# Essays on the econometric theory of rank regressions

Subbotin, Viktor

Northwestern University

December 2008

NORTHWESTERN UNIVERSITY

# Essays on the Econometric Theory of Rank Regressions

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Viktor Yevgenyevich Subbotin

EVANSTON, ILLINOIS

December 2008

# ABSTRACT

Essays on the Econometric Theory of Rank Regressions

Viktor Yevgenyevich Subbotin

Several semiparametric estimators recently developed in the econometrics literature are based on the rank correlation between the dependent and explanatory variables. Examples include the maximum rank correlation estimator (MRC) of Han [**1987**], the monotone rank estimator (MR) of Cavanagh and Sherman [**1998**], the pairwise-difference rank estimators (PDR) of Abrevaya [**2003**], and others. These estimators apply to various monotone semiparametric single-index models, such as the binary choice models, the censored regression models, the nonlinear regression models, and the transformation and duration models, among others, without imposing functional form restrictions on the unknown functions and distributions. This work provides several new results on the theory of rank-based estimators. In Chapter 2 we prove that the quantiles and the variances of their asymptotic distributions can be consistently estimated by the nonparametric bootstrap. In Chapter 3 we investigate the accuracy of inference based on the asymptotic normal and bootstrap approximations, and provide bounds on the associated error. In the case of MRC and MR, the bound is a function of the sample size of order close to $n^{-1/6}$. The

PDR estimators, however, belong to a special subclass of rank estimators for which the bound is vanishing with the rate close to $n^{-1/2}$. In Chapter 4 we study the efficiency properties of rank estimators and propose weighted rank estimators that improve efficiency. We show that the optimally weighted MR attains the semiparametric efficiency bound in the nonlinear regression model and the binary choice model. Optimally weighted MRC has the asymptotic variance close to the semiparametric efficiency bound in single-index models under independence when the distribution of the errors is close to normal, and is consistent under practically relevant deviations from the single index assumption. Under moderate nonlinearities and nonsmoothness in the data, the efficiency gains from weighting are likely to be small for MRC in the transformation model and for MRC and MR in the binary choice model, and can be large for MRC and MR in the monotone regression model. Throughout, the theoretical results are illustrated with Monte-Carlo experiments and real data examples.

## Acknowledgements

I am grateful to my advisors, Professor Joel L. Horowitz, Professor Rosa L. Matzkin and Professor Elie T. Tamer, for their support, guidance and constructive criticism.

# Contents

# List of Tables

# List of Figures

CHAPTER 1

# Introduction

Several semiparametric estimators recently developed in the econometrics literature are based on the rank correlation between the dependent and explanatory variables. The first was introduced by Han [**1987**], and is called the maximum rank correlation estimator (MRC). It applies to models in which the joint distribution of the data, $(Y, X)$, satisfies the condition: for two independent realizations, $(Y_1, X_1)$ and $(Y_2, X_2)$,

$$P\{Y_1 > Y_2 | X_1, X_1\} > P\{Y_1 > Y_2 | X_1, X_1\}, \tag{1.1}$$
$$\Longrightarrow X_1'\beta_0 > X_2'\beta_0,$$

for a vector of unknown coefficients $\beta_0$. This condition holds, in particular, in the generalized regression model of Han [**1987**], in which the outcome $Y$ and the vector of covariates $X$ are related according to the equation:

$$Y = D \circ F(X'\beta_0, \varepsilon), \tag{1.2}$$

for independent $X$ and $\varepsilon$, a nondecreasing function $D$, and a function $F$ which is strictly increasing in both arguments. This model itself nests such important estimation models as the binary choice models, the ordered discrete response models, the censored regression models, the transformation models, proportional and additive hazard models, and nonlinear regression models under the independence assumption and monotonicity constraints.

Relations (1.1) and (1.2) identify the vector $\beta_0$, up to scale, even if the other elements of the models (e.g. the functions $D$, $F$, and the distribution of the error term $\varepsilon$) are not specified. If a sample $\{X_i, Y_i\}$, $i = 1, ..., n$, of i.i.d. observations is available, $\beta_0$ can be estimated (up to scale) by the MRC estimator, a vector $\beta_n$ that maximizes the criterion function

$$\sum_{i \neq j} \mathbf{1}\{Y_i > Y_j\}\mathbf{1}\{X_i'\beta > X_j'\beta\}. \tag{1.3}$$

Cavanagh and Sherman [**1998**] considered the monotone regression model:

$$Y = f\left(X'\beta_0\right) + \varepsilon, \tag{1.4}$$

where $f$ is a nondecreasing function, and the error term satisfies the conditions:

$$\begin{aligned} E\left[\varepsilon|X\right] &= 0, \\ E\left[\varepsilon^2|X\right] &\equiv \sigma_0^2\left(X\right) < \infty. \end{aligned} \tag{1.5}$$

They have shown that $\beta_0$ can be consistently estimated, up to scale, by the monotone rank estimator (MR) which maximizes the criterion function

$$\sum_{i \neq j} Y_i \mathbf{1}\{X_i'\beta > X_j'\beta\}. \tag{1.6}$$

They also suggested an alternative estimator for $\beta_0$ in the generalized regression model, a solution of the maximization problem with the objective function

$$\sum_{i,j,k \text{ distinct}} \mathbf{1}\{Y_i > Y_j\}\, \mathbf{1}\{X_i'\beta > X_k'\beta\}. \tag{1.7}$$

Abrevaya [**2003**] considered a special case of the generalized regression model, the transformation model:

$$h(Y) = X'\beta_0 + \varepsilon, \tag{1.8}$$

where $h$ is a strictly increasing, unknown function. He proposed two pairwise-difference rank estimators of $\beta$. The PDR3 estimator maximizes the objective function

$$\sum_{i,j,k \text{ distinct}} \left(\mathbf{1}\{Y_i > Y_j\} - \mathbf{1}\{Y_j > Y_k\}\right) \mathbf{1}\left\{(X_i - X_j)'\beta > (X_j - X_k)'\beta\right\},$$

and the PDR4 estimator maximizes the objective function

$$\sum_{i,j,k,l \text{ distinct}} \left(\mathbf{1}\{Y_i > Y_j\} - \mathbf{1}\{Y_k > Y_l\}\right) \mathbf{1}\left\{(X_i - X_j)'\beta > (X_k - X_l)'\beta\right\}.$$

Estimators with a similar structure have been proposed for $\beta$ in the transformation model with observed or unobserved truncation (Abrevaya [**1999b**], Khan and Tamer [**2007**]), and in the binary response model with panel data (Lee [**1999**]), as well as for the link function $h$ in the transformation model (Chen [**2002**], Han [**1987b**], and Asparouhova et al [**2002**]), among others.

Rank correlation estimators have several advantages. First, they are root-$n$-consistent and asymptotically normal (Sherman [**1993**], Arcones, Chen, Giné [**1994**]). Second, they

do not require a choice of any tuning parameters (bandwidths, trimming parameters, etc.), unlike any other presently known asymptotically normal semiparametric estimators (such as the average derivative method of Powell, Stock and Stoker [**1989**], the semiparametric least-squares estimator of Ichimura [**1993**], the sieve minimum-distance estimator of Ai and Chen [**2003**] or the semiparametric maximum likelihood method for binary response models of Klein and Spady [**1993**]). This property is useful for empirical work, as choosing bandwidths or other tuning parameters is not always easy in practice. Third, rank estimators can be applied to models with heavy-tailed distributions of the error term, when other semiparametric estimators of $\varepsilon$ may not be consistent.

In this work we provide new results concerning the properties of rank-based estimators. In Chapter 2 we prove that the standard errors and the confidence intervals for such estimators can be consistently estimated by the nonparametric bootstrap. This result is important as it allows to keep the bandwidth-free nature of the rank regression methodology not only in estimation, but also in inference. In Chapter 3 we characterize the accuracy of inference based on either the asymptotic normal approximations, or the bootstrap approximations of the finite-sample distributions of these estimators. In particular, we show that the estimators like PDR3 and PDR4 can be substantially more accurate than the estimators like MRC or MR. In Chapter 4 we consider the problem of efficiency of rank estimators. We show that under commonly made assumptions one can construct estimators with lower asymptotic variances by introducing weights into the criterion functions defining the rank estimators. We also evaluate the resulting efficiency gains in numerical examples, and compare the lowest variances achievable by this method

with the semiparametric efficiency bounds for single-index models. Finally, we provide a real data application of these techniques.

The statistical theory of rank-based estimators relies on the empirical process theory for $U$-processes. In Appendix, Section A.3, we review the currently known results from this theory, and provide extensions that are necessary for our work and that may be useful for other applications.

CHAPTER 2

# Consistency of the Bootstrap for Rank Regressions

## 2.1. Introduction

This chapter is concerned with inference about the finite dimensional parameter estimated by a rank-based estimator. Under appropriate regularity conditions, such estimators are root-$n$-consistent and asymptotically normal. Therefore, the test statistics, critical values, and confidence intervals for the estimated parameters can be constructed in the usual way based on the limiting normal distributions. In the case of rank estimators, however, the asymptotic variances depend on moments of random variables that are not directly observed (the first and second-order derivatives of certain conditional expectations), and special procedures are needed for their estimation. Two methods that are available at present are the numerical derivative method of Pakes and Pollard [**1989**], and the nonparametric method of Sherman [**1993**] and Cavanagh and Sherman [**1998**]. However, both have drawbacks. First, they depend on tuning parameters (step sizes for numerical differentiation or bandwidths for kernel regressions). No objective, data-driven mechanism has been developed to set these parameters in practical applications. The numerical derivative method involves a finite-difference approximation of the second-order derivatives and often produces unstable results. The nonparametric method, which avoids the direct estimation of the second-order derivatives, requires additional programming effort, as the analytical expressions for the variances that it uses are specific for

each particular estimator, and are sometimes complicated (as in the case of PDR3, for example). Finally, both methods can be numerically intensive in large samples, with the computational burden rising with the sample size as $O(n^2)$ for MRC and MR, and as $O(n^4)$ for PDR4[1].

Alternatively, the asymptotic distribution can be estimated by resampling methods, particularly, the nonparametric bootstrap of Efron [**1979**]. This approach is free of tuning parameters, and is straightforward to implement. Unlike in most other econometric settings, the bootstrap of rank correlation estimators can be less computationally demanding than a direct variance estimation, due to availability of fast algorithms for evaluating their objective functions. For example, one evaluation of the objective function can be reduced to $O(n \log n)$ operations for MR (Cavanagh and Sherman [**1998**]), and MRC (Abrevaya [**1999b**]), and to $O(n^2 \log n)$ operations for both PDR (Abrevaya [**2003**]). The same efficient algorithms can be used in the nonparametric bootstrap, making it feasible and possibly more attractive computationally than other alternatives.

The results on the consistency of the bootstrap exposed in this chapter are obtained for a general class of maximizers of a criterion function in the form of a $U$-process, of which the rank estimators are particular examples. Prior to our work, it has not been known if the nonparametric bootstrap consistently estimates the asymptotic distribution of such estimators (the fact that an estimator is root-$n$-consistent and asymptotically normal does not guarantee consistency of the bootstrap, see Abadie and Imbens (2006) for a counterexample). The regularity conditions that we require for the bootstrap are,

[1]These and the following estimates of the computational complexity assume that the full sample is used for inference. When $n$ is large, inference can be performed, at the expense of lower precision, using a randomly chosen subsample of data.

up to a minor qualification, the same as the assumptions of Sherman [**1993**] and Arcones, Chen, Giné [**1994**] for the asymptotic normality.

The chapter is organized as follows. Section 2.2 presents the asymptotic and bootstrap theory of rank estimators. In Section 2.3 we apply the bootstrap in a real data example. Section 2.4 concludes. Proofs of all theoretical results are given in Appendix A.

## 2.2. Asymptotic Theory

We first define a class of estimators that includes all rank estimators listed in Chapter 1. The following notation is used: $\mathcal{Z}$ is a vector space, and $P$ is a probability measure on $\mathcal{Z}$;

$$\mathcal{H} = \{h_\theta(z_1, ..., z_m) : \theta \in \Theta \subset \mathbb{R}^d\}$$

is a family of real-valued functions defined on $\mathcal{Z}^m = \mathcal{Z} \times ... \times \mathcal{Z}$ ($m \geq 2$ times), indexed by a vector of parameters $\theta$. It will be a matter of notational convenience to assume that the functions $h$ are symmetric in their $z$ arguments:

$$h_\theta(z_1, ..., z, ..., z', ..., z_m) = h_\theta(z_1, ..., z', ..., z, ..., z_m).$$

Write $P^{m-k}h$, $k = 0, ..., m$, for the partial integral, relative to $P$, over the last $m - k$ arguments of $h$:

$$\left(P^{m-k}h\right)(z_1, ..., z_k) = \int h\left(z_1, ..., z_k, Z_{k+1}, ..., Z_m\right) dP\left(Z_{k+1}\right) ... dP\left(Z_m\right)$$

(in particular, $P^0 h = h$).

Assume that the parameter of interest, $\theta_0$, is a global maximum on $\Theta$ of the expected value of $h_\theta$, $P^m h_\theta$. Given an i.i.d. sample of observations, $\{Z_1, ..., Z_n\}$, from the space

$(\mathcal{Z}, P)$, one can construct a sample analog of $P^m h_\theta$, a $U$-process of order $m$ indexed by $\theta$:

$$G_{n,\theta} = U_n^{(m)} h_\theta \equiv \frac{(n-m)!}{n!} \sum_{i_1,...,i_m,\text{ distinct}} h_\theta(Z_{i_1}, ..., Z_m) \tag{2.1}$$

(a $U$-process considered for a specific $\theta$ is called a $U$-statistic. See e.g. Serfling [**1980**] on the basic properties of $U$-statistics).

The parameter $\theta_0$ can be estimated by an approximate solution of the sample analog of the population problem:

$$G_{n,\theta_n} \geq \sup_{\theta \in \Theta} \left[ G_{n,\theta} - r_{n,\theta} \right], \tag{2.2}$$

where the remainder term $r_{n,\theta}$ is introduced to ensure measurability of $\theta_n$ as in Pakes and Pollard [**1989**] and may also represent the terms that do not have the structure studied below (e.g. the numerical error of solving the maximization problem)[2].

Under general conditions, $\theta_n$ is root-$n$-consistent for $\theta_0$ and asymptotically normal. Namely, let the following assumptions hold.

**Assumption 1.** *$\Theta$ is a compact set; $P^m h_\theta$, $m \geq 2$, is continuous on $\Theta$ and $\theta_0$ is its unique global maximum on $\Theta$.*

**Assumption 2.** *$\mathcal{H}$ is a Euclidean class[3] of symmetric functions for a $P^m$-square-integrable envelope $H$ ($H$ is called an envelope for the class $\mathcal{H}$ if $|h| \leq H$ for each $h \in \mathcal{H}$).*

**Assumption 3.** *Define $\tau_\theta(z) = \left(P^{m-1} h_\theta\right)(z)$. There is an open neighborhood $\mathcal{N} \subset \Theta$ of $\theta_0$ such that*

[2]Below we describe admissible orders of magnitude of $r_{n,\theta}$.

[3]See Appendix, Section A.3, for the definition and basic properties of Euclidean classes.

*(i) All mixed partial derivatives of $\tau_\theta(z)$ with respect to $\theta$ of orders 1 and 2 exist on $\mathcal{N}$.*

*(ii) There is a $P$-integrable function $M(z)$ such that for all $z$ and all $\theta$ in $\mathcal{N}$,*

$$\left\|\partial^2\tau_\theta(z) - \partial^2\tau_{\theta_0}(z)\right\| \leq M(z)\left\|\theta - \theta_0\right\|,$$

*where $\partial^2\tau_\theta$ is the Hessian matrix of $\tau$ with respect to $\theta$, and $\|\cdot\|$ denotes the Euclidean norm.*

*(iii) The gradient of $\tau_\theta(z)$ with respect to $\theta$ at $\theta_0$, $\partial\tau_{\theta_0}(z)$, has finite variance relative to $P$.*

*(iv) The matrix $A = -P\left[\partial^2\tau_{\theta_0}\right]$ is finite and positive definite.*

**Assumption 4.** *As $\theta \to \theta_0$, $P^2\left[\left(P^{m-2}h_\theta - P^{m-2}h_{\theta_0}\right)^2\right] \to 0$.*

These assumptions are a stylized version of assumptions of Sherman [**1993**]. Assumption 1 is standard for identification. Assumption 2 says that the class of functions over which maximization is performed is not too large, which is necessary for consistency. Assumptions 3 and 4 repeat the continuity and smoothness conditions of Sherman for asymptotic normality.

For example, in the case of MRC, let $\beta = (\theta, 1)' \in \mathbb{R}^{d+1}$ (to fix the scale, the last component of $\beta$ is set to 1). The function $h_\theta$ is a symmetric version of the kernel in (1.3) (note that symmetrization does not change the optimization problem):

$$\begin{aligned} h_\theta(z_1, z_2) \;=\; & \\ & \mathbf{1}\{y_1 > y_2\}\mathbf{1}\{(\theta', 1)(x_1 - x_2) > 0\} \\ + & \mathbf{1}\{y_2 > y_1\}\mathbf{1}\{(\theta', 1)(x_2 - x_1) > 0\}, \end{aligned} \tag{2.3}$$

where $z = (y, x)$. Han [**1987**] provided primitive conditions under which $h_\theta$ satisfies Assumption 1. Sherman [**1993**] verified that for a compact $\Theta$, the class of functions $\{h_\theta(z_1, z_2)\}$ is Euclidean for the envelope $H = 1$, and gave conditions on the primitives of the model (1.2) under which Assumptions 3 and 4 are satisfied[4]. In particular, Assumption 4 is satisfied if the last component of vector $X$, denoted $V$, is continuously distributed conditionally on the vector of the first $d$ components, $U$. Also, the following condition is sufficient for parts (i)-(iii) of Assumption 3: $V$ is continuously distributed conditionally on $U$ and $Y$; the conditional density, $\phi_{V|U,Y}$, is three times differentiable in $V$ for almost all $U$ and $Y$, and is uniformly bounded together with its derivatives up to order three; and $P\|U\|^3 < \infty$[5]. Below, we will refer to these, or similar, sufficient conditions repeatedly. Assumptions 1-4 were verified for the other rank estimators in the corresponding papers listed in Chapter 1. It is worth noting, however, that Assumptions 1-4 do not rely on the specific structure of rank estimators, but rather on the fact that they maximize a $U$-process with sufficiently smooth leading terms. The applicability of our theoretical results, therefore, extends beyond the scope of rank estimators.

Theorem 1, which is essentially due to Sherman [**1993**] and Arcones, Chen, Giné [**1994**], says that the estimator $\theta_n$, after a proper normalization and recentering, converges in distribution, uniformly, to a normal law.

**Theorem 1.** *Let Assumptions 1-4 hold, and* $\sup_{\theta\in\Theta}|r_{n,\theta}| = o_p(n^{-1})$. *Define* $\Gamma = m^2 A^{-1} Var(\partial\tau_{\theta_0}) A^{-1}$. *Then* $\theta_n$ *is consistent for* $\theta_0$ *in probability, and*

$$\sup_{A\in\mathcal{A}}\left|\int_A dF_{n^{1/2}(\theta_n-\theta_0)} - \int_A d\Phi_\Gamma\right| = o(1), \tag{2.4}$$

[4]With the exception of Assumption 3 (iv), see Chapter 4 for the discussion.
[5]Sherman assumed that $U$ has bounded support, but $P\|U\|^3 < \infty$ suffices.

*where* $F_{n^{1/2}(\theta_n-\theta_0)}$ *is the c.d.f. of the random vector* $n^{1/2}(\theta_n-\theta_0)$, $\Phi_\Gamma$ *is the c.d.f. of the normal distribution with mean zero and variance* $\Gamma$*, and* $\mathcal{A}$ *is the collection of all measurable convex sets in* $\mathbb{R}^d$.

To use the result of Theorem 1 for inference, one needs an estimate of the asymptotic variance $\Gamma$. The latter, however, depends on moments of the derivatives of the unknown function $\tau$. As explained in Introduction, estimation of these moments may be difficult in practice.

Alternatively, the limiting distribution can be estimated by the nonparametric bootstrap of Efron [**1979**]. Specifically, let $\left\{\hat{Z}_1, ..., \hat{Z}_n\right\}$ be the bootstrap sample, i.e. a collection of independent draws, with replacement, from the sample $\{Z_1, ..., Z_n\}$. The bootstrapped objective function $\hat{U}_n^{(m)} h_\theta$ is formed as in (2.1) using $\hat{Z}_i$ instead of $Z_i$:

$$\hat{G}_{n,\theta} = \hat{U}_n^{(m)} h_\theta \equiv \frac{(n-m)!}{n!} \sum_{i_1,...,i_m, \text{ distinct}} h_\theta(\hat{Z}_{i_1}, ..., \hat{Z}_m). \tag{2.5}$$

The bootstrapped estimator, $\hat{\theta}_n$, is an approximate solution to the corresponding maximization problem:

$$\hat{G}_{n,\hat{\theta}_n} \geq \sup_{\theta \in \Theta} \left[\hat{G}_{n,\theta} - \hat{r}_{n,\theta}\right], \tag{2.6}$$

with some remainder $\hat{r}_{n,\theta}$.

To prove consistency of the bootstrap, we make one more assumption. It arises because the bootstrap draws, unlike the sample observations, are statistically dependent unconditionally. Note that Assumptions 1-4 provide no bounds on moments of function

$h$ if its arguments are statistically dependent. The form of dependency that needs to be explicitly controlled in the bootstrap is that of drawing the same sample realization of vector $Z$ two or more times. To state the assumption formally, define the function

$$H_{\omega_m}(z_1, ..., z_m) = H\left(z_{\omega_m(1)}, ..., z_{\omega_m(m)}\right),$$

where $\omega_m$ is a permutation, with repetition, of numbers $\{1, ..., m\}$, and the function

$$h_\theta^{[m-2]}(z_1, ..., z_{m-2}) = \int h_\theta(z_1, ..., z_{m-2}, Z_m, Z_m)\, dP(Z_m).$$

**Assumption 5.** *(a) For all* $\omega_m$, $P^m H_{\omega_m}^2 < \infty$.
*(b) As* $\theta \to \theta_0$, $P^{m-2}\left[h_\theta^{[m-2]} - h_{\theta_0}^{[m-2]}\right] \to 0$.

Assumption 5 is not restrictive for rank estimators. The moment condition on $H_{\omega_m}$ is trivially satisfied for bounded functions $h$ (which is the case for the majority of rank estimators). MR is an example when $h$ may be unbounded, however, the condition $PY^2 < \infty$, required by Assumption 2, also entails the moment condition in Assumption 5 (a) for the envelope $H = |Y_1| + |Y_2|$. The continuity condition on $h_\theta^{[m-2]}$ is also not difficult to verify. For MR and MRC, for example, it is satisfied vacuously, because in this case $h_\theta^{[m-2]} \equiv 0$. For other estimators, e.g. pairwise-difference rank estimators, $h_\theta^{[m-2]} \neq 0$. However, Assumption 5, similarly to Assumption 4, holds if the last component of the vector of regressors, $V$, is distributed continuously conditionally on the first $d$ components, $U$.

We now give two results showing consistency of the bootstrap. The distribution of the test statistic, $n^{1/2}(\theta_n - \theta_0)$, can be approximated by the conditional (on the data sample)

distribution of the bootstrapped statistic, $n^{1/2}(\widehat{\theta}_n - \theta_n)$, or by the normal c.d.f. with zero mean and variance equal to the conditional variance of $n^{1/2}(\widehat{\theta}_n - \theta_n)$. Both approaches give consistent results, although the second relies on slightly stronger regularity conditions.

**Theorem 2.** *Let the assumptions of Theorem 1 and Assumption 5 hold, and assume that* $\sup_\theta |\hat{r}_{n,\theta}| = o_p(n^{-1})$. *Then the bootstrap estimator of the asymptotic distribution of* $n^{1/2}(\theta_n - \theta_0)$ *is consistent in probability:*

$$\sup_{A \in \mathcal{A}} \left| \int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n - \theta_n\right)} - \int_A d\Phi_\Gamma \right| = o_p(1), \tag{2.7}$$

*where* $\hat{F}$ *is the conditional c.d.f. of the bootstrapped estimator.*

**Theorem 3.** *Let the assumptions of Theorem 1 hold, and, additionally,* $P^m H^p < \infty$, $PM^p < \infty$, $P\|\partial^2 \tau_{\theta_0}\|^p < \infty$, *for a* $p > 2$, *and* $P\sup_\theta |r_{n,\theta}|^2 = o(n^{-1})$. *Then*

$$Var\left[n^{1/2}(\theta_n - \theta_0)\right] \to \Gamma.$$

*If also* $P^m H^p_{\omega_m} < \infty$ *for each* $\omega_m$, $P\sup_\theta |\hat{r}_{n,\theta}|^2 = o(n^{-1})$, *and Assumption 5 (b) holds, then the bootstrap estimator of the asymptotic variance of* $n^{1/2}(\theta_n - \theta_0)$ *is consistent in probability:*

$$\widehat{Var}\left[n^{1/2}(\widehat{\theta}_n - \theta_n)\right] \to^p \Gamma.$$

*Here* $Var$ *is the finite sample variance and* $\widehat{Var}$ *is the bootstrap variance conditional on the sample.*

### 2.3. Empirical Example

Here we apply the bootstrap to a real data set. Our main focus is the feasibility of the bootstrap procedure. We reestimate the standard errors in the wage-equation example studied by Abrevaya [**2003**]. The data set, constructed by Ruud [**2000**], is an extract from the March 1995 CPS, consisting of 1,289 observations. The dependent variable is an hourly wage (WAGE). The regressors are years of schooling (EDUC), years of potential work experience (EXPER) and its square (EXPSQ), a female indicator variable (FEMALE), a union indicator variable (UNION), and a nonwhite indicator variable (RACE equal to 0 if white, 1 if not). The model is specified as the transformation model with an unknown link function $h$:

$$\begin{aligned} h\left(WAGE\right) &= \beta_1 EDUC + \beta_2 EXPER + \beta_3 EXPSQ \\ &\quad + \beta_4 FEMALE + \beta_5 UNION + \beta_6 RACE + \varepsilon \end{aligned}$$

(the traditional choice of $h$ in such models is the logarithmic function). The identification assumption is that $h$ is a strictly increasing function and $\varepsilon$ is an i.i.d. error term distributed independently of the covariates.

The coefficients are estimated by MRC and PDR4, with a scale normalization $\beta_1 = 1$. Abrevaya [**2003**] computed the estimates of the coefficients. He also applied the nonparametric method to estimate their standard errors. It is worth noting that the number of (one-dimensional) kernel regressions that the nonparametric method involves increases with the dimension of the vector of covariates, $d+1$, as $3+d$ for MRC and as $4+\frac{d(d+1)}{2}$ for PDR (so that in the example considered here one has to run, respectively, eight and

nineteen one-dimensional kernel regressions). As the implementation of each of them requires a choice of a bandwidth and some other details (such as the form of the kernel and the trimming parameters in the denominator of the Nadaraya-Watson conditional expectation estimator), the method contains an element of subjectivity.

Here we provide alternative estimates of the standard errors obtained by the bootstrap. The computational burden of the bootstrap is of order $O(Bn \log n)$ for MRC and $O(Bn^2 \log n)$ for PDR4, where $B$ is the number of bootstrap iterations. We used $B = 1000$, although the estimates of the standard errors were stable after $B = 200$ iterations already. For this sample size, the computation of the MRC objective function is faster than that of the PDR4 objective function, but the associated maximization problem for the former is more difficult to solve numerically. In the case of MRC we used the Nelder-Mead algorithm of optimization with five different combinations of parameters and starting values (chosen in trial runs) in each bootstrap iteration. For PDR4 we used the standard MATLAB maximization routine fminsearch with default settings and one starting vector, the estimated vector of coefficients. The objective functions of the two estimators were computed using the fast algorithm of Abrevaya [**1999**] for MRC and a sorting-based algorithm described in Appendix, Section C.1, for PDR4, both programmed in C. The computational times for one thousand bootstrap iterations were 34 minutes for MRC and 6.5 hours for PDR4, on an AMD Opteron 2.8 GHz processor. The memory usage was 62 megabytes for MRC and about 400 megabytes for PDR4.

| | MRC | | | PDR4 | | |
|---|---|---|---|---|---|---|
| | coef. | std. error | | coef. | std. error | |
| | | nonpar. | boots. | | nonpar. | boots. |
| $EDUC$ | 1.0000 | - | - | 1.000 | - | - |
| $EXPER$ | .3590 | .0559 | .0502 | .4068 | .0487 | .0432 |
| $EXRSQ$ | -.5965 | .1251 | .1123 | -.6741 | .1140 | .0977 |
| $FEMALE$ | -2.2105 | .3187 | .2744 | -2.3252 | .3102 | .2747 |
| $RACE$ | -.9851 | .4537 | .2937 | -1.2828 | .4076 | .3492 |
| $UNION$ | 1.5178 | .4482 | .3022 | 1.8922 | .4103 | .3166 |

Table 2.1. Wage equation estimation

Table 2.1 reports the values of the estimated coefficients and the standard errors. One can see that the standard errors estimated by the bootstrap and by the nonparametric methods can be substantially different in practice[6].

## 2.4. Conclusion

The nonparametric bootstrap is a way of performing inference in rank regressions without relying on subjective choices of tuning parameters. In this chapter we have established consistency of the nonparametric bootstrap for rank estimators and other finite-dimensional estimators that maximize $U$-processes of order 2 and higher. The computational feasibility of the bootstrap has been demonstrated in an empirical example.

[6]Since the true distribution of the data is not known in this example, we cannot say which method gives a better estimate of the true asymptotic standard deviations. The finite-sample performance of the bootstrap and the nonparametric methods will be investigated in Monte-Carlo experiments in Chapter 3.

CHAPTER 3

# Rates of Convergence in the Central Limit Theorems for Rank Estimators

## 3.1. Introduction

Here we investigate the accuracy of inference in rank regressions. As explained in Chapter 2, one can use the normal distribution with an estimated variance or the bootstrap distribution as consistent approximations of the finite-sample distribution of a rank estimator. However, the results on the asymptotic normality, or consistency of the bootstrap, provide no insight on the potential magnitudes of the error in such approximations. If the error converges to zero slowly with the number of observations $n$, the confidence intervals and tests of hypotheses constructed using either approximation may have coverage probabilities and levels very different from the nominal ones in finite samples.

The problem of the accuracy of inference is well understood in the case of an estimator that is a smooth function of sample moments (see e.g. Bhattacharya and Rao [**1976**] and Hall [**1992**]). Then, confidence intervals based on the asymptotic normal distribution typically attain the desired coverage probability up to an error of order $O\left(n^{-1/2}\right)$ for one-sided confidence intervals and $O\left(n^{-1}\right)$ for two-sided symmetric confidence intervals. In the case of $M$-estimators with nonsmooth criterion functions, the exact order of the approximation error is known only in several special cases, such as the least absolute deviation estimator studied by De Angelis, Hall and Young [**1993**]. Some results are available

for nonparametric methods that are applicable to models (1.2) or (1.6). Nishiyama and Robinson [**2005**] studied the accuracy of inference for the normal and the bootstrap approximations for the average derivative estimator of Powell, Stock and Stoker [**1989**] and showed that it can be the same as in parametric methods. However, this conclusion relies on restrictive moment and smoothness conditions. Particularly, there is a hidden curse-of-dimensionality effect: the conditional expectation $E[Y|X'\beta]$ has to have progressively higher numbers of bounded derivatives in the single index $X'\beta$ as the dimension of $X$ grows, and progressively higher orders of kernels have to be used in associated nonparametric regressions[1].

Below we give an upper bound on the error of approximation of the finite-sample distributions of MRC, MR, and the other rank estimators listed in Chapter 1. The bound is the same for approximations by both the bootstrap distribution and the normal distribution with the true variance. In the case of MRC, the error converges to zero with the rate arbitrarily close to $n^{-1/6}$. The rate is slower for the MR estimator if the outcome $Y$ is not bounded, but it also approaches the order of $n^{-1/6}$ if $Y$ has sufficiently many finite moments. The result holds under mild regularity conditions and is not subject to the curse of dimensionality. We further show that, under somewhat stronger assumptions, the PDR3 and PDR4 estimators of Abrevaya [**2003**] have a much smaller approximation error, close to $n^{-1/2}$ in the case of PDR3 and exactly $n^{-1/2}$ for the case of PDR4. Therefore, in

[1]The same is true for conditions under which this estimator is root-$n$-consistent. The sieve minimum-distance estimator of Ai and Chen [**2003**] also has a hidden curse of dimensionality, since it requires progressively stronger smoothness properties of the unknown functions when the dimension of the vector $X$ grows. Other methods, such as the estimator by Ichimura [**1993**], may not have this problem. Unfortunately, the second-order asymptotic properties of Ichimura's estimator are not known. It is likely though that strong smoothness assumptions will be needed for it to have the rate of convergence of order $O\left(n^{-1/2}\right)$ for the error between the finite-sample distribution of the estimator and the asymptotic normal distribution.

one-sided tests and confidence intervals, the pairwise-difference rank estimators achieve the same order of accuracy as the classical parametric estimators. We are not aware of existence of smoothing-based nonparametric techniques applicable to model (1.8) that would achieve this degree of precision of inference under the same regularity conditions.

The rest of the chapter is organized as follows. The asymptotic results are given in Section 3.2 for the rank estimators in general, and in Section 3.3 for the pairwise-difference rank estimators. In Section 3.4 the convergence properties are illustrated in Monte-Carlo experiments. Conclusions are given in Section 3.5. The proofs of the results exposed in Sections 3.2 and 3.3 can be found in Appendix A.

### 3.2. Rates of Convergence: General Case

In this section we obtain bounds on the approximation errors in Theorems 1 and 2 of Chapter 2 in the special case of the rank estimators. Here we consider the rank estimators in general, and in the next subsection we give stronger results for the subclass of the pairwise-difference rank estimators.

To expose the asymptotic theory, we use the same framework as in Chapter 2. The bounds that we find are closely related to the continuity properties of the quantity appearing in Assumption 4. The first result is obtained for the maximizers of $U$-processes whose kernel functions $h$ satisfy the following condition.

**Assumption 6.** *There exist numbers $\delta, C > 0$ such that for all $\theta_1$, $\theta_2$ in the $\delta$-neighborhood of $\theta_0$,*

$$P^2\left[\left(P^{m-2}h_{\theta_1} - P^{m-2}h_{\theta_2}\right)^2\right] \leq C\left\|\theta_1 - \theta_2\right\|. \tag{3.1}$$

Note that for differentiable kernels $h$ one would normally have, by a Taylor expansion argument, that

$$P^2\left[\left(P^{m-2}h_{\theta_1} - P^{m-2}h_{\theta_2}\right)^2\right] = O\left(\|\theta_1 - \theta_2\|^2\right). \tag{3.2}$$

Assumption 6, therefore, reflects a degree of nonsmoothness of the criterion function. To see why it is relevant for rank estimators, note that for estimators like MRC or MR, both (3.1), and its reverse:

$$P^2\left[\left(P^{m-2}h_{\theta_1} - P^{m-2}h_{\theta_2}\right)^2\right] \geq c\,\|\theta_1 - \theta_2\|. \tag{3.3}$$

(for a constant $c > 0$) are generally true.

Consider, for example, MRC. One can see that

$$[h_{\theta_1}(z_1, z_2) - h_{\theta_2}(z_1, z_2)]^2 = |h_{\theta_1}(z_1, z_2) - h_{\theta_2}(z_1, z_2)|$$

(this is a consequence of a property of the indicator function: for any two sets $A$, $B$, $[\mathbf{1}\{A\} - \mathbf{1}\{B\}]^2 = |\mathbf{1}\{A\} - \mathbf{1}\{B\}|$). Let $X = (U, V)$, $x = (u, v)$ where $V$ ($v$) is the last component of the vector $X$ ($x$), and $U$ ($u$) is the vector of the first $d$ components of $X$ ($x$). Suppose that $V$ is continuously distributed conditionally on $U$. Then, except on a set of $P$-measure zero,

$$\begin{aligned} |h_{\theta_1}(z_1, z_2) - h_{\theta_2}(z_1, z_2)| &= \mathbf{1}\{y_1 \neq y_2\} \cdot \\ &|\mathbf{1}\{v_2 > v_1 + \theta_1'(u_1 - u_2)\} - \mathbf{1}\{v_2 > v_1 + \theta_2'(u_1 - u_2)\}|. \end{aligned} \tag{3.4}$$

Suppose, further, that the density $\phi_{V|U}$ is uniformly bounded and that components of $U$ are $P$-integrable. Then

$$P^2\left[\left(h_{\theta_1}-h_{\theta_2}\right)^2\right] \leq 2\left\|\theta_1-\theta_2\right\| P\left\|U\right\| \sup \phi_{V|U}. \tag{3.5}$$

This is inequality (3.1), because in the case of $m=2$, $P^{m-2}h_\theta = h_\theta$. The same inequality can be obtained without difficulty for all existing rank correlation estimators by similar considerations.

To prove the reverse inequality, (3.3), for MRC, assume that $V$ has a continuous density $\phi_{V|Y,U}$ conditionally on both $Y$ and $U$. Then

$$P^2\left[\left(h_{\theta_1}-h_{\theta_2}\right)^2\right] \geq \left\|\theta_1-\theta_2\right\| \cdot$$
$$\int \mathbf{1}\left\{y_1 \neq y_2\right\}\left|(u_1-u_2)'\mathbf{n}_{\theta_1-\theta_2}\right| \gamma dP(y_2,u_2)dP(y_1,x_1).$$

In this formula $\mathbf{n}_{\theta_1-\theta_2}$ is the unit vector in the direction of $\theta_1-\theta_2$, $P(y_2,u_2)$ and $P(y_1,x_1)$ are marginal c.d.f.s of, respectively, $(Y,U)$ and $(Y,X)$, and

$$\gamma(v_1,y_2,u_2) = \min_{|r|\leq\delta\|U_1-U_2\|} \phi_{V|Y,U}\ (v_1+r|y_2,u_2),$$

where $\delta>0$ is so large that the compact $\Theta$ lies in the ball of radius $\delta$ with center at zero. If $\phi_{V|Y,U}$ is everywhere positive (so that $\gamma>0$), and the set

$$\left\{y_1 \neq y_2,\ (u_1-u_2)'\mathbf{n}_{\theta_1-\theta_2} \neq 0\right\}$$

has a positive $P-$measure, then the reverse of (3.5) holds. For $m=2$, this is the same as inequality (3.3).

Inequality (3.3) can be verified by similar methods for the other rank estimators that maximize a $U$-processes of the second order (e.g. MR or the partial rank estimator of Khan and Tamer [**2007**]), and for the estimator of Cavanagh and Sherman [**1998**] that maximizes the $U$-statistic of order 3 given by (1.7) (we will refer to this estimator as MR3). However, (3.3) does *not* hold for the pairwise-difference estimators PDR3 and PDR4. For the latter, (3.2) holds instead (for small differences $\theta_1 - \theta_2$). As explained in the next subsection, this property lies at the origin of the higher accuracy of inference associated with the pairwise-difference rank estimators.

In the bootstrap problem, we also need to account for the unconditional statistical dependence between the bootstrap draws.

**Assumption 7.** *There exist $\delta, C > 0$ such that for all $\theta_1$, $\theta_2$ in the $\delta$-neighborhood of $\theta_0$,*

$$\left(P^{m-2}h_{\theta_1}^{[m-2]} - P^{m-2}h_{\theta_2}^{[m-2]}\right)^2 \leq C\left\|\theta_1 - \theta_2\right\|.$$

Again, this condition is immediately true for MRC. For the other rank estimators it can be verified in the same way as Assumption 6, under the same sufficient conditions.

For estimators satisfying Assumption 6 (and, for the bootstrap, Assumption 7), the following upper bound holds.

**Theorem 4.** *Let Assumptions 1-3 and 6 hold. Assume that $P\left|\sup_{\theta\in\Theta} r_{n,\theta}\right| = O\left(n^{-3/2}\right)$, $PM^2 < \infty$, $P\left\|\partial^2\tau_{\theta_0}\right\|^2 < \infty$, $P\left\|\partial\tau_{\theta_0}\right\|^4 < \infty$, and, for a $p \geq 6$, $P^m H^p < \infty$. Then*

$$\sup_{A\in\mathcal{A}}\left|\int_A dF_{n^{1/2}(\theta_n-\theta_0)} - \int_A d\Phi_\Gamma\right| = O\left(\left(n^{-1/6}\left(\log n\right)^{2/3}\right)^{\frac{1}{1+2/3p}}\right). \tag{3.6}$$

*If, additionally,* $P\left|\sup_{\theta\in\Theta}\hat{r}_{n,\theta}\right| = O\left(n^{-3/2}\right)$, $P^m H^p_{\omega_m} < \infty$ *for each permutation, with repetition,* $\omega_m$, *and Assumption 7 holds, then*

$$\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)} - \int_A d\Phi_\Gamma\right| = O_p\left(\left(n^{-1/6}\left(\log n\right)^{2/3}\right)^{\frac{1}{1+2/3p}}\right). \tag{3.7}$$

The upper bound for the error with which the bootstrap quantiles of $\hat{\theta}_n$ approximate the finite-sample quantiles of $\theta_n$ can be found from (3.6) and (3.7) by the triangle inequality. In the case of MRC ($P^m H^p$, $P^m H^p_{\omega_m} < \infty$ for all positive $p$), we have

$$\sup_{A\in\mathcal{A}}\left|\int_A dF_{n^{1/2}(\theta_n-\theta_0)} - \int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)}\right| = O_p\left(n^{-1/6+\varepsilon}\right),$$

where $\varepsilon > 0$ can be taken arbitrarily small.

### 3.3. Rates of Convergence: Pairwise-Difference Rank Estimators

The bound obtained in the previous subsection converges to zero slowly. The rate of convergence improves substantially, however, if the quantity on the left hand side of (3.1) has stronger continuity properties. Namely, let the following assumption hold.

**Assumption 8.** *For* $s = 2$ *or* $3$[2], *function* $f_\theta = P^{m-s}h_\theta$ *is three times continuously differentiable in a* $\delta_0$*-neighborhood of* $\theta_0$. *There exists a function* $L\left(z_1, ..., z_s\right)$, *satisfying the condition* $P^s L^4 < \infty$, *such that*

$$\left\|\partial f_{\theta_0}\right\|, \left\|\partial^2 f_{\theta_0}\right\|, \left\|\partial^3 f_{\theta_0}\right\| \leq L,$$

[2]If this condition is satisfied with $s > 3$, then it is also satisfied with $s = 3$, which is sufficient for our analysis.

*and, for all* $\theta_1, \theta_2$ *in the* $\delta_0$*-neighborhood of* $\theta_0$,

$$\left\|\partial^3 f_{\theta_1} - \partial^3 f_{\theta_2}\right\| \leq L \left\|\theta_1 - \theta_2\right\|.$$

*(Here* $\partial^k f_\theta$, $k \geq 3$, *is the array of all partial derivatives of* $f$ *of order* $k$ *at* $\theta$, *and* $\left\|\partial^k f_\theta\right\|$ *is the maximum in the absolute value over all elements of the array.)*

It is clear that Assumption 8 cannot hold for MRC or MR, for which inequality (3.3) is true. Nonetheless, it can be satisfied for certain rank estimators maximizing a $U$-process of order 3 or higher. Historically, the first such example is the estimator by Han [**1987b**] of the parameter of the transformation function in the Box-Cox transformation model. Another example is the estimator for the same problem proposed in Asparouhova et al [**2002**]. Abrevaya's PDR3 and PDR4 estimators also satisfy Assumption 8. Below we will focus on Abrevaya's estimators as they have broader applicability than the former two estimators.

To see why Assumption 8 holds for pairwise-difference rank estimators, consider, for example, the objective function of the PDR3 estimator. The symmetric version of the kernel of the corresponding $U$-process is:

$$\begin{aligned}(3.8)\qquad & h_\theta(z_1, z_2, z_3) = \\ & \left(\mathbf{1}\{y_1 < y_2\} - \mathbf{1}\{y_2 < y_3\}\right)\left(\mathbf{1}\{(x_1 - x_2)'\beta < (x_2 - x_3)'\beta\}\right) \\ & + \left(\mathbf{1}\{y_2 < y_3\} - \mathbf{1}\{y_3 < y_1\}\right)\left(\mathbf{1}\{(x_2 - x_3)'\beta < (x_3 - x_1)'\beta\}\right) \\ & + \left(\mathbf{1}\{y_3 < y_1\} - \mathbf{1}\{y_1 < y_2\}\right)\left(\mathbf{1}\{(x_3 - x_1)'\beta < (x_1 - x_2)'\beta\}\right),\end{aligned}$$

where the scale of $\beta$ is fixed by setting $\beta = (\theta, 1)$. We will now check Assumption 8 with $s = 2$. To compute the value of the function $P^{m-2}h_\theta$ one should integrate out the pair $(x_1, y_1)$ in every term in the above expression. Once the component $v_1$ of the vector $x_1$ is integrated out, the first term becomes

$$\left(\mathbf{1}\left\{y_1 < y_2\right\} - \mathbf{1}\left\{y_2 < y_3\right\}\right) \int_{-\infty}^{(2x_2 - x_3)'\beta - u_1'\theta} \phi_{V|Y,U}\left(x|y_1, u_1\right) dx.$$

The derivative of this expression with respect to $\theta$ is

$$\left(\mathbf{1}\left\{y_1 < y_2\right\} - \mathbf{1}\left\{y_2 < y_3\right\}\right)(2u_2 - u_3 - u_1)$$

$$\cdot\phi_{V|Y,U}\left((2u_2 - u_3 - u_1)'\theta + 2v_2 - v_3|y_1, u_1\right).$$

Similar expressions can be obtained for the other two terms in (3.8). The following conditions are sufficient for Assumption 8 to be satisfied: $\phi_{V|Y,U}$ is three times differentiable in $V$ for almost all $U$ and $Y$ and is uniformly bounded together with its derivatives of orders up to 3, and $P\left\|U\right\|^{12} < \infty$. By a similar derivation, the PDR4 estimator (as well as the estimators of Han and Asparouhova et al) satisfies Assumption 8, with $s = 3$, under the same sufficient conditions.

Not every rank estimator whose criterion function is a $U$-process of order 3 satisfies Assumption 8. Consider the MR3 estimator. After symmetrization,

$$\begin{aligned} h_\theta(z_1, z_2, z_3) &= \\ \mathbf{1}\{y_1 < y_3\}\mathbf{1}\{x_1'\beta &< x_2'\beta\} \\ +\mathbf{1}\{y_3 < y_2\}\mathbf{1}\{x_3'\beta &< x_1'\beta\} \\ +\mathbf{1}\{y_2 < y_1\}\mathbf{1}\{x_2'\beta &< x_3'\beta\}. \end{aligned}$$

The value of $P^{m-2}h_\theta$ is obtained by integrating out $(x_1, y_1)$. After that, the first two terms will become differentiable in $\theta$, while the last term will still contain the indicator function $\mathbf{1}\{x_2'\beta < x_3'\beta\}$. It is clear that under general conditions, inequality (3.3) will hold, which is incompatible with Assumption 8.

When Assumption 8 is satisfied, the components of $\theta_n$ that it controls decrease rapidly with $n$. The following condition is imposed to ensure a similar asymptotic behavior of the higher-order terms.

**Assumption 9.** *Either $m = s$ or there exist $\delta, C > 0$ such that for all $\theta_1$, $\theta_2$ in the $\delta$-neighborhood of $\theta_0$,*

$$P^{s+1}\left[\left(P^{m-(s+1)}h_{\theta_1} - P^{m-(s+1)}h_{\theta_2}\right)^2\right] \leq C\left\|\theta_1 - \theta_2\right\|.$$

Similarly to the previous cases, an extra condition is needed in the bootstrap problem.

**Assumption 10.** *(a) Assumption 8 is satisfied with a function $L$ such that, for every permutation, with repetition, $\omega_s$,*

$$P^s L_{\omega_s}^4 < \infty.$$

*(b) If $s = 2$ or $3$, and $m > s$, then there exist $\delta, C > 0$ such that for all $\theta_1$, $\theta_2$ in the $\delta$-neighborhood of $\theta_0$,*

$$P^{s-1}\left[\left(P^{m-(s+1)}h_{\theta_1}^{[m-2]} - P^{m-(s+1)}h_{\theta_2}^{[m-2]}\right)^2\right] \leq C\left\|\theta_1 - \theta_2\right\|.$$

*If $s = 3$ and $m > 3$, then, additionally,*

$$\left(P^{m-4}h_{\theta_1}^{[m-4]} - P^{m-4}h_{\theta_2}^{[m-4]}\right)^2 \leq C\left\|\theta_1 - \theta_2\right\|,$$

*where*

$$\begin{aligned} & h_\theta^{[m-4]}\left(z_1, ..., z_{m-4}\right) \\ = & \int h_\theta\left(z_1, ...z_{m-4}, Z_{m-1}, Z_{m-1}, Z_m, Z_m\right) dP\left(Z_{m-1}\right) dP\left(Z_m\right). \end{aligned}$$

For PDR3 and PDR4 (and Han's and Asparouhova et al estimators) these conditions can be checked, for, respectively, $s = 2$ and $s = 3$, by the same methods that were used to obtain (3.5). Moreover, in Assumption 9, generally the reverse inequality is also true, which can be verified by an argument similar to the proof of inequality (3.3) for MRC.

The next theorem gives the rates of convergence for rank estimators satisfying Assumptions 8 and 9. For brevity, only the case of uniformly bounded functions $h$ is considered.

**Theorem 5.** *Suppose that Assumptions 1-3 and 8, 9 hold,* $\sup_{Z,\theta}|r_{n,\theta}| = O\left(n^{-2}\right)$, *and* $H$ *is a constant. If Assumptions 8, 9 are satisfied with* $s = 2$, *let* $\varepsilon > 0$ *be arbitrarily small, and if they are satisfied with* $s = 3$, *let* $\varepsilon$ *be zero*[3]. *Then*

$$\sup_{A\in\mathcal{A}}\left|\int_A dF_{n^{1/2}(\theta_n-\theta_0)} - \int_A d\Phi_\Gamma\right| = O\left(n^{-1/2+\varepsilon}\right). \tag{3.9}$$

*If also* $\sup_{Z,\theta}|\hat{r}_{n,\theta}| = O\left(n^{-2}\right)$, *Assumptions 5 (a) and 10 hold, then*

$$\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)} - \int_A d\Phi_\Gamma\right| = O_p\left(n^{-1/2+\varepsilon}\right). \tag{3.10}$$

### 3.4. Monte-Carlo Experiments

In this section we investigate the bootstrap properties of rank estimators in finite samples. We consider two estimators, MRC and PDR4. As explained in the previous sections, MRC has a wider scope of applications (in particular, it can be applied to limited dependent variable models) and its criterion function is cheaper to compute. However, our asymptotic results suggest that inference with MRC may be inaccurate in small samples. PDR4, on the other hand, needs substantial computational capacity (the fastest available algorithm for computing its objective function requires $O\left(n^2\log n\right)$ operations and $O\left(n^2\right)$ memory cells). However, within the scope of its application, PDR4 can serve as a good complement to MRC in small samples, where it achieves higher precision of inference.

In the Monte-Carlo experiments, MRC is applied to the binary choice model:

[3]In this case the order of magnitude of the approximation error of the distribution of $\theta_n$ by the asymptotic normal distribution is exactly $O\left(n^{-1/2}\right)$, since one can obtain the Edgeworth expansion of $\theta_n$ in which the first-order term has the magnitude $O\left(n^{-1/2}\right)$ and the next term is bounded by $O\left(n^{-3/4+\varepsilon}\right)$. The same also holds in probability for the bootstrap. See the footnote in Section A.2.4.

$$Y_i = \mathbf{1}\left\{X_i^{(1)} + X_i^{(2)} + \sigma\varepsilon_i > 0\right\}.$$

(In this case MRC and MR are numerically equivalent (see Cavanagh and Sherman [**1998**]), so the evidence presented below illustrates the properties of MR as well.) Three distributions for the first regressor are considered: the standard normal (a continuous case), binomial with the probability of 1 equal to 0.5 (a discrete case), and the Student distribution with 1.5 degrees of freedom. In the latter case, the first moment of $X^{(1)}$ is finite, but its second moment is infinite. This is a situation where the nonparametric method of Cavanagh and Sherman [**1998**] for estimating the asymptotic variance is rather difficult to apply. In particular, the moment conditions of Theorem 4 in Sherman [**1993**], under which the method is known to be consistent, are violated. Also, the rule for choosing the bandwidths (proportionally to the sample standard deviation of the estimated index $X'\hat{\beta}$) suggested by Cavanagh and Sherman may result in arbitrarily large bandwidths and is not practical. The second regressor, $X^{(2)}$, is distributed as $N(0,1)$ independently of $X^{(1)}$. It plays the role of a continuously distributed regressor needed for point identification of $\beta$. The error term, $\varepsilon$, is also distributed as $N(0,1)$ independently of both regressors. The scaling parameter $\sigma$ determines the noise-to-signal ratio in the dataset. We consider two cases, $\sigma = 1$ and $\sigma = 0.1$.

PDR4 is applied to the linear model:

$$Y_i = X_i^{(1)} + X_i^{(2)} + \varepsilon_i.$$

The regressor $X^{(1)}$ can have the standard normal or the Student(1.5) distribution. Regressor $X^{(2)}$ is distributed as a standard normal random variable independently of $X^{(1)}$. The error term is independent of both regressors and is distributed as either a standard normal or a standard Cauchy random variable. The latter case serves to demonstrate the robustness properties of PDR4 with respect to heavy-tailed distributions of the error term. Note that for the Cauchy distributed errors, $P|\varepsilon| = +\infty$, so that the OLS or nonparametric minimum-square-distance methods are not consistent in this case.

In rank regressions, the point identification of $\beta$ is achieved by imposing a scale normalization. Here we set the coefficient at the second regressor to be 1. The estimated model is then

$$y_i = f\left(\theta X_i^{(1)} + X_i^{(2)} + \varepsilon_i\right),$$

where $f(x) = 1\{x > 0\}$ in the binary choice model, and $f(x) = x$ in the linear model (function $f$ does not have to be known for implementation of MRC or PDR4) and $\varepsilon_i$ is the error term. The value of $\theta$ is found by maximizing the corresponding criterion function. The objective function of MRC is rather nonsmooth for our sample sizes, and its maximization is more difficult than that of the PDR4 objective function. We used the Nelder-Mead simplex maximization algorithm with parameters adjusted in trial runs of the program. For PDR4 estimator the standard maximization MATLAB routine fminsearch with default settings was enough. Both algorithms are iterative procedures requiring a starting approximation of the solution. In the sample problem, we took the true value, $\theta_0 = 1$. This option, of course, is not available in real data applications, where a grid of initial values should be considered. In the bootstrap we used both $\theta_0$ and $\theta_n$.

There are several asymptotically equivalent methods for computing the bootstrap critical values for the test statistic $n^{1/2}(\theta_n - \theta_0)$ that do not need an explicit estimator of the asymptotic variance. In the percentile method the quantiles of the test statistic are approximated by the conditional quantiles of the bootstrapped recentered statistic $n^{1/2}\left(\hat{\theta}_n - \theta_n\right)$. In our experiments with MRC, however, recentering of the bootstrapped estimator at $\theta_n$ led to inaccurate results. One alternative, motivated by the symmetry of the normal distribution, is the other percentile method, see Hall [**1992**], in which the quantiles of the test statistics are approximated by the quantiles of the statistic $n^{1/2}\left(\theta_n - \hat{\theta}_n\right)$. The procedure effectively eliminates the estimated value $\theta_n$ from computing the confidence intervals and critical values ($\theta_n$ cancels out in the corresponding expressions). This method was used to build one-sided and double-sided (equal-tailed) confidence intervals. The rejection probabilities for the corresponding tests were similar, and, for the sake of brevity, we report them only for the double-sided case[4]. There are other procedures that do not require recentering at $\theta_n$. One can approximate the c.d.f. of the test statistic by the c.d.f. of the demeaned bootstrapped statistic, $n^{1/2}\left(\hat{\theta}_n - \hat{P}\left[\hat{\theta}_n\right]\right)$, or by the c.d.f. of the normal distribution with zero mean and variance estimated by the conditional variance of $n^{1/2}\hat{\theta}_n$. These two methods can be more convenient than the other percentile method for inference about multidimensional $\theta$. The results for both are similar to the case of the other percentile method and are omitted.

MRC was computed for sample sizes $n = 200, 500$, and 1000 (see Table 3.1). The coverage probabilities are reasonably accurate except in the case with $n = 200$ and $\sigma = 0.1$ where the bootstrap fails dramatically for all three distributions of $X^{(1)}$. This should serve

[4]For a description of how rejection probabilities are computed, see Hall and Horowitz [**1996**].

| | nominal level 5% | | | nominal level 10% | | |
|---|---|---|---|---|---|---|
| $n =$ | 200 | 500 | 1000 | 200 | 500 | 1000 |
| Normal $X^{(1)}$ | | | | | | |
| $\sigma = 1$ | 3.0 | 3.2 | 5.1 | 7.7 | 8.5 | 9.4 |
| $\sigma = 0.1$ | 25.0 | 5.9 | 3.6 | 31.2 | 10.6 | 8.1 |
| Binary $X^{(1)}$ | | | | | | |
| $\sigma = 1$ | 3.6 | 4.2 | 4.2 | 7.7 | 9.6 | 8.3 |
| $\sigma = 0.1$ | 31.9 | 5.5 | 2.8 | 35.0 | 9.2 | 5.9 |
| Student (1.5) $X^{(1)}$ | | | | | | |
| $\sigma = 1$ | 2.7 | 2.5 | 4.3 | 5.3 | 7.3 | 8.9 |
| $\sigma = 0.1$ | 33.8 | 6.8 | 3.5 | 38.6 | 10.1 | 7.2 |

Table 3.1. Bootstrap rejection probabilities for equal-tailed t-tests - MRC

as a caution against using the bootstrap when the signal-to-noise ratio is high and the sample size is moderate. In this case the simulated distribution of MRC appears to have a mass point at zero. The bootstrap gives a distribution with a much higher concentration of mass at zero, and underestimates the length of the confidence intervals and the variance of the estimator. The phenomenon has to be taken into account when MRC is used together with a specification search: the bootstrap may reject models with low noise more often than it should.

In the case of PDR4, the percentile method (involving recentering) and the other percentile method gave close values of rejection probabilities. For brevity we report only the values obtained for the equal-tailed tests based on the other percentile method, for sample sizes $n = 50$, 100, and 200 (Table 3.2). It can be seen that bootstrap performs well even when the sample includes only 50 observations.

Finally, we compute the rejection probabilities in the same tests using the normal approximation with the asymptotic variance estimated by the nonparametric method of

| | nominal level 5% | | | nominal level 10% | | |
|---|---|---|---|---|---|---|
| $n =$ | 50 | 100 | 200 | 50 | 100 | 200 |
| Normal $X^{(1)}, \varepsilon$ | 5.9 | 4.8 | 5.8 | 12.8 | 10.1 | 10.3 |
| Student (1.5) $X^{(1)}, \varepsilon$ | 6.4 | 4.9 | 4.6 | 12.5 | 10.8 | 10.7 |
| Normal $X^{(1)}$, Cauchy $\varepsilon$ | 6.1 | 5.0 | 6.2 | 13.9 | 9.6 | 9.5 |

Table 3.2. Bootstrap rejection probabilities for equal-tailed t-tests - PDR4

Cavanagh and Sherman. Because of the complicated nature of the nonparametric method for PDR4, we only consider the case of MRC. We used kernel regressions to estimate the nonparametric functions that are required by this method, with the standard normal kernel and the bandwidths given by the rule of thumb used in Cavanagh and Sherman [**1998**], according to the formula:

$$h = k \cdot \hat{\sigma}_n n^{-1/5}.$$

Here $\hat{\sigma}_n$ is the sample variance of the estimated single index, and $k$ is a scaling multiplier allowing for different choices of bandwidths. Figures 3.1 and 3.2 show the rejection probabilities for 5% tests for coefficient $\theta$ in the binary choice models described above, with the standard normal, and the binary regressors $X^{(1)}$, respectively, for the sample sizes $n = 200$ and $n = 500$, and the noise-to-signal ratio $\sigma = 1$ and 0.1. One can see that although the rejection probabilities are close to the nominal level for some values of $k$, they can also deviate from that level for other values of $k$. Figure 3.3 shows the rejection probabilities in the binary choice model with $X^{(1)}$ distributed according to the Student distribution with 1.5 degrees of freedom. This examples is problematic for the nonparametric method and the specified rule for choosing the bandwidth for the reasons made clear in the preceding discussion. One can see that the nonparametric method is

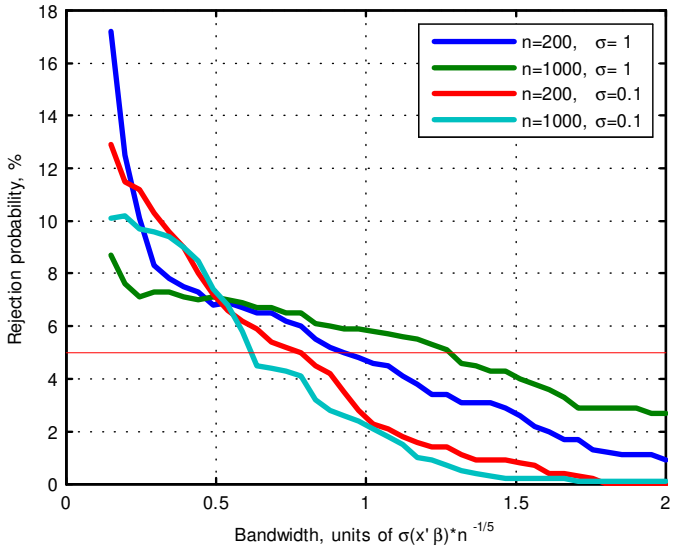


Figure 3.1. Rejection probabilities for the nonparametric method, the binary choice model, $X^{(1)} \sim N(0,1)$

particularly sensitive to the choice of $k$ here. Of course, the example is extreme and does not preclude the nonparametric method from practical use, but it underscores the necessity of developing objective and robust rules for choosing bandwidths in this method.

## 3.5. Conclusion

This chapter provides bounds on the approximation errors in the central limit theorems and the bootstrap consistency theorems for rank estimators, a class of methods that can be applied to popular semiparametric single-index models or used for robust estimation of parametric models. In the case of MRC and MR, the error is bounded by a function of the sample size of order close to $n^{-1/6}$, for both the sample and the bootstrap problem. Pairwise-difference rank estimators, such as PDR3 and PDR4, however, have a special structure due to which the bound is vanishing with the rate close to $n^{-1/2}$. Thus,

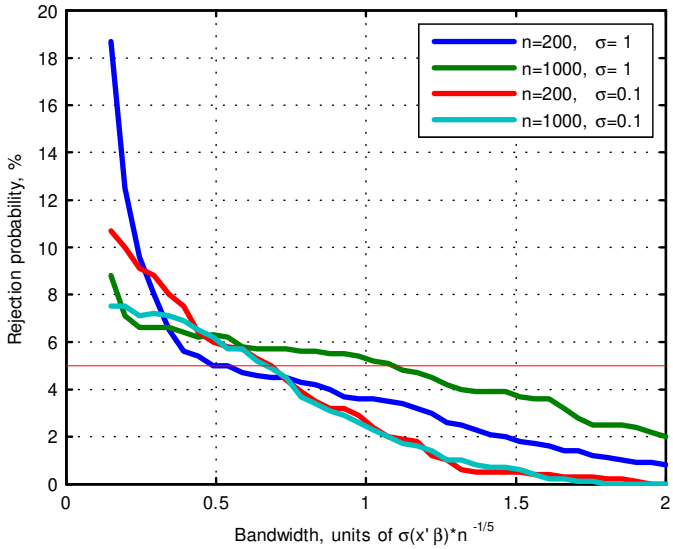


Figure 3.2. Rejection probabilities for the nonparametric method, the binary choice model, binomial $X^{(1)}$

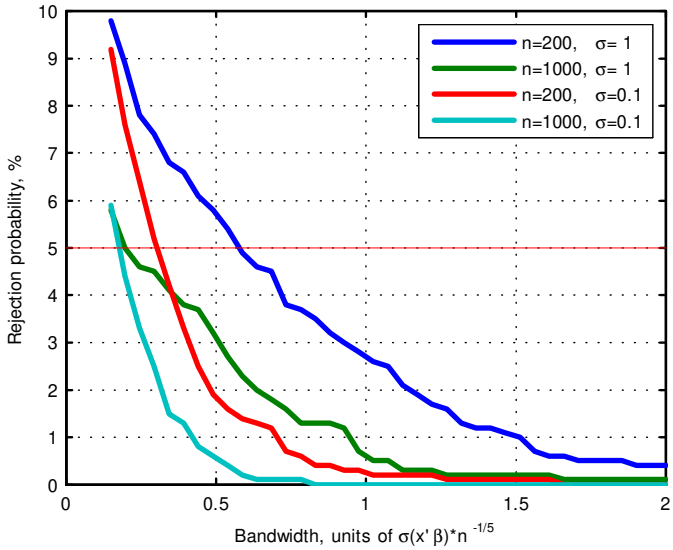


Figure 3.3. Rejection probabilities for the nonparametric method, the binary choice model, $X^{(1)} \sim Student\,(1.5)$

pairwise-difference estimators provide a remarkable example of a robust semiparametric method whose first- and second-order asymptotic properties approach those of parametric methods. The theoretical results have been illustrated with finite-sample Monte-Carlo experiments.

CHAPTER 4

# Weighted Rank Estimators

## 4.1. Introduction

In this chapter we consider the problem of efficiency of rank-based estimators. We focus on the rank estimators of order two, such as MRC, MR, the partial rank estimator of Khan and Tamer [**2007**], and others. The identification conditions on which these estimators are based can be written in the form: for two independent realizations of $(Y, X)$,

$$E\left[M\left(Y_1, Y_2\right)|X_1, X_2\right] > 0 \Longrightarrow X_1'\beta_0 > X_2'\beta_0. \tag{4.1}$$

Here $M$ is an antisymmetric function, such as $M\left(Y_1, Y_2\right) = Y_1 - Y_2$ for MR and $M\left(Y_1, Y_2\right) = sign\left(Y_1 - Y_2\right)$ for MRC. The estimator of the vector of parameters, $\beta_0$, up to scale, is the solution of the maximization problem

$$\max_{\beta=(\theta,1)} \sum_{i\neq j} M\left(Y_i, Y_j\right) \mathbf{1}\{X_i'\beta < X_j'\beta\}. \tag{4.2}$$

As shown by Sherman [**1993**] in the example of the binary choice model estimated by MR or MRC, such estimators do not attain the semiparametric efficiency bound for single-index models, in general.

To improve efficiency, we consider the estimation of $\beta_0$ by a maximizer of a weighted version of the criterion function in (4.2):

$$\max_{\beta=(\theta,1)} \sum_{i \neq j} w\left(X_i, X_j\right) M\left(Y_i, Y_j\right) \mathbf{1}\{X_i'\beta < X_j'\beta\}, \tag{4.3}$$

where $w \geq 0$ is a weighting function. Under regularity conditions, the resulting estimator is consistent and asymptotically normal. The optimal rank estimator is the weighted rank estimator that has the least asymptotic variance among weighted rank estimators. We look for conditions under which the optimal rank estimators exist, and verify if they attain the semiparametric efficiency bounds in various single-index models.

To evaluate the practical relevance of our weighting approach, as well as other methods seeking to improve efficiency, we compare the asymptotic variances of the unweighted and the optimal rank estimators, and semiparametric efficiency bounds, in models with exactly specified features (e.g. the distributions of the error term). These examples suggest that in the models with independent errors, the asymptotic variance of the optimal MRC is likely to be close to the semiparametric efficiency bound unless there are strong nonsmoothness effects in the distribution of the error term. The efficiency gains from weighting are likely to be small in the transformation model estimated by MRC, or the binary choice model estimated by MR or MRC. However they can be large in the monotone regression model estimated by MR or MRC. The performance of a feasible optimal MR and MRC in one such case is studied in a Monte-Carlo experiment, in which we obtain a substantial improvement in the finite-sample variances of these estimators over the unweighted estimators.

The rest of the chapter is organized as follows. Section 4.2 presents the asymptotic theory of the weighted rank estimators and the feasible optimal rank estimators. Section 4.3 provides numerical evidence. Section 4.4 concludes. In Appendices B and C, we give proofs of the theoretical results and discuss numerical algorithms for the proposed methods.

## 4.2. Asymptotic Theory of Weighted Rank Estimators

### 4.2.1. Identification, Consistency and Asymptotic Normality

In this section, we give conditions under which the estimator $\theta_n$ defined by (4.3) is consistent and asymptotically normal, and derive the expression for its asymptotic variance.

We make the following assumptions.

**Assumption 11.** *The observations* $(Y_i, X_i)$, $i = 1, ..., n$, $Y_i \in \mathbb{R}^{d_Y}$, $X_i \in \mathbb{R}^{d+1}$, *are i.i.d. across* $i$.

**Assumption 12.** *The function* $M(y_1, y_2)$ *satisfies the conditions:*

$$M(Y_1, Y_2) = -M(Y_2, Y_1)$$

*almost surely, and*

$$E|M(Y_1, Y_2)| < \infty.$$

*Condition (4.1) is satisfied for almost all* $X_1, X_2$ *and a* $\beta_0$ *such that its* $(d+1)$*-th component,* $\beta_0^{(d+1)}$, *is positive.*

Let $U$ be the vector of the first $d$ components of $X$, and $V$ be its $(d+1)$-th component.

**Assumption 13.** *Conditional on $U$, $V$ has a Lebesgue density, $g_{V|U}(v;u), v \in \mathbb{R}$, for almost all $u$, that is twice continuously differentiable in $v$ for almost all $(u,v)$. The conditional mean $E[M(Y,y)|X=(u,v)]$ is twice continuously differentiable in $v$ for almost all $y,u,v$, with an absolutely integrable first derivative.*

**Assumption 14.** *The weighting function $w(x_1,x_2)$ is nonnegative and is twice continuously differentiable in $v_1$ and $v_2$ for almost all $x_1,x_2$.*

In the remaining assumptions it is convenient to rearrange the components of $X$. Let $\theta_0 \in \mathbb{R}^d$ be the vector with the components $\theta_0^{(i)} = \beta_0^{(i)}/\beta_0^{(d+1)}, i = 1,...,d$, and $Z$ be the single index:

$$Z = U'\theta_0 + V.$$

Note that $Z$ has a density conditionally on $U$:

$$g_{Z|U}(z;u) = g_{V|U}(z - u'\theta_0; u).$$

Define the function

$$\lambda(y,u,z) = E[M(Y,y)|U=u, Z=z],$$

and the function

$$\mu(u_1,u_2,z) = \frac{\partial}{\partial\delta} E[M(Y_1,Y_2)|U_1=u_1, U_2=u_2, Z_1=z+\delta, Z_2=z]|_{\delta=0}.$$

By Assumption 13, the function $\mu$ is well defined for almost all $u_1,u_2,z$, and

$$\mu(u_1,u_2,z) = E[\lambda_z(Y,u_1,z)|U=u_2, Z=z],$$

where $\lambda_z$ is the partial derivative of the function $\lambda$ with respect to $z$.

Also, let

$$w_s(x_1, x_2) = w(x_1, x_2) + w(x_2, x_1)$$

be the symmetric version of the function $w$.

**Assumption 15.** *There is a set $\mathcal{U} \in \mathbb{R}^d$, and a nonempty open interval $I \subset \mathbb{R}$ such that:*

*(i) The support of the distribution of the vector*

$$(U_1 - U_2) 1\{U_1, U_2 \in \mathcal{U}\}$$

*does not lie in a proper linear subspace of $\mathbb{R}^d$.*

*(ii) The functions $g_{Z|U}(z, u_1)$, $\mu(u_1, u_2, z)$, $w_s(x(u_1, z), x(u_2, z))$, where*

$$x(u, z) = (u, z - u'\theta_0),$$

*are nonzero for any $z \in I$ and $u_1, u_2 \in \mathcal{U}$.*

**Assumption 16.** *(i)*

$$E\left[M^2(Y_1, Y_2) w^2(X_1, X_2)\right] < \infty. \tag{4.4}$$

*(ii) For a $\delta > 0$, and $s, l = 0, 1, 2,$*

$$E\|U_1 - U_2\|^{s+1} \sup_{\|\theta\| \leq \delta} \left|\alpha^{(s,l)}\left(Y_1, U_1, Z_1 + (U_1 - U_2)'\theta, X_2\right)\right| < \infty,$$

*where*

$$\alpha^{(s,l)}(y,u,z,x_2) = \partial_z^{s-l}\lambda(y,u,z)\,\partial_z^l\left[w_s(x(u,z),x_2)\,g_{Z|U}(z;u)\right],$$

*and* $\partial_z^l$ *denotes the partial derivative with respect to* $z$ *of order* $l$*.*

Let

$$\nabla_w(y,u,z) = g(z)\,E\left[(u-U)\,M(y,Y)\,w_s(x(u,z),X)\,|Z=z\right],$$

where $g(z)$ is the marginal density of $Z$.

**Assumption 17.** *The matrix*

$$V_w = E\left[\nabla_w(Y,U,Z)\,\nabla_w(Y,U,Z)'\right]$$

*is finite and positive definite.*

By Assumption 11, the estimator is applicable to cross-section data. Assumption 12 gives the identification condition. Assumption 13 imposes smoothness restrictions on the distribution of the data. One of the covariates, $V$, must be continuously distributed conditionally on the remaining covariates, $U$. This is similar to the original assumption of Han [**1987**] except that we do not require that $V$ have the full support conditionally on $U$. The full support condition is often unrealistic as the covariates may not take all values (e.g. may only be positive), or may not be observed for all values; so it is removed from our set of conditions. Additionally, we need a degree of smoothness of the conditional density of $V$ and the conditional distribution of $Y$ with respect to $V$, expressed via the smoothness of the conditional expectation $E[M(Y,y)\,|X]$. Assumption 14 states basic requirements to the weighting function. Note that the weighting function is allowed to

take zero values. Assumption 15 (i) imposes a non-collinearity restriction on the distribution of $U$. Assumption 15 (ii) says that the values of $U$ such that $U_1 - U_2$ form a basis in $\mathbb{R}^d$ should be observable together with all values of the single index $Z$ from a small open set $\left(g_{Z|U}(z,u) \neq 0,\ z \in I\right)$ and that they are not censored by the weighting function $(w_s > 0)$. Additionally, it requires that the derivative $\mu(u_1, u_2, z)$ be nonzero for such values of $u_1$, $u_2$ and $z$. Note that the identification condition (4.1) implies that the derivative in the definition of $\mu$, if it exists, cannot be negative ($\mu \geq 0$). The condition that $\mu > 0$ can be viewed as a local, differential form of (4.1) that excludes the trivial $M \equiv 0$. Previously, weaker conditions excluding the case $M \equiv 0$ were imposed. Our stronger requirement allows us to avoid the full support condition on $V$ discussed above. Assumptions 16 and 17 impose integrability conditions on the vector $U$ and the function $M$, as well as the conditional mean $E[M(Y,y)|X]$, the weighting function $w$, and the conditional density $g_{V|U}$ and their derivatives relative to $V$ (expressed using the single index $Z$). In particular, Assumption 16 (i) puts a bound on the random fluctuations of the criterion function in the sample optimization problem around its mean, which is needed for root-$n$-consistency of $\theta_n$.

The asymptotic variance of the estimator $\theta_n$ depends on the matrix $V_w$ defined in Assumption 17, and the matrix

$$\Delta_w = E\left[(U_1 - U_2)(U_1 - U_2)' g(Z_1)\, \mu(U_1, U_2, Z_1)\, w_s(X_1, X_2) \,|Z_2 = Z_1\right].$$

Assumption 17 requires that $V_w$ be finite and nonsingular. Assumption 16 (ii) for $s = 2$ and $l = 0$ ensures that the matrix $\Delta_w$ is finite. It is clear that for $\Delta_w$ to be nonsingular, a condition on the random vector $(U_1 - U_2)\mu(U_1, U_2, Z_1)$, the distribution of $Z$ and the

values of $w_s(X_1, X_2)$ is necessary. Here such condition is the one stated in Assumption 15[1].

The following theorem establishes consistency and asymptotic normality of the weighted rank estimator, and gives an expression for its asymptotic variance.

**Theorem 6.** *Let $\Theta$ be a compact set in $\mathbb{R}^d$, $\theta_0$ be its interior point, and $\theta_n$ solve the maximization problem (4.3) on $\Theta$. Under Assumptions 11-17,*
*(i) $\theta_n \to^p \theta_0$, and*
*(ii)*

$$\sup_{A \in \mathcal{A}} \left| \int_A dF_{n^{1/2}(\theta_n - \theta_0)} - \int_A d\Phi_{\Gamma_w} \right| \to 0,$$

*where $F_{n^{1/2}(\theta_n - \theta_0)}$ is the c.d.f. of the random vector $n^{1/2}(\theta_n - \theta_0)$, $\Phi_{\Gamma_w}$ is the c.d.f. of the normal distribution with mean zero and finite, nonsingular variance*

$$\Gamma_w = 4\Delta_w^{-1} V_w \Delta_w^{-1},$$

*and $\mathcal{A}$ is the collection of all measurable convex sets in $\mathbb{R}^d$.*

### 4.2.2. Optimal Choice of the Weighting Function

Among the consistent, asymptotically normal estimators of $\theta_0$, corresponding to different choices of $w$, it is natural to look for the one that achieves the lowest asymptotic variance. In general, such optimal estimator may not exist; however, it exists and can be found in a special case considered in this section. Namely, assume that the function $\lambda(y, u, z)$ does

[1]Sherman [**1993**] did not impose a differentiable form of the identification condition (4.1), such as positivity of $\mu$ in our Assumption 15. He argued that the matrix $\Delta_w$ is non-denerate because $\theta_0$ is uniquely identified. However, the fact that $\theta_0$ uniquely maximizes the population objective function does not imply that the Hessian of that objective function ($-2\Delta_w$) is nonsingular.

not depend on $u$:

$$\lambda(y,u,z) \equiv E[M(Y,y)|U=u,Z=z] = \lambda(y,z) \ \forall u. \tag{4.5}$$

This assumption is satisfied in the single-index models (the models in which $Y|X$ is distributed as $Y|Z$) and in the monotone regression model (1.4)-(1.5). However, the assumption is restrictive in that it does not fully cover some interesting models that can be estimated using the rank estimators. An example of such model will be given below.

As we show in Appendix, Lemma 21, the function $\mu(u_1,u_2,z)$ is symmetric in $u_1$, $u_2$. Therefore, under condition (4.5), it does not depend on either $u_1$, $u_2$: $\mu(u_1,u_2,z) = \mu(z)$. Then the matrices $V_w$ and $\Delta_w$ can be written as

$$\begin{aligned} V_w &= E[W_1 W_1'], \\ \Delta_w &= 2E\left[\frac{\mu(Z_1)U_1}{\sigma(X_1)} W_1'\right], \end{aligned} \tag{4.6}$$

where

$$\sigma^2(X) = E\left[\lambda^2(Y,Z)|X\right],$$

$$W_1 = \sigma(X_1)\, g(Z_1)\, E_2[(U_1 - U_2)\, w_s(X_1,X_2)|Z_2 = Z_1], \tag{4.7}$$

and $E_2$ is the expectation over the random variables labeled 2, holding the random variables labeled 1 constant.

**Theorem 7.** *Assume that the matrices $V_w, \Delta_w$ are given by (4.6) and (4.7), where $w_s, g, \sigma$ are nonnegative scalar functions, $w_s$ is symmetric in $x_1, x_2$, $U$ is a random vector taking values in $\mathbb{R}^d$;*

$$\begin{aligned} E\left[\left(1+\|U\|\right) w_s\left(x\left(U,z\right),x_2\right)\right] &< \infty, \\ E\left[\frac{1+\|U\|}{\sigma^2\left(X\right)}\right] &< \infty, \end{aligned}$$

*for almost every $x_2, z$; $V_w$ and $\Delta_w$ are finite, the matrix*

$$\Gamma_* = \left(E\left[\frac{\mu^2\left(Z\right)}{\sigma^2\left(X\right)}\left(U-\tilde{U}\left(Z\right)\right)\left(U-\tilde{U}\left(Z\right)\right)'\right]\right)^{-1} \tag{4.8}$$

*with*

$$\tilde{U}\left(z\right) = \frac{E\left[\frac{U}{\sigma^2(X)}\middle| Z=z\right]}{E\left[\frac{1}{\sigma^2(X)}\middle| Z=z\right]}$$

*is well defined as a nonsingular, finite matrix, and $\Delta_w$ is nonsingular. Then*

$$4\Delta_w^{-1}V_w\Delta_w^{-1} \geq \Gamma_*. \tag{4.9}$$

The variance $\Gamma_*$ is attained with the following weighting function:

$$w_1\left(x_1,x_2\right) = w_0^{1/2}\left(x_1\right) w_0^{1/2}\left(x_2\right), \tag{4.10}$$

where

$$w_0\left(x\right) = 1\left\{g\left(z\right)>0\right\} \cdot \frac{\mu\left(z\right)}{g\left(z\right)\sigma^4\left(x\right) E\left[\frac{1}{\sigma^2(X)}\middle| Z=z\right]}, \tag{4.11}$$

and for $x=(u,v)$, $z=u'\theta_0+v$.

If $\sigma^2(x) = \sigma^2(z)$, then the expression for $w_0(x)$ simplifies:

$$w_0(x) = 1\{g(z) > 0\} \cdot \frac{\mu(z)}{g(z)\sigma^2(z)}, \tag{4.12}$$

and the optimal variance is attained with either $w_1$, or

$$w_2(x_1, x_2) = w_0(z_1) + w_0(z_2). \tag{4.13}$$

If the support of $g(z)$ is unbounded, and the function $\mu(z)/\sigma^2(z)$ is bounded away from zero, the optimal weighting functions $w_1$ and $w_2$ will typically violate Assumption 16 (i) that was imposed for the root-$n$-consistency of $\theta_n$. Namely, in this case,

$$E\left[w^2(X_1, X_2)\right] = \infty,$$

while for Assumption 16 (i) to hold it is typically necessary that

$$E\left[w^2(X_1, X_2)\right] < \infty.$$

In this case, instead of the function $w_0$, one can use the function

$$w_{0\tau}(z) = \frac{\mu(z)}{\tau + g(z)\sigma^4(x) E\left[\frac{1}{\sigma^2(X)}\middle| Z = z\right]},$$

where $\tau > 0$ is a small number, to achieve approximate optimality. If the support of $Z$ is bounded, and the density $g(z)$ is bounded away from zero on the support, trimming the denominator in $w_0(z)$ is not needed.

### 4.2.3. Examples

**The Conditional Mean Model and MR.** In the monotone regression model (1.4)-(1.5), the vector of coefficients $\beta_0$ can be consistently estimated up to scale by the monotone rank estimator of Cavanagh and Sherman [**1998**]. For this method, the function $\lambda(y,u,z) = f(z) - y$ does not depend on $u$, so that the optimal weighting function exists. We have:

$$\begin{aligned} \mu(z) &= f'(z) = \partial_z E[Y|Z] \\ \sigma^2(x) &= \sigma_0^2(x) = Var(Y|X=x), \end{aligned}$$

where we assume, for notational simplicity, that $\beta_0^{(d+1)} = 1$, and where the second representation relates the functions $\mu$ and $\sigma^2$ directly to the observable data. The optimal weighting function is given by (4.10) with

$$w_0(x) = \left( \frac{f'(z)}{g(z)\,\sigma_0^4(x)\,E\left[\frac{1}{\sigma_0^2(X)}|Z=z\right]} \right)^{1/2},$$

and the corresponding variance of the estimator is

$$\left( E\left[ \frac{f'(Z)^2}{\sigma_0^2(X)} \left(U - \tilde{U}(Z)\right)\left(U - \tilde{U}(Z)\right)' \right] \right)^{-1},$$

which is the semiparametric efficiency bound for the nonlinear regression model (Newey and Stoker [**1993**], formula (4.4)).

A special case of the monotone regression model is the binary choice model:

$$Y = 1\{X'\beta - \varepsilon > 0\},$$

under the independence assumption between $\varepsilon$ and $X$. In this case $\lambda(y,z) = G_\varepsilon(z) - y$, and

$$\begin{aligned} \mu(z) &= g_\varepsilon(z) = \partial_z E[Y|Z=z], \\ \sigma^2(x) &= G_\varepsilon(z)(1-G_\varepsilon(z)) \\ &= E[Y|Z=z](1-E[Y|Z=z]), \end{aligned}$$

where $G_\varepsilon$ is the c.d.f. of $\varepsilon$ and $g_\varepsilon$ is its density. The variance of the optimal weighted MR is

$$\left(E\left[\frac{g_\varepsilon(Z)^2}{G_\varepsilon(Z)(1-G_\varepsilon(Z))}(U-E[U|Z])(U-E[U|Z])'\right]\right)^{-1}, \tag{4.14}$$

i.e. the semiparametric efficiency bound for this model (Chamberlain [**1992**], Cosslett [**1987**]).

**The Generalized Regression Model and MRC.** If the distribution of $Y$ depends on $X$ only through the single index $Z$, as e.g. in the generalized regression model with $\varepsilon$ independent of $X$, the function $\lambda(y,u,z)$ corresponding to MRC,

$$\lambda(y,u,z) = E[sign(Y-y)|X = x(u,z)],$$

does not depend on $u$, and one can find the optimal weighted MRC estimator. In this case,

$$\begin{aligned}
\lambda(y,z) &= 1-2\tilde{G}_{Y|Z}(y;z), \\
\mu(z) &= -2E\left[\partial_z \tilde{G}_{Y|Z}(Y;z)\,|Z=z\right], \\
\sigma^2(z) &= E\left[\left(2\tilde{G}_{Y|Z}(Y;z)-1\right)^2|Z=z\right],
\end{aligned}$$

where

$$\tilde{G}_{Y|Z}(y;z) = P\{Y<y|Z=z\} + P\{Y=y|Z=z\}/2.$$

The asymptotic variance of the optimal weighted MRC is given by (4.8), with $\tilde{U}(Z) = E[U|Z]$.

If the outcome $Y$ has a continuous distribution, $\tilde{G}_{Y|Z}(y;z)$ is just the c.d.f. of $Y$ conditional on $Z=z$, $G_{Y|Z}(y;z)$. In this case,

$$\sigma^2(z) = \int \left(2G_{Y|Z}(Y;z)-1\right)^2 dG(Y;z) = \frac{1}{3}$$

(this has been noticed by Sherman [**1993**]). The variance of the optimal MRC becomes

$$\Gamma_*^{MRC} = \left(E\left[12\left(E\left[\partial_z G_{Y|Z}(Y;Z)\right]\right)^2 (U-E[U|Z])(U-E[U|Z])'\right]\right)^{-1}.$$

Further simplification is possible in more specific models with continuously distributed $Y$. In the monotone regression model,

$$Y = f(X'\beta) + \varepsilon,$$

with $\varepsilon$ independent of $X$,

$$\mu(z) = E[g_\varepsilon(\varepsilon)] f'(z).$$

In the computation of the optimal MRC, the scalar multiplier $E[g_\varepsilon(\varepsilon)]$ does not have to be estimated, and one can use the same weighting function as in MR. The optimal variance is

$$\frac{1}{12(E[g_\varepsilon(\varepsilon)])^2}\left(E\left[f'(z)^2(U - E[U|Z])(U - E[U|Z])'\right]\right)^{-1}.$$

In the case of the transformation model:

$$f(Y) = X'\beta_0 + \varepsilon,$$

with $\varepsilon$ independent of $X$, the MRC estimator with the weighting function defined via

$$w_0(z) = \frac{1}{g(z)}$$

attains the optimal variance,

$$\frac{1}{12(E[g_\varepsilon(\varepsilon)])^2}\left(E\left[(U - E[U|Z])(U - E[U|Z])'\right]\right)^{-1}.$$

In the special case of the binary choice model, the expressions for $\mu(z)$ and $\sigma^2(z)$ are the same as for MR, and the weighted MRC attains the semiparametric efficiency bound given in (4.14) (for the binary choice model MRC and MR with the same weighting

function are numerically equivalent[2]). However, it generally does not attain the semiparametric efficiency bound in other models where the distribution of $Y$ is allowed to depend on $X$ only through the single index $Z = U'\theta_0 + V$, for an unknown $\theta_0$. The semiparametric efficiency bound for estimators of $\theta_0$ for this model is given by[3]

$$\Gamma^{eff} = \left( E\left[ \left( \frac{\partial_z g_{Y|Z}(Y;Z)}{g_{Y|Z}(Y;Z)} \right)^2 (U - E[U|Z])(U - E[U|Z])' \right] \right)^{-1}, \tag{4.15}$$

where $g_{Y|Z}(y;z)$ is the conditional density of $Y$ given $Z$ relative to a dominating measure (discrete, continuous, or mixed). This expression is different from the expression for the asymptotic variance of the optimal weighted MRC, in general.

**The Heteroskedastic Generalized Regression Model and MRC.** Let $D$, $F$, $X$ and $\varepsilon$ be as in the generalized regression model (1.2), and $Y$ is related to $X$ as either

$$Y = F(X'\beta_0, \varepsilon + \xi)$$

or

$$Y = D \circ F(X'\beta_0, \varepsilon) + \xi,$$

where, conditionally on $X$, $\xi$ is independent of $\varepsilon$ and symmetric around zero. Since the variance of $\xi|X$ is unrestricted, the first model can be used to introduce heteroskedasticity

[2]This has been shown by Cavanagh and Sherman [**1998**] for the unweighted MRC and MR, and can be shown for the weighted estimators by a similar derivation.

[3]For the model specified as the $g_{Y|X}(y;x) = g(y;z)$, where $z = u'\theta_0 + v$, $g$ is an unknown density function for each $z$, and $\theta_0$ is an unknown, finite dimensional parameter, the score for $\theta_0$ is $\frac{\partial_z g(Y;Z)}{g(Y;Z)} U$, the tangent set in the non-parametric dimension is $\{\varphi(Y;Z) | E[\varphi(Y;Z)|Z] = 0\}$ and the efficient score is $\frac{\partial_z g(Y;Z)}{g(Y;Z)} (U - E[U|Z])$.

into the error term $\varepsilon + \xi$. The second model can be used, for example, to represent heavy-tailed errors of measurement in $Y$[4]. Under these assumptions the distribution of $(Y, X)$ satisfies the identification condition (1.1) and, assuming that the regularity conditions given in Section 4.2.1 hold, the vector of coefficients $\beta_0$ can be consistently estimated, up to scale, by MRC. Thus, MRC allows for practically relevant deviations from the independence and single-index assumptions in the generalized regression model. Note also that, in general, the first model does not satisfy the identification condition of the monotone regression model, and cannot be consistently estimated by MR. In other words, MRC and MR are consistent under non-nested specifications.

In this example, the function $\lambda(y, u, z)$ generally depends on $u$ and the weighted MRC with the weighting function given in Section 4.2.2 may not have a smaller asymptotic variance than the unweighted MRC. As a practical matter, one can always compute both estimators and choose the one with the smaller estimated variance.

### 4.2.4. Feasible Weighted Rank Estimators

The optimal weighted rank estimators described above cannot be computed since they depend on $\theta_0$ and the unknown functions $\mu$, $\sigma^2$ and $g$. In a feasible procedure these objects have to be estimated. Under regularity conditions, the statistical uncertainty associated with such estimates does not affect the asymptotic distribution of the weighted rank estimators, as we show in this section.

[4]MR allows for non-symmetric $\xi$ as long as $E\left[\xi^2\right] < \infty$ . However, in the case of heavy-tailed symmetric distributions of $\xi$ (e.g. if $\xi$ models outliers), MRC is more efficient and has finite variance even if $E\left[\xi^2\right] = \infty$.

To obtain the feasible weighted rank estimator, one needs to compute a root-$n$-consistent estimator of $\theta_0$, $\theta_{0n}$, e.g. the unweighted rank estimator given by (4.2), and the estimated single index:

$$\hat{Z}_i = U_i'\theta_{0n} + V_i.$$

The optimal weighting functions given in (4.10)-(4.13) can be estimated using an appropriate representation for the functions $\mu$ and $\sigma^2$ obtained in Section 4.2.3 and an estimate of the probability density function of $Z$. In specific cases some of these functions need not be estimated altogether (such as in the case of MRC applied to continuous outcome models, where the function $\sigma^2(X)$ does not have to be estimated, or to the transformation model, where neither $\mu$ nor $\sigma^2$ have to estimated).

In this section we give a generic estimator of the function $w_0(X)$ for the case $\sigma^2(X) = \sigma^2(Z)$ and bounded $M$, without making further assumptions on the data (e.g. continuous vs. discrete outcomes), or specifying the function $M$ explicitly (so that our results are also applicable to the estimators other than MR and MRC)[5]. We will show that the corresponding feasible optimal rank estimators, with additive weights, are asymptotically equivalent to the optimal rank estimators. To avoid cumbersome notation, we do not present a proof of this result for other implementations of the feasible optimal rank estimators; however, one can obtain such proofs by replicating our techniques presented in Appendix, Section B.2.

The unknown functions will be estimated using kernel regressions. We impose the following regularity conditions on the kernel function $\phi(x)$ in these regressions.

[5]Among the presently existing rank estimators, the case of $\sigma^2(X) \neq \sigma^2(Z)$ and unbounded $M$ is only restrictive for MR, for which the estimation of the optimal weighting function is straightforward. We do not consider this case in order to simplify the regularity conditions and proofs.

**Assumption 18.** *The kernel function $\phi(x)$ has a finite support, is bounded in absolute value, twice continuously differentiable and symmetric around zero, has no more than a finite number of local maxima, and integrates to 1.*

These conditions are standard. In particular, they allow for higher order kernels. The condition that $\phi$ have a finite support is made here for simplicity and can be relaxed at the expense of more complicated proofs.

Write the function $w_0(z)$ as:

$$w_0(z) \equiv \frac{\mu(z)}{g(z)\sigma^2(z)} = \frac{g^2(z)\mu(z)}{g^3(z)\sigma^2(z)}.$$

The numerator can be estimated as

$$\widehat{g^2(z)\mu(z)} = \frac{1}{n^2 h_\mu^3} \sum_{i,j} M(Y_i, Y_j)\phi_{\mu i}(z)\phi'_{\mu j}(z),$$

with

$$\begin{aligned} \phi_{\mu i}(z) &= \phi\left(\frac{z - \hat{Z}_i}{h_\mu}\right), \\ \phi'_{\mu i}(z) &= \phi'\left(\frac{z - \hat{Z}_i}{h_\mu}\right). \end{aligned}$$

and a positive bandwidth $h_\mu$.

By the definition of the function $\sigma^2(z)$,

$$\sigma^2(z) = E\left[M(Y_1, Y_2) M(Y_1, Y_3) \,|X_1'\beta_0 = X_2'\beta_0 = X_3'\beta_0 = z\right],$$

where labels $1, 2, 3$ denote independent observations.

A consistent estimator of $\sigma^2(z) g^3(z)$ can be constructed as

$$\begin{aligned} & \widehat{\sigma^2(z) g^3}(z) \\ = & \frac{1}{n^3 h_\sigma^3} \sum_i \sum_{j \neq k} M(Y_i, Y_j) M(Y_i, Y_k) \phi_{\sigma i}(z) \phi_{\sigma j}(z) \phi_{\sigma k}(z), \end{aligned}$$

where

$$\phi_{\sigma i}(z) = \phi\left(\frac{z - \hat{Z}_i}{h_\sigma}\right),$$

and $h_\sigma$ is a bandwidth.

Therefore, the weighting function can be estimated by

$$\hat{w}_0(z) = \frac{\frac{1}{n^2 h_\mu^3} \sum_{i,j} M(Y_i, Y_j) \phi_{\mu i}(z) \phi'_{\mu j}(z)}{\tau + \frac{1}{n^3 h_\sigma^3} \sum_i \sum_{j \neq k} M(Y_i, Y_j) M(Y_i, Y_k) \phi_{\sigma i}(z) \phi_{\sigma j}(z) \phi_{\sigma k}(z)},$$

where a small constant $\tau > 0$ is introduced to ensure integrability of $w_0(z)$. It is worth noting that even though this expression involves double and triple sums, one can compute the function $\hat{w}_0(z)$ in $O(n \log n)$ operations for each $z$, and in $O(n^2)$ operations for the entire sample of $z$. In Appendix, Section C.3, we provide an example of such algorithm for MRC.

**Theorem 8.** *Let Assumptions 11-18 hold, the function $M(y_1, y_2)$ is bounded, $\theta_{0n}$ is a root-n-consistent estimator of $\theta_0$, the bandwidths satisfy the conditions $h_\sigma, h_\mu \to 0$, $n^{1/6} h_\sigma, n^{1/6} h_\mu \to \infty$, and $\tau > 0$. Denote by $\theta_n$ the solution to the optimization problem (4.3) with the weights $\hat{w}(X_i, X_j) = \hat{w}_0\left(\hat{Z}_i\right) + \hat{w}_0\left(\hat{Z}_j\right)$. Then*

$$\theta_n \xrightarrow{p} \theta_0,$$

*and*

$$n^{1/2}\left(\theta_n - \theta_0\right) \xrightarrow{d} N\left(0, 4\Delta_{w_\tau}^{-1} V_{w_\tau} \Delta_{w_\tau}^{-1}\right),$$

*where* $w_\tau\left(z_1, z_2\right) = \frac{g^2(z_1)\mu(z_1)}{\tau + g^3(z_1)\sigma^2(z_1)} + \frac{g^2(z_2)\mu(z_2)}{\tau + g^3(z_2)\sigma^2(z_2)}$.

## 4.3. Numerical Examples

Here we provide numerical examples illustrating the asymptotic theory exposed in the previous section. In the first set of examples we consider a submodel of the generalized regression model with continuously distributed outcomes and the error term independent of regressors. Both MRC and MR are consistent, but neither attains the semiparametric efficiency bound (computed assuming that $\varepsilon$ is independent of $X$) in this model. On the other hand, both MRC and MR are also consistent under certain (non-nested) deviations from the independence assumption. In these examples, we compare the asymptotic variances of the optimal MRC and MR, and the efficiency bound under independence, to identify situations in which the flexibility allowed by MRC and MR is costly. In the second set of examples we compare the asymptotic variance of the unweighted MRC (MR) with that of the optimal MRC (respectively, MR), for various models with exactly specified features, to find the conditions under which the optimal weighting can deliver tangible efficiency improvements. Next, we provide an example of the finite-sample performance of the feasible optimal MRC and MR. Finally, we illustrate our conclusions in a real data application.

### 4.3.1. Efficiency Comparisons for Exact Distributions

**Optimal Rank Estimators and a Semiparametric Efficiency Bound.** In our first illustration we compare the asymptotic variance of the optimal MRC with the semiparametric efficiency bound (4.15) in the transformation-monotone regression model:

$$h(Y) = f(Z) + \varepsilon,$$

where $f$ and $h$ are strictly monotone, differentiable functions, $Z = U'\theta_0 + V$ is the single index and $\varepsilon$ is independent of $X = (U, V)$ and has a differentiable Lebesgue density. In this case, the asymptotic variance of the optimal weighted MRC is

$$\Gamma_*^{MRC} = \frac{1}{12E\left[g_\varepsilon(\varepsilon)\right]^2}\left(E\left[f'(Z)^2\left(U - E[U|Z]\right)\left(U - E[U|Z]\right)'\right]\right)^{-1},$$

while the semiparametric efficiency bound is

$$\Gamma^{eff} = \frac{1}{E\left[\left(\frac{g_\varepsilon'(\varepsilon)}{g_\varepsilon(\varepsilon)}\right)^2\right]}\left(E\left[f'(Z)^2\left(U - E[U|Z]\right)\left(U - E[U|Z]\right)'\right]\right)^{-1}$$

(neither depends on the function $h$). Therefore,

$$\Gamma_*^{MRC} = \Gamma^{eff}\frac{E\left[\left(\frac{g_\varepsilon'(\varepsilon)}{g_\varepsilon(\varepsilon)}\right)^2\right]}{12\left(E\left[g_\varepsilon(\varepsilon)\right]\right)^2} \equiv \Gamma^{eff}\cdot\kappa_{MRC/eff}^2.$$

To give a numerical sense of the loss of efficiency involved, we computed the values of the coefficient $\kappa_{MRC/eff}$ (corresponding to the ratio of the standard deviations rather

than variances) assuming that $\varepsilon$ has the density from the Subbotin's [**1923**] family:

$$g_\alpha(e) \propto e^{-|e/\omega|^\alpha},\ e \in \mathbb{R},$$

where $\alpha$ is a positive parameter and $\omega$ is a positive scaling constant (related to the variance of the distribution). This family of distributions includes the normal distribution ($\alpha = 2$) and the double exponential distribution ($\alpha = 1$). It is worth noting that the coefficient $\kappa_{MRC/eff}$ does not depend on either the mean or the variance of $\varepsilon$, but only on the shape of its density function. In Figure 4.1 we plotted the densities $g_\alpha$ for various $\alpha$, choosing the scaling factor $\omega$ so that the variance of $\varepsilon$ is 1. One can see that for $\alpha$ close to zero the densities are steep near the origin, while for large values of $\alpha$ they are steep in the tails. The limit $\alpha \to \infty$ corresponds to the uniform distribution on a bounded support. In Figure 4.2 we plotted the densities $g_\alpha$ renormalized so that the value of the density at zero is $(2\pi)^{-1/2}$ (e.g. the same as for the standard normal distribution). It is apparent from this figure that the limit $\alpha \to 0$ also corresponds to heavy-tailed distributions.

The value of the coefficient $\kappa_{MRC/eff}$ for this family of distributions can be computed explicitly and is given by

$$\kappa_{MRC/eff}(\alpha) = \left(\frac{4^{\frac{1}{\alpha}}}{3}\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(2-\frac{1}{\alpha}\right)\right)^{1/2},$$

where $\Gamma(p)$ is the Gamma function. The function $\kappa_{MRC/eff}(\alpha)$ is plotted in Figure 4.3. The efficiency loss is small for the double exponential and the normal distributions. The efficiency loss is large when there are regions in the support of the distribution of $\varepsilon$ in which the density changes fast (the value of the ratio $\frac{g'_\varepsilon(\varepsilon)}{g_\varepsilon(\varepsilon)}$ is high). In this case

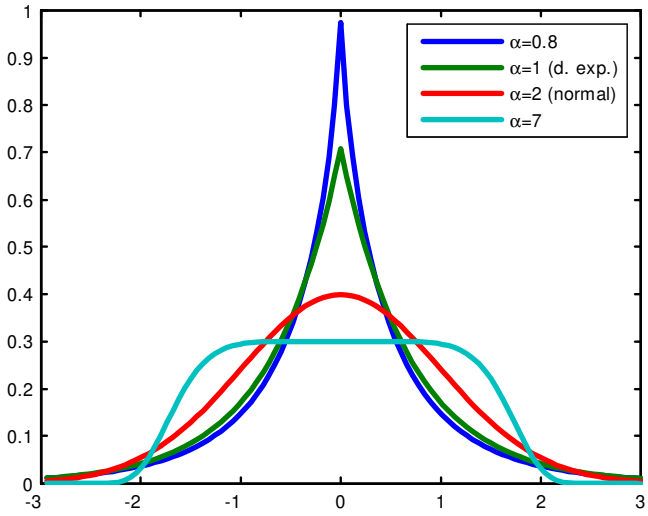


Figure 4.1. Probability densities $g_\alpha$, variances normalized to 1

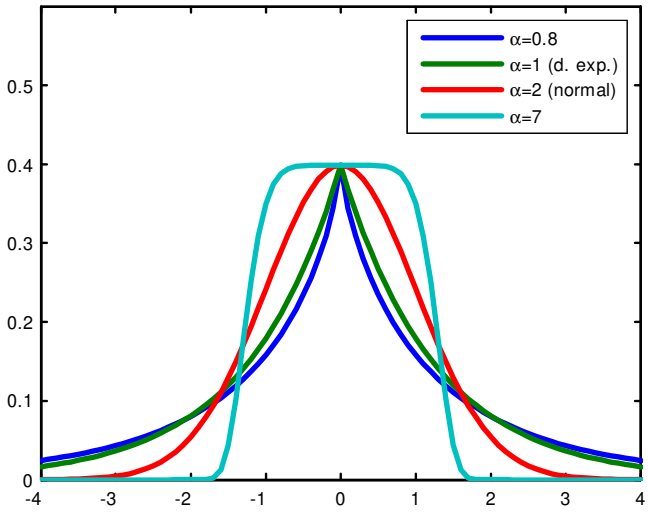


Figure 4.2. Probability densities $g_\alpha$, modes normalized to $(2\pi)^{-1/2}$

a maximum likelihood approach based on the model $Y|X \sim Y|Z$ can give substantial efficiency improvements over MRC (but will not be consistent from deviations of this model) as long as it correctly picks up this feature of the distribution of $\varepsilon$.

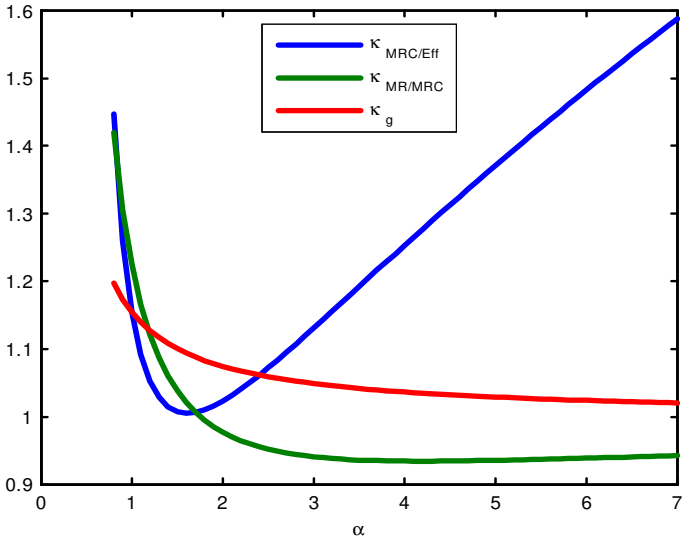


Figure 4.3. The functions $\kappa_{MRC/Eff}(\alpha)$, $\kappa_{MR/MRC}(a)$ and $\kappa_g(\alpha)$

Next, we compare the asymptotic variances of the optimal MRC and MR in the monotone regression model (i.e. $h \equiv 1$). The optimal variance of MR is

$$\Gamma_*^{MR} = Var(\varepsilon) \cdot \left(E\left[f'(Z)^2 (U - E[U|Z])(U - E[U|Z])'\right]\right)^{-1},$$

and, therefore,

$$\Gamma_*^{MR} = \Gamma_*^{MRC} \cdot 12E[g_\varepsilon(\varepsilon)]^2 Var(\varepsilon) \equiv \Gamma_*^{MRC} \cdot \kappa_{MR/MRC}^2.$$

For the family of densities introduced above,

$$\kappa_{MR/MRC}(\alpha) = \left(\frac{3\alpha^2}{4^{\frac{1}{\alpha}}} \frac{\Gamma\left(\frac{3}{\alpha}\right)}{\left(\Gamma\left(\frac{1}{\alpha}\right)\right)^3}\right)^{1/2}.$$

The graph of the function $\kappa_{MR/MRC}(\alpha)$ is shown in Figure 4.3. MRC has lower asymptotic variance if the distribution of $\varepsilon$ has heavy tails (low $\alpha$, e.g., $\alpha = 1$), and if there are regions

in the support of $\varepsilon$ where the density of $\varepsilon$ takes high values. However, one can also see that the asymptotic variance of MRC can be larger than that of MR ($\alpha \to \infty$ in the graph).

As noted above, in the binary choice model, the two estimators attain the semiparametric efficiency bound. As the binary choice model and the continuous outcome model can be considered as extreme cases of the models with censoring, one can expect that estimating the models with censoring by MRC or MR will lead to a loss of efficiency, of the order of magnitude comparable or smaller than the loss of efficiency in models with continuously distributed outcomes.

**Unweighted vs. Optimal Rank Estimators.** In this section we compare the asymptotic variances of unweighted rank estimators with those of the optimal rank estimators. Throughout, we maintain the following assumptions: the function $\lambda(y, u, z)$ does not depend on $u$; the vector of the first $d$ regressors, $U$, is distributed independently of the single index, $Z$, and the function $\sigma^2(X)$ depends on $X$ only through the single index $Z$. Then the asymptotic variances are given by

$$Var(U)^{-1} \frac{E\left[\sigma^2(Z) g^2(Z)\right]}{\left(E\left[\mu(Z) g(Z)\right]\right)^2}, \; Var(U)^{-1} \frac{1}{E\left[\frac{\mu^2(Z)}{\sigma^2(Z)}\right]}$$

for the unweighted and the optimal rank estimators, respectively. The variance of the unweighted estimator, therefore, is

$$\kappa^2_{g,\mu,\sigma} = \frac{E\left[\sigma^2(Z) g^2(Z)\right] E\left[\frac{\mu^2(Z)}{\sigma^2(Z)}\right]}{\left(E\left[\mu(Z) g(Z)\right]\right)^2}$$

times bigger than the variance of the optimal estimator.

As the first example, consider the case where the functions $\sigma^2(Z)$ and $\mu(Z)$ are constant (e.g. in the transformation model with independent errors estimated by MRC). In this case

$$\kappa_g^2 = \frac{E\left[g^2(Z)\right]}{E\left[g(Z)\right]^2}.$$

The coefficient $\kappa_g$ is bigger than 1 if the single index $Z$ is unevenly distributed over its support. To illustrate, we computed this coefficient for the Subbotin's family of densities, $g_\alpha(z) \propto e^{-|z|^\alpha}, z \in \mathbb{R}$, as a function of $\alpha$:

$$\kappa_g(\alpha) = \left(\frac{4}{3}\right)^{\frac{1}{2\alpha}}.$$

The graph of the function $\kappa_g(\alpha)$ is shown in Figure 4.3. One can see that substantial deviations from uniformity in the distribution of the single index are needed for noticeable efficiency gains from the optimal weighting.

Next, we consider the case where the functions $\mu(z)$ and $\sigma^2(z)$ are nonconstant. Examples include the transformation-monotone regression model with $\varepsilon$ independent of $X$, estimated by MRC ($\mu(z) \propto f'(z)$, constant $\sigma^2(z)$), other generalized regression models with $\varepsilon$ independent of $X$ estimated by MRC (constant $\sigma^2(z)$), and the monotone regression model estimated by MR ($\mu(z) = f'(z)$, $\sigma^2(z) = Var(Y|Z=z)$). We assume that the single index has a density from the Subbotin's family, $\mu(z) = |z|^{\beta-1}$ for a $\beta > 0$, and $\sigma^2(z) = |z|^\gamma$. Then

$$\kappa_{g,\mu,\sigma}(\alpha,\beta,\gamma) = \left(\frac{4^{\frac{\beta}{\alpha}}}{3^{\frac{2\gamma+1}{\alpha}}} \frac{\Gamma\left(\frac{2\beta-2\gamma-1}{\alpha}\right)\Gamma\left(\frac{2\gamma+1}{\alpha}\right)}{\Gamma\left(\frac{\beta}{\alpha}\right)^2}\right)^{1/2}$$

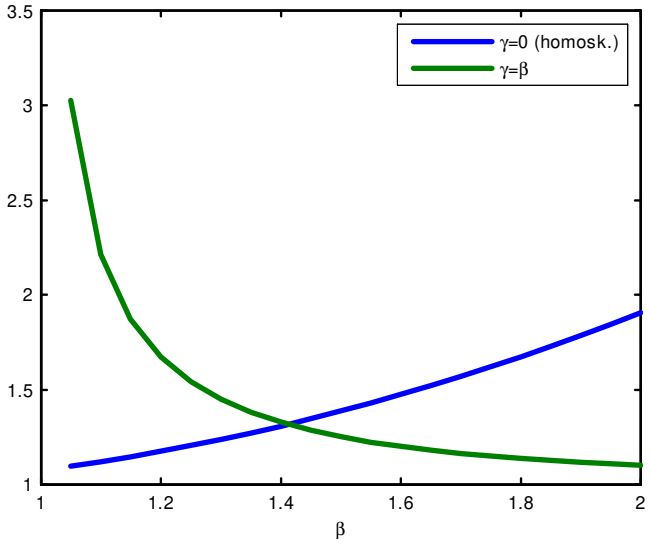


Figure 4.4. The functions $\kappa_{g,\mu,\sigma}(2,\beta,\gamma)$ for $\gamma = 0$ and $\gamma = \beta$

(note that for the power functions $\mu(z)$ and $\sigma^2(z)$, the coefficient $\kappa_{g,\mu}(\alpha,\beta)$ is invariant to changes in the scale of the single index, but not to the changes in the location of its distribution). Figure 4.4 shows the graph of the functions $\kappa_{g,\mu,\sigma}(2,\beta,0)$ and $\kappa_{g,\mu,\sigma}(2,\beta,\beta)$. One can see that, unlike in the previous cases, sizable gains in efficiency are possible with the optimally weighted estimators even with moderate nonlinearities in the data.

Finally, we consider an important special case of the monotone regression model with heteroskedasticity, the binary choice model. The ratio of the variance of the unweighted MR (MRC) estimator to the variance of the optimal estimator (i.e. the semiparametric efficiency bound for the binary choice model under independence) is given by

$$\kappa_{bin}^2 = \frac{E\left[G_\varepsilon(Z)\left(1-G_\varepsilon(Z)\right)g^2(Z)\right]\cdot E\left[\frac{g_\varepsilon(Z)^2}{G_\varepsilon(Z)(1-G_\varepsilon(Z))}\right]}{E\left[g_\varepsilon(Z)\,g(Z)\right]}.$$

Consider the case in which the densities $g_\varepsilon$, $g$ are bounded and have substantial overlap, or, formally, where the ratio

$$\frac{E\left[G_\varepsilon\left(Z\right)\left(1-G_\varepsilon\left(Z\right)\right)g^2\left(Z\right)\right]}{E\left[g_\varepsilon\left(Z\right)g\left(Z\right)\right]}$$

is bounded away from zero and infinity. Then one can expect the coefficient $\kappa_{bin}$ to be large if there are regions in the support of $Z$ where $G_\varepsilon\left(z\right)\left(1-G\left(z\right)\right)$ is close to zero and $g_\varepsilon\left(z\right)$ and $g\left(z\right)$ are not. For example, for $\varepsilon$ with a bounded support lying strictly inside the support of the single index, the coefficient $\kappa_{bin}$ can be large if the density of $\varepsilon$ is nonsmooth near the boundaries, e.g., $\kappa_{bin}=+\infty$ for uniformly distributed $\varepsilon$. However, $\kappa_{bin}$ is likely to be moderate if the density of $\varepsilon$ is Lipschitz near the boundaries of the support, in which case the function $\frac{g_\varepsilon(z)^2}{G_\varepsilon(z)(1-G_\varepsilon(z))}$ remains bounded.

In the numerical example, we consider $\varepsilon$ with the c.d.f. $G_{\varepsilon,\alpha}\left(e\right)=B_\alpha^{-1}\left(e\right)$, where

$$B_\alpha\left(t\right)=\int_{1/2}^{t}u^{-\alpha}\left(1-u\right)^{-\alpha}du,\ \alpha\geq 0,\ t\in\left(0,1\right).$$

The density of this distribution is given by

$$g_{\varepsilon,\alpha}\left(e\right)=G_{\varepsilon,\alpha}\left(e\right)^\alpha\left(1-G_{\varepsilon,\alpha}\left(e\right)\right)^\alpha.$$

The function $B_\alpha\left(t\right)$ has finite (infinite) range, and $G_{\varepsilon,\alpha}\left(e\right)$ has finite (infinite) support for $\alpha<1$ ($\alpha\geq 1$). The zero value of $\alpha$ corresponds to the uniform distribution of $\varepsilon$. The value $\alpha=1$ corresponds to the logistic distribution with the c.d.f. $\left(1+e^{-z}\right)^{-1}$. For $0<\alpha<\frac{1}{2}$, the density $g_{\varepsilon,\alpha}\left(e\right)$ converges to zero near the boundaries of the support, but its derivative diverges to $\infty$ in absolute value. For $\alpha\geq\frac{1}{2}$, the derivative of the density is

bounded. The density of the single index is specified as

$$g_{\alpha,\beta}(z) = c_{\alpha,\beta}^{-1} \cdot G_{\varepsilon,\alpha}(e)^{\beta} (1 - G_{\varepsilon,\alpha}(e))^{\beta}, \ \beta > \alpha - 1,$$

where $c_{\alpha,\beta} = \int G_{\varepsilon,\alpha}(e)^{\beta} (1 - G_{\varepsilon,\alpha}(e))^{\beta} de$ is a normalization constant. Therefore, in this example, the single index has the same support (finite or infinite) as the error term, while $\beta$ measures relative thickness of the two densities near the boundaries of the support (or at infinity). For these distributions,

$$\kappa_{bin}(\alpha, \beta) = \left( \frac{\tilde{B}(\alpha + \beta) \tilde{B}(3\beta + 2 - \alpha)}{\tilde{B}(2\beta + 1)^2} \right)^{1/2},$$

where $\tilde{B}(t) = B(t, t)$, and $B(s, t) = \int_0^1 u^{s-1} (1 - u)^{t-1} du$ is the Beta function. Figure 4.5 shows the graphs of the functions $\kappa_{bin}(\alpha, \beta)$ for fixed $\alpha = 0$, 0.5, 1, 1.5. One can see that the efficiency gains from optimal weighting are small except in the case of the uniform distribution of $\varepsilon$ and sufficiently thick density of the single index near the boundaries of the support.

### 4.3.2. Finite-Sample Performance of the Feasible Optimal Rank Estimators

We evaluate the finite-sample performance of the feasible optimal rank estimators in the model:

$$\begin{aligned} Y &= Z^2 sign(Z) + \varepsilon, \\ Z &= \theta_0 X^{(1)} + X^{(2)}, \end{aligned}$$

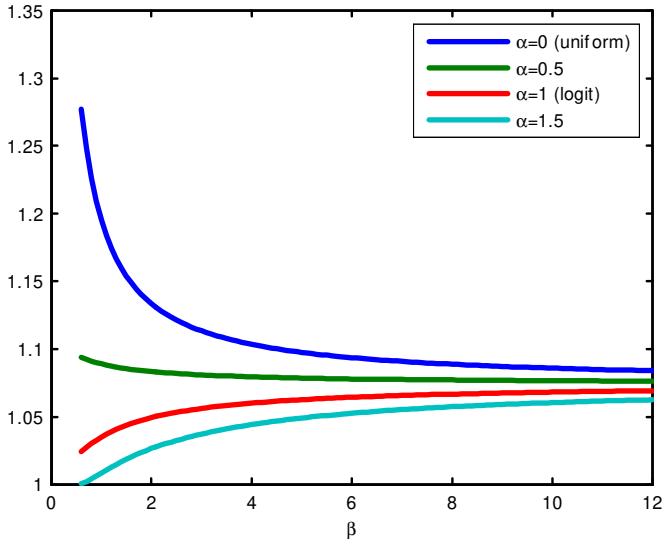


Figure 4.5. The functions $\kappa_{bin}(\alpha, \beta)$ for $\alpha = 0, 0.5, 1$ and 1.5

where the random vector $\left(X^{(1)}, X^{(2)}, \varepsilon\right)$ is distributed according to the standard normal distribution and $\theta_0 = 0$. We estimate the parameter $\theta_0$ by the unweighted and the feasible optimal MR and MRC on 1000 data samples of $n = 1000$ observations. Using the unweighted estimator, $\hat{\theta}_{0n}$, of $\theta_0$, we computed the estimated single index, $\hat{Z} = \hat{\theta}_{0n} X^{(1)} + X^{(2)}$, and estimated the density of the single index, $g(z)$, and the functions $\mu(z) = \partial_z E[Y|Z]$ and $\sigma^2(z) = Var(Y|Z)$ using the kernel regressions with the kernel function given by the density of the standard normal distribution and a bandwidth $h$. The corresponding estimators are denoted by $\hat{g}_h(z)$, $\hat{\mu}_h(z)$, and $\hat{\sigma}_h^2(z)$. The objective function in the optimization problem for the weighted MR is

$$\sum_{i \neq j} \left( \hat{w}_{0h}\left(\hat{Z}_i\right) + \hat{w}_{0h}\left(\hat{Z}_j\right) \right) Y_i 1\left\{ \theta X_i^{(1)} + X_i^{(2)} > \theta X_j^{(1)} + X_j^{(2)} \right\},$$

and the objective function for the weighted MRC is

$$\sum_{i\neq j}\left(\hat{w}_{0h}\left(\hat{Z}_i\right)+\hat{w}_{0h}\left(\hat{Z}_j\right)\right)1\left\{Y_i>Y_j\right\}1\left\{\theta X_i^{(1)}+X_i^{(2)}>\theta X_j^{(1)}+X_j^{(2)}\right\},$$

where $\hat{w}_{0h}(z)=\frac{\hat{\mu}_h(z)}{\hat{\tau}+\hat{g}_h(z)\hat{\sigma}_h^2(z)}$. For additive weights, the MRC and MR objective functions can be evaluated using $O(n\log n)$ operations, which is important for their practical use (the same is true for multiplicative weights; see Appendix, Section C.2, for the numerical algorithms). We computed the feasible optimal MR and MRC estimators for the bandwidths

$$h=kn^{-1/6}\left(\hat{q}_z(.95)-\hat{q}_z(.05)\right),$$

where $\hat{q}_z(.05)$ and $\hat{q}_z(.95)$ are the estimated 5% and 95%-quantiles of the distribution of the single index, and $k$ is a scaling factor, taking ten values in the interval 0.05:0.5. The truncation parameter $\hat{\tau}$ and the truncation parameters in the kernel estimators of the functions $g$, $\mu$, and $\sigma^2$ were set at the level of 0.01 of the mean value of the denominator over the sample. The maximization of the objective functions was performed on a grid of 20,000 points in the interval $[-5\cdot n^{-1/2},5\cdot n^{-1/2}]$. The simulated biases and standard deviations of the statistic $n^{1/2}(\theta_n-\theta_0)$ are reported in Table 4.1. One can see that the biases are low for both the unweighted and weighted estimators, for all considered values of the bandwidth. The weighted estimators provide substantial efficiency gains for all $k$, with the lowest standard deviation of the estimator attained near $k=0.2$ for MR and $k=0.25$ for MRC.

| | MR | | MRC | |
|---|---|---|---|---|
| | bias | st.dev. | bias | st.dev. |
| Unweighted | .070 | 1.009 | .041 | 1.023 |
| Optimal: | | | | |
| $k =$ | | | | |
| .05 | .048 | .809 | .046 | .849 |
| .10 | .031 | .684 | .046 | .734 |
| .15 | .029 | .653 | .039 | .695 |
| .20 | .036 | .649 | .040 | .686 |
| .25 | .037 | .651 | .040 | .682 |
| .30 | .040 | .670 | .040 | .687 |
| .35 | .047 | .683 | .038 | .698 |
| .40 | .047 | .699 | .043 | .723 |
| .45 | .052 | .709 | .039 | .737 |
| .50 | .057 | .721 | .037 | .753 |

Table 4.1. Biases and standard deviations of the unweighted and feasible optimal rank estimators

### 4.3.3. Empirical Example

In this example we study a model of choice between private and public schools in Chile. We use the census data SIMCE 2006 collected by the Ministry of Education of Chile[6], which contains the information on the type of schools attended by grade four students, along with various demographic and income characteristics of the students and their families. Our purpose is to estimate the vector of coefficients $\beta$ in the binary choice model:

$$Y = 1\left\{X'\beta + \varepsilon > 0\right\},$$

where $Y$ is the choice of school (1 if private, 0 if public), and $X$ contains the following variables: $NUP$, the number of people living in the household excluding the child, $INC$, the income category of the family, $CHS$, the binary variable showing if the child has ever received a poverty subsidy from the government, $GEN$, the gender of the child (1 if

[6]I am grateful to Professor Sergio Urzúa at Northwestern University for providing me with this data set.

female), $FIN$ and $MIN$, the binary variables showing, respectively, if the father and the mother belong to an Indian tribe, and $EDU$, the combined years of education attained by the farther and the mother of the child. The estimation sample is restricted to children from urban families, and contains 160,998 observations.

To estimate the model using MR, we need a continuously distributed regressor with a smooth density. Because all variables in this example are discrete, the coefficient $\beta$ is not pointwise identified, up to scale, without further assumptions. To resolve this problem, we adopt a simplified approach. We assume that the education of the parents affects the choice of school via an unobservable, continuously distributed variable $EDU^*$, related to $EDU$ according to the equation:

$$EDU^* = EDU + \eta,$$

with the random term $\eta$ being distributed with a known distribution independently of $EDU$, the other regressors in the model, and $\varepsilon$. Since the discretization step in $EDU$ is equal to 1 unit (year), we specify the distribution function of $\eta$ as $N\left(0, \frac{1}{9}\right)$ to allow for a small overlap between the supports of the distributions $N\left(EDU, \frac{1}{9}\right)$ for consecutive values of $EDU$[7]. Under these assumptions, $\beta_0 = (\theta_0, 1)$ can be consistently estimated, up to scale, from the model:

$$Y = 1\left\{\theta_0' U + EDU + \xi + \varepsilon > 0\right\},$$

[7]As a robustness check, we also estimated the model with $\eta$ distributed uniformly in the interval $\left(-\frac{1}{4}, \frac{1}{4}\right)$. In this specification, the density of $EDU^*$ is discontinuous, and its support consists of non-overlapping intervals around the integer values. The estimated coefficients and their standard errors were very similar to those obtained under the normally distributed $\eta$.

where the vector $U$ contains all regressors except $EDU$, and $\xi$ is a randomly generated noise having the same distribution as $\eta$ independently of $U$, $EDU$ and $\varepsilon$. In this setup the continuously distributed regressor with the coefficient fixed to 1 is $EDU + \xi$.

As a benchmark, we first found the standard logit estimator of $\beta_0$, $\beta_{Logit}$, computed the vector $\theta_{Logit}$ by dividing the components of $\beta_{Logit}$ by the coefficient at $EDU$, and estimated the variance of $\theta_{Logit}$ from the variance of $\beta_{Logit}$ by the delta-method. Next, we found the unweighted MR estimator of $\theta_0$, and the single index, $\hat{Z}$. To perform numerical optimization we used the Nelder-Mead algorithm as described in Appendix, Section C.4. To find the feasible optimal MR, we first estimated the weighting function, which depends on the density of the index, $g(z)$, the conditional mean of the outcome given the index, $E[Y|Z]$, and the function $\mu(z) = \frac{\partial}{\partial z} E[Y|Z = z]$, which in the binary choice model is the density of the error term[8]. We computed these functions, on a grid of $2,000$ values of the single index, for bandwidths $h = 3, 4, 5$ and the truncation parameter in the denominator, $\tau$, equal to $\tau_0$ times the mean of the denominator, for $\tau_0 = 0.02$, $0.05$, and $0.1$. Figure 4.6 shows the density of the single index and the density of the error term (the function $\mu(z)$) for $h = 4$ and $\tau_0 = 0.05$. One can see, in particular, that the estimated density is not symmetric, thus deviating from the logit assumption. To illustrate the range of the weighting functions resulting from our choices of bandwidths and trimming parameters, we plotted them in Figure 4.7 for $h = 3, 5$ and $\tau_0 = 0.02$, $0.1$. Once the weighting functions were found, we computed the weighted MR estimators. Finally, the standard errors of the coefficients were found using the $M$ out of $N$ bootstrap with $M = 10,000$.

[8]In the model with an added noise, the error term $\varepsilon$ is contaminated by the term $\eta - \xi \sim N\left(0, \frac{2}{9}\right)$.

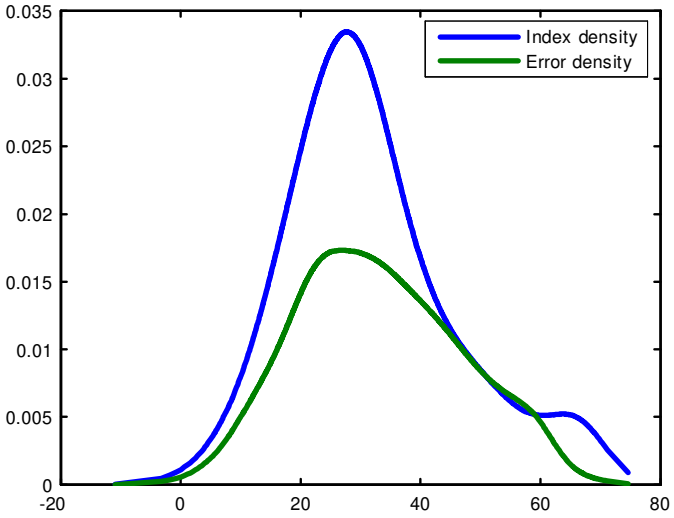


Figure 4.6. School choice equation, the densities of the index and the error term

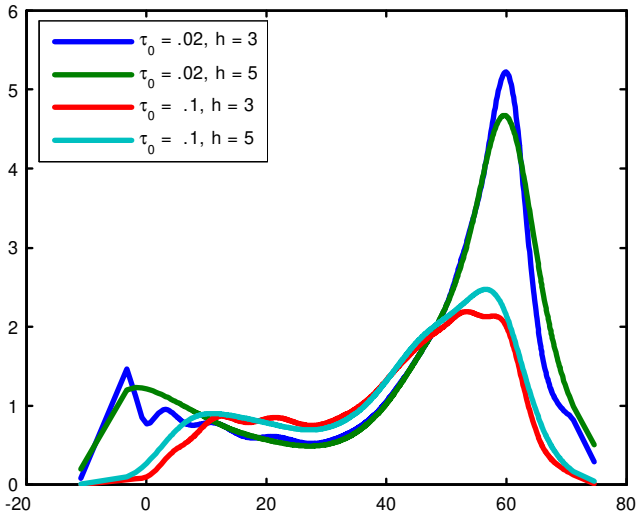


Figure 4.7. School choice equation, the weighting functions

Tables 4.2 and 4.3 report, respectively, the estimates and the standard errors of the coefficients for the logit, unweighted MR, and the feasible optimal MR estimators. It is apparent that estimates obtained by the weighted and unweighted MR agree with each other. However, the logit appears to overestimate the effect of the family income on the

| | NUP | FIN | MIN | INC | CHS | GEN |
|---|---|---|---|---|---|---|
| Logit | -1.2548 | -.4302 | -.6134 | 3.5605 | -3.2781 | .3222 |
| Unweighted MR | -1.1472 | -.6231 | -.8845 | 2.8074 | -3.6328 | .2096 |
| Weighted MR: | | | | | | |
| $\tau = .02$, $h = 3$ | -1.1973 | -.8256 | -1.1200 | 2.6743 | -3.6476 | .2505 |
| $\tau = .02$, $h = 4$ | -1.2003 | -.7287 | -1.2290 | 2.6746 | -3.6599 | .2478 |
| $\tau = .02$, $h = 5$ | -1.2002 | -.7266 | -1.2185 | 2.6735 | -3.6629 | .2467 |
| $\tau = .05$, $h = 3$ | -1.1965 | -.7215 | -1.1032 | 2.6734 | -3.5466 | .2502 |
| $\tau = .05$, $h = 4$ | -1.1972 | -.7059 | -1.1155 | 2.6743 | -3.5908 | .2524 |
| $\tau = .05$, $h = 5$ | -1.2009 | -.7289 | -1.1080 | 2.6758 | -3.6476 | .2474 |
| $\tau = .10$, $h = 3$ | -1.1945 | -.6968 | -1.1094 | 2.6830 | -3.5390 | .2422 |
| $\tau = .10$, $h = 4$ | -1.1939 | -.6965 | -1.1049 | 2.6810 | -3.5651 | .2425 |
| $\tau = .10$, $h = 5$ | -1.1983 | -.7038 | -1.0972 | 2.6803 | -3.5464 | .2398 |

Table 4.2. School choice equation, coefficients

choice of school. The standard errors of the weighted estimators are lower by just a small fraction of the standard errors of the unweighted MR. This suggests that the variance of the unweighted MR estimator is near the semiparametric efficiency bound, the property that has already been observed in our previous examples. Interestingly, the logit estimator has little to offer in terms of reducing the standard errors for most coefficients relative to the more robust semiparametric estimators.

## 4.4. Conclusion

Rank-based estimators are important tools of robust estimation in popular semiparametric models under monotonicity constraints. Using weights in their criteria functions can lead to lower asymptotic variances. We provided conditions under which the optimally weighted rank estimators exist and studied the associated gains in efficiency. Optimal monotone rank estimator exists and attains the semiparametric efficiency bound in the

| | NUP | FIN | MIN | INC | CHS | GEN |
|---|---|---|---|---|---|---|
| Logit | .0526 | .4494 | .4259 | .0847 | .3221 | .1501 |
| Unweighted MR | .0572 | .4897 | .4837 | .1174 | .3556 | .1375 |
| Weighted MR: | | | | | | |
| $\tau = .02$, $h = 3$ | .0551 | .4784 | .4732 | .1021 | .3434 | .1301 |
| $\tau = .02$, $h = 4$ | .0550 | .4833 | .4740 | .1029 | .3451 | .1316 |
| $\tau = .02$, $h = 5$ | .0551 | .4809 | .4746 | .1027 | .3465 | .1323 |
| $\tau = .05$, $h = 3$ | .0557 | .4800 | .4731 | .1041 | .3453 | .1322 |
| $\tau = .05$, $h = 4$ | .0553 | .4800 | .4746 | .1040 | .3454 | .1323 |
| $\tau = .05$, $h = 5$ | .0555 | .4857 | .4760 | .1037 | .3469 | .1325 |
| $\tau = .10$, $h = 3$ | .0559 | .4809 | .4771 | .1057 | .3460 | .1326 |
| $\tau = .10$, $h = 4$ | .0558 | .4815 | .4765 | .1058 | .3461 | .1323 |
| $\tau = .10$, $h = 5$ | .0554 | .4799 | .4775 | .1056 | .3473 | .1323 |

Table 4.3. School choice equation, standard errors

nonlinear regression model and the binary choice model. Optimal maximum rank correlation estimator exists in single-index models with independent errors, has the asymptotic variance close to the semiparametric efficiency bound when the distribution of the errors is close to normal, and is consistent under practically relevant deviations from the single index assumption. Under moderate nonlinearities and nonsmoothness in the data, the efficiency gains from weighting are likely to be small for MRC in the transformation model and for MRC and MR in the binary choice model, and can be large for MRC and MR in the monotone regression model.

# References

Abadie A., and Imbens, G. (2006). On the failure of the bootstrap for matching estimators, mimeo.

Abrevaya, J. (1999). Abrevaya, Jason. Computation of the maximum rank correlation estimator. Econom. Lett. 62, no. 3, 279–285.

Abrevaya, J. (1999a). Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. Journal of Econometrics, 93, 203-228.

Abrevaya, J. (1999b). Rank estimation of a transformation model with observed truncation. Econometric J., 2, 292-305.

Abrevaya, J. (2003). Pairwise-difference rank estimation of the transformation model. J. Bus. Econom. Statist. 21, no. 3, 437–447.

Ai, C.; Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, no. 6, 1795–1843.

Arcones, M. A. (1995). The asymptotic accuracy of the bootstrap of U-quantiles. Ann. Statist. 23, no. 5, 1802–1822.

Arcones, M. A.; Chen, Z.; Giné, E. (1994). Estimators related to U-processes with applications to multivariate medians: asymptotic normality. Ann. Statist. 22, no. 3, 1460–1477.

Arcones, M. A.; Giné, E. (1992). On the bootstrap of U and V-statistics. Ann. Statist. 20, no. 2, 655–674.

Arcones, M. A.; Giné, E. (1993). Limit theorems for U-processes. Ann. Probab. 21, no. 3, 1494–1542.

Arcones, M. A.; Giné, E. (1994). U-processes indexed by Vapnik-Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. Stochastic Process. Appl. 52, no. 1, 17–38.

Asparouhova, E.; Golanski, R.; Kasprzyk, K.; Sherman, R. P.; Asparouhov, T. (2002). Rank estimators for a transformation model. Econometric Theory 18, no. 5, 1099–1120.

Bhattacharya, R. N.; Ranga Rao, R. (1976). Normal approximation and asymptotic expansions. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-London-Sydney. xiv+274 pp.

Bickel, Peter J.; Freedman, David A. (1981). Some asymptotic theory for the bootstrap. Ann. Statist. 9, no. 6, 1196–1217.

Bolthausen, E.; Götze, F. (1993). The rate of convergence for multivariate sampling statistics. Ann. Statist. 21, no. 4, 1692–1710.

Cavanagh, Christopher; Sherman, Robert P. (1998). Rank estimators for monotonic index models. J. Econometrics 84, no. 2, 351–381.

Chamberlain, Gary (1992). Efficiency bounds for semiparametric regression. Econometrica 60, no. 3, 567–596.

Chen, S. (2002). Rank estimation of transformation models. Econometrica 70, no. 4, 1683–1697.

Cosslett, Stephen R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. Econometrica 55, no. 3, 559–585.

De Angelis, D.; Hall, Peter; Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in $L^1$ regression. J. Amer. Statist. Assoc. 88, no. 424, 1310–1316..

De la Peña, V. H. (1992). Decoupling and Khintchine's inequalities for U-statistics. Ann. Probab. 20, no. 4, 1877–1892.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Ann. Statist. 7, no. 1, 1–26.

Giné, E.; Mason, D. M. (2007). On local U–statistic processes and the estimation of densities of functions of several sample variables. Annals of Statistics, to appear.

Giné E. and Zinn J. (1990). Bootstrapping general empirical measures. Annals of Probability 18, 851-869.

Giné, Evarist; Zinn, Joel (1992). On Hoffmann-Jørgensen's inequality for $U$-processes. Probability in Banach spaces, 8 (Brunswick, ME, 1991), 80–91, Progr. Probab., 30, Birkhäuser Boston, Boston, MA.

Hall, P. (1992). The bootstrap and Edgeworth expansion. Springer Series in Statistics. Springer-Verlag, New York.

Hall, Peter; Horowitz, Joel L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. Econometrica 64, no. 4, 891–916.

Han, Aaron K. (1987). Nonparametric analysis of a generalized regression model. The maximum rank correlation estimator. J. Econometrics 35, no. 2-3, 303–316.

Han, Aaron K. (1987b). A nonparametric analysis of transformations. J. Econometrics 35, no. 2-3, 191–209.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J. Econometrics 58, no. 1-2, 71–120.

Khan, Shakeeb; Tamer, Elie (2007). Partial rank estimation of duration models with general forms of censoring. J. Econometrics 136, no. 1, 251–280.

Klein, Roger W.; Spady, Richard H. (1993). An efficient semiparametric estimator for binary response models. Econometrica 61, no. 2, 387–421.

Lee, M. (1999) A root-N consistent semiparametric estimator for related-effect binary response panel data. Econometrica, 67, 427–433.

Nishiyama, Y.; Robinson, P. M. (2005). The bootstrap and the Edgeworth correction for semiparametric averaged derivatives. Econometrica 73, no. 3, 903–948.

Newey, Whitney K.; Stoker, Thomas M. Efficiency of weighted average derivative estimators and index models. Econometrica 61 (1993), no. 5, 1199–1223.

Nolan, Deborah; Pollard, David (1987). U-processes: rates of convergence. Ann. Statist. 15, no. 2, 780–799.

Pakes, Ariél; Pollard, David (1989). Simulation and the asymptotics of optimization estimators. Econometrica 57, no. 5, 1027–1057.

Pollard, D. (1985). New ways to prove central limit theorems. Econometric Theory, 1, 295–313.

Pollard, David (1989). Asymptotics via empirical processes. With comments and a rejoinder by the author. Statist. Sci. 4, no. 4, 341–366.

Powell, James L.; Stock, James H.; Stoker, Thomas M. (1989). Semiparametric estimation of index coefficients. Econometrica 57, no. 6, 1403–1430.

Ruud, P. A. (2000). An introduction to classical econometric theory. Oxford, U.K.: Oxford University Press.

Serfling, Robert J. (1980). Approximation theorems of mathematical statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. xiv+371 pp.

Sherman, Robert P. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica 61, no. 1, 123–137.

Sherman, Robert P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. Ann. Statist. 22, no. 1, 439–459.

Subbotin, M.T. (1923). On the law of frequency of errors. Matematicheskii Sbornik, 31, 296-301.

van der Vaart, Aad W.; Wellner, Jon A. (1996). Weak convergence and empirical processes. With applications to statistics. Springer Series in Statistics. Springer-Verlag, New York. xvi+508 pp.

APPENDIX A

# Appendix to Chapters 2 and 3

Here we provide the proofs of Theorems 1-5 presented in Chapters 2 and 3. Due to mathematical complexity and length, the derivation is divided into four steps. First, we discuss the structure and the principal ingredients of our analysis (Section A.1). The actual proofs are given in Section A.2. Section A.3 provides an overview of the empirical process theory for $U$-processes, with necessary extensions. Section A.4 contains an auxiliary lemma on the Berry-Esséen bound for $M$-estimators with a criterion function in the form of a smooth $U$-process.

## A.1. Main Tools of Proof

This subsection describes the main ideas underlying the proofs of Theorems 1-5. The essence of the analysis is to separate a smooth and a nonsmooth components of the objective function. The estimator $\theta_n$ is approximated by a maximizer of the smooth component whose properties can be studied using the Taylor expansion and the Berry-Esséen bounds. Then the empirical process theory for $U$-processes is used to show that the effect of the nonsmooth remainder in the objective function on the distribution of $\theta_n$ is negligible.

To simplify notation, we assume, without loss of generality, that $\theta_0 = 0$, and that the function $h_0$ is identically zero.

### A.1.1. Approximation

Consider an estimator, $\theta_n$, that solves the problem

$$G_{n,\theta_n} \geq \sup_{\theta \in \Theta} \left[ G_{n,\theta} - r_{n,\theta} \right],$$

and assume that the objective function $G_{n,\theta}$ admits the representation

$$G_{n,\theta} = G^0_{n,\theta} + \zeta_{n,\theta}, \tag{A.1}$$

where $\theta \in \Theta \subset \mathbb{R}^d$, $G^0_{n,\theta}$ is a smooth random function of $\theta$, and $\zeta_{n,\theta}$ is a remainder. An approximation to $\theta_n$, denoted by $\eta_n$, solves the problem

$$\eta_n \in \arg\max_{\theta \in \Theta} G^0_{n,\theta}. \tag{A.2}$$

If the remainder terms $\zeta_{n,\theta}$, $r_{n,\theta}$ are small in an appropriate sense, then the difference $n^{1/2}(\theta_n - \eta_n)$ will also be small. The following theorems formalize this idea.

The first theorem is useful for establishing the asymptotic normality of $\theta_n$ (part (a)), and estimating its variance (part (b)). Here it is enough to consider the representation (A.1) with

$$G^0_{n,\theta} \equiv \theta' W_n - \frac{1}{2}\theta' A \theta,$$

where $W_n$ is a $d \times 1$ random vector, not depending on $\theta$, and $A$ is a matrix of constants. Then $\eta_n = A^{-1} W_n$, as long as the vector on the right-hand side is an element of $\Theta$. The first part of the theorem is a variant of Pollard's [**1985**] asymptotic normality theorem (see also Sherman [**1993**] and Arcones, Chen, Giné [**1994**]), and the second part is a simple extension.

**Theorem 9.** *Assume that* $0$ *is an interior point of* $\Theta$*, and* $A$ *is a symmetric, positive definite, constant matrix. (a) If* $\theta_n \to^p 0$*,* $W_n = O_p\left(n^{-1/2}\right)$*, and for every sequence of numbers* $\delta_n \to +0$*,*

$$\sup_{\|\theta\| \le \delta_n} \frac{\left|\zeta_{n,\theta}\right| + |r_{n,\theta}|}{n^{-1} + \|\theta\|^2} \to^p 0, \tag{A.3}$$

*then*

$$n^{1/2}\left(\theta_n - A^{-1}W_n\right) \to^p 0.$$

*(b) If, additionally,* $\Theta$ *is a bounded set,* $P\|W_n\|^2 = O\left(n^{-1}\right)$*, and for every* $\varepsilon > 0$*, and every sequence of numbers* $\delta_n \to +0$*,*

$$P\left\{\|\theta_n\| > \varepsilon\right\} = o\left(n^{-1}\right),$$

$$P\left\|W_n^2\right\| 1\left\{\|W_n\| > \varepsilon\right\} = o\left(n^{-1}\right),$$

*and*

$$P\left\{\sup_{\|\theta\| \le \delta_n} \frac{\left|\zeta_{n,\theta}\right| + |r_{n,\theta}|}{n^{-1} + \|\theta\|^2} > \varepsilon\right\} = o\left(n^{-1}\right), \tag{A.4}$$

*then*

$$P\left\|n^{1/2}\left(\theta_n - A^{-1}W_n\right)\right\|^2 \to 0. \tag{A.5}$$

**Proof.** Denote $t_n = n^{1/2}\left(\theta_n - A^{-1}W_n\right)$. When $A^{-1}W_n \in \Theta$, by the defining property of $\theta_n$ and (A.1),

$$G_{n,\theta_n} \geq G_{n,A^{-1}W_n} - r_{n,A^{-1}W_n},$$

$$\theta_n' W_n - \frac{1}{2}\theta_n' A\theta_n + \zeta_{n,\theta_n} \geq \frac{1}{2}W_n' A^{-1}W_n + \zeta_{n,A^{-1}W_n} - r_{n,A^{-1}W_n},$$

$$t_n' A t_n \leq 2n\left[\zeta_{n,\theta_n} - \zeta_{n,A^{-1}W_n} - r_{n,A^{-1}W_n}\right]. \tag{A.6}$$

Note that this implies that

$$\|t_n\|^2 \leq \frac{2n}{\lambda}\left[\zeta_{n,\theta_n} - \zeta_{n,A^{-1}W_n} - r_{n,A^{-1}W_n}\right],$$

where $\lambda > 0$ is the smallest eighen-value of $A$.

(a) Fix $\varepsilon > 0$, and let $E_{\varepsilon,n}$ be the event that

$$\|t_n\|^2 \leq \varepsilon\left(1 + \|t_n\|^2 + n\|W_n\|^2\right).$$

We next show that, by the assumptions of the theorem and (A.6), $P\left(\overline{E_{\varepsilon,n}}\right) = o(1)$. Without a loss of generality, assume that $A = I$. First, note that the fact that $\theta_n \to^p 0$ implies that there exists a deterministic sequence $\delta_n \to +0$ such that

$$P\left\{\|\theta_n\| > \delta_n\right\} = o(1).$$

Now fix $\mu_0 > 0$. Take $\delta_n \to +0$ such that $P\left\{\|\theta_n\| > \delta_n\right\} = o(1)$ and $P\left\{\|W_n\| > \delta_n\right\} = o(1)$ (note that $W_n \to^p 0$). For all $n \geq N_1$, $P\left\{\|\theta_n\| > \delta_n\right\} \leq \mu_0$ and $P\left\{\|W_n\| > \delta_n\right\} \leq \mu_0$. Since 0 is in interior of $\Theta$, this also implies that $P\left\{W_n \notin \Theta\right\} \leq \mu_0$ for all $n \geq N_2$. By

(A.3), for all $n \geq N_3$,

$$P\left\{\sup_{\|\theta\|\leq\delta_n} \frac{\left|\zeta_{n,\theta}\right| + |r_{n,\theta}|}{n^{-1} + \|\theta\|^2} > \frac{\varepsilon}{6}\right\} < \mu_0.$$

Thus, with probability at least $1 - 4\mu_0$ (when none of the above events is true), for all $n \geq N_4 = \max\{N_1, N_2, N_3\}$,

$$\begin{aligned} \|t_n\|^2 &\leq 2n\left[\zeta_{n,\theta_n} - \zeta_{n,A^{-1}W_n} - r_{n,A^{-1}W_n}\right] \\ &\leq \frac{\varepsilon}{6} 2n\left[2n^{-1} + \|\theta_n\|^2 + \|W_n\|^2\right] \\ &\leq \varepsilon\left[1 + \|t_n\|^2 + n\|W_n\|^2\right], \end{aligned}$$

i.e. $P\left(\overline{E_{\varepsilon,n}}\right) = o(1)$.

Next, since $W_n = O_p\left(n^{-1/2}\right)$, there exists $K > 1$ such that for $n > N_5$,

$$P\left\{n\|W_n\|^2 > K\right\} < \mu_0,$$

therefore,

$$P\left\{n\|W_n\|^2 \leq K \text{ and } E_{\frac{\varepsilon}{K},n}\right\} = 1 - o(1).$$

On this latter event, $\|t_n\|^2 \leq \varepsilon\left[2 + \|t_n\|^2\right]$, or $\|t_n\|^2 \leq \frac{2\varepsilon}{1-\varepsilon}$, which implies that $t_n = o_p(1)$.

(b) Fix $\varepsilon > 0$, and let $E_{\varepsilon,n}$ be the event defined above. Assumptions of the theorem and (A.6) imply that $P\left(\overline{E_{\varepsilon,n}}\right) = o\left(n^{-1}\right)$. To show this, we first prove that there is a deterministic sequence $\delta_n \to +0$, such that

$$P\left\{\|\theta_n\| > \delta_n\right\} = o\left(n^{-1}\right).$$

Define

$$\delta_n = \inf\left\{\delta > 0 : P\left\{\|\theta_n\| > \delta\right\} \leq n^{-1}\delta\right\}.$$

Note that $\delta_n \leq n$ is always finite and so well defined. Since $P\left\{\|\theta_n\| > \delta\right\} \to 1$ as $\delta \to 0$, $\delta_n > 0$ for all $n$. Also, by continuity of probability,

$$P\left\{\|\theta_n\| > \delta_n\right\} \leq n^{-1}\delta_n.$$

It remains to show that $\delta_n \to 0$. Take a $\mu > 0$. We have: $P\left\{\|\theta_n\| > \mu\right\} = o\left(n^{-1}\right)$, that is, there exists $n_\mu$ such that $P\left\{\|\theta_n\| > \mu\right\} < \mu n^{-1}$ for all $n \geq n_\mu$. By definition of $\delta_n$, $\delta_n \leq \mu$ for all $n \geq n_\mu$. In other words, $\delta_n \to +0$.

Note that the condition

$$P\left\|W_n^2\right\| 1\left\{\|W_n\| > \mu\right\} = o\left(n^{-1}\right)$$

implies, by Chebyshev inequality, that for any $\mu > 0$, $P\left\{\|W_n\| > \mu\right\} = o\left(n^{-1}\right)$, therefore, there is a sequence $\delta_n \to +0$ such that $P\left\{\|\theta_n\| + \|W_n\| > \delta_n\right\} = o\left(n^{-1}\right)$.

Fix $\mu > 0$ and take $\delta_n$ as above. Take $N$ so large that the union of the events

$$\left\{\|\theta_n\| + \|W_n\| > \delta_n\right\}$$

$$\left\{\sup_{\|\theta\| \leq \delta_n} \frac{\left|\zeta_{n,\theta}\right| + \left|r_{n,\theta}\right|}{n^{-1} + \|\theta\|^2} > \frac{\varepsilon\lambda}{6}\right\}$$

and

$$\left\{W_n \notin \Theta\right\}$$

has the probability less than $\mu$. On the complement event,

$$\|t_n\|^2 \leq \varepsilon \left[1 + \|t_n\|^2 + n \|W_n\|^2\right]$$

by the same argument as in (a).

Choose $\varepsilon < 1$. Then, taking into account that $\theta_n \in \Theta$ is bounded,

$$\begin{aligned}
& P \|t_n\|^2 \\
\leq\ & P \|t_n\|^2 1_{E_{\varepsilon,n}} + P \|t_n\|^2 1_{\overline{E_{\varepsilon,n}}} \\
\leq\ & \frac{\varepsilon}{1-\varepsilon} P \left[1 + n \|W_n\|^2\right] + P \|t_n\|^2 1_{\overline{E_{\varepsilon,n}}} \\
\leq\ & \frac{\varepsilon}{1-\varepsilon} O(1) + 2Pn \|\theta_n\|^2 1_{\overline{E_{\varepsilon,n}}} + 2Pn \|W_n\|^2 1_{\overline{E_{\varepsilon,n}}} \\
\leq\ & \frac{\varepsilon}{1-\varepsilon} O(1) + 2 \sup_{\Theta} \|\theta\|^2 nP\left(\overline{E_{\varepsilon,n}}\right) \\
& + 2nP \|W_n\|^2 1\left\{\|W_n\|^2 > 1\right\} + 2nP\left\{\overline{E_{\varepsilon,n}}\right\} \\
\leq\ & \frac{\varepsilon}{1-\varepsilon} O(1) + o(1).
\end{aligned}$$

Therefore, $P \|t_n\|^2 = o(1)$. □

To assess the accuracy of the normal approximation, one needs to investigate the nature of the difference between $\theta_n$ and $\eta_n$ more closely.

**Theorem 10.** . *Suppose that equations (A.1) and (A.2) hold. Assume that there exists a sequence of numbers $a_n \geq 1$, and numbers $\lambda, \delta_0 > 0$ and $\alpha \in [0, 2)$ such that the ball with center zero and radius $\delta_0$ is in $\Theta$, and*
*(i) For any $\delta > 0$, $P\{\|\eta_n\| + \|\theta_n\| > \delta\} = O(a_n^{-1})$.*

*(ii)*

$$\begin{aligned} & P\left\{\begin{array}{c} \textit{Matrix } \partial^2 G^0_{n,\theta} \textit{ exists and is continuous, and} \\ -\partial^2 G^0_{n,\theta} \geq \lambda I \textit{ for all } \|\theta\| \leq \delta_0 \end{array}\right\} \\ = \;& 1 - O\left(a_n^{-1}\right). \end{aligned}$$

*(iii) For any* $0 < \delta \leq \delta_0$,

$$\begin{aligned} & P\left\{\sup_{\|\theta\| \leq \delta} \frac{\zeta_{n,\eta_n+\theta} - \zeta_{n,\eta_n} + r_{n,\eta_n}}{n^{-1}a_n^{-2} + \delta\|\theta\|^2 + \left(n^{-1/2}a_n^{-1}\right)^{2-\alpha}\|\theta\|^{\alpha}} \leq \frac{1}{\delta_0}\right\} \\ = \;& 1 - O\left(a_n^{-1}\right). \end{aligned}$$

*Then there exists a constant* $K$ *such that*

$$P\left\{n^{1/2}\|\theta_n - \eta_n\| > K a_n^{-1}\right\} = O\left(a_n^{-1}\right).$$

**Proof.** Let $\delta^* = \min\left\{\delta_0, \frac{\lambda\delta_0}{4}\right\}$. Let $E_n$ be the union of event

$$\left\{\|\theta_n\|, \|\eta_n\| < \delta^*\right\},$$

the event in condition (ii), and the event in condition (iii) for $\delta = \delta^*$. Conditions (i)-(iii) imply that $P\left(\overline{E_n}\right) = O\left(a_n^{-1}\right)$. Define $t_n = n^{1/2}a_n\left(\theta_n - \eta_n\right)$. Since $\eta_n \in \Theta$, we have

$$G^0_{n,\theta_n} - G^0_{n,\eta_n} \geq -r_{n,\theta} + \zeta_{n,\eta_n} - \zeta_{n,\theta_n}.$$

When on $E_n$, $\eta_n$ is an interior maximum and so the F.O.C., $\partial G^0_{n,\eta_n} = 0$, is satisfied. Use this to expand the left-hand side around $\eta_n$: for some $\left\|\tilde{\theta}_n\right\| \leq \delta^*$,

$$\begin{aligned} G^0_{n,\theta_n} - G^0_{n,\eta_n} &= \frac{1}{2} n^{-1} a_n^{-2} t_n' \partial^2 G^0_{n,\tilde{\theta}_n} t_n \\ &\leq -\frac{\lambda}{2} n^{-1} a_n^{-2} \|t_n\|^2 . \end{aligned}$$

Therefore,

$$\begin{aligned} \|t_n\|^2 &\leq \frac{2}{\lambda} n a_n^2 \left[ r_{n,\theta} + \zeta_{n,\theta_n} - \zeta_{n,\eta_n} \right] \\ &\leq \frac{2}{\lambda \delta_0} \left[ 1 + \delta^* \|t_n\|^2 + \|t_n\|^\alpha \right] , \end{aligned}$$

or (recall that $\delta^* \leq \frac{\lambda \delta_0}{4}$)

$$\|t_n\|^2 \leq \frac{4}{\lambda \delta_0} \left( 1 + \|t_n\|^\alpha \right) .$$

Because $\alpha \in [0, 2)$, this implies that for some constant $K = K(\alpha, \delta_0, \lambda) > 0$,

$$\|t_n\| \leq K.$$

Taking into account the possibility of the event $\overline{E_n}$, we have

$$P\left\{ n^{1/2} \|\theta_n - \eta_n\| > K a_n^{-1} \right\} = P\left( \overline{E_n} \right) = O\left( a_n^{-1} \right) .$$

□

### A.1.2. Hoeffding Decomposition and Its Bootstrap Version

When the estimator maximizes a $U$-process, representation (A.1) can be obtained by the so-called Hoeffding decomposition (or the $U$-decomposition). Let $h_\theta : \mathcal{Z}^m \to \mathbb{R}$ be a symmetric, $P$-measurable function. Denote by $\pi_{k.m}h_\theta$ the projection of $h_\theta$ onto the space of functions of $k$ arguments that are *degenerate* with respect to the measure $P$, in the sense that their expectation relative to $P$ over any one argument, holding the other arguments constant, is zero:

$$(\pi_{k,m}h_\theta)(z_1, ..., z_k) = (\delta_{z_1} - P) ... (\delta_{z_k} - P) P^{m-k}h_\theta$$

(where $\delta_{z_1}h_\theta = h_\theta(z_1, \cdot)$). Then

$$U_n^{(m)}h_\theta = P^m h_\theta + mP_n\pi_{1,m}h_\theta + \sum_{k=2}^{m} \binom{m}{k} U_n^{(k)} (\pi_{k,m}h)_\theta , \tag{A.7}$$

where $P_n$ is the sample mean, i.e. the $U$-process of order 1 (see e.g. Arcones and Giné [**1992**] for the $U$-decomposition in this notation).

The importance of the Hoeffding decomposition is that it isolates terms of progressively higher order in $n^{-1/2}$. The first term is the expectation of $h_\theta$ and has the order $O(1)$. The second term is the sample mean of a random variable with zero mean; it has the order $O_p\left(n^{-1/2}\right)$ by the Central Limit Theorem. The following terms are of the order $O_p\left(n^{-k/2}\right)$. Representation (A.1) can be obtained if the first few terms in (A.7) are twice differentiable in $\theta$, so that they admit a Taylor expansion with leading terms given by $G_{n,\theta}^0$, while the error term $\zeta_{n,\theta}$ will collect the remainder from the Taylor expansion and the higher-order $U$-processes in (A.7). Specific decompositions will be considered below.

A similar decomposition is also needed for the bootstrap problem. In the literature on the bootstrap of $U$-statistics, it is common to write the Hoeffding decomposition of the bootstrapped process, $\hat{U}_n^{(m)} h_\theta$, conditionally on the sample of data $\{Z_i\}_{i=1}^n$, i.e. relative to the empirical measure $P_n$ in place of $P$. This approach makes the analysis of the higher-order processes no more difficult in the bootstrap problem than in the sample problem. It is inconvenient for $M$-estimators, however, because the leading terms of the $U$-decomposition relative to $P_n$ may not have the smoothness properties of the leading terms in (A.7). For example, the first term will be:

$$P_n^m h_\theta \equiv \frac{1}{n^2} \sum_{i_1,...,i_m} h_\theta \left( Z_{i_1}, Z_{i_2}, ..., Z_{i_m} \right)$$

which is not a differentiable function of $\theta$ for the rank estimators. Thus, the Taylor expansion arguments leading to representation (A.1) for the sample problem will not be directly applicable to the bootstrap problem.

Here we suggest a different approach. Write the Hoeffding decomposition of the bootstrapped process in terms of the same functions $\pi_{k.m} h_\theta$ (integrals of $h_\theta$ relative to $P$) that appear in (A.7):

$$\hat{U}_n^{(m)} h_\theta = P^m h_\theta + m \hat{P}_n \left( \pi_{1,m} h_\theta \right) + \sum_{k=2}^{m} \binom{m}{k} \hat{U}_n^{(k)} \left( \pi_{k.m} h_\theta \right). \tag{A.8}$$

(To obtain this formula, apply the summation operator $\hat{U}_n^{(m)}$ to formula (2.5) in Arcones and Giné [**1992**].) Now, the functional form, and therefore, smoothness properties with respect to $\theta$, of the leading terms in $G_{n,\theta}$ and $\hat{G}_{n,\theta}$ are the same, and only the sample of data on which they are evaluated differ.

### A.1.3. Bounds on the Higher-Order $U$-Processes

To apply the approximation theorems, we need to check their equicontinuity assumptions for the components of $\zeta_{n,\theta}$ and $\hat{\zeta}_{n,\theta}$ given by the higher-order $U$-processes in the Hoeffding decomposition. This is the most challenging part of the proof, which is mostly deployed in Section A.3. Here we give only the final results relevant to our problem.

Given a function $h(z_1, ..., z_m)$, define the function

$$h^{[m-2s]}(z_1, ..., z_{m-2s}) =$$
$$\int h(z_1, ..., z_{m-2s}, Z_{m-s+1}, Z_{m-s+1}, ..., Z_m, Z_m)\, dP(Z_{m-s+1}) ... dP(Z_m).$$

For the sample problem, the following two bounds hold.

**Lemma 11.** *(a) Let $\mathcal{H} = \{h_\theta : \mathcal{Z}^m \to \mathbb{R}\}$, $m \geq 1$, be a class of $P$-degenerate symmetric functions, which is Euclidean for an envelope $H$ satisfying $P^m H^{p\vee 2} < \infty$ for $p \geq 1$, and $\mathcal{H}_n$ be its subclasses. Then, as $n \to \infty$,*

$$n^{m/2} \left( P \sup_{h\in\mathcal{H}} \left| U_n^{(m)} h \right|^p \right)^{1/p} = O(1).$$

*(b) If, additionally, $\sup_{h\in\mathcal{H}_n} P^m h^2 \to 0$, then*

$$n^{m/2} \left( P \sup_{h\in\mathcal{H}_n} \left| U_n^{(m)} h \right|^p \right)^{1/p} = o(1).$$

*(c) If, additionally to conditions in (a), $P^m H_{\omega_m}^{p\vee 2} < \infty$ for each permutation, with repetition, $\omega_m$, then*

$$n^{m/2} \left( P \sup_{h\in\mathcal{H}} \left| \hat{U}_n^{(m)} h \right|^p \right)^{1/p} = O(1).$$

*(d) If, additionally to conditions in (b) and (c), for each* $s$, $1 \leq s \leq \frac{m}{2}$,

$$\sup_{h \in \mathcal{H}_n} P^{m-2s} \left(h^{[m-2s]}\right)^2 \to 0,$$

*then*

$$n^{m/2} \left( P \sup_{h \in \mathcal{H}_n} \left| \hat{U}_n^{(m)} h \right|^p \right)^{1/p} = o\left(1\right).$$

**Lemma 12.** *Let* $\mathcal{H} = \{h_\theta : \mathcal{Z}^m \to \mathbb{R}\}$, $m \geq 2$, *be a class of symmetric,* $P$*-degenerate functions, Euclidean for an envelope* $H$*. Assume that there exist constants* $\delta_0$, $C > 0$ *such that for all* $\theta_1$, $\theta_2$ *in the* $\delta_0$*-neighborhood of* $0$,

$$P^m \left[ \left( h_{\theta_1} - h_{\theta_2} \right)^2 \right] \leq C \left\| \theta_1 - \theta_2 \right\|. \tag{A.9}$$

*Then*

$$P \left\{ \sup_{\left\|\tilde{\theta}\right\|,\, \|\theta\| \leq \delta_0/2} \frac{\left| U_n^{(m)} \left( h_{\tilde{\theta}+\theta} - h_{\tilde{\theta}} \right) \right|}{n^{-1} a_n^{-2} + \left( n^{-1/2} a_n^{-1} \right)^{3/2} \left\| \theta \right\|^{1/2}} > 1 \right\} = O\left( a_n^{-1} \right), \tag{A.10}$$

*with any* $a_n \geq 1$ *satisfying*

$$a_n \leq \left( n^{1/6} \left( \log n \right)^{-2/3} \right)^{1/(1+2/3p)}, \text{ if } m = 2 \text{ and } P^m H^6 < \infty;$$

$$a_n \leq n^{(m-1)/4-\varepsilon}, \text{ if } m \geq 3 \text{ and } P^m H^p < \infty \text{ for all } p.$$

*In the last expression, $\varepsilon > 0$ can be arbitrarily small*[1].

*(b) If, additionally, the integrability conditions imposed on function $H$ also hold for functions $H_{\omega_m}$, for all permutations, with repetition, $\omega_m$; and for all $\theta_1$, $\theta_2$, and for all $s$, $1 \leq s \leq \frac{m}{2}$, in the $\delta_0$-neighborhood of* 0,

$$P^{m-2s}\left[\left(h_{\theta_1}^{[m-2s]} - h_{\theta_2}^{[m-2s]}\right)^2\right] \leq C\left\|\theta_1 - \theta_2\right\|,$$

*then inequality (A.10) also holds (with the same rates $a_n$) with $U_n^{(m)}$ changed to $\hat{U}_n^{(m)}$.*

## A.2. Proofs of the Main Results

### A.2.1. Asymptotic Normality and Consistency of the Bootstrap

Only the proof of Theorem 2 is provided. The proof of Theorem 1 is analogous (and simpler), and is close to the proofs in Sherman [**1993**] and Arcones, Giné and Chen [**1994**].

First, we obtain a quadratic approximation for the bootstrap objective function $\hat{U}_n h_\theta$. Define $\tau_\theta = P^{m-1}h_\theta$ and $A = -P\left[\partial^2 \tau_0\right]$. By Assumptions 1 and 3, $A$ is a symmetric, positive definite matrix; $P\left[\partial \tau_0\right] = 0$ (this is the first-order condition in the population maximization problem), and $P\left\|\partial \tau_0\right\|^2 < \infty$. Define

$$\begin{aligned} R_\theta(z) &= \left[P^m h_\theta + m\pi_{1,m} h_\theta\right](z) - m\theta' \partial \tau_0(z) + \frac{1}{2}\theta' A \theta \\ &= P\tau_\theta + m\left(\tau_\theta(z) - P\tau_\theta - \theta' \partial \tau_0(z)\right) + \frac{1}{2}\theta' A \theta. \end{aligned}$$

[1]Results for other combinations of $m$ and $p$ can be easily deduced from the proof. We omit them for brevity.

Using this and the Hoeffding decomposition for the bootstrapped $U$-statistic, we obtain

$$\hat{U}_n^{(m)} h_\theta = \theta' \hat{W}_n - \frac{1}{2}\theta' A \theta + \hat{\zeta}_{n,\theta}, \tag{A.11}$$

where $\hat{W}_n = m\hat{P}_n \partial \tau_0$, and $\hat{\zeta}_{n,\theta}$ is the remainder:

$$\hat{\zeta}_{n,\theta} = \hat{P}_n R_\theta + \sum_{k=2}^{m} \binom{m}{k} \hat{U}_n^{(k)} \left(\pi_{k,m} h_\theta\right). \tag{A.12}$$

Let $\delta_0 > 0$ be such that the neighborhood $\mathcal{N}$ in Assumption 3 contains the ball of radius $\delta_0$ with the center at zero. By Assumptions 3 (i), (ii), conditions $h_0 \equiv 0$, $P\partial\tau_0 = 0$, and the second-order Taylor expansion around zero,

$$\left|\hat{P}_n R_\theta\right| \leq m\left(PM + \hat{P}_n M\right)\|\theta\|^3 + m\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0\right\|\|\theta\|^2 \tag{A.13}$$

for all $\|\theta\| \leq \delta_0$.

Now we check conditions of Theorem 9. By the bootstrap Hoeffding decomposition, Assumptions 2, 5 and Lemma 11 (c),

$$P \sup_{\theta\in\Theta} \left|\hat{U}_n^{(m)} h_\theta - P^m h_\theta\right| \to 0,$$

which together with the identification Assumption 1 implies consistency of $\hat{\theta}_n$ for 0, by the standard argument: for every $\delta > 0$, there is $\eta > 0$ such that

$$\begin{aligned} P\left\{\left\|\hat{\theta}_n\right\| > \delta\right\} &\leq P\left\{\sup_{\theta\in\Theta}\left|\hat{U}_n^{(m)} h_\theta - P^m h_\theta\right| > \eta\right\} \\ &\leq \eta^{-1} P \sup_{\theta\in\Theta}\left|\hat{U}_n^{(m)} h_\theta - P^m h_\theta\right| = o(1). \end{aligned}$$

(the last line follows from the Chebyshev Inequality).

Clearly, $\hat{W}_n = O_p\left(n^{-1/2}\right)$: note that $\hat{W}_n = \left(\hat{W}_n - W_n\right) + W_n$, where $W_n = mP_n\partial\tau_0$. The second moment of both terms relative to $P$ is $m^2 Var\left(\partial\tau_0\right)/n$, so by the Chebyshev inequality, $\hat{W}_n = O_p\left(n^{-1/2}\right)$. Next, use (A.13) and integrability conditions imposed in Assumption 3 to argue that $\hat{P}_n R_\theta$ satisfies condition (A.3), actually, the stronger condition

$$\sup_{\|\theta\|\leq\delta_n} \frac{\left|\hat{P}_n R_\theta\right|}{\|\theta\|^2} \to^p 0$$

whenever $\delta_n \to +0$. It is enough to show that $\hat{P}_n M = O_p(1)$ and $\left(\hat{P}_n - P\right)\partial^2\tau_0 = o_p(1)$ under conditions $PM < \infty$ and $P\|\partial^2\tau_0\| < \infty$. Both follow from the following: if $P|f| < \infty$ then $\left|\hat{P}_n f - Pf\right| = o_p(1)$. In fact, $\left(\hat{P}_n - P\right) f = \left(\hat{P}_n - P_n\right) f + \left(P_n - P\right) f$. The second term is $o_p(1)$ by the Law of Large Numbers, and the first term is $o_p(1)$ by the bootstrap weak law of large numbers given e.g. in Theorem 3.5 in Giné and Zinn [**1990**]. Condition $P|f| < \infty$ is sufficient for condition (i) of that theorem. Then, $\hat{P}\left|\left(\hat{P}_n - P_n\right) f\right| = o_p(1)$. By the Chebyshev inequality, for any $\varepsilon > 0$, $\hat{P}\left\{\left|\left(\hat{P}_n - P_n\right) f\right| > \varepsilon\right\} = o_p(1)$. The left-hand side is bounded by 1. Integrate over $P$ to obtain

$$P\left\{\left|\left(\hat{P}_n - P_n\right) f\right| > \varepsilon\right\} = o(1).$$

It remains to verify condition (A.3) for the higher-order $U$-processes in (A.12). Use the maximal inequality, Lemma 11 (c) (with $p = 1$). For $k \geq 3$,

$$P\sup_{\theta\in\Theta}\left|\hat{U}_n^{(k)}\left(\pi_{k,m}h_\theta\right)\right| = O\left(n^{-3/2}\right).$$

For $k = 2$, take a sequence $\delta_n \to +0$ and apply the maximal inequality from Lemma 11 (d) (with $m = 2$ in that lemma) to classes $\mathcal{H}_n = \{\pi_{2,m} h_\theta : \|\theta\| \leq \delta_n\}$:

$$P \sup_{\|\theta\| \leq \delta_n} \left| \hat{U}_n^{(2)} (\pi_{2,m} h) \right| = o\left(n^{-1}\right).$$

Conclude that condition (A.3) is satisfied for all $k \geq 2$.

By Theorem 9,

$$n^{1/2} \left( \hat{\theta}_n - \hat{P}_n A^{-1} \partial \tau_0 \right) = o_p(1).$$

A similar derivation for the sample problem gives

$$n^{1/2} \left( \theta_n - P_n A^{-1} \partial \tau_0 \right) = o_p(1).$$

Therefore,

$$\nu_n \equiv n^{1/2} \left( \hat{\theta}_n - \theta_n \right) - n^{1/2} \left( \hat{P}_n - P_n \right) A^{-1} \partial \tau_0 = o_p(1). \tag{A.14}$$

By Theorem 2.2 of Bickel and Freedman [**1981**], for almost all sequences $\{Z_1, Z_2, ...\}$

$$n^{1/2} \left( \hat{P}_n - P_n \right) A^{-1} \partial \tau_0 \to N(0, \Gamma).$$

Weak convergence to the multivariate normal distribution is always uniform (Corollary 2.6, Theorem 3.1 and Corollary 3.2 of Bhattacharya and Rao [**1976**]); therefore, for almost all sequences $\{Z_1, Z_2, ...\}$,

$$\sup_{A \in \mathcal{A}} \left| \int_A d\hat{F}_{n^{1/2}\left(\hat{P}_n - P_n\right) A^{-1} \partial \tau_0} - \int_A d\Phi_\Gamma \right| \to 0. \tag{A.15}$$

This and (A.14) imply the conclusion of Theorem 2:

$$\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)}-\int_A d\Phi_\Gamma\right|=o_p\left(1\right), \tag{A.16}$$

as follows. For $\varepsilon>0$, and a set $A\in\mathcal{A}$, define $A^\varepsilon=\cup\left\{B\left(x,\varepsilon\right),x\in A\right\}$, where $B\left(x,\varepsilon\right)$ is the open ball with center $x$ and radius $\varepsilon$, and $A^{-\varepsilon}=\mathbb{R}^d\backslash\left(\mathbb{R}^d\backslash A\right)^\varepsilon$. Both sets are in $\mathcal{A}$ (both are convex, the first is open and the second is closed, so both are measurable), and $A^{-\varepsilon}\subset A\subset A^\varepsilon$. It is known that

$$\sup_{A\in\mathcal{A}}\int_{A^\varepsilon\backslash A^{-\varepsilon}}d\Phi_\Gamma\leq K\left(d,\Gamma\right)\varepsilon,$$

see formula (3) and Corollary 3.2 in Bhattacharya and Rao [**1976**]. We have

$$\int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)}\leq\int_{A^\varepsilon}d\hat{F}_{n^{1/2}\left(\hat{W}_n-W_n\right)}+P\left\{\left\|\nu_n\right\|\geq\varepsilon\right\}$$

and

$$\int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)}\geq\int_{A^{-\varepsilon}}d\hat{F}_{n^{1/2}\left(\hat{W}_n-W_n\right)}-P\left\{\left\|\nu_n\right\|\geq\varepsilon\right\}.$$

Then,

$$\begin{aligned}&\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)}-\int_A d\Phi_\Gamma\right|\\ \leq\ &\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{n^{1/2}\left(\hat{P}_n-P_n\right)A^{-1}\partial\tau_0}-\int_A d\Phi_\Gamma\right|\\ &+\sup_{A\in\mathcal{A}}\int_{A^\varepsilon\backslash A^{-\varepsilon}}d\Phi_\Gamma+P\left\{\left\|\nu_n\right\|\geq\varepsilon\right\}.\end{aligned}$$

Therefore, (A.16) holds.

### A.2.2. Estimation of the Variance

Here we prove consistency of the asymptotic variance estimators given in Theorem 3. We consider only the bootstrap problem, while the (simpler) proof for the sample problem can be reconstructed using the same steps. We check conditions of part (b) of Theorem 9. By condition $P^m H^p < \infty$, for $p > 2$, the bootstrap Hoeffding decomposition, and Lemma 11 (c), for each $\varepsilon > 0$ there is $\eta > 0$ such that

$$\begin{aligned} P\left\{\left\|\hat{\theta}_n\right\| > \varepsilon\right\} &\leq \eta^{-p} P \sup_{\theta\in\Theta} \left|\hat{U}_n^{(m)} h_\theta - P^m h_\theta\right|^p \\ &= \eta^{-p} O\left(n^{-p/2}\right) = o\left(n^{-1}\right). \end{aligned}$$

Next note that conditions $P^m H^p < \infty$, $PM^p < \infty$, $P\left\|\partial^2 \tau_0\right\|^p < \infty$, and the Taylor expansion, imply that $P\left\|\partial \tau_0\right\|^p < \infty$. Then by Rosenthal inequality, $P\left\|\hat{W}_n\right\|^p = O\left(n^{-p/2}\right)$, and, therefore,

$$\begin{aligned} & nP\left\|\hat{W}_n\right\|^2 1\left\{\left\|\hat{W}_n\right\|^2 > \varepsilon\right\} \\ = & \int_{n\varepsilon^2}^{\infty} x^2 dF_{n\left\|\hat{W}_n\right\|^2} \leq \frac{1}{\left(n\varepsilon^2\right)^{p-2}} \int_{n\varepsilon^2}^{\infty} x^p dF_{n\left\|\hat{W}_n\right\|^2} \\ \leq & \frac{1}{\left(n\varepsilon^2\right)^{p-2}} P\left\|\hat{W}_n\right\|^p = o\left(n^{-1}\right). \end{aligned}$$

The extra integrability assumptions of Theorem 3 ensure that $\hat{P}_n R_{n,\theta}$ satisfies condition (A.4). (For example, since $PM^p < \infty$,

$$\begin{aligned} P\left\|\left(\hat{P}_n - P\right) M > \varepsilon\right\| &\leq \frac{P\left\|\left(\hat{P}_n - P\right) M\right\|^p}{\varepsilon^p} \\ &= O\left(n^{-p/2}\right) = o\left(n^{-1}\right). \end{aligned}$$

Here we used the Rosenthal inequality, applied conditionally for the bootstrap and integrated over $P$, and applied unconditionally for the sample mean. Similarly,

$$P\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0 > \varepsilon\right\| = o\left(n^{-1}\right)$$

if $P\left\|\partial^2\tau_0\right\|^p < \infty$.)

To check (A.4) for the higher-order U-processes, invoke Lemma 11 (c) with $p > 2$. Theorem 3 implies

$$P\left\|n^{1/2}\left(\hat{\theta}_n - A^{-1}\hat{P}_n\partial\tau_0\right)\right\|^2 \to 0.$$

By Chebyshev inequality,

$$\hat{P}\left\|n^{1/2}\left(\hat{\theta}_n - A^{-1}\hat{P}_n\partial\tau_0\right)\right\|^2 \to^p 0.$$

By Theorem 2.2 of Bickel and Freedman [**1981**],

$$\widehat{Var}\left(n^{1/2}A^{-1}\hat{P}_n\partial\tau_0\right) - \Gamma \to^{a.s.} 0$$

so,

$$\widehat{Var}\left(n^{1/2}\hat{\theta}_n\right) - \Gamma \to^p 0.$$

### A.2.3. Generic Bound for Rank Estimators

Here we prove Theorem 4 for the bootstrap problem. The proof for the sample problem follows the same steps. We use the same representation (A.11), but check conditions of Theorem 10. The rate in Theorem 4, $a_n$, is determined by the rate of convergence to zero of the $U$-process of order 2 in the remainder $\hat{\zeta}_{n,\theta}$, $\hat{U}_n^{(2)}\left(\pi_{2,m}h_\theta\right)$. It is given in Lemma 12. To

apply it, consider the class of functions $\left\{\breve{h} = \pi_{2,m} h\right\}$. The class consists of $P$-degenerate functions of two arguments. Note that by Jensen inequality, the condition on $P^{m-2} h_\theta$, $P^{m-2} h_\theta^{[m-2]}$ in Assumptions 6, 7 imply the same condition for functions $\breve{h}_\theta$, $\breve{h}_\theta^{[m-2]}$. If the class $\{h\}$ is Euclidean, then so is the class $\left\{\breve{h}\right\}$ (see the properties of the Euclidean classes in Section A.3). Also, the class $\left\{\breve{h}\right\}$ inherits from the class $\{h\}$ its integrability properties (finiteness of moments). Lemma 12 (b) gives the rate, $a_n$, with which condition (iii) of Theorem 10 is satisfied for $\hat{U}_n^{(2)}(\pi_{2,m} h_\theta)$: $a_n = \left(n^{1/6} (\log n)^{-2/3}\right)^{1/(1+2/3p)}$ if $P^m H_{\omega_m}^p < \infty$ for $p \geq 6$ and all permutations, with repetition, $\omega_m$. It now suffices to check that the other conditions of Theorem 10 are satisfied with this rate and the probability $1 - O\left(n^{-1/6}\right)$.

First, check condition (i). For $\hat{\theta}_n$, as in the previous subsection, for $p = 6$ (this is the minimal integrability assumption imposed in Theorem 4),

$$\begin{aligned} P\left\{\left\|\hat{\theta}_n\right\| > \delta\right\} &\leq \eta_\delta^{-p} P \sup_{\theta \in \Theta} \left|\hat{U}_n^{(m)} h_\theta - P^m h_\theta\right|^p \\ &= \eta_\delta^{-p} O\left(n^{-p/2}\right) = O\left(n^{-3}\right). \end{aligned}$$

Since by the Rosenthal inequality, $P\left\|\hat{W}_n\right\|^4 = O\left(n^{-2}\right)$, under condition $P\left\|\partial \tau_0\right\|^4 < \infty$,

$$P\left\{\left\|\hat{W}_n\right\| > \delta\right\} = O\left(n^{-2}\right).$$

Condition (ii) is trivial here because $A$ is assumed to be a constant positive definite matrix. As a consequence, $\hat{\eta}_n = \hat{W}_n$, except on an event of probability $O\left(n^{-2}\right)$. We, therefore, can neglect the distinction between $\hat{\eta}_n$ and $\hat{W}_n$.

Condition (iii) for higher-order $U$-processes in $\hat{\zeta}_{n,\theta}$, $\hat{U}_n^{(k)}(\pi_{k,m}h_\theta)$, for $a_n \leq n^{1/6}$, is trivial because $\sup_\Theta \left|\hat{U}_n^{(k)}(\pi_{k,m}h_\theta)\right| = O_p\left(n^{-k/2}\right)$ by Lemma 11 (c), and the rate $n^{k/2}$, $k \geq 3$, dominates the rate $na_n^2$, which is at most $n^{4/3}$.

Condition (iii) for $\hat{P}_n R_\theta$ can be checked using the extra integrability assumptions on $M(z)$, $\partial^2\tau_0$, and $\partial\tau_0$ made in Theorem 4.

[Let us verify, for example, that for a sufficiently small $\delta_0 > 0$, for all $0 < \delta < \delta_0$,

$$P\left\{\sup_{\|\theta\|\leq\delta} \frac{\left|\hat{P}_n R_{\hat{W}_n+\theta}\right|}{n^{-1}a_n^{-2} + \delta\|\theta\|^2} > \frac{1}{\delta_0}\right\} = O\left(n^{-1/6}\right).$$

Use the bound in (A.13):

$$\left|\hat{P}_n R_\theta\right| \leq m\left(PM + \hat{P}_n M\right)\|\theta\|^3 + m\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0\right\|\|\theta\|^2.$$

Clearly, it is enough to check the following: for all $\varepsilon > 0$,

$$P\left\{na_n^2\left\|\hat{W}_n\right\|^3 > \varepsilon\right\} = O\left(n^{-1/6}\right),$$

$$P\left\{\left(\hat{P}_n - P\right)M > \varepsilon\right\} = O\left(n^{-1/6}\right),$$

$$P\left\{\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0\right\| > \varepsilon\right\} = O\left(n^{-1/6}\right),$$

$$P\left\{na_n^2\left\|\hat{W}_n\right\|^2\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0\right\| > \varepsilon\right\} = O\left(n^{-1/6}\right).$$

The first three follow from the Chebyshev and Rosenthal inequalities (the latter bounds the moments of bootstrapped means). Note that the latter requires a finite second population moment. So, for example,

$$\begin{aligned} & P\left\{na_n^2\left\|\hat{W}_n\right\|^3 > \varepsilon\right\} \\ \leq\ & P\left\{n^{4/3}\left\|\hat{W}_n\right\|^3 > \varepsilon\right\} = P\left\{\left\|n^{1/2}\hat{W}_n\right\|^3 > \varepsilon n^{1/6}\right\} \\ =\ & O\left(n^{-1/6}\right), \end{aligned}$$

since $P\left\|n^{1/2}\hat{W}_n\right\|^3 = O(1)$. For the last one, we have:

$$\begin{aligned} & P\left\{na_n^2\left\|\hat{W}_n\right\|^2\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0\right\| > \varepsilon\right\} \\ \leq\ & P\left\{n^{11/12}\left\|\hat{W}_n\right\|^2 > \varepsilon^{1/2}\right\} \\ & +P\left\{n^{5/12}\left\|\left(\hat{P}_n - P\right)\partial^2\tau_0\right\| > \varepsilon^{1/2}\right\} \\ =\ & P\left\{\left\|n^{1/2}\hat{W}_n\right\|^2 > \varepsilon^{1/2}n^{1/12}\right\} \\ & +P\left\{\left\|n^{1/2}\left(\hat{P}_n - P\right)\partial^2\tau_0\right\| > \varepsilon^{1/2}n^{1/12}\right\} \\ =\ & O\left(n^{-1/6}\right), \end{aligned}$$

because $P\left\|n^{1/2}\hat{W}_n\right\|^4, P\left\|n^{1/2}\left(\hat{P}_n - P\right)\partial^2\tau_0\right\|^2 = O(1)$.]

We, therefore, have

$$P\left\{\left\|n^{1/2}\left(\hat{\theta}_n - A^{-1}\hat{P}_n\partial\tau_0\right)\right\| > Ka_n^{-1}\right\} = O\left(a_n^{-1}\right).$$

A similar derivation gives

$$P\left\{\left\|n^{1/2}\left(\theta_n - A^{-1}P_n\partial\tau_0\right)\right\| > Ka_n^{-1}\right\} = O\left(a_n^{-1}\right).$$

Therefore,

$$\begin{aligned} \text{(A.17)} \qquad & P\left\{\nu_n \equiv \left\|n^{1/2}\left(\hat{\theta}_n - \theta_n\right) - n^{1/2}A^{-1}\left(\hat{P}_n - P_n\right)\partial\tau_0\right\| > Ka_n^{-1}\right\} \\ = \; & O\left(a_n^{-1}\right) \end{aligned}$$

for some $K > 0$.

Next we use the multivariate Berry-Esséen Theorem (Corollary 18.3 in Bhattacharya and Rao [**1976**]). For the sample problem, under conditions that $Var\left(\partial\tau_0\right)$ is a positive definite matrix and $P\left\|\partial\tau_0\right\|^3 < \infty$, we have:

$$\sup_{A\in\mathcal{A}}\left|\int_A dF_{A^{-1}P_n\partial\tau_0} - \int_A d\Phi_\Gamma\right| \leq n^{-1/2}c\left(d\right)P\left\|\Gamma^{-1/2}A^{-1}\partial\tau_0\right\|^3,$$

where $c\left(d\right)$ is an absolute constant for each $d$.

For the bootstrap problem, let $C_0$ be a constant such that

$$\limsup_{n\to\infty} P_n\left\|\Gamma_n^{-1/2}A^{-1}\partial\tau_0\right\|^3 < C_0(P-a.s),$$

where $\Gamma_n = \widehat{Var}\left(n^{1/2}A^{-1}\hat{P}_n\partial\tau_0\right)$. Such finite constant exists by the law of large numbers under conditions that $\Gamma$, $A$ are positive definite and $P\left\|\partial\tau_0\right\|^3 < \infty$. Apply the Berry-Esséen Theorem conditionally on sequences of data for which this condition is satisfied.

Then, $P-a.s.$,

$$\lim_{n\to\infty}\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{A^{-1}\left(\hat{P}_n-P_n\right)\partial\tau_0}-\int_A d\Phi_{\Gamma_n}\right|\leq n^{-1/2}c\left(d\right)C_0$$

Integrate over $P$ and take into account that the integrand is a sequence of bounded functions, apply the Lebesgue dominated convergence theorem:

$$\lim_{n\to\infty}P\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{A^{-1}\left(\hat{P}_n-P_n\right)\partial\tau_0}-\int_A d\Phi_{\Gamma_n}\right|\leq n^{-1/2}c\left(d\right)C_0$$

or, by the Chebyshev inequality,

$$\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{A^{-1}\left(\hat{P}_n-P_n\right)\partial\tau_0}-\int_A d\Phi_{\Gamma_n}\right|=O_p\left(n^{-1/2}\right).$$

Finally, condition $P\left\|\partial\tau_0\right\|^4<\infty$, implies that $\Gamma_n-\Gamma=O_p\left(n^{-1/2}\right)$. Namely,

$$\begin{aligned}
& P\left\{\left\|\Gamma_n-\Gamma\right\|>Kn^{-1/2}\right\}\\
\leq\ & P\left\{\left\|A^{-1}\right\|^2\left\|P_n\left[\partial\tau_0\partial\tau_0'\right]-\left(P_n\partial\tau_0\right)\left(P_n\partial\tau_0\right)'-\Gamma\right\|>Kn^{-1/2}\right\}\\
\leq\ & P\left\{\left\|A^{-1}\right\|^2\left\|n^{1/2}\left(P_n\left[\partial\tau_0\partial\tau_0'\right]-\Gamma\right)\right\|>\frac{K}{2}\right\}\\
& +P\left\{\left\|A^{-1}\right\|^2\left\|n^{1/2}P_n\partial\tau_0\right\|^2>\frac{K}{2}n^{1/2}\right\}\\
\leq\ & \frac{2\left\|A^{-1}\right\|^2P\left\|n^{1/2}\left(P_n\left[\partial\tau_0\partial\tau_0'\right]-\Gamma\right)\right\|^2}{K}\\
& +\frac{2\left\|A^{-1}\right\|^2P\left\|n^{1/2}\left(P_n\left[\partial\tau_0\partial\tau_0'\right]-\Gamma\right)\right\|^2}{Kn^{1/2}}\\
\to\ & 0\text{ as }K\to\infty,
\end{aligned}$$

where we used the Rosenthal inequality (the fourth moment of $\partial\tau_0$ is needed because the Rosenthal inequality requires the second moment).

Then it follows from the properties of the normal distribution that

$$\sup_{A\in\mathcal{A}}\left|\int_A d\Phi_{\Gamma_n}-\int_A d\Phi_{\Gamma}\right|=O_p\left(n^{-1/2}\right).$$

[To see that, use the Taylor expansion:

$$\begin{aligned}&\int_A d\Phi_{\Gamma_n}-\int_A d\Phi_{\Gamma}\\ =\;&\left(\int_A\left(\frac{\partial\phi}{\partial\Gamma}\right)_{\tilde{\Gamma}_n}dX\right)'(\Gamma_n-\Gamma)\end{aligned}$$

and the fact that

$$\sup_{A\in\mathcal{A}}\left|\int_A\left(\frac{\partial\phi}{\partial\Gamma}\right)_{\tilde{\Gamma}_n}dX\right|\leq\int_{\mathbb{R}^d}\left|\left(\frac{\partial\phi}{\partial\Gamma}\right)_{\tilde{\Gamma}_n}\right|dX<\infty.]$$

So, we have

$$\sup_{A\in\mathcal{A}}\left|\int_A d\hat{F}_{A^{-1}\left(\hat{P}_n-P_n\right)\partial\tau_0}-\int_A d\Phi_{\Gamma}\right|=O_p\left(n^{-1/2}\right). \tag{A.18}$$

Now we obtain the uniform result of Theorem 4. We show it for the bootstrap. Use (A.17) and (A.18), and the logic of the proof of uniformity in consistency theorems. Let

$\varepsilon_n = Ka_n^{-1}$. We have:

$$\begin{aligned}
& \sup_{A\in\mathcal{A}} \left| \int_A d\hat{F}_{n^{1/2}\left(\hat{\theta}_n-\theta_n\right)} - \int_A d\Phi_\Gamma \right| \\
\leq\ & \sup_{A\in\mathcal{A}} \left| \int_A d\hat{F}_{n^{1/2}\left(\hat{P}_n-P_n\right)A^{-1}\partial\tau_0} - \int_A d\Phi_\Gamma \right| \\
& + \sup_{A\in\mathcal{A}} \int_{A^{\varepsilon_n}\setminus A^{-\varepsilon_n}} d\Phi_\Gamma + P\left\{\|\nu_n\| \geq \varepsilon_n\right\} \\
=\ & O_p\left(n^{-1/2}\right) + O\left(\varepsilon_n\right) + O_p\left(a_n^{-1}\right) = O_p\left(a_n^{-1}\right).
\end{aligned}$$

### A.2.4. Better Rates under Additional Smoothness Assumptions

Under additional Assumption 8, the degenerate $U$-processes of order up to $s \geq 2$ in the Hoeffding decomposition of the criterion function $G_{n,\theta}$ are all smooth functions of $\theta$. Then one can approximate $\theta_n$ by the random vector $\eta_n$ which solves the problem

$$\eta_n \in \arg\max_{\theta\in\Theta} G_{n,\theta}^0 \equiv U_n^{(s)} h_\theta^*,$$

where

$$h_\theta^* = \sum_{k=0}^{s} \binom{m}{k} \pi_{k,m} h_\theta = \sum_{k=0}^{s} \binom{m}{k} \pi_{k,s} f_\theta.$$

The bootstrapped estimator, $\hat{\theta}_n$ can be approximated by

$$\hat{\eta}_n \in \arg\max_{\theta\in\Theta} \hat{U}_n^{(s)} h_\theta^*.$$

The properties of $\eta_n$ and $\hat{\eta}_n$ can be found by powerful methods based on the Taylor expansion and Berry-Esséen bounds for higher-order $U$-statistics. Note first that by the

Hoeffding decomposition, maximal and Chebyshev inequalities, for any $\delta > 0$,

$$P\left\{\|\eta_n\| > \delta\right\} = O\left(n^{-1/2}\right),$$

and

$$P\left\{\|\hat{\eta}_n\| > \delta\right\} = O\left(n^{-1/2}\right).$$

In particular, with probability at least $1- O\left(n^{-1/2}\right)$, $\eta_n - \theta_0$ coincides with the solution to the first order condition:

$$U_n^{(2)} g_{\theta_0+\theta} = \xi_{n,\theta_0+\theta},$$

where $g_\theta = \left(P^m + m\pi_{1,s} + \frac{m(m-1)}{2}\pi_{2,s}\right)\partial f_\theta$, and $\xi_{n,\theta} = \sum_{k=3}^{s} \binom{m}{k} U_n^{(k)} \pi_{k,s} \partial f_\theta$. Functions $g_\theta$, $\xi_{n,\theta}$, and $\eta_n$ satisfy the assumptions of Lemma 20 (in particular, $P^m \partial g_0 = P^m \partial f_0 = 0$, by the first-order condition in the population problem), and, therefore, the following Berry-Esséen bound holds:

$$\sup_{A\in\mathcal{A}} \left| \int_A dF_{n^{1/2}(\eta_n-\theta_0)} - \int_A d\Phi_\Gamma \right| = O\left(n^{-1/2}\right). \tag{A.19}$$

To obtain a similar bound for the bootstrap problem, apply Lemma 20 conditionally on the sample. With probability at least $1- O\left(n^{-1/2}\right)$, the random vector $\hat{\eta}_n - \eta_n$ coincides with the solution to the equation:

$$\hat{U}_n^{(2)} g_{\eta_n+\theta} = \hat{\xi}_{n,\eta_n+\theta}$$

where $\hat{\xi}_{n,\theta} = \sum_{k=3}^{s} \binom{m}{k} \hat{U}_n^{(k)} \pi_{k,s} \partial f_\theta$. Note, in particular, that

$$\begin{aligned} P_n^2 g_{\eta_n+\theta} &= \frac{n-1}{n} U_n g_{\eta_n} + \frac{1}{n^2} \sum_{i=1}^{n} g_{\eta_n}(Z_i, Z_i) \\ &= \xi_{n,\eta_n} + \frac{1}{n^2} \sum_{i=1}^{n} g_{\eta_n}(Z_i, Z_i); \end{aligned}$$

and, therefore, satisfies Assumption (ii) of Lemma 20 with $K = O_p(1)$. Also note that the conditional moments required to apply Lemma 20 are bounded for almost all sequences of data $\{Z_1, Z_2 ...\}$, by the moment conditions on $L$ in Assumption 10 (a); therefore, $c_d$ in Lemma 20 will be $O_p(1)$. Thus,

$$\sup_{A \in \mathcal{A}} \left| \int_A d\hat{F}_{n^{1/2}(\hat{\eta}_n - \eta_n)} - \int_A d\Phi_{\Gamma_n} \right| = O_p\left(n^{-1/2}\right).$$

Under the assumption that $P \|\partial f_\theta\|^4 < \infty$, we can rewrite the last bound as

$$\sup_{A \in \mathcal{A}} \left| \int_A d\hat{F}_{n^{1/2}(\hat{\eta}_n - \eta_n)} - \int_A d\Phi_{\Gamma} \right| = O_p\left(n^{-1/2}\right). \tag{A.20}$$

The objective function for the estimators $\theta_n$, $\hat{\theta}_n$, contains additional terms given by:

$$\zeta_{n,\theta} = \sum_{k=s+1}^{m} \binom{m}{k} U_n^{(k)}(\pi_{k,m} h_\theta) + r_{n,\theta}$$

and

$$\hat{\zeta}_{n,\theta} = \sum_{k=s}^{m} \binom{m}{k} \hat{U}_n^{(k)}(\pi_{k,m} h_\theta) + \hat{r}_{n,\theta}.$$

To estimate the differences $\theta_n - \eta_n$ and $\hat{\theta}_n - \hat{\eta}_n$, use Theorem 10. Under Assumption 8, by the Taylor expansion,

$$U_n^{(s)} h_\theta = U_n^{(s)} h_\theta^* = \theta' W_n - \frac{1}{2} \theta' A_{n,\theta} \theta,$$

where

$$W_n = U_n^{(s)} \partial h_0^*,$$

and, for $A = -\partial^2 P h_0 = -\partial^2 P h_0^*$,

$$\partial^2 G_{n,\theta}^0 = -A + U_n^{(s)} \partial^2 \left( h_0^* - P h_0^* \right) + O\left( L \left\| \theta \right\| \right).$$

Condition (i) of Theorem 10 is satisfied with $a_n = n^{1/2}$, for both $\theta_n$ and $\eta_n$. Condition (ii) follows from the previous display, positive definiteness of $A$, and the moment conditions on function $L$. If Assumptions 8, 9 are satisfied with $s = 2$, then condition (iii) of the theorem is satisfied with $a_n = n^{1/2-\varepsilon}$, where $\varepsilon > 0$ is arbitrarily small, by Lemma 12 (a) (for the degenerate $U$-process of order 3), and Lemma 11 (a) with sufficiently high $p$ (for the degenerate $U$-processes of order 4 and higher). If $s = 3$, then condition (iii) is satisfied with $a_n = n^{3/4-\varepsilon}$, by the same lemmas (Lemma 12 now should be used for the degenerate $U$-processes of order 4)[2]. Conditions (i-iii) can be verified for the bootstrap (i.e. relative

[2]The result for $s = 3$ can be used to obtain the Edgeworth expansion for the distribution functions of $n^{1/2}\theta_n$ and $n^{1/2}\left(\hat{\theta}_n - \theta_n\right)$ with the error term of order $O\left(n^{-3/4+\varepsilon}\right)$ ($O_p\left(n^{-3/4+\varepsilon}\right)$ for the bootstrap), which implies that the *symmetric* confidence intervals for $\theta_n$ constructed using the bootstrap are $O_p\left(n^{-3/4+\varepsilon}\right)$-accurate. The last bound may not be tight, even with $\varepsilon$ omitted. For the parametric estimators, the symmetric confidence intervals are also more accurate than one-sided ones, and have the error of coverage probability $O_p\left(n^{-1}\right)$ (see Hall [**1992**]). The derivation is tedious because it requires further terms in the Edgeworth expansion for $\eta_n$ and $\hat{\eta}_n$, and is omitted.

to the unconditional distribution of the bootstrap draws) in the same way. Particularly, condition (iii) directly follows from Lemmas 11 (c) and 12 (b).

It follows that for some constant $K > 0$,

$$P\left\{n^{1/2}\left\|\theta_n - \eta_n\right\| > Ka_n^{-1}\right\} = O\left(a_n^{-1}\right) \tag{A.21}$$

and

$$P\left\{n^{1/2}\left\|\hat{\theta}_n - \hat{\eta}_n\right\| > Ka_n^{-1}\right\} = O\left(a_n^{-1}\right).$$

Combining the last to bounds we have

$$P\left\{n^{1/2}\left\|\left(\hat{\theta}_n - \theta_n\right) - \left(\hat{\eta}_n - \eta_n\right)\right\| > Ka_n^{-1}\right\} = O\left(a_n^{-1}\right). \tag{A.22}$$

The sample version of Theorem 5 follows from (A.19) and (A.21), while its bootstrap counterparts follows from (A.20) and (A.22).

### A.3. Bounds on Oscillations of U-Processes

Here we provide a brief discussion of the empirical process theory for $U$-processes, and extensions to it, that eventually lead to Lemmas 11, 12. The bounds listed here are relevant for the $U$-processes indexed by a Euclidean class of functions. For the convenience of the reader we remind the definition. Call function $H$ an envelope of a class of functions $\mathcal{H}$ if $|h| \leq H$ for each $h \in \mathcal{H}$.

**Definition 13.** *(Nolan and Pollard* [**1987**]*) Let $\mathcal{H}$ be a class of real-valued functions defined on the same set. Call $\mathcal{H}$ Euclidean for the envelope $H$ if there exist positive constants (referred to as Euclidean numbers in the sequel) $A$ and $V$ such that for any*

*measure* $\mu$, *for which* $0 < \mu H < \infty$,

$$N_2(\varepsilon, d_\mu) \leq A\varepsilon^{-V}, \ 0 < \varepsilon \leq 1.$$

*Here, for* $h_1, h_2 \in \mathcal{H}$, $d_\mu(h_1, h_2) = \mu|h_1 - h_2|^2/\mu H^2$ *and* $N_2(\varepsilon, d_\mu)$ *is the packing number of* $\mathcal{H}$ *with respect to the pseudometric* $d_\mu$, *i.e. the largest number* $N$ *such that there exist functions* $h_1, ..., h_N$ *with the property* $d_\mu(h_i, h_j) > \rho$ *for* $i \neq j$.

A detailed review of the properties of Euclidean classes of functions can be found in Nolan and Pollard [**1987**] and Pakes and Pollard [**1989**]. In particular, if $\mathcal{H}_1$ and $\mathcal{H}_2$ are two Euclidean classes for the envelopes, respectively, $H_1$ and $H_2$, then the class $\mathcal{H}_1 + \mathcal{H}_2 \equiv \{h_1 + h_2 : h_i \in \mathcal{H}_i\}$ is Euclidean for the envelope $H_1 + H_2$ and the class $\mathcal{H}_1 \cdot \mathcal{H}_2 \equiv \{h_1 \cdot h_2 : h_i \in \mathcal{H}_i\}$ is Euclidean for the envelope $H_1 \cdot H_2$. If $\mathcal{H} = \{h : \mathcal{Z}^m \to \mathbb{R}\}$ is $A, V$- Euclidean for the envelope $H$, then the class $\{|h| : h \in \mathcal{H}\}$ is $A, V$-Euclidean for the envelope $H$, and for any probability distribution $\mu$, acting on variables $z_1, ..., z_k$, the class

$$\{\mu h(\cdot, z_{k+1}, ..., z_m) : h \in \mathcal{H}\}$$

is $A, V$-Euclidean for the envelope $\mu H$ (in particular, $\mu$ may put mass 1 on a value of $(z_1, ..., z_k)$).

It is convenient to introduce extra notation for the rest of this subsection. Throughout $\lesssim$ will denote inequality up to a multiplicative constant. The constant may depend on certain parameters of the model (typically, the Euclidean numbers $A$ and $V$, the order of the process $m$ and so on), but not on $n$ or the sample data $\{Z_1, ..., Z_n\}$. In particular, we will often use the inequality $(a + b)^p \lesssim a^p + b^p$, $a, b \geq 0$, $p > 0$, where the constant depends

on $p$ only (for $p \in (0,1)$ the constant is 1). Symbol $\|\cdot\|_{\mathcal{H}}$ will stand for the supremum over a class of functions $\mathcal{H}$.

Lemma 14 gives bounds for the first moment of the suprema of the degenerate empirical and $U$-processes.

**Lemma 14.** *Let $\mathcal{H}$ be a class of $P$-degenerate symmetric functions which is Euclidean for an envelope $H$ with $P^m H > 0$. Then*

$$\begin{aligned} & P \left\| U_n^m h \right\|_{\mathcal{H}} \\ \lesssim \; & n^{-m/2} P \left[ \left( U_n^m H^2 \right)^{1/2} \int_0^{\left( \left\| U_n^m h^2 \right\|_{\mathcal{H}} / U_n^m H^2 \right)^{1/2}} (1 - \log \varepsilon)^{m/2} \, d\varepsilon \right] \end{aligned}$$

*Here the multiplicative constant depends on $m$ and the Euclidean numbers $A, V$ only.*

**Proof.** Cases $m = 1, 2$ were considered in Pollard [**1989**], Theorem 4.2 (i), and Nolan and Pollard [**1987**]. For $m \geqslant 1$, the inequalities follow from Propositions 2.1, 2.2 and 2.6 in Arcones and Giné [**1993**]; see also the calculations in Arcones and Giné [**1994**]. □

**Remark 15.** *The integral that appears in Lemma 14 (with $\frac{\left\| U_n^m h^2 \right\|_{\mathcal{H}}}{U_n^m H^2} \equiv x \in (0,1]$) can be bounded from above and from below by multiples of function*

$$J_m(x) = x^{1/2} \left( 1 - \frac{1}{m} \log x \right)^{m/2},$$

*which is increasing, concave, and bounded on $x \in (0,1]$. Furthermore, $J_m(x)$, $m \geq 1$, satisfies*

$$(m/2)^{m/2} (\log n)^{-m/2} J_m(x) \leq x^{1/2} \vee \left( n^{-1} \log n \right)^{1/2}$$

*for all* $x \in (0,1]$ *and* $n \geq e^m$ *(particularly, if* $x \leq n^{-1}\log n$*,* $J_m(x) \leq J(n^{-1}\log n)$ *by monotonicity).*

The bound on $P\|U_n^m h\|_{\mathcal{H}}$ is related to the "continuity modulus" of the class $\mathcal{H}$, $\|P^m h^2\|_{\mathcal{H}}^{1/2}$.

**Lemma 16.** *Let* $\mathcal{H} = \{h : \mathcal{Z}^m \to \mathbb{R}\}, m \geq 1$*, be a Euclidean class of symmetric,* $P$*-degenerate functions with envelope* 1*. Then for all* $n$*,*

$$P\left\|U_n^{(m)} h\right\|_{\mathcal{H}} \lesssim \left(n^{-1}\log n\right)^{m/2}\left\|P^m h^2\right\|_{\mathcal{H}}^{1/2} + \left(n^{-1}\log n\right)^{(m+1)/2}, \tag{A.23}$$

*where the multiplicative constant depends on* $m$ *and the Euclidean numbers of the class only.*

**Proof.** Follows from Theorem 8 in Giné and Mason [**2007**]. □

**Lemma 17.** *Let* $\mathcal{H} = \{h : \mathcal{Z}^m \to \mathbb{R}\}$ *be a class of symmetric,* $P$*-degenerate functions, Euclidean for an envelope* $H$*. If for* $p \geq 2$*,* $P^m H^p < \infty$*, then*

$$P\left\|U_n^{(m)} h\right\|_{\mathcal{H}} \lesssim \left(n^{-1}\log n\right)^{m/2}\left\|P^m h^2\right\|_{\mathcal{H}}^{1/2} + \left(n^{-1}\log n\right)^{(m+1)/2-1/p}.$$

*In these inequalities, the multiplicative constants depend on* $m$*,* $P^m H^p$ *and the Euclidean numbers of the class only.*

**Proof.** First, we obtain

$$P\left\|U_n^{(m)} h^2\right\|_{\mathcal{H}} \lesssim \left\|P^m h^2\right\|_{\mathcal{H}} + \left(n^{-1}\log n\right)^{1-2/p}. \tag{A.24}$$

Let $\mathcal{H}_L$, $L \geq 1$, be the class of functions $\{h \cdot 1\{|h| \leq L\} : h \in \mathcal{H}\}$. Note that $\mathcal{H}_L$ is Euclidean for the envelope $L$. Consider the case $L = 1$. By the Hoeffding decomposition, Lemma 14 and Remark 15,

$$\begin{aligned} P\left\|U_n^{(m)}h^2\right\|_{\mathcal{H}_1} &\lesssim \left\|P^m h^2\right\|_{\mathcal{H}_1} + n^{-1/2} P J_1\left(\left\|P_n\left(\pi_{1,m}h^2\right)^2\right\|_{\mathcal{H}_1}\right) + n^{-1} \\ &\lesssim \left\|P^m h^2\right\|_{\mathcal{H}_1} \\ &\quad + n^{-1/2} P\left\|P_n\left(\pi_{1,m}h^2\right)^2 \log n\right\|_{\mathcal{H}_1}^{1/2} + n^{-1}\log n. \end{aligned}$$

Note that

$$\begin{aligned} \left\|P_n\left(\pi_{1,m}h^2\right)^2\right\|_{\mathcal{H}_1} &\lesssim \left\|P_n\left(P^{m-1}h^2\right)^2\right\|_{\mathcal{H}_1} + \left\|P^m h^4\right\|_{\mathcal{H}_1} \\ &\leq \left\|P_n P^{m-1}h^2\right\|_{\mathcal{H}_1} + \left\|P^m h^2\right\|_{\mathcal{H}_1}. \end{aligned}$$

Therefore, (also using $2|xy| \leq x^2 + y^2$)

$$\begin{aligned} P\left\|U_n^{(m)}h^2\right\|_{\mathcal{H}_1} &\lesssim \left\|P^m h^2\right\|_{\mathcal{H}_1} + n^{-1}\log n \\ &\quad + \left(n^{-1}\log n\right)^{1/2} P\left\|P_n P^{m-1}h^2\right\|_{\mathcal{H}_1}^{1/2}. \end{aligned} \tag{A.25}$$

Apply this inequality to the process $P_n\left(P^{m-1}h^2\right)$; denoting by $X$ the expression

$$P\left\|P_n\left(P^{m-1}h^2\right)\right\|_{\mathcal{H}_1},$$

and by $C > 0$ the multiplicative constant,

$$X \leq C\left\|P^m h^2\right\|_{\mathcal{H}_1} + CX^{1/2}\left(n^{-1}\log n\right)^{1/2} + Cn^{-1}\log n.$$

One possibility is that $X > 4C^2 n^{-1} \log n$, in which case the previous inequality gives

$$X \leq C \left\| P^m h^2 \right\|_{\mathcal{H}_1} + \frac{1}{2} X + C n^{-1} \log n,$$

so that

$$X \lesssim \left\| P^m h^2 \right\|_{\mathcal{H}_1} + n^{-1} \log n.$$

The other possibility is that $X \geq 4C^2 n^{-1} \log n$. In both cases,

$$P \left\| P_n \left( P^{m-1} h^2 \right) \right\|_{\mathcal{H}_1} \leq \left\| P^m h^2 \right\|_{\mathcal{H}_1} + n^{-1} \log n.$$

Substitute this into (A.25):

$$P \left\| U_n^{(m)} h^2 \right\|_{\mathcal{H}_1} \lesssim \left\| P^m h^2 \right\|_{\mathcal{H}_1} + n^{-1} \log n.$$

For an arbitrary $L \geq 1$, by rescaling,

$$P \left\| U_n^{(m)} h^2 \right\|_{\mathcal{H}_L} \lesssim \left\| P^m h^2 \right\|_{\mathcal{H}_L} + L^2 n^{-1} \log n.$$

Next, as

$$\begin{aligned} h^2 &= h^2 1\{|h| \leq L\} + h^2 1\{|h| \geq L\} \\ &\leq h^2 1\{|h| \leq L\} + H^2 1\{H \geq L\}, \end{aligned}$$

we have

$$\begin{aligned} & P\left\|U_n^{(m)}h^2\right\|_{\mathcal{H}} \\ \lesssim\ & \left\|P^m h^2\right\|_{\mathcal{H}} + L^2\left(n^{-1}\log n\right) + PH^2 1\left\{H > L\right\} \\ =\ & \left\|P^m h^2\right\|_{\mathcal{H}} + L^2\left(n^{-1}\log n\right) + o\left(L^{-p+2}\right). \end{aligned}$$

Taking $L = \left(n^{-1}\log n\right)^{-1/p}$ gives (A.24).

For a $U$-statistic of order $m$, use first Lemma 14 and Remark 15:

$$\begin{aligned} P\left\|U_n^{(m)}h\right\|_{\mathcal{H}} \lesssim\ & \left(n^{-1}\log n\right)^{m/2} P\left\|U_n^{(m)}h^2\right\|_{\mathcal{H}} \\ & + \left(P^m H^2\right)^{1/2}\left(n^{-1}\log n\right)^{(m+1)/2}. \end{aligned}$$

Now use (A.24). □

**Lemma 18.** *(Hoffmann-Jørgensen inequality for $U$-processes indexed by Euclidean classes of functions). Let $\mathcal{H} = \{h : \mathcal{Z}^m \to \mathbb{R}\}$ be a class of $P$-degenerate symmetric functions which is Euclidean for a $P^m$-square-integrable envelope $H$. Then for every $p \geq 2$*

$$P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}^p \lesssim \left(P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}\right)^p + n^{-p(m+1)/2+1}P^m H^p,$$

*with a constant depending on $m, p$ and Euclidean constants $A, V$ of the class only.*

**Proof.** For $m = 1$ this inequality is well-known: it holds without constraints on the capacity of the class $\mathcal{H}$, see van der Vaart and Wellner [**1996**], Theorem 2.14.5. For $m \geq 2$, Giné and Zinn [**1992**], Corollary 4, obtained the following bound (also without

capacity restrictions on $\mathcal{H}$):

$$P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}^p \lesssim \left(P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}\right)^p$$
$$+P\max_{i_m\leq n}\left\|\binom{n}{m}^{-1}\sum_{i_1,...,i_{m-1}:(i_1,...,i_{m-1})\in I_n^{(m)}}^{n} h(Z_{i_1},...,Z_{i_m})\right\|_{\mathcal{H}}^p.$$

The second term can be bounded by

$$P\sum_{j=1}^{n}\left\|\binom{n}{m}^{-1}\sum_{i_1,...,i_{m-1}:(i_1,...,i_{m-1})\in I_n^{(m)}}^{n} h(Z_{i_1},...,Z_{i_m})\right\|_{\mathcal{H}}^p$$
$$\lesssim \quad n^{-p+1}P'P\left\|U_{n-1}^{(m-1)}h(\cdot,Z')\right\|_{\mathcal{H}}^p,$$

where $Z'$ is an independent copy of $Z_i$, and $P'$ integrates over $Z'$. Using the same argument for $U_{n-1}^{(m-1)}h(\cdot,z)$, with fixed $z$, we have:

$$P\left\|U_{n-1}^{(m-1)}h(\cdot,z)\right\|_{\mathcal{H}}^p \lesssim \left(P\left\|U_{n-1}^{(m-1)}h(\cdot,z)\right\|_{\mathcal{H}}\right)^p$$
$$+n^{-p+1}P'P\left\|U_{n-2}^{(m-2)}h(\cdot,Z',z)\right\|_{\mathcal{H}}^p.$$

Euclidean property of the class $\mathcal{H}$ gives an upper bound for the first term:

$$\left(P\left\|U_{n-1}^{(m-1)}h(\cdot,z)\right\|_{\mathcal{H}}\right)^p \lesssim n^{-(m-1)p/2}\left(PH\left(\cdot,z\right)^2\right)^{p/2},$$

where the multiplicative constant is the same for all $z$.

Continue by induction, and use eventually the Hoffmann-Jørgensen inequality for $m = 1$ for the remaining $P$-process:

$$\begin{aligned} P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}^p &\lesssim \left(P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}\right)^p + \sum_{s=1}^{m-1} n^{(-p+1)s-(m-s)p/2} P^m H^p \\ &\quad + n^{(-p+1)(m-1)} n^{-1+1/p} P^m H^p \\ &\lesssim \left(P\left\|U_n^{(m)}h\right\|_{\mathcal{H}}\right)^p + n^{-(m+1)p/2+1} P^m H^p. \end{aligned}$$

□

Now consider the bootstrap version of the $U$-process. As in the preceding literature (e.g. Theorem 2.2 in Arcones and Giné [**1994**]) the goal is to relate the moments of the bootstrapped process $\hat{U}_n^{(m)}h$ to the moments of a modified sample process, by using the symmetrization and poissonization techniques suggested in Giné and Zinn [**1990**]. Note, however, that decomposition (A.8) requires the result under the assumption that $h$ is $P$-degenerate, rather than $P_n$-degenerate, as it was assumed by previous authors.

We need extra notation. Let $Q_i^{(j)}, i = 1, 2, ...;\ j = 1, ..., m$, be i.i.d. (across $i$ and $j$) random variables, independent of all $Z_i$, and having the Poisson distribution with parameter $1/2$. Define random vectors

$$\tilde{Z}_i = \left(Z_i, Q_i^{(1)}, ..., Q_i^{(m)}\right),$$

and let $\tilde{P}$ be the distribution of each $\tilde{Z}_i$, and $\tilde{h}\left(\tilde{z}_1, ..., \tilde{z}_m\right)$ be a symmetric version of the function

$$\tilde{h}^0\left(\tilde{z}_1, ..., \tilde{z}_m\right) = h\left(z_1, ..., z_m\right) q_1^{(1)} \cdot ... \cdot q_m^{(m)},$$

where $\tilde{z} = \left(z, q^{(1)}, ..., q^{(m)}\right)$. Note that functions $\tilde{h}$ are degenerate relative to the distribution $\tilde{P}$. The usefulness of the following lemma stems from the fact that the class of functions $\tilde{\mathcal{H}} = \left\{\tilde{h} : h \in \mathcal{H}\right\}$ inherits the capacity and integrability properties (relative to $\tilde{P}$) from those of the class $\mathcal{H}$. In particular, if $\mathcal{H}$ is Euclidean for an envelope $H(z_1, ..., z_m)$, then $\tilde{\mathcal{H}}$ is Euclidean for a symmetric version of the envelope $H(z_1, ..., z_m) \cdot q_1^{(1)} \cdot ... \cdot q_m^{(m)}$, denoted $\tilde{H}$. Also, since all moments of $Q_i^{(j)}$ are finite, $\{Z_i\}$ and $\left\{Q_i^{(j)}\right\}$ are independent, $\tilde{H}$ has as many finite moments relative to $\tilde{P}$, as $H$ does relative to $P$.

**Lemma 19.** *Let $\mathcal{H} = \{h : \mathcal{Z}^m \to \mathbb{R}\}$ be a class of $P$-degenerate real symmetric functions. Assume that $\mathcal{H}$ has an envelope $H$, and $P^m H^p < \infty$. Then*

$$P \left\| \hat{U}_n^{(m)} h \right\|_{\mathcal{H}}^p \lesssim P \left\| \frac{1}{n^m} \sum_{i_1, ..., i_m} \tilde{h}\left(\tilde{Z}_{i_1}, ..., \tilde{Z}_{i_m}\right) \right\|_{\tilde{\mathcal{H}}}^p,$$

*where the constant depends on $m$ and $p$ only.*

**Proof.** Use Hoeffding decomposition of the bootstrapped statistic relative to $P_n$ (i.e. conditionally on the sample):

$$\hat{U}_n^{(m)} h = \sum_{k=0}^{m} \binom{m}{k} \hat{U}_n^{(k)} \left(\pi_{k,m}^{P_n} h\right),$$

where

$$\left(\pi_{k,m}^{P_n} h_\theta\right)(z_1, ..., z_k) = \left(\delta_{z_1} - P_n\right) ... \left(\delta_{z_k} - P_n\right) P_n^{m-k} h_\theta.$$

Next we show that for each $k = 0, ..., m$,

$$\begin{aligned} P\left\|\hat{U}_n^{(k)}\left(\pi_{k,m}^{P_n} h\right)\right\|_{\mathcal{H}}^p &\lesssim P\left\|\frac{1}{n^m}\sum_{i_1,...,i_m}\tilde{h}^0\left(\tilde{Z}_{i_1},...,\tilde{Z}_{i_m}\right)\right\|_{\tilde{\mathcal{H}}}^p \\ &= P\left\|\frac{1}{n^m}\sum_{i_1,...,i_m}\tilde{h}\left(\tilde{Z}_{i_1},...,\tilde{Z}_{i_m}\right)\right\|_{\tilde{\mathcal{H}}}^p. \end{aligned}$$

Denote by $E$ the expectation conditional on the sample $Z_1, ..., Z_n$. Let $\left\{\hat{Z}_1^{(j)}, ..., \hat{Z}_n^{(j)}\right\}$ be i.i.d. samples from $P_n$, independent across $j = 1, ..., k$; denote by $\hat{P}_n^{(j)}$ the bootstrap empirical measure that puts mass $1/n$ on each $\hat{Z}_i^{(j)}$. Let $N_1^{(j)}, ..., N_n^{(j)}$ be i.i.d. across $i$ and $j$, independent from all $Z_i$, $\hat{Z}_i^{(j)}$, and each distributed as a difference between two independent Poisson r.v. with parameter $1/2$. Then

$$\begin{aligned} & E\left\|\hat{U}_n^{(k)}\left(\pi_{k,m}^{P_n} h\right)\right\|_{\mathcal{H}}^p \\ \lesssim\ & E\left\|\frac{1}{n^k}\sum_{i_1,...,i_k \text{ distinct}}\left(\pi_{k,m}^{P_n} h\right)\left(\hat{Z}_{i_1}^{(1)},...,\hat{Z}_{i_k}^{(k)}\right)\right\|_{\mathcal{H}}^p \\ \lesssim\ & E\left\|\frac{1}{n^k}\sum_{i_1,...,i_k}\left(\pi_{k,m}^{P_n} h\right)\left(\hat{Z}_{i_1}^{(1)},...,\hat{Z}_{i_k}^{(k)}\right)\right\|_{\mathcal{H}}^p \\ =\ & E\left\|\left(\hat{P}_n^{(1)} - P_n\right)\left(\hat{P}_n^{(k)} - P_n\right)P_n^{m-k} h\right\|_{\mathcal{H}}^p =: (*). \end{aligned}$$

Here the first inequality follows by the decoupling inequality of de la Peña [**1992**], applied conditionally on $Z_1, ..., Z_n$. In the second inequality the LHS is different from the RHS in that the latter includes summation over coinciding indices $i_1, ..., i_k$. The second inequality follows from the following observation: for any r.v. $X_h, Y_h$, if $E[Y_h|X_h] = 0$, then, by the convexity inequality, $E\|X_h + Y_h\|_{\mathcal{H}}^p \geq E\|X_h + E[Y_h|X_h]\|_{\mathcal{H}}^p = E\|X_h\|_{\mathcal{H}}^p$. Apply this to

obtain

$$E\left\|\sum_{\substack{i_1,...,i_s \text{ distinct},\\ i_{s+1},...,i_k \text{ unrestricted}}}\left(\pi_{k,m}^{P_n}h\right)\left(\hat{Z}_{i_1}^{(1)},...,\hat{Z}_{i_k}^{(k)}\right)\right\|_{\mathcal{H}}^{p}$$

$$\lesssim E\left\|\sum_{\substack{i_1,...,i_{s-1} \text{ distinct},\\ i_s,...,i_k \text{ unrestricted}}}\left(\pi_{k,m}^{P_n}h\right)\left(\hat{Z}_{i_1}^{(1)},...,\hat{Z}_{i_k}^{(k)}\right)\right\|_{\mathcal{H}}^{p}$$

(call $X_h$ the first sum, $Y_h$ the difference between the second and the first sums, then one can see that $E\left[Y_h|X_h\right]=0$ by degeneracy of $\pi_{k,m}^{P_n}h$ and independence of $\hat{Z}_i^{(j)}$ across both $i$ and $j$). Apply the last inequality sequentially in $s=m,m-1,...,2$ to obtain that the unrestricted sum dominates the sum over distinct indices.

Next we apply a poissonization argument. Define

$$\hat{X}_{i_1}=\delta_{z_1}\left(\hat{P}_n^{(2)}-P_n\right)...\left(\hat{P}_n^{(k)}-P_n\right)P_n^{m-k}h\Big|_{z_1=\hat{Z}_{i_1}^{(1)}},$$

and

$$X_{i_1}=\delta_{z_1}\left(\hat{P}_n^{(2)}-P_n\right)...\left(\hat{P}_n^{(k)}-P_n\right)P_n^{m-k}h\Big|_{z_1=Z_{i_1}}.$$

Let $\hat{\mathbf{Z}}:=\left\{\hat{Z}_{i_2}^{(2)},...,\hat{Z}_{i_k}^{(k)}|i_2,...,i_k=1,...,n\right\}$. Note, that conditionally on $\hat{\mathbf{Z}}$, $\hat{X}_{i_1}$ are the bootstrap drops from the sample $\{X_1,...,X_n\}$, *and* $E\left[\hat{X}_{i_1}|\hat{\mathbf{Z}}\right]=\frac{1}{n}\sum_{i_1}X_{i_1}$. Apply the symmetrization inequality of Proposition 2.1 in Arcones and Giné [**1993**], conditionally on $\hat{\mathbf{Z}}$; it gives:

$$\begin{aligned}(*) &= E\left\|n^{-1}\sum\nolimits_{i_1}\left(\hat{X}_{i_1}-E_{|\hat{\mathbf{Z}}}\hat{X}_{i_1}\right)\right\|_{\mathcal{H}}^{p} \\ &\lesssim E\left\|n^{-1}\sum\nolimits_{i_1}\varepsilon_{i_1}\hat{X}_{i_1}\right\|_{\mathcal{H}}^{p},\end{aligned}$$

where $\{\varepsilon_{i_1}\}$ is a Rademacher sequence independent of all other r.v. in the model. Next by the proof of Lemma 2.1 and Proposition 2.2 of Giné and Zinn [**1990**], applied to $\|\cdot\|^p$ rather than $\|\cdot\|$, we obtain:

$$\begin{aligned}& E\left\|n^{-1}\sum\nolimits_{i_1}\varepsilon_{i_1}\hat{X}_{i_1}\right\|_{\mathcal{H}}^{p} \\ \lesssim\ & E\left\|n^{-1}\sum\nolimits_{i_1}Q_{i_1}^{(1)}X_{i_1}\right\|_{\mathcal{H}}^{p} \\ =\ & E\left\|n^{-1}\sum\nolimits_{i_1}Q_{i_1}^{(1)}\delta_{Z_{i_1}}\left(\hat{P}_n^{(2)}-P_n\right)...\left(\hat{P}_n^{(k)}-P_n\right)P_n^{m-k}h\right\|_{\mathcal{H}}^{p}\end{aligned}$$

(where the result that we use give the inequality for $Q_{i_1}^{(1)}$ being distributed as a difference of two independent Poisson r.v. (with parameter 1/2). Use the triangle inequality to obtain the inequality for $Q_{i_1}^{(1)}$ being just the poisson r.v. with parameter 1/2).

Sequential application of this logic to the other arguments (with conditioning on previously introduced poisson r.v.), and integrating over the distribution of the sample lead to the inequality

$$(*)\lesssim E\left\|n^{-m}\sum_{i_1,...,i_m}Q_{i_1}^{(1)}...Q_{i_k}^{(k)}h\left(Z_{i_1},Z_{i_2},...,Z_{i_m}\right)\right\|_{\mathcal{H}}^{p}.$$

[Here we show the second step in the sequence.

$$\begin{aligned} & E\left\|n^{-1}\sum\nolimits_{i_1} Q_{i_1}^{(1)} X_{i_1}\right\|_{\mathcal{H}}^{p} \\ = \; & E\left\|n^{-1}\sum\nolimits_{i_1} Q_{i_1}^{(1)} \delta_{Z_{i_1}} \left(\hat{P}_n^{(k)} - P_n\right) P_n^{m-k} h\right\|_{\mathcal{H}}^{p} \end{aligned}$$

Define

$$\hat{X}_{i_2}^{(2)} = Q_{i_1}^{(1)} \delta_{Z_{i_1}} \delta_{\hat{Z}_{i_2}} \left(\hat{P}_n^{(3)} - P_n\right) ... \left(\hat{P}_n^{(k)} - P_n\right) P_n^{m-k} h,$$

$$X_{i_2}^{(2)} = Q_{i_1}^{(1)} \delta_{Z_{i_1}} \delta_{Z_{i_2}} \left(\hat{P}_n^{(3)} - P_n\right) ... \left(\hat{P}_n^{(k)} - P_n\right) P_n^{m-k} h,$$

and

$$\hat{\mathbf{Z}}^{(2)} := \left\{ Q_{i_1}^{(1)}, \hat{Z}_{i_k}^{(3)}, ..., \hat{Z}_{i_k}^{(k)} | i_2, ..., i_k = 1, ..., n \right\}.$$

Then, conditionally on the data sample and $\hat{\mathbf{Z}}^{(2)}$, $\left\{\hat{X}_{i_2}^{(2)}\right\}_{i_2=1,...,n}$ is the bootstrap sample (i.e. i.i.d. draws with replacement) from $\left\{X_{i_2}^{(2)}\right\}_{i_2=1,...,n}$. Applying the poissonization technique to this bootstrap problem, we obtain:

$$(*) \lesssim E \left\| n^{-2} \sum_{i_1, i_2} Q_{i_1}^{(1)} Q_{i_2}^{(k)} \delta_{Z_{i_1}} \delta_{Z_{i_2}} \left(\hat{P}_n^{(3)} - P_n\right) ... \left(\hat{P}_n^{(k)} - P_n\right) P_n^{m-k} h \right\|_{\mathcal{H}}^{p}$$

etc.]

Note that

$$\begin{aligned} & E\left\|n^{-m}\sum_{i_1,...,i_m} Q_{i_1}^{(1)}...Q_{i_k}^{(k)}h\left(Z_{i_1},Z_{i_2},...,Z_{i_m}\right)\right\|_{\mathcal{H}}^{p} \\ \lesssim\ & E\left\|n^{-m}\sum_{i_1,...,i_m} Q_{i_1}^{(1)}...Q_{i_k}^{(k)}...Q_{i_m}^{(m)}h\left(Z_{i_1},Z_{i_2},...,Z_{i_m}\right)\right\|_{\mathcal{H}}^{p} \\ =\ & E\left\|\frac{1}{n^m}\sum_{i_1,...,i_m}\tilde{h}\left(\tilde{Z}_{i_1},...,\tilde{Z}_{i_m}\right)\right\|_{\tilde{\mathcal{H}}}^{p}. \end{aligned}$$

by the Jensen inequality:

$$\begin{aligned} & E\left\|n^{-m}\sum_{i_1,...,i_m} Q_{i_1}^{(1)}...Q_{i_k}^{(k)}...Q_{i_m}^{(m)}h\left(Z_{i_1},Z_{i_2},...,Z_{i_m}\right)\right\|_{\mathcal{H}}^{p} \\ \geq\ & E\left\|n^{-m}\sum_{i_1,...,i_m} Q_{i_1}^{(1)}...Q_{i_k}^{(k)}h\left(Z_{i_1},Z_{i_2},...,Z_{i_m}\right)E\left[Q_{i_{k+1}}^{(k+1)}...Q_{i_m}^{(m)}\right]\right\|_{\mathcal{H}}^{p}, \end{aligned}$$

and the fact that $E\left[Q_{i_{k+1}}^{(k+1)}...Q_{i_m}^{(m)}\right] > 0$.

To complete the proof, integrate the bound over the sample measure. □

Finally, we prove Lemmas 11 and 12.

**Proof.** (Lemma 11.) (a) For $p = 1$, see Corollary 4(i) in Sherman [**1994**]. For $p \geq 2$ use also the Hoffmann-Jørgensen inequality, Lemma 18 (b). For $p = 1$, see the proof of Corollary 8 in Sherman [**1994**] (only straightforward notational changes are required). For $p \geq 2$ use also the Hoffmann-Jørgensen inequality, Lemma 18.

(c) By Lemma 19 (see the construction of function $\tilde{h}$, $\tilde{z}$, and $\tilde{P}$ there; in particular, $\tilde{h}$ is $\tilde{P}$-degenerate),

$$P\left\|\hat{U}_n^{(m)}h\right\|_{\mathcal{H}}^p \lesssim P\left\|\frac{1}{n^m}\sum_{i_1,...,i_m}\tilde{h}\left(\tilde{Z}_{i_1},...,\tilde{Z}_{i_m}\right)\right\|_{\tilde{\mathcal{H}}}^p.$$

Let $\tilde{U}_n^{(k)}$ denote the $U$-statistic based on the sample $\left\{\tilde{Z}_{i_1},...,\tilde{Z}_{i_m}\right\}$. Also, for $s \leq m-s$, let $\omega_s$ be a permutation, with repetition, having $s$ elements from the set $\{1,...,s\}$. The permutation $\{1,...,m-s,\omega_s(1),...,\omega_s(s)\}$, therefore, contains $m-s$ distinct elements. Denote by $e_m(\omega_s) = m-s-\#\{\omega_s(1),...,\omega_s(s)\} \geq m-2s$, the number of its non-repeating elements. Denote by $\tilde{h}_{\omega_s}$ the symmetric version of the function

$$\tilde{h}_{\omega_s}(\tilde{z}_1,...,\tilde{z}_s) = \tilde{h}\left(\tilde{z}_1,...,\tilde{z}_{m-s},\tilde{z}_{\omega_s(1)},...,\tilde{z}_{\omega_s(s)}\right).$$

We can write:

$$\begin{aligned}&\left\|\frac{1}{n^m}\sum_{i_1,...,i_m}\tilde{h}\left(\tilde{Z}_{i_1},...,\tilde{Z}_{i_m}\right)\right\|_{\tilde{\mathcal{H}}}^p\\ \lesssim\ &\left\|\tilde{U}_n^{(m)}\tilde{h}\right\|_{\tilde{\mathcal{H}}}^p + \sum_{1\leq s\leq m/2}\sum_{\omega_s}\left\|n^{-s}\tilde{U}_n^{(m-s)}\tilde{h}_{\omega_s}\right\|_{\tilde{\mathcal{H}}}^p.\end{aligned}$$

Note that functions $\tilde{h}_{\omega_s}$ satisfy the condition

$$\tilde{P}^{\#\{\omega_s(1),...,\omega_s(s)\}+1}\tilde{h}_{\omega_s} = 0$$

(because integrating out any $\#\{\omega_s(1),...,\omega_s(s)\}+1$ variables in the function $\tilde{h}_{\omega_s}$ necessarily involves integrating out at least one non-repeating $\tilde{Z}_i$). Apply operators $\tilde{P}^k$,

$k = m-s, m-s+1..., \#\{\omega_s(1), ..., \omega_s(s)\}+1$, consecutively to both sides of the Hoeffding decomposition of $\tilde{U}_n^{(m-s)}\tilde{h}_{\omega_s}$ relative to the measure $\tilde{P}$ (the corresponding projections are denoted $\pi^{\tilde{P}}_{k,m-s}$), and conclude that its elements of order $k = 0, 1, ..., e_m(\omega_s) - 1$, are zero:

$$\tilde{U}_n^{(s)}\tilde{h}_{\omega_s} = \sum_{k=e_m(\omega_s)}^{m-s} \binom{m-s}{k} \tilde{U}_n^{(k)} \pi^{\tilde{P}}_{k,m-s}\tilde{h}_{\omega_s}.$$

Next note that every function $\pi^{\tilde{P}}_{k,m-s}\tilde{h}_{\omega_s}$ satisfies the assumptions of part (a) of the theorem, so that

$$\left\| n^{k/2}\tilde{U}_n^{(k)}\pi^{\tilde{P}}_{k,m-s}\tilde{h}_{\omega_s} \right\|^p_{\tilde{\mathcal{H}}} = O(1).$$

We then have

$$\begin{aligned} & \left\| n^{\frac{m}{2}}\hat{U}_n^{(m)}h \right\|^p_{\mathcal{H}} \\ \lesssim & \left\| n^{\frac{m}{2}}\tilde{U}_n^{(m)}\tilde{h} \right\|^p_{\tilde{\mathcal{H}}} + \sum_{1\le s\le m/2} \sum_{\omega_s} \sum_{k=e_m(\omega_s)}^{m-s} n^{\frac{m-2s-k}{2}} \left\| n^{k/2}\tilde{U}_n^{(k)}\pi^{\tilde{P}}_{k,m-s}\tilde{h}_{\omega_s} \right\|^p_{\tilde{\mathcal{H}}}. \end{aligned}$$

Next note that in the above sum $m-s-k \le 0$, and the equality can only be achieved when $k = e_m(\omega_s) = m-2s$, that is when all elements of $\omega_s$ are distinct. Finally, note that by the $\tilde{P}$-degeneracy of $\tilde{h}$, for $\omega_s = \{1, 2, ..., s\}$, $\pi^{\tilde{P}}_{m-2s,s}\tilde{h}_{\omega_s}$ is a constant multiple of the function $\tilde{h}^{[m-2s]}(\tilde{z}_1, ..., \tilde{z}_{m-2s})$ (because the other integrals in $\pi^{\tilde{P}}_{m-2s,s}\tilde{h}_{\omega_s}$ involve integrating out non-repeating $\tilde{Z}_i$; also note that $\tilde{h}^{[m-2s]}(\tilde{z}_1, ..., \tilde{z}_{m-2s})$ is $\tilde{P}$-degenerate).

Therefore,

$$\begin{aligned} & P\left\| n^{\frac{m}{2}}\hat{U}_n^{(m)}h \right\|^p_{\mathcal{H}} \\ \lesssim & P\left\| n^{\frac{m}{2}}\tilde{U}_n^{(m)}\tilde{h} \right\|^p_{\tilde{\mathcal{H}}} + \sum_{1\le s\le \frac{m}{2}} P\left\| n^{(m-2s)/2}\tilde{U}_n^{(m-2s)}\tilde{h}^{[m-2s]} \right\|^p_{\tilde{\mathcal{H}}} + O\left(n^{-1/2}\right). \end{aligned}$$

By part (a), the RHS is $O(1)$.

(d) The inequality in the previous display still holds. We check that for $0 \leq s \leq \frac{m}{2}$

$$P \left\| n^{(m-2s)/2} \tilde{U}_n^{(m-2s)} \tilde{h}^{[m-2s]} \right\|_{\tilde{\mathcal{H}}}^p = o(1)$$

(where $\tilde{h}^{[m]} = \tilde{h}$). This will follow from part (b) if we show that

$$\left\| \tilde{P}^{m-2s} \left( \tilde{h}^{[m-2s]} \right)^2 \right\|_{\tilde{\mathcal{H}}_n} \to 0.$$

This follows from the extra condition in (d) and the construction of $\tilde{h}$ from $h$. □

**Proof.** (Lemma 12.) (a) Define the class of functions

$$\mathcal{H}_n = \left\{ h^{\theta,t} = \frac{h_{\theta + n^{-1/2} a_n^{-1} t} - h_\theta}{1 + \|t\|^{1/2}} : \|\theta\|, \ n^{-1/2} a_n^{-1} \|t\| \leq \delta_0 \right\}.$$

Note that $\mathcal{H}_n$ is Euclidean for the envelope $2H$ because it is a subclass of the Euclidean class

$$\left\{ \frac{h_{\theta+t} - h_\theta}{1 + \left\|\tilde{t}\right\|^{1/2}} : \|\theta\|, \ \ \|t\| \leq \delta_0, \ \tilde{t} \in \mathbb{R}^d \right\}.$$

To prove the lemma, it is enough to show that

$$P \left\{ n a_n^2 \left\| U_n^{(m)} h^{\theta,t} \right\|_{\mathcal{H}_n} > 1 \right\} = O\left(a_n^{-1}\right). \tag{A.26}$$

By the Chebyshev inequality,

$$P \left\{ n a_n^2 \left\| U_n^{(m)} h^{\theta,t} \right\|_{\mathcal{H}_n} > 1 \right\} \leq \left(n a_n^2\right)^p P \left\| U_n^{(m)} h^{\theta,t} \right\|_{\mathcal{H}_n}^p.$$

The continuity modulus of class $\mathcal{H}_n$ satisfies

$$\left\| P^m \left( h^{\theta,t} \right)^2 \right\|_{\mathcal{H}_n} \leq C n^{-1/2} a_n^{-1}.$$

By Lemmas 17 and 18,

$$P \left\| U_n^{(m)} h^{\theta,t} \right\|_{\mathcal{H}_n}^p \lesssim \left( n^{-1} \log n \right)^{pm/2} \left( n^{-1/2} a_n^{-1} \right)^{p/2} + \left( n^{-1} \log n \right)^{p(m+1)/2-1}.$$

Therefore, (A.26) is satisfied if

$$n a_n^2 \left( n^{-1} \log n \right)^{m/2} \left( n^{-1/2} a_n^{-1} \right)^{1/2} \leq a_n^{-1/p}$$

and

$$n a_n^2 \left( n^{-1} \log n \right)^{(m+1)/2-1/p} \leq a_n^{-1/p}.$$

These inequalities give

$$a_n \leq \left( n^{m/3-1/2} \left( \log n \right)^{-m/3} \right)^{\frac{1}{1+2/3p}}$$

and

$$a_n \leq \left( n^{\frac{m-1}{4}-\frac{1}{2p}} \left( \log n \right)^{\frac{1}{2p}-\frac{m+1}{4}} \right)^{\frac{1}{1+1/2p}}.$$

from which the result follows immediately.

(b) Let $\mathcal{H}_n$ be as above. Use the inequality obtained in the proof of Lemma 19 (c), rewritten as:

$$\begin{aligned} & P \left\| \hat{U}_n^{(m)} h^{\theta,t} \right\|_{\mathcal{H}}^p \\ \lesssim \; & P \left\| \tilde{U}_n^{(m)} \tilde{h}^{\theta,t} \right\|_{\tilde{\mathcal{H}}}^p + \sum_{1 \leq s \leq \frac{m}{2}} P \left\| n^{-s} \tilde{U}_n^{(m-2s)} \tilde{h}^{\theta,t[m-2s]} \right\|_{\tilde{\mathcal{H}}}^p + O\left( n^{-(m+1)/2} \right). \end{aligned}$$

From the additional assumptions made in part (b) of the lemma, and by construction of functions $\tilde{h}$, we have:

$$\left\| P^m \left( \tilde{h}^{\theta,t} \right)^2 \right\|_{\mathcal{H}_n} \leq C n^{-1/2} a_n^{-1}, \tag{A.27}$$

and, for each $s$, $1 \leq s \leq m/2$,

$$\left\| P^{m-2s} \left( \tilde{h}^{\theta,t[m-2s]} \right)^2 \right\|_{\mathcal{H}_n} \leq C n^{-1/2} a_n^{-1}. \tag{A.28}$$

The result now follows from part (a). In particular, notice that we will have

$$\begin{aligned} & P \left\| n^{-s} \tilde{U}_n^{(m-2s)} \tilde{h}^{\theta,t[m-2s]} \right\|_{\tilde{\mathcal{H}}_n}^p \\ \lesssim \; & n^{-sp} \left( n^{-1} \log n \right)^{p(m-2s)/2} \left( n^{-1/2} a_n^{-1} \right)^{p/2} \\ & + n^{-sp} \left( n^{-1} \log n \right)^{p(m-2s+1)/2-1} \\ = \; & n^{pm/2} \left( n^{-1} \log n \right)^{p(m-2)/2} \left( n^{-1/2} a_n^{-1} \right)^{p/2} \\ & + \left( n^{-1} \right)^{p(m+1)/2-1} (\log n)^{p(m-1)/2-1}. \end{aligned}$$

which is dominated by the bound for $P \left\| \tilde{U}_n^{(m)} \tilde{h}^{\theta,t} \right\|_{\tilde{\mathcal{H}}_n}^p$ obtained in part (a) under condition (A.27). □

## A.4. A Berry-Esséen Bound

**Lemma 20.** *Let $Z_1, ..., Z_n$ be i.i.d. random variables taking values in a probability space $(\mathcal{Z}, P)$ ($P$ may depend on $n$), and let $g_\theta : \mathcal{Z}^2 \to \mathbb{R}^d$, $\theta \in \mathbb{R}^d$ be a vector-function,*

*symmetric in* $z_1, z_2$. *Let* $\theta_n$ *solve the system of equations*

$$U_n^{(2)} g_\theta^{(l)} = \xi_n^{(l)}(\theta),$$

$l = 1, ..., d$. *Assume that there are numbers* $\delta_0, K > 0$ *such that for all* $n \geq 1$*:*

*(i)* $P\left\{\sup_{\|\theta\| \leq \delta_0} \|\xi_n(\theta)\| > n^{-1}\right\} \leq K n^{-1/2}$.

*(ii)* $\|P g_0\| \leq K n^{-1}$.

*(iii)*$P\|\pi_{1,2} g\|^4 < \infty$, $P\|\pi_{2,2} g\|^2 < \infty$.

*(iv)* $g_\theta$ *is twice continuously differentiable in the* $\delta_0$*-neighborhood of* $0$, $P-a.e.$, $P\|\partial^2 g_0\|^4 < \infty$, $P\|\partial g_0\|^3 < \infty$, *and there is* $L(z_1, z_2)$ *with* $P^2 L^3 < \infty$ *such that*

$$\left\|\partial^2 g_{\theta_1} - \partial^2 g_{\theta_2}\right\| \leq L\|\theta_1 - \theta_2\|,$$

*for all* $\|\theta_1\|, \|\theta_2\| \leq \delta_0$.

*(v) The* $d \times d$ *matrix* $\Gamma = [\partial P g_0]^{-1} Var(2\pi_{1,2} g_0) [\partial P g_0]^{-1}$ *is well defined and is positive definite (her* $Var$ *is the variance relative to* $P$*).*

*Then for all* $n \geq 1$,

$$\sup_{A \in \mathcal{A}} \left| \int_A dF_{n^{1/2}\theta_n} - \int_A d\Phi_\Gamma \right| \leq c_d n^{-1/2} + c_d P\{\|\theta_n\| > \delta_d\},$$

*where, for each* $d$, $c_d$ *and* $\delta_d < \delta_0$ *are continuous functions of* $K$, $P\|\pi_{1,2} g_0\|^4$, $P\|\pi_{2,2} g_0\|^2$, $P\|\partial^2 g_0\|^3$, $P\|\partial g_0\|^4$, $PL^3$.

**Proof.** Within the proof, $\lesssim$ denotes an inequality up to a multiplicative constant that may depend on $d$ only, and $c_d$ and $\delta_d$ satisfy the conditions of the theorem (but may

change from line to line in the proof). To reduce notation, we assume, without loss of generality, that $\delta_0 \leq 1$ and $K \geq 1$. It is enough to consider the case of $\|\theta_n\| < \delta_0$.

**Step 1.** Here we prove that, for all $n \geq 1$,

$$P\left\{\|\theta_n\| > n^{-1/3}\right\} \leq c_d n^{-1/2}.$$

Without loss of generality, assume in this step that $\partial P g_0 = I$ (identity matrix). Use the Taylor expansion around $\theta = 0$ and the Hoeffding decomposition (we omit the index $l$):

$$\begin{aligned} \xi_n(\theta) &= U_n^{(2)} g_\theta \\ &= U_n^{(2)} g_0 + \left\{I + P_n \partial \pi_{1,2} g_{\tilde{\theta}}\right\} \theta, \end{aligned}$$

where $\tilde{\theta}$ lies between 0 and $\theta$. By our assumptions, the class $\{P_n \partial \pi_{1,2} g_\theta,\ \|\theta\| \leq \delta\}$ is Euclidean class for the envelope

$$M(z_1, z_2) = 1 + \|\partial g_0(z_1, z_2)\| + \left\|\partial^2 g_0(z_1, z_2)\right\| + 2\sqrt{d} L(z_1, z_2)$$

(see Lemma (2.13) of Pakes and Pollard [**1989**] and use the identity, for any $f$,

$$\partial_l f_\theta = \partial_l f_0 + \int_0^\theta \partial^2_{l,l} f_{\tilde{\theta}} d\theta^{(l)} = \partial_l f_0 + \partial^2_{l,l} f_0 + \int_0^\theta \left(\partial^2_{l,l} f_{\tilde{\theta}} - \partial^2_{l,l} f_0\right) d\theta^{(l)}.$$

The envelope is made bigger than necessary to simplify notation later on). Therefore,

$$P\left\{\|P_n \partial \pi_{1,2} g_{\tilde{\theta}}\| \geq \frac{1}{2}\right\} \lesssim n^{-1} P M^2$$

(to see this, use the fact that $P\partial\pi_{1,2}g_\theta = 0$ for each non-random $\theta$, because $P\pi_{1,2}g_\theta = 0$ and $P\|\partial g_0\| < \infty$; then apply the bound for the suprema for the second moment to deal with the randomness in $\tilde{\theta}$).

Then

$$\begin{aligned} & P\left\{\|\theta_n\| > n^{-1/3}\right\} \\ \lesssim\ & P\left\{4\left\|U_n^{(2)}g_0\right\| > n^{-1/3}\right\} + P\left\{2\|\xi_n(\theta)\| > n^{-1/3}\right\} + n^{-1}PM^2 \\ \lesssim\ & Kn^{-1/2}, \end{aligned}$$

by the Hoeffding decomposition, Maximal and Chebyshev inequalities. Namely,

$$\begin{aligned} & P\left\{4\left\|U_n^{(2)}g_0\right\| > n^{-1/3}\right\} = P\left\{4n^{1/2}\left\|U_n^{(2)}g_0\right\| > n^{1/6}\right\} \\ \lesssim\ & n^{-1/2}P\|\pi_{1,2}g_0\|^3 + n^{-1/3}n^{-2/3}P\|\pi_{2,2}g_0\|^2 + Kn^{-1/2} \end{aligned}$$

(note that here we have used the fact that $\|P^2 g_0\| \leq Kn^{-1}$). Note also that

$$P\left\{\|2\xi_n(\theta)\| > n^{-1/3}\right\} \leq P\left\{2\|\xi_n(\theta)\| > n^{-1}\right\} \lesssim Kn^{-1/2}.$$

This gives the estimate.

**Step 2.** Obtain the representation: for all $n \geq 1$,

$$\theta^{(l_1)} = U_n^{(2)}g_*^{(l_1)} + C_{l_1,l_2,l_3}\theta^{(l_1)}\theta^{(l_2)} + \zeta_n^{(l_1)}, \tag{A.29}$$

where $C_{l_1,l_2,l_3}$ are constants, function $g_*^{(l_1)}(z_1,z_2)$ does not depend on $\theta$ and satisfies $P^2 g_* = 0$, $Var\left[2\pi_{1,2}g_*\right] = \Gamma$, $P^2\left\|\pi_{1,2}g_*\right\|^3 < \infty$, $P^2\left\|\pi_{2,2}g_*\right\|^2 < \infty$; and

$$P\left\{\|\zeta_n\| > c_d n^{-1}\right\} \leq n^{-1/2}c_d + P\left\{\|\theta_n\| > \delta_d n^{-1/3}\right\}.$$

In this step, assume, without loss of generality, that $\partial P g_0 = I$. By the Hoeffding decomposition and the Taylor expansion, for each component $l_1$,

$$\begin{aligned}(\text{A.30}) \qquad 0 &= U_n^{(2)} g_\theta^{(l_1)} - \xi_{n,\theta}^{(l_1)} \\ &= U_n^{(2)}\left(g_0^{(l_1)} - P^2 g_0^{(l_1)}\right) + \left\{\left(I + B_n\right)\theta\right\}^{(l_1)} - C_{l_1,l_2,l_3}\theta^{(l_2)}\theta^{(l_3)} \\ &\quad + \tilde{\zeta}_n^{(l_1)}(\theta),\end{aligned}$$

where $(B_n)_{l_1,l_2} = 2P_n\partial_{l_2}\pi_{1,2}g_0^{(l_1)}$, $C_{l_1,l_2,l_3} = -\frac{1}{2}\partial_{l_2,l_3}P^2 g_0^{(l_1)}$, and

$$\begin{aligned}\tilde{\zeta}_n^{(l_1)}(\theta) &= \frac{1}{2}\left(\partial_{l_2,l_3}P^2 g_{\tilde{\theta}}^{(l_1)} - \partial_{l_2,l_3}P^2 g_0^{(l_1)}\right)\theta^{(l_2)}\theta^{(l_3)} \\ &\quad + P_n\partial_{l_2,l_3}\pi_{1,2}g_{\tilde{\theta}}^{(l_1)}\theta^{(l_2)}\theta^{(l_3)} \\ &\quad + U_n^{(2)}\partial_{l_2}\pi_{2,2}g_{\tilde{\theta}}^{(l_1)}\theta^{(l_2)} - \xi_{n,\theta}^{(l_1)} + P^2 g_0^{(l_1)},\end{aligned}$$

and $\tilde{\theta}$ is in between 0 and $\theta$ (in fact, $\tilde{\theta}$ is different for each $l_1$ and each term above, but we will ignore this distinction). (Note that by the Chebyshev inequality

$$P\left\{\|B_n\| \geq \frac{1}{2}\right\} \lesssim n^{-1}P\left\|\partial\pi_{1,2}g_0\right\|^2,$$

and, therefore, it is enough to restrict attention to the event $\left\{\|B_n\| < \frac{1}{2}\right\}$.)

Using the identity

$$\begin{aligned}(I+B_n)^{-1} &= I - B_n (I+B_n)^{-1} \\ &= I - B_n + B_n^2 (I+B_n)^{-1},\end{aligned}$$

we can rewrite (A.30) as

$$\theta^{(l_1)} = U_n^{(2)} g_*^{(l_1)} + C_{l_1,l_2,l_3}\theta^{(l_1)}\theta^{(l_2)} + \zeta_n^{(l_1)}(\theta),$$

where

$$g_*^{(l_1)}(z_1,z_2) = -\left(g_0^{(l_1)}(z_1,z_2) - P^2 g_0^{(l_1)}\right) - 2\partial_{l_2}\pi_{1,2}g_0^{(l_1)}(z_1)\cdot\pi_{1,2}g_0^{(l_2)}(z_2),$$

(in particular, the random vector $g_*$ satisfies the above properties), and

$$\begin{aligned}\zeta_n^{(l_1)}(\theta) &= U_n^{(2)}\left\{-(I-B_n)\left(g_0 - P^2 g_0\right) - g_*\right\}^{(l_1)} \\ &\quad - \left\{B_n^2 (I+B_n)^{-1} U_n^{(2)}\left(g_0 - P^2 g_0\right)\right\}^{(l_1)} \\ &\quad + \left[B_n (I+B_n)^{-1}\right]_{l_1,l_2} C_{l_2,l_3,l_4}\theta^{(l_3)}\theta^{(l_4)} \\ &\quad - \left\{(I+B_n)^{-1}\tilde{\zeta}_n(\theta)\right\}^{(l_1)}.\end{aligned}$$

It remains to obtain the estimate for the remainder term $\zeta_n^{(l_1)}(\theta)$. It is enough to restrict attention to the event that $\|\theta_n\| \le n^{-1/3}$. First, consider the expression for $\tilde{\zeta}_n^{(l_1)}(\theta)$. Its first term has the order $PL\cdot\|\theta\|^3$, from which the bound follows. For the second term,

we have

$$\begin{aligned} & P\left\{\left\|P_n\partial_{l_2,l_3}\pi_{1,2}g_{\tilde{\theta}}^{(l_1)}\right\|\|\theta\|^2 > n^{-1}\right\} \\ \leq\ & P\left\{\left\|P_n\partial_{l_2,l_3}\pi_{1,2}g_{\tilde{\theta}}^{(l_1)}\right\| > n^{-1/3}\right\} \\ =\ & P\left\{n^{1/2}\left\|P_n\partial_{l_2,l_3}\pi_{1,2}g_{\tilde{\theta}}^{(l_1)}\right\| > n^{1/6}\right\} \\ \lesssim\ & n^{-1/2}\left(P^2\left\|\pi_{1,2}g\right\|^3 + P^2\left\|\pi_{2,2}g\right\|^2 + P^2M^2\right), \end{aligned}$$

where we have used the fact that the classes $\left\{\partial_{l_1,l_2}\pi_{1,2}g_\theta(z),\ \theta\in\mathbb{R}^d, \|\theta\|\leq\delta_0\right\}$ and $\left\{\partial_{l_1,l_2}\pi_{2,2}g_\theta(z),\ \theta\in\mathbb{R}^d, \|\theta\|\leq\delta_0\right\}$, for each $l_1, l_2$, are Euclidean for the envelope

$$M(z_1,z_2) = 1 + \|\partial g_0(z_1,z_2)\| + \left\|\partial^2 g_0(z_1,z_2)\right\| + 2\sqrt{d}L(z_1,z_2)$$

(the envelope is made bigger than necessary to reduce notation here and below). For the third,

$$\begin{aligned} & P\left\{\left\|U_n^{(2)}\partial_{l_2}\pi_{2,2}g_{\tilde{\theta}}^{(l_1)}\right\|\|\theta\| > n^{-1}\right\} \\ \leq\ & P\left\{\left\|U_n^{(2)}\partial_{l_2}\pi_{2,2}g_{\tilde{\theta}}^{(l_1)}\right\| > n^{-2/3}\right\} \\ \leq\ & n^{-1/2}\left(P\left\|\pi_{1,2}g\right\|^3 + P\left\|\pi_{2,2}g\right\|^2 + PM^2\right). \end{aligned}$$

The bounds for the last two terms are assumed in the theorem.

Finally, we show that $P\left\{\|\zeta_n\| > c_d n^{-1}\right\} \leq c_d n^{-1/2}$. For the first line in the expression for $\zeta_n$, we have

$$\begin{aligned} & U_n^{(2)}\left\{(I - B_n)\left(g_0 - P^2 g_0\right) + g_*\right\} \\ = \; & -n^{-1} P_n h - B_n U_n^{(2)} \pi_{2,2} g_0 + n^{-1} U_n^{(2)}(g_* - g_0), \end{aligned}$$

where $h(z) = 2\partial_{l_2} \pi_{1,2} g_0(z) \cdot \pi_{1,2} g_0^{(l_2)}(z)$. Note that

$$\begin{aligned} & P\left\{n^{-1}\|P_n h\| \geq (P\|h\| + 1) n^{-1}\right\} \\ \leq \; & P\left\{\left\|n^{1/2}(P_n - P) h\right\| \geq n^{1/2}\right\} \\ \leq \; & n^{-1} P\|h\|^2 \lesssim n^{-1} P\|\partial \pi_{1,2} g_0\|^4 \cdot P\|\pi_{1,2} g_0\|^4; \end{aligned}$$

$$\begin{aligned} & P\left\{\left\|B_n U_n^{(2)} \pi_{2,2} g_0\right\| > n^{-1}\right\} \\ \leq \; & P\left\{\left\|n^{1/2} B_n \cdot n U_n^{(2)} \pi_{2,2} g_0\right\| > n^{1/2}\right\} \\ \lesssim \; & n^{-1/2} P\|\partial \pi_{1,2} g_0\|^2 + n^{-1/2} P\|\pi_{2,2} g_0\|^2; \end{aligned}$$

$$P\left\{n^{-1}\left\|U_n^{(2)}(g_* - g_0)\right\| > n^{-1}\right\} \lesssim n^{-2}\|\partial \pi_{1,2} g_0\|^2 \|\pi_{1,2} g_0\|^2.$$

Now consider the second through fourth lines. By the Chebyshev inequality,

$$P\left\{\|B_n\| \geq n^{-1/3}\right\} \lesssim n^{-1/2} P\|\partial \pi_{1,2} g_0\|^3.$$

From the bound on $P\left\{\|B_n\| \geq \frac{1}{2}\right\}$ (where $\|\cdot\|$ is the spectral norm for matrices),

$$P\left\{\left\|(I + B_n)^{-1}\right\| \geq 2\right\} \lesssim n^{-1} P\|\partial \pi_{1,2} g_0\|^2.$$

Therefore,

$$\begin{aligned} & P\left\{\left\|\zeta_n^{(l_1)}(\theta_n)\right\| > n^{-1}\right\} \\ \lesssim\ & P\left\{\|B_n\| \geq n^{-1/3}\right\} + P\left\{U_n^{(2)} g_0 > n^{-1/3}\right\} \\ & + \|C_{l_1,l_2,l_3}\| * \\ & \left\{P\left\{\|B_n\| \geq n^{-1/3}\right\} + P\left\{\|\theta\| > n^{-1/3}\right\} + P\left\{\left\|(I+B_n)^{-1}\right\| \geq 2\right\}\right\} \\ & + P\left\{\left\|(I+B_n)^{-1}\right\| \geq 2\right\} + P\left\{\left\|\tilde{\zeta}_n^{(l_1)}(\theta_n)\right\| > n^{-1}\right\}. \end{aligned}$$

Combining these estimates gives the result.

**Step 3.** Now use representation (A.29) to obtain the Berry-Esséen bound.

Consider the system of equations

$$\theta^{(l_1)} = \gamma^{(l_1)} + C_{l_1 l_2 l_3} \theta^{(l_2)} \theta^{(l_3)}. \tag{A.31}$$

By the Implicit Function Theorem and the Taylor expansion, there are numbers $\delta^* > 0$, $K_1 > 0$, and $b_{l_1 l_2 l_3}$, continuously depending on $C_{l_1 l_2 l_3}$, such that if $\|\gamma\| \leq \delta^*$, and $\theta$ is the solution of (A.31) satisfying $\|\theta\| \leq \delta^*$, then

$$\theta^{(l_1)} = \gamma^{(l_1)} + b_{l_1 l_2 l_3}^{(2)} \gamma^{(l_2)} \gamma^{(l_3)} + \phi^{(l_1)}(\gamma)$$

and

$$\|\phi(\gamma)\| \leq K_1 \|\gamma\|^3.$$

Let $\gamma_n = U_n^{(2)} g_* + \zeta_n$. By the Hoeffding decomposition, the properties of $g_*$ and $\zeta_n$, and the Chebyshev inequality,

$$P\left\{\|\gamma_n\| > \delta^*\right\} \lesssim c_d n^{-1/2} + P\left\{\|\theta\| > \delta_d n^{-1/3}\right\}, \tag{A.32}$$

and

$$\begin{aligned}
& P\left\{\|\phi(\gamma_n)\| \geq n^{-1}, \|\gamma_n\| \leq \delta^*\right\} \\
\leq\ & P\left\{K_1^{1/3}\|\gamma_n\| \geq n^{-1/3}\right\} \\
\lesssim\ & P\left\{2K_1^{1/3}\left\|n^{1/2}P_n\pi_{1,2}g_*\right\| \geq n^{1/6}\right\} \\
& + P\left\{2K_1^{1/3}\left\|nU_n^{(2)}\pi_{2,2}g_*\right\| \geq n^{2/3}\right\} \\
\lesssim\ & n^{-1/2}\left(P\|\pi_{1,2}g_*\|^3 + P\|\pi_{2,2}g_*\|^2\right).
\end{aligned}$$

Next, consider the statistic $T_n$ defined as follows:

$$T_n^{(l_1)} = U_n^{(2)} g_* + b_{l_1 l_2 l_3} \frac{1}{n^2} \sum_{i \neq j} \left\{\pi_{1,2} g_{*i}^{(l_2)} \cdot \pi_{1,2} g_{*j}^{(l_3)}\right\},$$

where $\pi_{1,2} g_{*i}^{(l_2)} \equiv \pi_{1,2} g_*^{(l_2)}(Z_i)$.

Note that

$$
\begin{aligned}
\text{(A.33)} \qquad & n^{1/2} P\left\{\left|\theta_n^{(l_1)} - T_n^{(l_1)}\right| \geq n^{-1}, \|\gamma_n\| \leq \delta^*, \|\theta_n\| \leq \delta^*\right\} \\
\lesssim\ & n^{1/2} P\left\{3\left|b_{l_1 l_2 l_3}\left(U_n^{(2)} g_*^{(l_2)} U_n^{(2)} g_*^{(l_3)} - P_n \pi_{1,2} g_*^{(l_2)} P_n \pi_{1,2} g_*^{(l_3)}\right) \geq n^{-1}\right|\right\} \\
& + n^{1/2} P\left\{3\left|b_{l_1 l_2 l_3} \frac{1}{n^2} \sum_i \pi_{1,2} g_{*i}^{(l_2)} \cdot \pi_{1,2} g_{*i}^{(l_3)}\right| \geq n^{-1}\right\} \\
& + P\left\|\pi_{1,2} g_*\right\|^3 + P\left\|\pi_{2,2} g_*\right\|^2 \\
\lesssim\ & P\left\|\pi_{1,2} g_* \cdot \pi_{2,2} g_*\right\| + P\left\|\pi_{1,2} g_*\right\|^2 + P\left\|\pi_{1,2} g_*\right\|^3 + P\left\|\pi_{2,2} g_*\right\|^2 \\
\lesssim\ & P\left\|\pi_{1,2} g_*\right\|^2 + P\left\|\pi_{1,2} g_*\right\|^3 + P\left\|\pi_{2,2} g_*\right\|^2 .
\end{aligned}
$$

Using (A.32) and (A.33), we conclude that

$$\theta_n = T_n + \xi_n$$

$$\text{(A.34)} \qquad P\left\{\|\xi_n\| > n^{-1}\right\} \leq c_d n^{-1/2} + P\left\{\|\theta\| > \delta_d n^{-1/3}\right\}.$$

$T_n$ has a form of a $U$-statistic of order 2 with zero mean. Its variance is $n^{-1}\Gamma$ up to a term of order $O\left(n^{-2}\right)$.

The Berry-Esséen bound for $T_n$ follows from Theorem 2 of Bolthausen and Götze [**1993**]. To check the conditions of the theorem, let $T_0 = P_n \pi_{1,2} g_*$. Then, in the notation of the theorem, and using the Cauchy-Schwartz inequality,

$$\beta_3\left(n^{1/2} T^0\right) = nP\left\|n^{-1/2} \pi_{1,2} g_*\right\|^3 = n^{-1/2} P\left\|\pi_{1,2} g_*\right\|^3,$$

$$
\begin{aligned}
& \delta\left(n^{1/2}T_n, n^{1/2}T_0\right) \\
= \;& n^{1/2}P\left[\left|\left(T-T^0\right)(Z_1, Z_2..., Z_n) - \left(T-T^0\right)(Z_{n+1}, Z_2..., Z_n)\right|\right] \\
\lesssim \;& n^{-1}P\left\|n^{-1/2}\sum_{j=2}^{n}\pi_{2,2}g_*(Z_1, Z_j)\right\| \\
& + n^{-1}\left(P\left\|\pi_{1,2}g_*\right\|^2\right)^{1/2}\left(P\left\|n^{1/2}P_n\pi_{1,2}g_*\right\|^2\right)^{1/2} \\
\leq \;& n^{-1}\left(P\left\|\pi_{2,2}g_*\right\|^2 + P\left\|\pi_{1,2}g_*\right\|^2\right),
\end{aligned}
$$

and, by the Maximal inequality for a degenerate $U$-statistic of order 2,

$$
n^{1/2}E\left|\left(T-T^0\right)(Z_1, Z_2..., Z_n)\right| = n^{-1/2}\left(P\left\|\pi_{2,2}g_*\right\|^2 + P\left\|\pi_{1,2}g_*\right\|^2\right)^{1/2}.
$$

Then by the above-mentioned theorem applied to $n^{1/2}T_n$,

$$
\sup_{A\in\mathcal{A}}\left|\int_A dF_{n^{1/2}T_n} - \int_A d\Phi_\Gamma\right| \leq c_d n^{-1/2}.
$$

The conclusion of the theorem follows from the last result and (A.34) by an argument similar to that at the end of Section A.2.3. □

APPENDIX B

# Appendix to Chapter 4

## B.1. Proofs of Theorems

**Proof.** (Theorem 6.)

(1) Identification. We check that $\theta_0$ is the unique maximizer of the population objective function:

$$S(\theta) = E\left[M(Y_1, Y_2)\, 1\left\{X_1'\beta(\theta) > X_2'\beta(\theta)\right\} w_s(X_1, X_2)\right]$$

where $\beta(\theta) = (\theta, 1)$ (note that changing $w$ to $w_s$ does not affect the optimization problem by the antisymmetry of $M$ and the continuity of the distribution of $X'\beta_0$). Define

$$\begin{aligned}\Delta S(\theta; X_1, X_2) &= E\left[M(Y_1, Y_2)\,|X_1, X_2\right] w_s(X_1, X_2) \cdot \\ &\quad \left(1\left\{X_1'\beta(\theta_0) > X_2'\beta(\theta_0)\right\} - 1\left\{X_1'\beta(\theta) > X_2'\beta(\theta)\right\}\right).\end{aligned}$$

Condition (4.1) and antisymmetry of $M$ imply that

$$E\left[M(Y_1, Y_2)\,|X_1, X_2\right] < 0 \Longrightarrow X_1'\beta_0 < X_2'\beta_0 \tag{B.1}$$

Together, (4.1) and (B.1) imply that $\Delta S(\theta; X_1, X_2) \geq 0$ for all $\theta$ and almost all $X_1, X_2$. In particular, $\theta_0$ is a maximizer of $S(\theta)$.

Write $S(\theta_0) - S(\theta)$ as an iterated integral:

$$\begin{aligned} & S(\theta_0) - S(\theta) \\ = & \int \Delta S(\theta; x_1, x_2)\, w_s(x_1, x_2)\, g_{Z|U}(z_1, u_1)\, g_{Z|U}(z_2, u_2)\, dz_1 dz_2 dG(u_1)\, dG(u_2), \end{aligned}$$

where $x_i = x(u_i, z_i)$, and $G(u)$ is the marginal c.d.f. of $U$. Assume that $\theta_*$ is another maximizer of $S(\theta)$, so that $S(\theta_0) - S(\theta_*) = 0$. Since $\Delta S(\theta; x_1, x_2) \geq 0$, the integrated expressions must be zero almost surely. Also, condition (4.1) implies that $\mu(u_1, u_2, z_0) \geq 0$ a.s. It follows that there are $d$ pairs

$$u_{1,k}, u_{2,k} \in \mathcal{U},\ k = 1, ..., d,$$

a number $z_0 \in I$ and a countable set of numbers $z_t \in I$, dense in $I$, such that the $d$ vectors $(u_{1,k} - u_{2,k})$ are linearly independent, the function $\mu$ is well-defined and is positive on the vectors $(u_{1,k}, u_{2,k}, z_0)$, and

$$\Delta S(\theta; x(u_{1,k}, z_t), x(u_{2,k}, z_0)) = 0,\ k = 1, ..., d,\ \forall t.$$

As $\mu(u_{1,k}, u_{2,k}, z_0) > 0$, for $z_t$ sufficiently close to $z_0$,

$$E[M(Y_1, Y_2) | X_1 = x(u_{1,k}, z_t), X_2 = x(u_{2,k}, z_0)]$$

has the same sign as $z_t - z_0$. By conditions (4.1), (B.1), for such $z_t$,

$$sign\left\{(u_{1,k} - u_{2,k})'(\theta_* - \theta_0) + z_t - z_0\right\} = sign\{z_t - z_0\},$$

so that

$$(u_{1,k} - u_{2,k})' (\theta_* - \theta_0) = 0.$$

Since the $d$ vectors $u_{1,k} - u_{2,k}$ are linearly independent, $\theta_* = \theta_0$.

(2) Consistency. Under the conditions of the theorem, the population objective function is continuous in $\theta$. Note also that the class of functions $\left\{1\left\{X_1'\beta > X_2'\beta\right\}, \beta \in \mathbb{R}^{d+1}\right\}$ is Euclidean for a constant envelope as shown by Sherman (1993). Therefore, the class of functions

$$M(Y_1, Y_2)\, 1\left\{X_1'\beta(\theta) > X_2'\beta(\theta)\right\} w_s(X_1, X_2)$$

is Euclidean for the square-integrable envelope $|M(Y_1, Y_2)|\, w_s(X_1, X_2)$ (for the latter, the square-integrability follows from (4.4) and antisymmetry of $M$). By the maximal inequalities for $U-$processes (see Lemma 11), the sample objective function converges to the population objective function in probability uniformly in $\theta \in \Theta$. Therefore, $\theta_n \to^p \theta_0$ by a standard argument.

(3) To prove the asymptotic normality we check the conditions of Theorem 1. Assumptions 1 and 2 have already been verified. We now check the smoothness properties of the function $\tau_\theta(y_1, z_1)$ defined as

$$\tau_\theta(y, u, z) = E\left[M(y, Y)\, sign\left(z - Z + (u - U)'(\theta - \theta_0)\right) w_s(x, X)\right].$$

A calculation as in Sherman (1993) shows that the gradient and the Hessian of the function $\tau_\theta$ with respect to $\theta$ are:

$$\partial_\theta \tau_\theta(y, u, z) = 2E\left[(u - U)\, \lambda(y, U, Z_\theta)\, w_s(x(u, z), x(U, Z_\theta))\, g_{V|U}(V_\theta)\right]$$

and

$$\begin{aligned} & \partial_\theta^2 \tau_\theta (y,u,v) \\ = \; & 2E\left[(u-U)(u-U)'\left(\partial_z \lambda (y,U,Z_\theta)\right) w\left(x(u,z), x(U,Z_\theta)\right) g_{Z|U}(Z_\theta)\right] \\ & +2E\left[(u-U)(u-U) S(y,U,Z_\theta)\, \partial_z \left(w\left(x(u,z), x(U,Z_\theta)\right) g_{Z|U}(Z_\theta)\right)\right] \end{aligned}$$

where

$$Z_\theta = z + (\theta - \theta_0)'(u-U).$$

Then required properties follow from our Assumptions 13, 14 and 16. At $\theta = \theta_0$,

$$\partial_{\theta_0} \tau_\theta (y,u,z) = 2\nabla_w (y,u,z)$$

Note also that by the continuity of the function $\lambda(y,u,z)$ in $z$ and (4.1),

$$E\left[\lambda(Y,u,z) \,|X = x(u_2,z)\right] = 0,$$

therefore,

$$E\left[\partial_{\theta_0}^2 \tau_\theta (Y,U,Z)\right] = -2\Delta_w$$

By our Assumptions 16, 17, the variance of $\partial_{\theta_0}\tau_\theta$ is a finite, positive definite matrix, and the expected value of $\partial_{\theta_0}^2 \tau_\theta (Y,U,V)$ is a finite matrix. By Assumption 15, $\Delta_w$ is positive definite. Assumption 4 of Theorem 1 can be checked in the same way as in Sherman (1993), so that the conclusion of part (b) follows. □

**Lemma 21.** *Under Assumptions 11-17, the function $\mu(u_1,u_2,z)$ is symmetric in $u_1$ and $u_2$ for almost all $u_1, u_2, z$.*

**Proof.** By the antisymmetry of $M$,

$$\begin{aligned} & E\left[M\left(Y_1, Y_2\right) | U_1 = u_2, U_2 = u_1, Z_1 = z + \delta, Z_2 = z\right] \\ = \; & -E\left[M\left(Y_2, Y_1\right) | U_1 = u_2, U_2 = u_1, Z_1 = z + \delta, Z_2 = z\right] \\ = \; & -E\left[M\left(Y_1, Y_2\right) | U_1 = u_1, U_2 = u_2, Z_1 = z, Z_2 = z + \delta\right] \end{aligned}$$

where in the second equality we exchanged the labels 1 and 2 of the variables that are integrated out. Note that the function

$$\varphi\left(z_1, z_2\right) = E\left[M\left(Y_1, Y_2\right) | U_1 = u_2, U_2 = u_1, Z_1 = z, Z_2 = z\right]$$

(or fixed $u_1$ and $u_2$) is differentiable in $z_1$ and $z_2$ and satisfies $\varphi\left(z, z\right) = 0$. Therefore, $\varphi'_{z_1}\left(z, z\right) + \varphi'_{z_2}\left(z, z\right) = 0$. This implies that the derivative of the right-hand side of the previous display at $\delta = 0$ is equal to the derivative of

$$E\left[M\left(Y_1, Y_2\right) | U_1 = u_1, U_2 = u_2, Z_1 = z + \delta, Z_2 = z\right];$$

therefore, $\mu\left(u_1, u_2, z\right) = \mu\left(u_2, u_1, z\right)$. □

**Proof.** (Theorem 7.)

Define

$$A_1 = \frac{\mu\left(Z_1\right)}{\sigma\left(X_1\right) E\left[\frac{1}{\sigma^2(X_2)} \middle| Z_2 = Z_1\right]} E_2\left[\frac{U_1 - U_2}{\sigma^2\left(X_2\right)} \middle| Z_2 = Z_1\right].$$

Then

$$E\left[A_1 W_1'\right] = E\left[W_1 A_1'\right] = \Delta_w$$

(note that we have used the symmetry of $w_s$). The desired inequality then follows from the matrix version of the Cauchy-Schwartz inequality:

$$E\left[W_1 W_1'\right] \geq E\left[W_1 A_1'\right] E\left[A_1 A_1'\right]^{-1} E\left[A_1 W_1'\right].$$

□

## B.2. Estimation of the Optimal Weighting Functions

Here we show that the feasible optimal rank estimators are asymptotically equivalent to the (unfeasible) rank estimators with the theoretical optimal weighting functions. The preliminary results, stated first, can also be used to prove the asymptotic equivalence under the conditions different from those imposed in Theorem 8.

Make the following assumptions.

**Assumption 19.** *$\Theta$ is a finite-dimensional set. $\{m_\theta(z_1, z_2) : \theta \in \Theta\}$ is a Euclidean class of symmetric functions for a square-integrable envelope.*

**Assumption 20.** *For some $m \geq 0$, and each $s = 1, ..., S$, and $n = 1, 2, ...,$*

$$\left\{\psi_{\gamma n}^{(s)}(z_1, z_2; x_{11}, ..., x_{1m}) : \gamma \in \Gamma\right\}$$

*is a Euclidean class, with the Euclidean constants not depending on $n$, of functions bounded by a constant $L_n$, symmetric in $z_1, z_2$ and $x_{11}..., x_{1m}$, and such that for each $\gamma \in \Gamma$ and almost all $z_1, z_2$,*

$$E\left[\psi_{\gamma n}(z_1, z_2; X_{11}..., X_{1m})\right] = 0. \tag{B.2}$$

*The numbers $L_n$ satisfy the condition:*

$$n^{-1}L_n^2 \to 0.$$

**Assumption 21.** *The set $\Theta$ is a convex open set. Define the function*

$$l_{\gamma,\theta,n} = E\left[m_\theta\left(z_1, Z_2\right)\prod_{s=1}^{S}\psi_{\gamma n}^{(s)}\left(z_1, Z_2; x_{s1}, x_{s2}, x_{s3}, ..., x_{sm}\right)\right].$$

*For each $n$, $l_{\gamma,\theta,n}$ is continuously differentiable in $\theta$, and the class of functions*

$$\{\partial_\theta l_{\gamma,\theta,n},\ \theta \in \Theta, \gamma \in \Gamma\}$$

*is Euclidean for an envelope $F\left(z_1\right)\cdot L_n^S$, satisfying $EF^2 < \infty$, with the Euclidean numbers not depending on $n$.*

**Assumption 22.** *The function*

$$E_{Z_1}\left[l_{\gamma,\theta,n}\right] \equiv E\left[l_{\gamma,\theta,n}\left(Z_1; x_{11}, ..., x_{Sm}\right)\right]$$

*is twice continuously differentiable in $\theta$, and satisfies:*

$$\partial_{\theta_0} E_{Z_1}\left[l_{\gamma,\theta,n}\right] = 0. \tag{B.3}$$

*For each $n$, the class of functions $\left\{\partial_\theta^2 E_{Z_1}\left[l_{\gamma,\theta,n}\right],\ \theta \in \Theta, \gamma \in \Gamma\right\}$ is Euclidean for the constant envelope $L_n^S$, with the Euclidean numbers not depending on $n$.*

Consider the sum

$$\begin{aligned} & S_{\gamma,\theta,n} \\ = \; & n^{-2}\sum_{i\neq j} m_\theta\left(Z_i,Z_j\right)\prod_{s=1}^{S}\left(n^{-m}\sum_{k_1,...,k_m}\psi_{\gamma n}^{(s)}\left(Z_i,Z_j;X_{k_1},...,X_{k_m}\right)\right). \end{aligned}$$

**Theorem 22.** *(a) Under Assumptions 19-20,*

$$\sup_{\gamma\in\Gamma,\theta\in\Theta}\left|S_{\gamma,\theta,n}\right| \to^p 0$$

*(b) Under Assumptions 19-22,*

$$\begin{aligned} & \sup_{\gamma\in\Gamma,\theta\in\Theta}\left|S_{\gamma,\theta,n}\right| \\ = \; & const_n + \left\|\theta-\theta_0\right\|^2 o_p\left(1\right) + \left\|\theta-\theta_0\right\| o_p\left(n^{-1/2}\right) + o_p\left(n^{-1}\right), \end{aligned} \tag{B.4}$$

*where $const_n$ is a random term that does not depend on $\theta$.*

**Proof.** The proof is provided for the more difficult part (b). Note that the functions

$$\varphi_{\gamma,\theta,n} = m_\theta\left(z_1,z_2\right)\prod_{s=1}^{S}\psi_{\gamma,n}^{(s)}\left(z_1,z_2;x_{s1},...,x_{sm}\right),\ \theta\in\Theta,\gamma\in\Gamma,$$

form a Euclidean class of functions with the Euclidean numbers that do not depend on $n$. The sum $S_{\gamma,\theta,n}$ can be represented as the sum of $U$-statistics of order up to $Sm+2$:

$$S_{\gamma,\theta,n} = n^{-Sm-2}\sum_{a=0}^{Sm+1}\sum_{a \text{ indices coincide}}\varphi_{\theta,\gamma,n}\left(Z_i,Z_j;X_{k_s^1},...,X_{k_s^m}\right) \tag{B.5}$$

(note that since in $S_{\gamma,\theta,n}$ $i\neq j$, it does not contain the term with all indices of $\varphi_{\theta,\gamma,n}$ coinciding). Consider first the sum over the non-coinciding indices ($a=0$). Write its

Hoeffding decomposition. By condition (B.2), its first (lowest order, nonzero) term is a degenerate $U$-statistic of order $S$, consisting of the sums of the form:

$$n^{-S} \sum_{k_1,...,k_S \text{ distinct}} l^{(0)}_{\gamma,\theta,n}\left(X_{k_1},...X_{k_S}\right),$$

where $l^{(0)}_{\gamma,\theta,n}\left(x_{11},...,x_{1m}\right)$ is obtained from $l_{\gamma,\theta,n}\left(z_1,x_{11},...,x_{1m}\right)$ by integrating out $z_1$ and all $x_{ks}$ for $k=2,...m$, $s=1,...,S$. By Assumption 22 and the Taylor expansion,

$$\begin{aligned} n^{-S}\sum l^{(0)}_{\gamma,\theta,n} &= n^{-S}\sum l^{(0)}_{\gamma,\theta_0,n} \\ &\quad + (\theta-\theta_0)'\left(n^{-S}\sum \partial^2_\theta l^{(0)}_{\gamma,\theta,n}\right)(\theta-\theta_0), \end{aligned}$$

where $\tilde{\theta}\in\Theta$. By (B.2), for each $\gamma,\theta$, the function $\partial^2_\theta l^{(0)}_{\gamma,\theta,n}$ is degenerate of order $S$:

$$E\left[\partial^2_\theta l^{(0)}_{\gamma,\theta,n}\left(X_1,x_2,...,x_S\right)\right]=0.$$

By the Euclidean property for the class of functions $\left\{\partial^2_\theta l^{(0)}_{\gamma,\theta,n},\ \theta\in\Theta,\gamma\in\Gamma\right\}$ (see e.g. Lemma 11), we have

$$E\sup_{\gamma,\theta}\left|n^{-S}\sum \partial^2_\theta l^{(0)}_{\gamma,\theta,n}\right| = O\left(L_n^S n^{-S/2}\right)=o\left(1\right).$$

The next term in the Hoeffding decomposition is a degenerate $U$-statistics of order $S+1$. It consists of the following terms:

$$n^{-S-1}\sum_{k_1,k_2}\begin{pmatrix} E\left[l^{(s_0)}_{\gamma,\theta,n}\left(Z_1;x_{11},x_{21}...,x_{S1};x_{s_02}\right)\right] \\ -E\left[l^{(s_0)}_{\gamma,\theta,n}\left(Z_1;x_{11},x_{21}...,x_{S1};X_{s_02}\right)\right] \\ -E\left[l^{(s_0)}_{\gamma,\theta,n}\left(Z_1;x_{11},...,X_{s_01},...,x_{S1};x_{s_02}\right)\right] \end{pmatrix}$$

and

$$\frac{1}{n\left(n-1\right)}\sum_{k_1,i}\left(\begin{array}{c} E\left[l_{\gamma,\theta,n}^{(s_0)}\left(z_1;x_{11},x_{21}...,x_{S1};X_{s_02}\right)\right] \\ -E\left[l_{\gamma,\theta,n}^{(s_0)}\left(Z_1;x_{11},x_{21}...,x_{S1};X_{s_02}\right)\right] \end{array}\right),$$

where $l_{\gamma,\theta,n}^{(s_0)}\left(z_1;x_{11},x_{21}...,x_{S1};x_{s_02}\right)$ is obtained from $l_{\gamma,\theta,n}$ by integrating out all $x_{ks}$ for $k=3,...m$, $s=1,...,S$, and also all $x_{2s}$ for $s\neq s_0$. By Assumption 21, the Taylor expansion and the Maximal Inequality (Lemma 11), both terms are of the order

$$const_n+\left\|\theta-\theta_0\right\|O_p\left(n^{-S/2-1/2}L_n^S\right)=const_n+\left\|\theta-\theta_0\right\|o_p\left(n^{-1/2}\right).$$

The remaining terms in the Hoeffding decomposition are of order at most $O_p\left(n^{-S/2-1}L_n^S\right)=o_p\left(n^{-1}\right)$.

We now consider the terms in (B.5) corresponding to ties between indices ($a>0$). The ties result in a smaller number of indices being summed up, so that the sums contain $O\left(n^a\right)$ times less terms than the sum over non-coinciding indices. Secondly, a tie in the index reduces the order of the first nonzero term in the Hoeffding decomposition because condition (B.2) becomes irrelevant in the presence of dependencies between the $X$ variables. When the $a$ ties occur only between the $k$-indices, the first nonzero order of the Hoeffding decomposition depends on $\max\left\{S-2a,0\right\}$ independent copies of $X$. Therefore the order of this term is at most $O_p\left(n^{-a}n^{-(S-2a)/2}L_n^S\right)=O_p\left(n^{-S/2}L_n^S\right)$. In this term both $Z_1$ and $Z_2$ are integrated out, therefore, we can use the smoothness of the function $l_{\gamma,\theta,n}$ and condition (B.3) to show that it has the representation given in (B.4). The next term in the Hoeffding decomposition, which has the order of magnitude $O_p\left(n^{-S/2-1/2}L_n^S\right)$, can also be treated as above, and the remaining terms are of order $O_p\left(n^{-S/2-1}L_n^S\right)=o_p\left(n^{-1}\right)$.

Finally, we need to consider the ties between one or both indices $i$, $j$ with some of the $k$-indices (possibly, in the presence of the ties within the $k$-indices). The result of such tie is that the condition (B.3) can no longer be used to remove the linear term in the Taylor expansion of $l_{\gamma,\theta,n}$ after integrating out $Z_i, Z_j$. In the case of the tie with one of the indices $i, j$, the first nonzero term in the Hoeffding decomposition, in accordance with condition (B.2), contains $\max\{S-2(a-1)+1,0\} = \max\{S-2a+1,0\}$ independent copies of the $X$-variable. Therefore, its order is at most $O_p\left(n^{-S/2-1/2}L_n^S\right)$. Using smoothness, we can find that this term has the order

$$const_n + O_p\left(n^{-S/2-1/2}L_n^S\right)\|\theta-\theta_0\| = const_n + o_p\left(n^{-1/2}\right)\|\theta-\theta_0\|.$$

The higher-order terms of the Hoeffding decomposition can be neglected. If there are ties with both $i, j$ we cannot use the smoothness in $\theta$ any longer. However, condition (B.2) now implies that the order of the first nonzero term in the Hoeffding decomposition is $O_p\left(n^{-1}n^{-S/2}L_n^S\right) = o_p\left(n^{-1}\right)$, and the remaining terms are of even smaller order. □

Now we use the above result to prove Theorem 8.

**Proof.** (Theorem 8.)

Consistency of $\theta_n$ follows from the fact that the objective function converges to its expectation with $w_3(z)$ in place of $\hat{w}_3(z)$ uniformly in $\theta$, which proof we omit. We now prove the result on the asymptotic normality.

Let

$$a(z,h,\theta_{0n}) = \frac{1}{h^3}E\left[M(Y_1,Y_2)\,\phi\left(\frac{z-X'\beta_{0n}}{h_\mu}\right)\phi'\left(\frac{z-X'\beta_{0n}}{h_\mu}\right)\right]$$

and

$$\begin{aligned} & b\left(z,h,\theta_{0n}\right) \\ = \; & \frac{1}{h^3}E\left[M\left(Y_1,Y_2\right)M\left(Y_1,Y_3\right)\phi\left(\frac{z-X_1'\beta_{0n}}{h_\sigma}\right)\phi\left(\frac{z-X_2'\beta_{0n}}{h_\sigma}\right)\phi\left(\frac{z-X_3'\beta_{0n}}{h_\sigma}\right)\right] \end{aligned}$$

Rewrite the function $\hat{w}_3$ as

$$\hat{w}_3 = \frac{a\left(z,h,\theta_{0n}\right)+\psi^{(\mu)}_{\theta_{0n}n}\left(\tau+b\left(z,h,\theta_{0n}\right)\right)}{\left(\tau+b\left(z,h,\theta_{0n}\right)\right)\left(1-\psi^{(\sigma)}_{\theta_{0n}n}\right)}$$

(this defines the functions $\psi^{(\mu)}_{\gamma n}$, $\psi^{(\sigma)}_{\gamma n}$).

Note that the function $m_\theta = M\left(y_1,y_2\right) sign\left(\left(X_1-X_2\right)'\left(\theta,1\right)\right)$ and the symmetrized functions $\psi^{(\mu)}_{\gamma n}$, $\psi^{(\sigma)}_{\gamma n}$ (where $\gamma\in\Theta$ is the placeholder for $\theta_{0n}$), satisfy Assumptions 19 and 20 for $L_n = \frac{1}{\max\left\{h_\mu^3,h_\sigma^3\right\}}$. Next, note that by the maximal inequalities for the U-processes (see Lemmas 16 and 18 for $m=1$, and Lemma 11 for the higher-order degenerate $U$-processes), for $p\geq 2$,

$$\left(E\sup_{\gamma,\theta,z}\left|\psi^{(\sigma)}_{\theta_{0n}n}\right|^p\right)^{1/p} = O\left(n^{-1/2}h_\mu^{3/2}\log n\right) = o\left(n^{-1/4}\log n\right)$$

so that $\left(E\sup_{\gamma,\theta,z}\left|\psi^{(\sigma)}_{\theta_{0n}n}\right|^{10}\right)^{1/2} = o\left(n^{-1}\right)$. Since

$$\left(E\sup_{\gamma,\theta,z}\left|g^2\left(z\right)\mu\left(z\right)+\psi^{(\mu)}_{\theta_{0n}n}\right|^2\right)^{1/2} = O\left(1\right),$$

we have, by the Cauchy-Schwartz inequality,

$$\hat{w}_3 = \left(\frac{a\left(z,h,\theta_{0n}\right)}{\tau+b\left(z,h,\theta_{0n}\right)}+\psi^{(\mu)}_{\theta_{0n}n}\right)\left(1+\sum_{s=1}^{4}\left(\psi^{(\sigma)}_{\theta_{0n}n}\right)^s\right)+o_p\left(n^{-1}\right)$$

where the last term is $o_p\left(n^{-1}\right)$ uniformly in $\theta, \gamma$ and $z$.

Let

$$\begin{aligned} & S_{\gamma,\theta,n} \\ = \; & n^{-2}\sum_{i\neq j} m_\theta\left(Y_i, X_i, Y_j, Z_j\right)\left(\hat{w}_3\left(Z_i\right) - w_3\left(Z_i\right)\right). \end{aligned}$$

Once we check that Assumptions 21 and 22 hold for the functions

$$\begin{aligned} \frac{a\left(z,h,\gamma\right)}{\tau + b\left(z,h,\gamma\right)}\left(\psi^{(\sigma)}_{\theta_{0n}n}\right)^s, \; s &= 1,...,4 \\ \psi^{(\mu)}_{\theta_{0n}n}\left(\psi^{(\sigma)}_{\theta_{0n}n}\right)^s, \; s &= 0,...,4, \end{aligned}$$

Theorem 22 will imply that

$$\sup_{\theta\in\Theta}\left|S_{\theta_{0n},\theta,n}\right| \to^p 0$$

and, with probability approaching 1,

$$\sup_{\|\theta-\theta_0\|\to 0}\left|S_{\theta_{0n},\theta,n}\right| = const_n + \|\theta-\theta_0\|^2 o_p\left(1\right) + \|\theta-\theta_0\| o_p\left(n^{-1/2}\right) + o_p\left(n^{-1}\right).$$

Therefore, the sum $S_{\gamma,\theta,n}$ has the order that does not affect consistency, asymptotic normality and the asymptotic variance of $\theta_n$ (see conditions on $\zeta_{n,\theta}$ in Theorem 9).

Let now $\Theta_0$ be an open ball in $\Theta$ containing $\theta_0$. Since $\theta_n \to \theta_0$, with probability approaching to one $\theta_n$ is in $\Theta_0$. It is clear that Assumption 21 is satisfied for the same numbers $L_n$ as above.

Consider the function

$$\sum_{i \neq j} \frac{a\left(z, h_{\mu}, \theta_{0n}\right)}{\tau + b\left(z, h_{\sigma}, \theta_{0n}\right)} m_{\theta}\left(Y_i, Y_j; X_i, X_j\right).$$

Use the Taylor expansion of the functions $a\left(z, h, \theta_{0n}\right)$ and $b\left(z, h, \theta_{0n}\right)$ in the powers of $h$ and $\left(\theta_{0n} - \theta_0\right)$ keeping the terms of order 6 and lower. Because $h_{\mu}, h_{\sigma} = o\left(n^{-1/6}\right)$ and $\left(\theta_{0n} - \theta_0\right) = O_p\left(n^{-1/2}\right)$, the remainder of the expansion has the order $o_p\left(n^{-1}\right)$ uniformly in $z$. The zero order term of the expansion is $w_3\left(z\right)$. It is easy to see that for the terms of the expansion other than the zero-order term, the corresponding $U$-statistic can be represented as the LHS of (B.4), and so these terms do not affect consistency, asymptotic normality and the asymptotic variance of $\theta_n$. □

APPENDIX C

# Computational Algorithms

## C.1. PDR4 Criterion Function

Here we provide a brief description of our algorithm for computing Abrevaya's [**2003**] PDR4 criterion function:

$$\sum_{i,j,k,l \text{ distinct}} \left(\mathbf{1}\left\{Y_i > Y_j\right\} - \mathbf{1}\left\{Y_k > Y_l\right\}\right) \mathbf{1}\left\{Z_i > Z_j\right\},$$

where $Z_i = X_i'\beta$. The PDR4 criterion function is a $U$-statistic of order four, and its brute-force computation requires $O\left(n^4\right)$ operations. The number of operations can be reduced to $O\left(n^2 \log n\right)$ by using sorting (Abrevaya [**2003**]). Additionally, one can exploit the pairwise-difference structure of the criterion function to reduce the amount of computations by a fixed proportion. One such algorithm is presented below.

It is easier to compute a form of the PDR4 criterion function in which the summation is done over the set of indices $i, j, k, l \in \mathring{I}_n^{(4)}$ where $\mathring{I}_n^{(4)}$ excludes the following coincidences of indices: $i = j$, $k = l$, $\{i = k, j = l\}$ and $\{i = l, j = k\}$. Note that the effect of (the remaining) coinciding indices is analogous to the effect of ties in the bootstrap, and it can be ignored asymptotically under Assumptions 5 and 10. Summation over the set $\mathring{I}_n^{(4)}$ can be performed using $O\left(n^2 \log n\right)$ operations after sorting the vector of differences $Z_i - Z_j$. An efficient algorithm exploits the fact that only the positive differences of $Z_i - Z_j$ need to be sorted. Secondly, if the vector $\{Z\}_{i=1}^n$ is itself sorted before computing the differences,

the stacked vector of differences $\{Z_i - Z_j\}$ consists of sorted segments of known lengths, and a more efficient sorting algorithm (relative to all-purpose algorithms such as quicksort or heapsort) can be applied to it. Taking into account these two features allows us to speed up the computation of the PDR4 criterion function by about three times.

To explain the algorithm, we first rewrite the objective function in terms of nonnegative differences of $Z_i - Z_j$. Assume that $\{Z_i\}$ are ordered in a nondecreasing order: $Z_i - Z_j \geq 0$ if $i > j$. The sums are over the indices in $\mathring{I}_n^{(4)}$ and the additional restrictions on indices shown explicitly:

$$\begin{aligned}
(*) \quad :\quad & = \sum \left(1\left\{Y_i > Y_j\right\} - 1\left\{Y_k > Y_l\right\}\right) 1\left\{Z_i - Z_j > Z_k - Z_l\right\} \\
& = \sum 1\left\{Y_i > Y_j\right\} 1\left\{Z_i - Z_j > Z_k - Z_l\right\} \\
& \quad - \sum 1\left\{Y_k > Y_l\right\} 1\left\{Z_i - Z_j > Z_k - Z_l\right\} \\
& = \sum sign\left(Y_i - Y_j\right) 1\left\{Z_i - Z_j > Z_k - Z_l\right\}
\end{aligned}$$

Here we exchanged labels in the second sum as follows: $k \leftrightarrow j$, $l \leftrightarrow i$ (note that this does not change the summation set $\mathring{I}_n^{(4)}$), and used the definition of $sign\left(Y_i - Y_j\right)$.

$$
\begin{aligned}
(*) &= \left(\sum_{i>j} + \sum_{i<j}\right) sign\left(Y_i - Y_j\right) 1\left\{Z_i - Z_j > Z_k - Z_l\right\} \\
&= \sum_{i>j} sign\left(Y_i - Y_j\right) \cdot \\
&\quad \left[1\left\{Z_i - Z_j > Z_k - Z_l\right\} - 1\left\{-\left(Z_i - Z_j\right) > Z_k - Z_l\right\}\right] \\
&= \sum_{i>j} sign\left(Y_i - Y_j\right) \cdot \\
&\quad 1\left\{Z_i \neq Z_j\right\}\left[1\left\{Z_i - Z_j > Z_k - Z_l\right\} - 1\left\{Z_i - Z_j < Z_k - Z_l\right\}\right]
\end{aligned}
$$

Here we first exchanged labels $i, j$ in the sum over $i < j$, noted that the summation term is zero if $Z_i = Z_j$ and then exchanged labels $k, l$ in the second term in line 3. Summation over $k \neq l$ of the last term can be written as

$$
\begin{aligned}
& 1\left\{Z_i \neq Z_j\right\}\left(\sum_{k>l} + \sum_{k<l}\right)\left[\begin{array}{c} 1\left\{Z_i - Z_j > Z_k - Z_l\right\} \\ +1\left\{Z_i - Z_j \geq Z_k - Z_l\right\} - 1 \end{array}\right] \\
= \; & 1\left\{Z_i \neq Z_j\right\} \sum_{k>l}\left[1\left\{Z_i - Z_j > Z_k - Z_l\right\} + 1\left\{Z_i - Z_j \geq Z_k - Z_l\right\}\right]
\end{aligned}
$$

because the remaining terms sum up to zero due to our assumption that $\{Z_i\}$ are ordered in the nondecreasing order. Therefore,

$$
\begin{aligned}
(*) &= \sum_{i>j} sign\left(Y_i - Y_j\right) 1\left\{Z_i \neq Z_j\right\} \cdot \\
&\quad \sum_{k>l}\left[2 \cdot 1\left\{Z_i - Z_j > Z_k - Z_l\right\} + 1\left\{Z_i - Z_j = Z_k - Z_l\right\}\right]
\end{aligned}
$$

The last summation should be performed with the additional restriction $(i, j) \neq (k, l)$, while the restriction $(j, i) \neq (k, l)$ follows from the restrictions $i > j$, $k > l$.

This suggests the following algorithm:

(i) Sort $\{Z_i\}_{i=1}^n$ in the ascending order, and rearrange $\{Y_i\}_{i=1}^n$ accordingly.

(ii) For $i > j$, compute the differences $Z_i - Z_j$ and stack them into a vector $\xi = \{\xi_q\}_{q=1}^{n(n-1)/2}$. Use the following order of stacking the elements of vector $\xi$: first the element with $i = 2$, $j = 1$, then the two elements with $i = 3, j = 2, 1$; then the three elements with $i = 4, j = 3, 2, 1$, and so on. The resulting vector $\xi$ consists of segments of lengths $1, 2, 3, ..., n-1$, that are each sorted in the ascending order.

(iii) For $i > j$, compute the values of $sign(Y_i - Y_j) 1\{Z_i \neq Z_j\}$ and stack them in a vector $\eta = \{\eta_q\}_{q=1}^{n(n-1)/2}$ in the same order as in step (ii).

(iv) Sort the vector $\xi$ in the ascending order and rearrange vector $\eta$ accordingly. To exploit the structure of $\xi$, use the mergesort algorithm iteratively. In each iteration, choose a pair of non-overlaping sorted segments (e.g. take adjacent segments, whose lengths are known) and merge them into one sorted segment. Repeat this step until the entire vector is sorted.

(v) The value of sums $\sum_{q'} 1\{\xi_q > \xi_{q'}\}$ and $\sum_{q'} 1\{\xi_q = \xi_{q'}\}$ over $q \neq q'$ can now be determined from the order of element $\xi_q$ in the sorted vector $\xi$ with a correction on coinciding elements of the vector. After finding these sums for every $q$, compute the sum $(*)$.

## C.2. Weighted Criterion Functions

An important aspect of the weighted rank estimators is that their criterion function, for a known weighting function (multiplicative or additive in the two observations), depends on the number of observations as $O(n \log n)$, just as for the unweighted rank estimators.

Given the large amount of the criterion function evaluations required to perform the maximization, this property is crucial for the practical usefulness of these estimators. Here we describe the corresponding algorithms for the two specific estimators considered above, MR and MRC.

In the case of MR, the weighted criterion function can be computed as follows. Consider the case when the weighting function is multiplicative in the two observations, i.e.

$$w\left(z_1, z_2\right) = w_0\left(z_1\right) w_0\left(z_2\right).$$

As discussed in the previous subsection, under Assumption 5, the ties in the single index do not affect the asymptotic distribution of the estimator. With the ties ignored, the criterion function can be rewritten as

$$\sum_{i \neq j} w_i w_j Y_i 1\left\{X_i'\left(\theta, 1\right) > X_j'\left(\theta, 1\right)\right\},$$

where $w_i$ is an estimator of $w_0\left(X_i'\left(\theta_0, 1\right)\right)$. To compute the double sum for a candidate $\theta$ using $O\left(n \log n\right)$ observations, (1) compute the vector of the modified $Y$-values: $\tilde{Y}_i = w_i Y_i$, (2) sort the values $X_i'\left(\theta, 1\right)$ in the ascending order, and rearrange the vectors $\{w_i\}$ and $\left\{\tilde{Y}_i\right\}$ accordingly. After reordering, the sum takes the form:

$$\begin{aligned} \sum_{i>j} w_j \tilde{Y}_i &= \sum_{i=2}^{n} W_i \tilde{Y}_i \\ W_i &= \sum_{j=1}^{i-1} w_j. \end{aligned}$$

(3) Compute the vector of values $\{W_i\}$ by recursion:

$$\begin{aligned} W_2 &= w_1, \\ W_i &= W_{i-1} + w_{i-1}, i = 2, ..., n, \end{aligned}$$

and compute the double sum $\sum_{i=2}^{n} W_i \tilde{Y}_i$.

In the case of the additive weights, $w(z_1, z_2) = w_0(z_1) + w_0(z_2)$, the computation algorithm is similar. The criterion function is

$$\begin{aligned} & \sum_{i \neq j} (w_i + w_j) Y_i 1\{Z_i > Z_j\} \\ = & \sum_{i \neq j} \left( \tilde{Y}_i + w_j Y_i \right) 1\{Z_i > Z_j\} \\ = & \sum_{i} \tilde{Y}_i Rank(Z_i) + \sum_{i \neq j} W_i Y_i, \end{aligned}$$

where $Rank(Z_i) = \sum_{j \neq i} 1\{Z_j < Z_i\}$, $\tilde{Y}_i = w_i Y_i$, and $W_i = \sum_j w_j 1\{Z_i > Z_j\}$ can be computed recursively after sorting the vector $\{Z_i\}$.

For the weighted MRC with additive weights, the computation of the criterion function:

$$\begin{aligned} & \sum_{i \neq j} (w_i + w_j) \{Y_i > Y_j\} 1\{Z_i > Z_j\} \\ = & \sum_{i \neq j} w_i \left( \{Y_i > Y_j\} 1\{Z_i > Z_j\} + \{Y_i < Y_j\} 1\{Z_i < Z_j\} \right) \\ = & \sum_{i} w_i \left( S_i + S_i' \right), \end{aligned}$$

where

$$\begin{aligned} S_i &= \sum_j \{Y_i > Y_j\} 1\{Z_i > Z_j\}, \\ S_i' &= \sum_j \{Y_i < Y_j\} 1\{Z_i < Z_j\}, \end{aligned}$$

is straightforward, since the entire vectors of sums $\{S_i\}_{i=1}^n$, $\{S_i'\}_{i=1}^n$, can be computed in $O(n \log n)$ operations using the algorithm of Abrevaya [**1999**] (for sorted $Z_i$, $S_i'$ can also be expressed through $S_i$ after a correction for ties).

## C.3. Optimal Weighting Functions

Here we explain how the estimated weighting function,

$$w_3(z) = \frac{\frac{1}{n^2 h_\mu^3} \sum_{i,j} M(Y_i, Y_j) \phi_{\mu i} \phi'_{\mu j}}{\tau + \frac{1}{n^3 h_\sigma^3} \left\{ \sum_i \left( \sum_j M(Y_i, Y_j) \phi_{\sigma j} \right)^2 \phi_{\sigma i} - \sum_{i,j} M^2(Y_i, Y_j) \phi_{\sigma j}^2 \phi_{\sigma i} \right\}}$$

can be computed using at most $O(n \log n)$ operations for each $z$. In the case of MR, this follows from the multiplicative form of the involved sums (requiring $O(n)$ computations for each $z$). For other rank estimators, the only non-multiplicative term is $1\{Y_i > Y_j\}$, however, using sorting such sums can be computed in $O(n \log n)$ operations.

Specifically, consider MRC: $M(Y_1, Y_2) = sign(Y_1 - Y_2)$ (defined as zero at zero). First sort $Y_i$ in the nondecreasing order in $i$ (this requires at most $O(n \log n)$ operations, and needs to be done only once), and rearrange the other vectors accordingly). Let $\{Y_{0k}\}$ be the distinct values of $Y_i$ in the sample ordered in the increasing order. Consider first the

numerator and define

$$\begin{aligned}\Phi_{\mu k} &= \sum_{i:Y_i=Y_{0k}} \phi_{\mu i},\\ \Phi'_{\mu k} &= \sum_{i:Y_i=Y_{0k}} \phi'_{\mu i}.\end{aligned}$$

The entire vectors $(\Phi_{\mu k})$, $\left(\Phi'_{\mu k}\right)$ can be computed in $O(n)$ operations. The sum in the numerator is

$$\begin{aligned}&\sum_{k,l} M(Y_{0k}, Y_{0l})\,\Phi_{\mu k}\Phi'_{\mu j}\\ =\;& 2\sum_{l<k}\Phi_{\mu k}\Phi'_{\mu l} - \sum_{k\neq l}\Phi_{\mu k}\Phi'_{\mu l}\\ =\;& 2\sum_{l<k}\Phi_{\mu k}\Phi'_{\mu l} - \sum_{k}\Phi_{\mu k}\sum_{l}\Phi'_{\mu l} + \sum_{k}\Phi_{\mu k}\Phi'_{\mu k}\end{aligned}$$

which can be computed in $O(n)$ operations using recursion.

To compute the denominator, let $\Phi_{\sigma k} = \sum_{i:Y_i=Y_{0k}} \phi_{\sigma i}$, and $\Phi_{2\sigma k} = \sum_{i:Y_i=Y_{0k}} \phi^2_{\sigma i}$. Vectors $(n_k)$, $(\Phi_{\sigma k})$, $(\Phi_{2\sigma k})$ can be computed in $O(n)$ operations. The sums in the denominator are equal to

$$\begin{aligned}&\sum_{k}\Phi_{\sigma k}\left(\sum_{l}\Phi_{\sigma l}M(Y_{0k}, Y_{0l})\right)^2 - \sum_{l,k}\Phi_{2\sigma l}\Phi_{\sigma k}M^2(Y_{0l}, Y_{0k})\\ =\;& \sum_{k}\Phi_{\sigma k}\left(2\sum_{l:l<k}\Phi_{\sigma l} - \sum_{l:l\neq k}\Phi_{\sigma l}\right)^2 - \sum_{k\neq l}\Phi_{\sigma k}\Phi_{2\sigma l}\\ =\;& \sum_{k}\Phi_{\sigma k}\left(2\sum_{l:l<k}\Phi_{\sigma l} - \sum_{l}\Phi_{\sigma l} + \Phi_{\sigma k}\right)^2 - \sum_{k}\Phi_{\sigma k}\sum_{l}\Phi_{2\sigma l}\\ &+\sum_{k}\Phi_{\sigma k}\Phi_{2\sigma k}\end{aligned}$$

which can be computed in $O(n)$ operations using recursion. Similar algorithms can be constructed for the other rank estimators.

### C.4. Maximization Procedure in the School Choice Example

To maximize the MR criterion function we used the Nelder-Mead (NM) algorithm, with three initial simplices, the identity matrix, $I$, and the matrices $\lambda \cdot I$, and $\lambda^2 \cdot I$, with $\lambda = 0.4$. For each initial simplex, the NM algorithm is run until the values of the objective function on the vertices of the simplex are within $10^{-9}$ of each other. As starting points for the algorithms we take $\theta_{(0)} = \theta_{Logit}$, and

$$\theta_{(k)} = \theta^*_{(k-1)} + \zeta_k, \ k = 1, ..., 50$$

where $\theta^*_{(k-1)}$ is the convergence point of the NM algorithm from the starting point $\theta_{(k-1)}$, and $\zeta_k$ is a random draw from the distribution

$$N\left(\theta_{Logit}, 4 * diag(Var(\theta)_{Logit})\right).$$

The draws $\zeta_k$ are the same for all experiments (unweighted MR, weighted MR and the bootstrapped estimators). To reduce the computational burden of the bootstrap, we took only the first 30 draws of $\zeta_k$, and, in the weighted MR, used the same weighting functions as on the estimation stage.

VIKTOR YEVGENYEVICH SUBBOTIN

E-mail: viktor.subbotin@u.northwestern.edu

DOCTORAL STUDIES

Ph.D., Economics, Northwestern University, Evanston, Illinois, U.S.A.

PREDOCTORAL STUDIES

B.A.: Applied Mathematics and Physics, Moscow Institute of Physics and Technology, Moscow, Russia, 1998.

M.A.: Applied Mathematics and Physics, Moscow Institute of Physics and Technology, Moscow, Russia, 2000.

M.A.: Economics, New Economic School, Moscow, Russia, 2002.

WORK EXPERIENCE

Consultant, Economic Expert Group, Moscow, Russia, 2002-2004.

PUBLICATIONS

Subbotin V.Ye. Estimating the tax burden in the Russian oil sector under the price parity hypothesis. *Ekonomika i Matematicheskie Metody* 41 (2005), no 3.

Vasilieva A.V., Gurvich E.T., Subbotin V. Ye. The economic analysis of the tax reform. *Voprosy Ekonomiki (2003)*, no 6.

Vdovichenko A., Voronina V., Dynnikova O., Subbotin V., and Ustinov A. Inflation and exchange rate policy. *Problems of Economic Transition* 47 (2004), no 5, pp. 6–31.