Q^{1/2} - \hat{Q}^{1/2}) \) with unit length eigenvector \( v_\lambda \). Since \( (Q^{1/2} - \hat{Q}^{1/2}) \) is symmetric either \( \lambda \) or \( -\lambda \) is an eigenvalue of \( (Q^{1/2} - \hat{Q}^{1/2}) \) with eigenvector \( v_\lambda \). It follows that
\[
\sup_{\|v\|=1} |v'(Q - \hat{Q})v| \geq |v'_\lambda(Q - \hat{Q})v_\lambda| = |v'_\lambda Q^{1/2}(Q^{1/2} - \hat{Q}^{1/2})v_\lambda + v'_\lambda(Q^{1/2} - \hat{Q}^{1/2})\hat{Q}^{1/2}v_\lambda| = |\lambda||v'_\lambda Q^{1/2}v_\lambda + v'_\lambda\hat{Q}^{1/2}v_\lambda| \geq |\lambda| \left( \lambda_{\min}(Q^{1/2}) + \lambda_{\min}(\hat{Q}^{1/2}) \right)
\]

36
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
\[ \|(Q - \hat{Q})v\| = \|Q^{1/2}(Q^{1/2} - \hat{Q}^{1/2})v + (Q^{1/2} - \hat{Q}^{1/2})\hat{Q}^{1/2}v\| \leq \left(\lambda_{\max}(Q^{1/2}) + \lambda_{\max}(\hat{Q}^{1/2})\right)\|Q^{1/2} - \hat{Q}^{1/2}\|_S. \]

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

\[ \square \]

**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 \( \lambda_{\min}(\Sigma) \) is bounded and bounded away from 0 with probability approaches 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 \( P \in \mathcal{P} \)
\[ \frac{\lambda_{\min}(\hat{\Omega})}{\lambda_{\min}(\Omega)} - 1 \xrightarrow{p} 0. \]

Hence \( \lambda_{\min}(\hat{\Omega}) > 0 \) with probability approaching 1 and, uniformly over \( P \in \mathcal{P} \),
\[ \frac{\lambda_{\min}(\Omega)}{\lambda_{\min}(\hat{\Omega})} - 1 \xrightarrow{p} 0. \]

Take \( Z_n \) as defined in Assumption 2 and \( \Lambda_n(\theta_0) = \{\lambda \in \mathbb{R}^K : \lambda = \kappa_n(\theta - \theta_0) \text{ for some } \theta \in \Theta_R\} \) and define
\[ Z_n(\theta_0, \Sigma, \Omega) = \arg\min_{\lambda \in \Lambda_n(\theta_0)} \|\lambda - Z_n\|_\Omega. \]
By Assumptions 5 there exists a constant $C$ such that with probability approaching 1
\[
|T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) - T(Z_n(\theta_0, \Sigma, \Omega), \Sigma)|
\leq |T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) - T(Z_n(\theta_0, \Sigma, \Omega), \hat{\Sigma})| + |T(Z_n(\theta_0, \Sigma, \Omega), \hat{\Sigma}) - T(Z_n(\theta_0, \Sigma, \Omega), \Sigma)|
\leq C \kappa_n(\hat{\theta}_r - \theta_0) - Z_n(\theta_0, \Sigma, \Omega) + C Z_n(\theta_0, \Sigma, \Omega) \|\hat{\Sigma} - \Sigma\|_S
\leq C \kappa_n(\hat{\theta}_r - \theta_0) - Z_n(\theta_0, \Sigma, \hat{\Omega}) + C \kappa_n(\theta_0, \Sigma, \Omega) - Z_n(\theta_0, \Sigma, \hat{\Omega})
\leq + C Z_n(\theta_0, \Sigma, \Omega) \|\hat{\Sigma} - \Sigma\|_S.
\]

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
\[
\|\kappa_n(\hat{\theta}_r - \theta_0) - Z_n(\theta_0, \Sigma, \hat{\Omega})\|
\leq \left\| \arg \min_{\lambda \in \Lambda_n(\theta_0)} \|\lambda - \kappa_n(\hat{\theta}_r - \theta_0)\|^2_{\hat{\Omega}} - Z_n(\theta_0, \Sigma, \hat{\Omega}) \right\| + \|R_n\|
\leq \lambda_{\min}(\hat{\Omega})^{-1/2} \left\| \arg \min_{\lambda \in \Lambda_n(\theta_0)} \|\lambda - \kappa_n(\hat{\theta}_r - \theta_0)\|^2_{\hat{\Omega}} - Z_n(\theta_0, \Sigma, \hat{\Omega}) \right\| + \|R_n\|
\leq \lambda_{\min}(\hat{\Omega})^{-1/2} \|\kappa_n(\hat{\theta}_r - \theta_0) - Z_n\|_{\hat{\Omega}} + \|R_n\|
\leq \sqrt{\frac{\lambda_{\max}(\hat{\Omega})}{\lambda_{\min}(\hat{\Omega})}} \|\kappa_n(\hat{\theta}_r - \theta_0) - Z_n\| + \|R_n\|.
\]

Also notice that $\sqrt{\lambda_{\max}(\hat{\Omega})} = O_p(1)$ and $\left| \frac{\lambda_{\min}(\hat{\Omega})}{\lambda_{\min}(\hat{\Omega})} - 1 \right| = o_p(1)$ 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
\[
\|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\Omega} \leq \|Z_n\|_{\Omega}
\]
and thus,
\[
\sqrt{\lambda_{\min}(\Omega)} \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\| \leq \sqrt{\lambda_{\max}(\Omega)} \|Z_n\|.
\]

It follows that
\[
\|Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n\|_{\hat{\Omega}} \leq \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\Omega}^2
= \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\hat{\Omega}}^2 + \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\hat{\Omega}}^2 - \|Z_n\|_{\hat{\Omega}}^2
\leq \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\hat{\Omega}}^2 + \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\hat{\Omega}}^2 - \|\hat{\Omega} - \Omega\|_S
\leq \|Z_n(\theta_0, \Sigma, \Omega) - Z_n\|_{\hat{\Omega}}^2 + \lambda_{\max}(\Omega) \lambda_{\min}(\Omega) \|Z_n\|_2^2 = \|\hat{\Omega} - \Omega\|_S.
\]
Let
\[ \hat{V}_1 = \frac{\lambda_{\text{max}}(\Omega)}{\lambda_{\text{min}}(\Omega)} \| Z_n \|^2 \| \hat{\Omega} - \Omega \|_S. \]

Analogously, we get
\[ \| Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n \|^2 \leq \| Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n \|^2 + \hat{V}_2, \]
where
\[ \hat{V}_2 = \frac{\lambda_{\text{max}}(\hat{\Omega})}{\lambda_{\text{min}}(\hat{\Omega})} \| Z_n \|^2 \| \hat{\Omega} - \Omega \|_S. \]

Since \( \Lambda_n(\theta_0) \) is convex it follows that for any \( \gamma \in (0, 1) \)
\[ \| Z_n(\theta_0, \Sigma, \Omega) - Z_n \|^2 \leq \| \gamma Z_n(\theta_0, \Sigma, \Omega) + (1 - \gamma) Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n \|^2 \]
\[ = \gamma \| Z_n(\theta_0, \Sigma, \Omega) - Z_n \|^2 + (1 - \gamma) \| Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n \|^2 \]
\[ - \gamma (1 - \gamma) \| Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega) \|^2 \]
\[ \leq \| Z_n(\theta_0, \Sigma, \Omega) - Z_n \|^2 + (1 - \gamma)(\hat{V}_1 + \hat{V}_2) \]
\[ - \lambda_{\text{min}}(\Omega) \gamma (1 - \gamma) \| Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega) \|^2. \]

Therefore,
\[ \| Z_n(\theta_0, \Sigma, \hat{\Omega}) - Z_n(\theta_0, \Sigma, \Omega) \|^2 \leq \frac{1}{\lambda_{\text{min}}(\Omega) \gamma} (\hat{V}_1 + \hat{V}_2) \]
\[ = \frac{1}{\lambda_{\text{min}}(\Omega) \gamma} \left( \frac{\lambda_{\text{max}}(\Omega)}{\lambda_{\text{min}}(\Omega)} + \frac{\lambda_{\text{max}}(\hat{\Omega})}{\lambda_{\text{min}}(\hat{\Omega})} \right) \| Z_n \|^2 \| \hat{\Omega} - \Omega \|_S \]
\[ \leq \frac{\lambda_{\text{max}}(\Sigma)}{\lambda_{\text{min}}(\Omega) \gamma} \left( \frac{\lambda_{\text{max}}(\Omega)}{\lambda_{\text{min}}(\Omega)} + \frac{\lambda_{\text{max}}(\hat{\Omega})}{\lambda_{\text{min}}(\hat{\Omega})} \right) \| \Sigma^{-1/2} Z_n \|^2 \| \hat{\Omega} - \Omega \|_S. \]

Since
\[ \left| \frac{\lambda_{\text{min}}(\Omega)}{\lambda_{\text{min}}(\hat{\Omega})} - 1 \right| \overset{p}{\to} 0, \]
\( \lambda_{\text{max}}(\hat{\Omega}) \) 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
\[ C\|Z_n(\theta_0, \Sigma, \Omega)\|\|\hat{\Sigma} - \Sigma\|_S \leq 2C\|\hat{\Sigma} - \Sigma\|_S \sqrt{\frac{\lambda_{\text{max}}(\Omega)}{\lambda_{\text{min}}(\Omega)}} \|Z_n\| \]
and by Assumption 3 and \(\|Z_n\| = o_p(\sqrt{\lambda_{\text{max}}(\Sigma)K_n})\)
\[ C\|Z_n(\theta_0, \Sigma, \Omega)\|\|\hat{\Sigma} - \Sigma\|_S = o_p(\varepsilon_n) \]
uniformly over \(P \in \mathcal{P}\).

Next define
\[ B_n = C\left\|\kappa_n(\hat{\theta}_r - \theta_0) - Z_n(\theta_0, \Sigma, \hat{\Omega})\right\| + C\left\|Z_n(\theta_0, \Sigma, \Omega) - Z_n(\theta_0, \Sigma, \hat{\Omega})\right\|
+ C\|Z_n(\theta_0, \Sigma, \Omega)\|\|\hat{\Sigma} - \Sigma\|_S, \]
The previous derivations imply that
\[ \sup_{P \in \mathcal{P}} P\left(B_n \geq \frac{1}{2}\varepsilon_n\right) \to 0. \]
Therefore
\[ P\left(T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega})\right) \]
\[ \geq P\left(T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) - B_n\right) - o(1) \]
\[ \geq P\left(T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) - \frac{1}{2}\varepsilon_n, B_n \leq \frac{1}{2}\varepsilon_n\right) - o(1) \]
\[ \geq P\left(T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) - \frac{1}{2}\varepsilon_n\right) - P\left(B_n \geq \frac{1}{2}\varepsilon_n\right) - o(1), \]
where the \(o(1)\) term belongs to \(P(B_l \leq \lambda_{\text{min}}(\hat{\Sigma}) \leq B_u, B_l \leq \lambda_{\text{max}}(\hat{\Omega}) \leq B_u)\) and it converges to 0 uniformly over \(P \in \mathcal{P}\). Similarly, we get
\[ P\left(T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega})\right) \]
\[ \leq P\left(T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \frac{1}{2}\varepsilon_n\right) + P\left(B_n \geq \frac{1}{2}\varepsilon_n\right) + o(1). \]
We next show that for any sufficiently small \(\delta_q \in (0, \alpha)\) it holds that
\[ \sup_{P \in \mathcal{P}} P\left(c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \geq c_{1-\alpha-\delta_q,n}(\theta_0, \Sigma, \Omega) - \frac{1}{2}\varepsilon_n\right) - 1 \to 0 \]
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
\[
\inf_{p \in \mathcal{P}} P \left( T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) + \varepsilon_n \right) 
\geq \inf_{p \in \mathcal{P}} P \left( T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha-\delta,q,n}(\theta_0, \Sigma, \Omega) \right) - o(1)
\]
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 \( \delta_q \) 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 \( \delta_q \) sufficiently small
\[
P \left( T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \right) 
\geq P \left( T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha-\delta,q,n}(\theta_0, \Sigma, \Omega) - \varepsilon_n \right) - o(1)
\]
which implies that
\[
P \left( T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \right) - (1 - \alpha) 
\geq P \left( T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha-\delta,q,n}(\theta_0, \Sigma, \Omega) - \varepsilon_n \right) - (1 - \alpha - \delta_q) - \delta_q - o(1).
\]
Analogously,
\[
P \left( T(\kappa_n(\hat{\theta}_r - \theta_0), \hat{\Sigma}) \leq c_{1-\alpha,n}(\theta_0, \hat{\Sigma}, \hat{\Omega}) \right) - (1 - \alpha) 
\leq P \left( T(Z(\theta_0, \Sigma, \Omega), \Sigma) \leq c_{1-\alpha+\delta,q,n}(\theta_0, \Sigma, \Omega) + \varepsilon_n \right) - (1 - \alpha + \delta_q) + \delta_q + o(1).
\]
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 \( \delta_q \) 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 \( \delta_q \), and define the set \( \mathcal{H}_n \) as all \( (\hat{\Sigma}, \hat{\Omega}) \) on the support of \( (\hat{\Sigma}, \hat{\Omega}) \) such that \( B_l \leq \lambda_{\min}(\hat{\Sigma}), \lambda_{\max}(\hat{\Omega}) \leq B_u \),
\[
\sqrt{K_n} \frac{\sqrt{\lambda_{\max}(\Sigma)}}{\sqrt{\lambda_{\min}(\Omega)}} \| \hat{\Sigma} - \Sigma \|_S \leq \delta_\varepsilon \varepsilon_n
\]
and
\[ K_n \lambda_{\max}(\Sigma) \left( \frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\tilde{\Omega})}{\lambda_{\min}(\Omega)} \right) \|\tilde{\Omega} - \Omega\|_S \leq \delta^2 \varepsilon_n^2. \]

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
\[ \tilde{Z}_n(\theta_0, \Sigma, \Omega) = \arg\min_{\lambda \in \Lambda_n(\theta_0)} \|\lambda - \Sigma^{1/2} \tilde{Z}_n\|^2. \]

For any \((\Sigma^*, \Omega^*) \in \mathcal{H}_n\) we get by Assumption 5 that
\[
\begin{align*}
&\left| T(\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*), \Sigma^*) - T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma) \right|
\leq \left| T(\tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*), \Sigma^*) - T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma^*) \right| \\
&+ \left| T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma^*) - T(\tilde{Z}_n(\theta_0, \Sigma, \Omega), \Sigma) \right|
\leq C \left\| \tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*) - \tilde{Z}_n(\theta_0, \Sigma, \Omega^*) \right\| \\
&+ C \left\| \tilde{Z}_n(\theta_0, \Sigma, \Omega) - \tilde{Z}_n(\theta_0, \Sigma, \Omega^*) \right\| + C \|\tilde{Z}_n(\theta_0, \Sigma, \Omega)\| \|\Sigma - \Sigma^*\|_S.
\end{align*}
\]

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

where the last line follows from Lemma 2. Also by the previous results
\[
\begin{align*}
\|\tilde{Z}_n(\theta_0, \Sigma, \Omega^*) - \tilde{Z}_n(\theta_0, \Sigma, \Omega)\|^2 &\leq \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)^2} \left( \frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\Omega^*)}{\lambda_{\min}(\Omega^*)} \right) \|\bar{\tilde{Z}}_n\|^2 \|\Omega^* - \Omega\|_S.
\end{align*}
\]

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

Let
\[
H_n = C \left\| \tilde{Z}_n(\theta_0, \Sigma^*, \Omega^*) - Z_n(\theta_0, \Sigma, \Omega^*) \right\| + C \|Z_n(\theta_0, \Sigma, \Omega) - Z_n(\theta_0, \Sigma, \Omega^*)\| \\
+ C \|Z_n(\theta_0, \Sigma, \Omega)\| \|\Sigma - \Sigma^*\|_S.
\]

42
Then, for constants $M_1$ and $M_2$ that do not depend on $P$ or $\delta_\varepsilon$

$$H_n^2 \leq 4C^2 \left\| Z_n(\theta_0, \Sigma^*, \Omega^*) - Z_n(\theta_0, \Sigma, \Omega) \right\|^2 + 4C^2 \left\| Z_n(\theta_0, \Sigma, \Omega) \right\|^2 \left\| \Sigma - \Sigma^* \right\|^2_S$$

$$+ 4C^2 \left\| Z_n(\theta_0, \Sigma, \Omega) - Z_n(\theta_0, \Sigma, \Omega^*) \right\|^2$$

$$\leq M_1 \left\| \Sigma - \Sigma^* \right\|^2 \left\| \bar{Z}_n \right\|^2 + M_1 \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)} \left\| \bar{Z}_n \right\|^2 \left\| \Sigma - \Sigma^* \right\|^2_S$$

$$+ M_1 \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Omega)} \left( \frac{\lambda_{\max}(\Omega)}{\lambda_{\min}(\Omega)} + \frac{\lambda_{\max}(\Omega^*)}{\lambda_{\min}(\Omega^*)} \right) \left\| \bar{Z}_n \right\|^2 \left\| \Omega^* - \Omega \right\|_S$$

$$\leq M_2 \varepsilon_n^2 \delta_\varepsilon^2 \frac{\left\| \bar{Z}_n \right\|^2}{K_n}.$$ 

Since $E(\left\| \bar{Z}_n \right\|^2) = K_n$ 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

$$1 - \alpha = P \left( T(\bar{Z}_n(\theta_0, \Sigma^*, \Omega^*)) \leq c_{1,\alpha,n}(\theta_0, \Sigma^*, \Omega^*) \right)$$

$$\leq P \left( T(\bar{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1,\alpha,n}(\theta_0, \Sigma^*, \Omega^*) + H_n \right)$$

$$\leq P \left( T(\bar{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1,\alpha,n}(\theta_0, \Sigma^*, \Omega^*) + \frac{1}{2}\varepsilon_n \right) + 4M_2\delta_\varepsilon^2.$$ 

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

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

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

$$1 - \alpha = P \left( T(\bar{Z}_n(\theta_0, \Sigma^*, \Omega^*)) \leq c_{1,\alpha,n}(\theta_0, \Sigma^*, \Omega^*) \right)$$

$$\geq P \left( T(\bar{Z}_n(\theta_0, \Sigma, \Omega)) \leq c_{1,\alpha,n}(\theta_0, \Sigma^*, \Omega^*) - H_n \right)$$

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

43
It follows that we can pick \( \delta \) 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


