

CUBE-ROOT-N AND FASTER CONVERGENCE, LAPLACE ESTIMATORS, AND UNIFORM INFERENCE\*

Sung Jae Jun<sup>†</sup>

Joris Pinkse<sup>‡</sup>

Yuanyuan Wan<sup>§</sup>

Center for the Study of Auctions, Procurements, and Competition Policy

Department of Economics

The Pennsylvania State University

We extend the Chernozhukov and Hong (2003) Laplace estimation procedure to Kim and Pollard (1990)-type objective functions (e.g. the maximum score estimator (Manski, 1975) and the least median of squares estimator (Rousseeuw, 1984)). Scaling the non-regular objective functions by a sample-size ( $n$ )-dependent input parameter  $\alpha_n^2$ , we establish that the Laplace estimation procedure can improve the  $\sqrt[3]{n}$ -convergence rate of Kim and Pollard (1990) to a rate arbitrarily close to  $\sqrt{n}$ . We further show that the proposed estimator has three different types of limiting distribution, depending on the rate at which  $\alpha_n$  diverges. We provide a simple-to-implement uniform inference method which yields (asymptotically) correct inference irrespective of which of the three types applies. Generally, a slower rate of increase of  $\alpha_n$  translates to a faster convergence rate for our estimator  $\hat{\theta}$ , albeit that the convergence rate of  $\hat{\theta}$  is never worse than  $\sqrt[3]{n}$  and that to achieve faster than  $n^{2/5}$ -convergence a bias-correction procedure is needed. We provide two such procedures: the bias can be estimated and subtracted or one can use a special ‘prior’ resembling the Jeffreys (1946) prior. Computation is straightforward and can typically be accomplished using Gibbs sampling (Geman and Geman, 1984). A limited simulation study yields encouraging results.

---

\*We thank the Human Capital Foundation for their support of CAPCP. We thank Bo Honoré and Haiqing Xu for helpful suggestions.

<sup>†</sup>sjun@psu.edu (corresponding author), 608 Kern Graduate Building, University Park, PA 16802

<sup>‡</sup>joris@psu.edu

<sup>§</sup>yxw162@psu.edu

## 1. MOTIVATION

We extend the Chernozhukov and Hong (2003) Laplace estimation procedure to Kim and Pollard (1990)–type objective functions (e.g. the maximum score estimator (Manski, 1975) and the least median of squares estimator (Rousseeuw, 1984)). Scaling the non–regular objective functions by a sample–size ( $n$ )–dependent input parameter  $\alpha_n^2$ , we establish that the Laplace estimation procedure can improve the  $\sqrt[3]{n}$ –convergence rate of Kim and Pollard (1990) to a rate arbitrarily close to  $\sqrt{n}$ . We further show that the proposed estimator has three different types of limiting distribution, depending on the rate at which  $\alpha_n$  diverges. We provide a simple–to–implement uniform inference method which yields (asymptotically) correct inference irrespective of which of the three types applies. Generally, a slower rate of increase of  $\alpha_n$  translates to a faster convergence rate for our estimator  $\hat{\theta}$ , albeit that the convergence rate of  $\hat{\theta}$  is never worse than  $\sqrt[3]{n}$  and that to achieve faster than  $n^{2/5}$ –convergence a bias–correction procedure is needed. We provide two such procedures: the bias can be estimated and subtracted or one can use a special ‘prior’ resembling the Jeffreys (1946) prior. Computation is straightforward and can typically be accomplished using Gibbs sampling (Geman and Geman, 1984). A limited simulation study yields encouraging results.

Chernozhukov and Hong (2003) proposed integrating instead of optimizing the (exponential of an) objective function of extremum estimators to obtain an estimator of an unknown parameter vector  $\theta_0$ ; they labelled their estimator a *Laplace* estimator. The objective in Chernozhukov and Hong (2003) is primarily to facilitate the computation of some difficult–to–compute estimators by using the Markov Chain Monte Carlo (MCMC) method. Under the assumption that the objective function of interest admits a classical quadratic expansion, Chernozhukov and Hong (2003) show that their Laplace–type estimators are generally  $\sqrt{n}$ –consistent and have a limiting normal distribution. One version of their estimator can be interpreted as the ‘posterior’ mean based on a ‘prior’ and a ‘pseudo–likelihood,’ where the latter is formed using the extremum estimator objective function of interest.<sup>1</sup> Chernozhukov and Hong (2003) derive conditions under which their estimator is efficient and in which its limiting distribution coincides with the (pseudo) posterior distribution.

The results of Chernozhukov and Hong (2003) do not extend to the class of estimators considered by Kim and Pollard (1990) because the Kim and Pollard objective functions do not admit the quadratic expansion needed. Computation of this class of estimators and their confidence regions

---

<sup>1</sup>Despite the use of terminology which has a Bayesian ring to it, the Chernozhukov and Hong (2003) procedure — and indeed ours — is entirely classical.

can be cumbersome (see e.g. Manski and Thompson, 1986; Pinkse, 1993; Florios and Skouras, 2008), so a Laplace-type procedure would be valuable.

The extension of the Laplace-procedure to Kim and Pollard (1990)-type objective functions is not merely a generalization of the Chernozhukov and Hong (2003) conditions. Indeed, whereas for the class of estimators admitting the standard quadratic expansion, the estimator is always  $\sqrt{n}$ -consistent and the limiting distribution is always normal regardless of the scaling of the objective function, in the case studied in this paper both the convergence rate of our estimator and the nature of the limiting distribution depend on the input parameter  $\alpha_n^2$  that scales the objective function.

If  $\alpha_n$  diverges at a rate no slower than  $\sqrt[3]{n}$  then our estimator  $\hat{\theta}$  is  $\sqrt[3]{n}$ -convergent. If  $\alpha_n$  diverges faster than  $\sqrt[3]{n}$  then  $\hat{\theta}$  moreover has the same limiting distribution as the corresponding Kim and Pollard (1990) extremum estimator even though our estimator is always a singleton and the Kim and Pollard (1990) estimator can be set-valued as in e.g. the case of the maximum score estimator. When  $\alpha_n$  increases at the rate of  $\sqrt[3]{n}$ , the limiting distribution of  $\hat{\theta}$  is characterized by the ratio of two integrals of a certain Gaussian process. The third type of limiting distribution, namely the normal, arises when  $\alpha_n$  increases at a rate slower than  $\sqrt[3]{n}$ . Indeed, we show that subject to sufficient smoothness the convergence rate of  $\hat{\theta}$  can be as good as  $\sqrt{n/\alpha_n}$ ; since  $\alpha_n$  must increase to infinity with  $n$ , a convergence rate of  $\sqrt{n}$  is not achievable. If  $\alpha_n$  increases more slowly (i.e. no faster than  $\sqrt[5]{n}$ ) then asymptotic bias becomes an issue; the asymptotic bias is discussed further below.

This trichotomy of limiting distributions is interesting, but since in practice one only chooses a value of  $\alpha_n$ , not a rate, deciding which limit distribution to use is problematic. So, we develop a simple-to-execute simulation-based inference procedure which automatically adapts to the correct limit distribution. So inference is uniform in the choice of  $\alpha_n$ .<sup>2</sup>

Uniform inference procedures exist in numerous other environments, including weak identification (Staiger and Stock, 1997), roots near unity (Mikusheva, 2007), subsampling (Andrews and Guggenberger, 2008), kernel estimation (Guerre and Lavergne, 2005), HAC estimation (Kiefer and Vogelsang, 2005), and average derivative estimation (Cattaneo, Crump, and Jansson, 2008). Whereas in Staiger and Stock (1997) uniformity is achieved in a nuisance parameter and in Mikusheva (2007) in the parameter of interest, here it is achieved in a sample-size-dependent input parameter  $\alpha_n$ . In contrast to e.g. Guerre and Lavergne (2005), in our case uniformity is in the rate of the input parameter instead of a constant multiplying the rate. Like Mikusheva (2007), but unlike e.g. Cattaneo,

---

<sup>2</sup>Directly using the quantiles of the pseudo-posterior to conduct inference does not appear to be feasible in our case, although Chernozhukov and Hong (2003) derive conditions under which this can be done in their, regular, case.

Crump, and Jansson (2008), our limiting distributions differ not just in parameter values, but also in type, and here there are three distinct types instead of two as in Mikusheva (2007).

As we mentioned earlier, a slow rate of increase of  $\alpha_n$  (i.e. no faster than  $\sqrt[5]{n}$ ) introduces an asymptotic bias issue. We provide two methods for removing the asymptotic bias, namely to estimate it and to use a convenient prior. Indeed, we show that using the equivalent of the Jeffreys (1946) prior in the current context, or an estimated version thereof, removes the  $n^{-2/5}$ -order bias. This is the only instance that we are aware of in which the choice of prior matters in large samples when there is point-identification.<sup>3</sup>

For consistency,  $\alpha_n$  must increase to infinity with the sample size; for consistency the rate at which  $\alpha_n$  increases is immaterial. If one were to let  $\alpha_n$  decrease to zero, however, then our estimator converges to the mean of the prior, which would also be true for the regular Chernozhukov and Hong (2003) estimator. A consequence of this is that for small choices of  $\alpha_n$ , the estimator is biased towards the mean of the prior and that this type of bias cannot be corrected using asymptotic methods. A similar issue arises in kernel regression estimation if one lets the bandwidth go to infinity. We investigate both types of bias in our simulation study.

We illustrate our methodology using the maximum score estimator (Manski, 1975) as a leading example. We emphasize the binary choice version of the maximum score estimator, albeit that the discussion applies equally to the ordered response and multinomial choice cases (Lee, 1992; Manski, 1975). It continues to be applied in empirical work, including Bajari, Fox, and Ryan (2008); Bajari and Fox (2009); Fox (2007, 2009).

The maximum score estimator (MSE) is  $\sqrt[3]{n}$ -consistent for the vector of regression coefficients in a binary choice model using little more than a conditional median restriction on the errors in the latent variable equation. In particular, unlike other parametric estimators such as probit or logit, the error distribution need not be specified. Unlike other semiparametric single-index models (e.g. Klein and Spady, 1993), the error distribution need not be independent of the regressors, and heteroskedasticity of unknown form is permitted. Despite this desirable generality, the MSE has not been especially popular because of its computational difficulty, its slow convergence rate, and its non-standard limiting distribution. Horowitz (1992) has shown that by replacing the maximum score objective function by a smoothed version thereof, a rate arbitrarily close to  $\sqrt{n}$  is attainable, subject to sufficient smoothness. Subsequently de Jong and Woutersen (2007) and Kotlyarova and Zinde-Walsh (2003) have extended the Horowitz results. Although the best attainable convergence

---

<sup>3</sup>When the model is not fully identified, the choice of prior generally matters in large samples in Bayesian analysis (e.g. Moon and Schorfheide, 2009).

rate of the smoothed MSE (Horowitz, 1992) is the same as ours under similar conditions, our estimator is entirely different from Horowitz's, to our knowledge the Horowitz estimator has not been extended to other Kim and Pollard (1990)-type estimators, due to the nonconcavity of the objective function it is more difficult to compute than ours, and no uniform inference procedures are available for it.

As noted at the beginning of this section, for the maximum score case our estimator can be computed using Gibbs sampling. It turns out that computation is fast, simple, and accurate. If one chooses a prior for which a closed form solution exists for the integrated prior of each coefficient conditional on the others then computation of our estimator involves nothing more complicated than drawing uniform random numbers, sorting, and averaging. Likewise, irrespective of the type of Kim and Pollard (1990) estimator, uniform inference requires little more than drawing random numbers from a multivariate normal.

We study the properties of our estimator in a limited simulation study. The behavior of the proposed estimator reflects what one would expect on the basis of the theory. First, there is a tradeoff between bias and variance, with higher values of  $\alpha_n$  resulting in less bias, but more variance. Further, the asymptotic bias can be corrected using the Jeffreys prior but if  $\alpha_n$  is chosen very small then the estimator is biased towards the mean of the prior. Finally, the uniform inference procedure moves smoothly from the limiting distribution of the maximum score estimator to the normal with the value of  $\alpha_n$ .

There are several issues of potential interest that are not studied in this paper. First, one could look at statistics other than the posterior mean, as do Chernozhukov and Hong (2003). Further, one could use the bootstrap (Efron and Tibshirani, 1997) to remove the bias or indeed for inference, but because the bootstrap has been shown to be inconsistent in the case of the maximum score estimator (Abrevaya and Huang, 2005),<sup>4</sup> and because the bootstrap would offer no asymptotic refinement here even if it were consistent since the limiting distribution is nonpivotal, we have not pursued the use of the bootstrap or similar resampling schemes either.

The paper is organized as follows. In section 2 we discuss our estimation method and derive convergence results in section 3, which also includes a discussion of both of our bias correction methods. Our uniform inference procedure is described in section 4. Finally, the simulation study is contained in section 5.

---

<sup>4</sup>Subsampling (Politis, Romano, and Wolf, 1999) has been shown to be consistent for a class of  $\sqrt[3]{n}$ -consistent estimators by Delgado, Rodriguez-Poo, and Wolf (2001), but is less attractive than the bootstrap for reasons of efficiency.

## 2. ESTIMATION METHOD

**2.1. Set Up.** Let  $L_n(\theta) = n^{-1} \sum_{i=1}^n g_i(\theta)$  be the (renormed) objective function of a  $\sqrt[3]{n}$ -consistent estimator of an unknown parameter vector  $\theta_0 \in \Theta \subset \mathfrak{R}^d$ . The functions  $g_i = g(\zeta_i, \cdot)$  are defined such that  $g_i(\theta_0) = 0$  a.s.. For the specific case of the maximum score estimator with regressor vector  $x_i = [a_i, z_i^\top]^\top$  where the coefficient on  $a_i$  is normalized to equal minus one we get for  $\zeta_i = [y_i, x_i^\top]^\top$ ,

$$g_i(\theta) = (2y_i - 1)(I(a_i \leq z_i^\top \theta) - I(a_i \leq z_i^\top \theta_0)), \quad (1)$$

but the results below apply to general  $g_i$ , provided that our assumptions are satisfied.

We consider Laplace-type estimators of the form

$$\hat{\theta} = \frac{\int \theta \pi(\theta) \exp(\alpha_n^2 L_n(\theta)) d\theta}{\int \pi(\theta) \exp(\alpha_n^2 L_n(\theta)) d\theta}, \quad (2)$$

where  $\{\alpha_n\}$  is some sequence for which  $\alpha_n \rightarrow \infty$  as  $n \rightarrow \infty$ ,<sup>5</sup> we will call  $\pi$  a *prior*, even though for most of our results we do not require it to be nonnegative. If the prior is nonnegative everywhere, then  $\hat{\theta}$  can be interpreted as the mean of a *posterior* distribution; we will use the term ‘posterior’ regardless of whether the prior is nonnegative. If  $L_n$  were the objective function of a  $\sqrt{n}$ -consistent estimator instead of the one considered here and  $\alpha_n = \sqrt{n}$ , we would have the Laplace estimator of Chernozhukov and Hong (2003).

Provided that  $\theta_0$  is a unique maximizer of  $Q(\theta) = \mathbb{E}[g_i(\theta)]$ , consistency is straightforward to establish; see e.g. Robert and Casella (2004), corollary 5.11. The purpose of this paper is to study the effect of the choice of  $\{\alpha_n\}$  on the asymptotic properties of  $\hat{\theta}$ . If  $\alpha_n$  increases faster than  $\sqrt[3]{n}$  then it turns out that  $\hat{\theta}$  is asymptotically equivalent to the estimator maximizing  $L_n$ , for which Kim and Pollard (1990) derived limit results.<sup>6</sup> Although we emphasize the case in which  $\alpha_n$  increases no faster than  $\sqrt[3]{n}$ , we provide results for  $\alpha_n$  that increase faster, also. In subsequent sections we show that the best achievable convergence rate is  $\kappa_n = \max(\sqrt[3]{n}, \sqrt{n/\alpha_n})$ , albeit that to achieve a convergence rate better than  $n^{2/5}$  requires bias correction; this is in line with the properties derived in Horowitz (1992) for the smoothed maximum score estimator.

**2.2. Intuition for our results.** We now provide some intuition for our results when  $\alpha_n$  increases no faster than  $\sqrt[3]{n}$ , but please note that our results also cover the case in which  $\alpha_n$  increases faster than

<sup>5</sup>In Chernozhukov and Hong (2003)  $\alpha_n = \sqrt{n}$ , but its choice does not affect the limiting distribution provided that  $\alpha_n$  diverges to infinity.

<sup>6</sup>For the case in which  $L_n$  is the maximum score objective function, if  $\alpha_n$  increases faster than  $\sqrt{n}$  (not  $\sqrt[3]{n}$ ),  $\hat{\theta}$  is for large  $n$  in fact a prior-weighted average over the maximum score estimator, i.e. the (possibly noncontiguous) set of values maximizing the maximum score objective function.

that. Let  $\mathbf{S}_n(\theta) = n^{-1} \sum_{i=1}^n \tilde{\mathbf{g}}_i(\theta)$  with  $\tilde{\mathbf{g}}_i(\theta) = \mathbf{g}_i(\theta) - Q(\theta)$ . Then (2) can be rewritten as

$$\hat{\boldsymbol{\theta}} = \frac{\int \theta \pi(\theta) \exp(\alpha_n^2 \mathbf{S}_n(\theta) + \alpha_n^2 Q(\theta)) d\theta}{\int \pi(\theta) \exp(\alpha_n^2 \mathbf{S}_n(\theta) + \alpha_n^2 Q(\theta)) d\theta}. \quad (3)$$

Let  $\beta_n = \sqrt{\alpha_n^3/n}$  and define  $\tilde{\alpha}_n$  as  $\alpha_n$  if  $\beta_n$  has a finite limit and as  $\sqrt[3]{n}$  if  $\beta_n$  diverges to infinity.<sup>7</sup> By applying the substitution  $t = \alpha_n(\theta - \theta_0)$  to (3) we obtain

$$\sqrt{\frac{n}{\alpha_n}}(\hat{\boldsymbol{\theta}} - \theta_0) = \frac{1}{\beta_n} \frac{\int t \pi_n(t) \exp(\beta_n \tilde{\mathbf{S}}_n(t) + Q_n(t)) dt}{\int \pi_n(t) \exp(\beta_n \tilde{\mathbf{S}}_n(t) + Q_n(t)) dt}, \quad (4)$$

where  $\pi_n(t) = \pi(\theta_0 + t/\tilde{\alpha}_n)$ ,  $Q_n(t) = \tilde{\alpha}_n^2 Q(\theta_0 + t/\tilde{\alpha}_n)$  and  $\tilde{\mathbf{S}}_n(t) = \sqrt{n\tilde{\alpha}_n} \mathbf{S}_n(\theta_0 + t/\tilde{\alpha}_n)$ .

For large  $n$ ,  $Q_n(t) \approx -t^\top V t/2$  for  $V = -Q''(\theta_0)$  and  $\pi_n(t) \approx \pi_0$ . Replacing  $Q_n, \pi_n$  in the right hand side of (4) with their respective approximations yields

$$\frac{1}{\beta_n} \frac{\int t \exp(\beta_n \tilde{\mathbf{S}}_n(t) - t^\top V t/2) dt}{\int \exp(\beta_n \tilde{\mathbf{S}}_n(t) - t^\top V t/2) dt} = \frac{1}{\beta_n} \frac{\int t \exp(\beta_n \tilde{\mathbf{S}}_n(t)) \phi_V(t) dt}{\int \exp(\beta_n \tilde{\mathbf{S}}_n(t)) \phi_V(t) dt}, \quad (5)$$

where  $\phi_V(t)$  is the multivariate mean zero, variance  $V^{-1}$ , normal density function.

We establish in an appendix that  $\tilde{\mathbf{S}}_n \xrightarrow{w} \mathbf{G}$ , where  $\mathbf{G}$  is a tight Gaussian process defined on the entire Euclidean space. So if  $\alpha_n = c_\alpha^2 \sqrt[3]{n}$  then (5) converges in distribution to

$$\frac{1}{c_\alpha^3} \frac{\int t \exp(c_\alpha^3 \mathbf{G}(t)) \phi_V(t) dt}{\int \exp(c_\alpha^3 \mathbf{G}(t)) \phi_V(t) dt},$$

which is indeed the result of theorem 3 below. If  $\alpha_n = o(\sqrt[3]{n})$  then  $\beta_n = o(1)$  and the right hand side in (5) is approximately

$$\int t \tilde{\mathbf{S}}_n(t) \phi_V(t) dt,^8$$

which has a limiting normal distribution since  $\tilde{\mathbf{S}}_n$  is a renormed sample average of a sequence of i.i.d. mean zero variates. This is a result to be established in theorem 4, albeit that the result of theorem 4 contains a bias term  $\mathcal{B}_n$ . This bias term arises from the approximation of  $Q_n(t)$  by  $-t^\top V t/2$ . It is asymptotically negligible if  $\alpha_n$  diverges at a rate faster than  $\sqrt[5]{n}$ . If  $\alpha_n = c_\alpha^2 \sqrt[5]{n}$  then the limiting distribution will be a normal with nonzero mean, as shown in theorem 6, much like in the case of nonparametric kernel estimation of a function of a single argument and indeed like Horowitz (1992). We derive an expansion for the bias in theorem 5 and show that — subject to additional smoothness conditions — the bias can be removed. A bias-corrected estimator then can have a convergence rate arbitrarily close to  $\sqrt{n}$ .

<sup>7</sup>We implicitly restrict  $\beta_n$  to have a (possibly infinite) limit.

<sup>8</sup>This approximation uses  $\exp(\beta_n \tilde{\mathbf{S}}_n(t)) \approx 1 + \beta_n \tilde{\mathbf{S}}_n(t)$  and the fact that  $\int \phi_V(t) dt = 1, \int t \phi_V(t) dt = 0$ .

So, letting  $\alpha_n$  increase at a rate slower than  $\sqrt[3]{n}$  implicitly smoothes out the discontinuities in  $L_n$ , much like the smoothed maximum score estimator of Horowitz (1992) does explicitly.

### 3. CONVERGENCE RESULTS

We now proceed to state our main results. We first state our main assumptions, followed by a discussion of  $\sqrt[3]{n}$ -consistent estimators, followed by a discussion of estimators that converge faster. We will use the maximum score estimator as an example to motivate the assumptions.

**3.1. Assumptions.** The first of our assumptions is standard in the literature and, for the case of the maximum score estimator, is also found in Horowitz (1992).

**Assumption A.**  $\theta_0$  is in the interior of some compact set  $\Theta$ .

Assumption B is a condition necessary for identification. Indeed, note that  $Q(\theta_0) = 0$  by construction.

**Assumption B.**  $\forall \theta \in \Theta : \theta \neq \theta_0 \Rightarrow Q(\theta) < 0$ .

Let  $p(a, z) = \mathbb{P}[y = 1 | a = a, z = z]$  and  $f(a|z)$  denote the conditional density of  $a$  at  $a$  given  $z = z$ . For the specific case of the maximum score estimator Manski (1985) requires that the support of  $x$  is not contained in any proper linear subspace of  $\mathbb{R}^d$ , that  $0 < p(a, z) < 1$  for almost all  $a, z$ , and that for almost all  $z$ ,  $f(a|z) > 0$  for all  $a \in \mathfrak{R}$ , which is sufficient for assumption B. Assumption B is less primitive than the identification conditions of Manski (1985), but it applies to all estimators satisfying our conditions.

**Assumption C.**  $\pi(\theta_0) \neq 0$ ,  $\pi$  is bounded in absolute value on  $\Theta$  by  $\bar{\pi}$ , zero outside of  $\Theta$ , and integrates to one.

Since we can choose the prior, assumption C is innocuous. The same cannot be said for assumption D.

**Assumption D.** The function  $Q$  is continuous on  $\Theta$ . Further, for some  $q \geq 0$ ,  $Q$  is  $\Delta = q + 2$  times continuously differentiable at  $\theta_0$ ;  $\pi$  is  $q$  times continuously differentiable at  $\theta_0$ .<sup>9</sup>  $V = -Q''(\theta_0)$  is positive definite.

<sup>9</sup>If  $q = 0$ ,  $\pi$  is merely assumed continuous at  $\theta_0$ .

Manski (1985) does not derive a convergence rate for the maximum score estimator. Horowitz (1992) obtains a  $n^{2/5}$ -consistent estimator subject to assumptions including that (i)  $f(a|z)$  has a uniform upper bound in  $a$  and (almost all)  $z$ , (ii) for almost all  $z$ ,  $f'(a|z)$  is continuous in  $a$  in a neighborhood of  $a = z^\top \theta_0$  with a uniform bound over  $z$ , (iii) for almost all  $z$ , the second partial derivative of  $p$  with respect to  $a$ ,  $p_{aa}(a, z)$ , is a continuous function of  $a$  in a neighborhood of  $a = z^\top \theta_0$ , (iv)  $\mathbb{E}\|z_i\|^4 < \infty$ , (v)  $V$  is positive definite. To achieve the same convergence rate we need  $\Delta = 3$  (see assumption H below), which implies the existence of three moments on  $z_i$ , two partial derivatives of  $p$  with respect to  $a$  at  $z^\top \theta_0$ , and two derivatives of  $f(\cdot|z)$  at  $z^\top \theta_0$ . So the conditions are different from but similar to those in Horowitz (1992); neither set of assumptions implies the other. Again, our conditions are less primitive in order to accommodate other estimators. Finally, lemma G.1 shows that for the maximum score estimator  $V = -2\mathbb{E}[z_i z_i^\top p_a(z_i^\top \theta_0, z_i) f(z_i^\top \theta_0 | z_i)]$ , which by lemma G.2 is positive definite under weak conditions.

**Assumption E.**  $\mathbb{E}[\sup_{\theta \in \mathbb{R}^d} |g_i(\theta)|] < \infty$  and for some  $\iota, \iota^* > 0$  with  $\iota + \iota^* > 1$ , some function  $v$  for which  $\int \|t\|^{2+\iota} v(t) \phi_V(t) dt < \infty$ , and all  $t \in \mathbb{R}^d$ ,  $\lim_{\alpha \rightarrow \infty} \alpha^\iota \mathbb{E}|\tilde{g}_i(\theta_0 + t/\alpha)|^{2+\iota} \leq v(t)$ .

Assumption E is trivially satisfied for the maximum score estimator since it only involves indicator functions. The second part of assumption E is implied by assumption F below if  $g$  is a bounded function for  $\iota = \iota^* = 1$  since the limit in assumption E is then bounded by a constant times  $H(t, t)$ . Assumption F is also used in Kim and Pollard (1990).

**Assumption F.** For all  $t, s$ ,  $H(t, s) = \lim_{\alpha \rightarrow \infty} \alpha \mathbb{E}[g_i(\theta_0 + t/\alpha) g_i(\theta_0 + s/\alpha)]$  exists and is finite.<sup>10</sup>

We show in lemma G.3 that for the maximum score estimator using the normalization adopted earlier,  $H(t, s) = \mathbb{E}[|M(z_i^\top t, z_i^\top s, 0)| f(z_i^\top \theta_0 | z_i)]$  with  $M$  the median of its arguments. Assumption G is needed to establish weak convergence of the process  $\tilde{S}_n$  to  $G$ .

**Assumption G.** Let  $\mathcal{F} = \{g(\cdot; \theta)\}_{\theta \in \mathbb{R}^d}$  and for  $1 = o(\tilde{\alpha}_n)$ , let  $\mathcal{F}_n = \{\sqrt{\tilde{\alpha}_n} g(\cdot; \theta_0 + t/\tilde{\alpha}_n)\}_{t \in \mathcal{T}}$ , where  $\mathcal{T}$  is an arbitrary compact subset of  $\mathbb{R}^d$ . Then

- (i) For all  $\xi$  in the support  $\Xi$  of  $\xi_i$ ,  $g(\xi, \theta)$  is right- (or left-) continuous at  $\theta_0$ .
- (ii) There exists an envelope function  $F_n$  such that for all  $\xi \in \Xi$ :  $\sup_{t \in \mathcal{T}} \sqrt{\tilde{\alpha}_n} |g(\xi; \theta_0 + t/\tilde{\alpha}_n)| \leq F_n(\xi)$  and  $\mathbb{E}[F_{ni}^2] = O(1)$  where  $F_{ni} = F_n(\xi_i)$ .
- (iii) For any  $\epsilon > 0$ ,  $\mathbb{E}[F_{ni}^2 I(F_{ni} > \epsilon \sqrt{n})] = o(1)$ .
- (iv) For any  $\epsilon_n \downarrow 0$ ,  $\sup_{\|t-s\| < \epsilon_n} \tilde{\alpha}_n \mathbb{E}[(g_i(\theta_0 + t/\tilde{\alpha}_n) - g_i(\theta_0 + s/\tilde{\alpha}_n))^2] = o(1)$ .

<sup>10</sup>Note that this implies that  $\limsup_{\|t\| \rightarrow \infty} H(t, t)/\|t\| < \infty$ .

(v) Let  $\mathcal{N}(\epsilon, \mathcal{F}_n, \mathbb{L}_2(\mathcal{P}))$  be the  $(\mathbb{L}_2^-)$  covering number for  $\mathcal{F}_n$  with respect to the probability measure  $\mathcal{P}$ . Then for every  $\epsilon_n \downarrow 0$ ,  $\sup_{\mathcal{Q}} \int_0^{\epsilon_n} \sqrt{\log(\mathcal{N}(\epsilon \|F_n\|_{\mathcal{Q}, 2}, \mathcal{F}_n, \mathbb{L}_2(\mathcal{Q})))} d\epsilon = o(1)$ .

Assumption G is common, but is not always straightforward to verify. It ensures weak convergence of  $\tilde{\mathcal{S}}_n$  to  $\mathbf{G}$  and is satisfied by the maximum score estimator as theorems 1 and 2 show.

**Theorem 1.** Under assumptions A–G,  $\tilde{\mathcal{S}}_n \xrightarrow{w} \mathbf{G}$  on  $\mathcal{L}^\infty(\mathcal{T}_1, \mathcal{T}_2, \dots)$  for any increasing sequence of compact sets  $\mathcal{T}_j$  such that  $0 \in \mathcal{T}_1$  and whose union is  $\mathbb{R}^d$ , where  $\mathcal{L}^\infty(\mathcal{T}_1, \mathcal{T}_2, \dots)$  is the space of functions which are uniformly bounded on each  $\mathcal{T}_j$ .

*Proof.* The proofs of all theorems are in appendix H. □

**Theorem 2.** For the maximum score case, under assumptions A–E, if  $\tilde{\alpha}_n = o(n)$  then assumption G is satisfied if  $\mathbb{E}[\sup_s f(s|\mathbf{z}_i)|\|\mathbf{z}_i\|] < \infty$ .

We are now in a position to state our limit results.

**3.2. Cube–Root– $n$ –Convergence.** We are now in a position to state our main results. The first of these deals with the case  $\alpha_n = c_\alpha^2 \sqrt[3]{n}$ . Let  $\mathbf{G}$  be a mean zero Gaussian process on  $\mathbb{R}^d$  with covariance kernel  $H$ .

**Theorem 3.** (i) If  $\alpha_n = c_\alpha^2 \sqrt[3]{n}$  for some  $0 < c_\alpha < \infty$  and assumptions A–G hold with  $\tilde{\alpha}_n = \alpha_n$  then

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \frac{1}{c_\alpha^2} \frac{\int t \exp(c_\alpha^3 \mathbf{G}(t)) \phi_V(t) dt}{\int \exp(c_\alpha^3 \mathbf{G}(t)) \phi_V(t) dt}. \quad (6)$$

(ii) If  $\sqrt[3]{n} = o(\alpha_n)$ ,  $\underline{\pi} = \min_{\theta \in \Theta} \pi(\theta) > 0$ , and assumptions A–G hold with  $\tilde{\alpha}_n = \sqrt[3]{n}$  then for  $\tilde{\mathbf{G}}(t) = \mathbf{G}(t) - t^\top V t / 2$ ,

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \underset{t}{\operatorname{argmax}} \tilde{\mathbf{G}}(t). \quad (7)$$

Theorem 3 establishes that our estimator has the same rate of convergence as the Kim and Pollard (1990) estimator, i.e. the equivalent extremum estimator, if one lets  $\alpha_n$  increase at a rate no slower than  $\sqrt[3]{n}$ . In fact, part (ii) of theorem 3 demonstrates that if  $\alpha_n$  increases faster than  $\sqrt[3]{n}$  then the limit distributions of the Kim and Pollard (1990) estimator and ours coincide. There is continuity between parts (i) and (ii) of theorem 3 since if one lets  $c_\alpha \rightarrow \infty$  after  $n \rightarrow \infty$ , then the limit distribution of (i) converges to that of (ii). To see this, consider a different representation of the same limit distribution

that arises if  $c_\alpha^4$  is incorporated into  $\mathbf{g}_i$  instead of into  $\alpha_n^2$ .<sup>11</sup> Then

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \frac{\int t \exp\{c_\alpha^4(\mathbf{G}(t) - t^\top V t/2)\} dt}{\int \exp\{c_\alpha^4(\mathbf{G}(t) - t^\top V t/2)\} dt}. \quad (8)$$

For large values of  $c_\alpha$ , the right hand side of (8) is close to the value at which  $\mathbf{G}(t) - t^\top V t/2$  is maximized, whose distribution is exactly the limit distribution of the Kim and Pollard (1990) estimator.

Likewise, it follows from (6) and l'Hôpital's rule that  $\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  converges to zero as  $c_\alpha \downarrow 0$ . More interestingly, noting that  $\sqrt{n/\alpha_n} = \sqrt[3]{n}/c_\alpha$ , (8) suggests that  $\sqrt{n/\alpha_n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  has an approximate  $\int t\mathbf{G}(t)\phi_V(t)dt$ -distribution, which is indeed the normal distribution of theorem 4 below.

Intuitively, then, the accuracy of  $\hat{\boldsymbol{\theta}}$  is decreasing in the value of  $c_\alpha$  with the Kim and Pollard (1990) estimator the least efficient possibility. In numerical results reported in section 5 we find that the limiting distribution indeed becomes more dispersed as  $c_\alpha$  increases. But please note that these are asymptotic results; for small values of  $c_\alpha$  the small sample bias can be substantial. Nevertheless, the Kim and Pollard (1990) estimator (or indeed our estimator with  $\alpha_n = \infty$ ), and hence the maximum score estimator, is unlikely to be the optimal choice in the class of estimators studied here, even in samples of finite size.

**3.3. Faster convergence.** We now proceed with the case in which  $\alpha_n$  increases more slowly than  $\sqrt[3]{n}$ . Let

$$\begin{cases} \mathcal{N}_n &= \alpha_n^d \sqrt{n/\alpha_n} \int \pi(\theta)(\theta - \boldsymbol{\theta}_0) \exp(\alpha_n^2 \mathcal{S}_n(\theta) + \alpha_n^2 \mathcal{Q}(\theta)) d\theta, \\ \mathcal{D}_n &= \alpha_n^d \int \pi(\theta) \exp(\alpha_n^2 \mathcal{S}_n(\theta) + \alpha_n^2 \mathcal{Q}(\theta)) d\theta, \\ \mathcal{B}_n &= \alpha_n^d \sqrt{n/\alpha_n} \int \pi(\theta)(\theta - \boldsymbol{\theta}_0) \exp(\alpha_n^2 \mathcal{Q}(\theta)) d\theta. \end{cases} \quad (9)$$

**Theorem 4.** *If  $\alpha_n = o(\sqrt[3]{n})$  and assumptions A–H are satisfied with  $\tilde{\alpha}_n = \alpha_n$ , then*

$$\sqrt{\frac{n}{\alpha_n}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \frac{\mathcal{B}_n}{\mathcal{D}_n} = \frac{\mathcal{N}_n - \mathcal{B}_n}{\mathcal{D}_n} \xrightarrow{d} N(0, \mathcal{V}), \quad (10)$$

where  $\mathcal{V} = \iint t s^\top H(t, s) \phi_V(t) \phi_V(s) dt ds$ .

In lemma G.4 we show that for the maximum score case in which the regressor vector includes a constant,<sup>12</sup>

$$\mathcal{V} = \frac{1}{2\sqrt{\pi}} V^{-1} \mathbb{E} \left[ \frac{\mathbf{z}_i \mathbf{z}_i^\top}{\sqrt{\mathbf{z}_i^\top V^{-1} \mathbf{z}_i}} f(\mathbf{z}_i^\top \boldsymbol{\theta}_0 | \mathbf{z}_i) \right] V^{-1}, \quad (11)$$

<sup>11</sup>Alternatively, one can carry out the substitution  $\tilde{t} = t/c_\alpha^2$  and note that the process  $\mathbf{G}^*$  with  $\mathbf{G}^*(\tilde{t}) = \mathbf{G}(c_\alpha^2 \tilde{t})/c_\alpha$  has the same statistical properties as  $\mathbf{G}$ .

<sup>12</sup>What is needed is in fact weaker: for some  $j = 1, \dots, d$ :  $\mathbb{P}[z_{ij} = 0] = 0$ .

which suggests that using observation-specific weights can reduce (some scalar-valued function of) the asymptotic variance.<sup>13</sup> Some scale normalization of the weights should be imposed lest they assume the role of  $\alpha_n$ . We do not pursue a weighting procedure in this paper.

The bias term  $\mathcal{B}_n$  can affect the limiting distribution. As will become apparent below, if  $Q$  is sufficiently smooth then  $\mathcal{B}_n$  decreases at a rate of  $1/\alpha_n\beta_n = \sqrt{n/\alpha_n^5}$ , meaning that  $\alpha_n \sim \sqrt[5]{n}$  yields the typical nonparametric convergence rate of  $n^{-2/5}$ , also found in Horowitz (1992) (for his  $h = 2$ ). We first state our assumption requiring additional smoothness of  $Q$ .

**Assumption H.**  $q \geq 1$ .

We can now obtain a simple expansion for the bias, followed by an expression of the asymptotic distribution of  $n^{2/5}(\hat{\theta} - \theta_0)$  when  $\alpha_n = c_\alpha^2 \sqrt[5]{n}$ .

**Theorem 5.** *Let assumptions A–H be satisfied. For finite weights  $b_{q\tau}^*$  independent of  $n$  and which are zero for even  $\tau$ ,  $\mathcal{B}_n = \beta_n^{-1} \sum_{\tau=0}^q b_{q\tau}^* / \alpha_n^\tau + o(\alpha_n^{-q} \beta_n^{-1})$ .*

An expression for the values of  $b_{q\tau}^*$  is provided in the proof of theorem 5. As is shown in the proof of theorem 6 below,  $b_{q1}^* = \int (\pi_0 D_{Q3}(t) + D_{\pi1}(t)) t \phi_V(t) dt$ , where  $\pi_0 = \pi(\theta_0)$ ,  $D_{\pi1}(t)$  is the first term in a Taylor expansion of  $\pi(\theta_0 + t)$  about  $\pi(\theta_0)$ , i.e.  $t^\top \pi_\theta(\theta_0)$ , and  $D_{Q3}(t)$  is the third term in a Taylor expansion of  $Q(\theta_0 + t)$  about  $Q(\theta_0)$ . This then leads to the following result.

**Theorem 6.** *For any  $0 < c_\alpha < \infty$ , if  $\alpha_n = c_\alpha^2 \sqrt[5]{n}$  and assumptions A–H are satisfied, then*

$$n^{2/5}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\frac{\int (\pi_0 D_{Q3}(t) + D_{\pi1}(t)) t \phi_V(t) dt}{\pi_0 c_\alpha^4}, c_\alpha^2 \mathcal{V}\right).$$

So the bias depends on the choice of prior as well as the values of the second and third partial derivatives of  $Q$ . These derivatives are estimable and two methods for removing the bias are proposed in section 3.4. As noted in theorem 5,  $b_{q\tau}^* = 0$  for all even values of  $\tau$ . Consequently, provided that  $\hat{b}_{q1}^*$  converges to  $b_{q1}^*$  sufficiently fast and  $\alpha_n = c_\alpha^2 \sqrt[5]{n}$ , we can obtain the result

$$n^{4/9}(\hat{\theta} - \theta_0) - \frac{n^{2/9} \hat{b}_{q1}^*}{c_\alpha^5 \mathcal{D}_n} \xrightarrow{d} N(b_{q3}^* / c_\alpha^8, c_\alpha^2 \mathcal{V}). \quad (12)$$

This process can be repeated to obtain an estimator of  $\theta_0$  which converges at a rate arbitrarily close to  $\sqrt{n}$ , assuming sufficient smoothness.

<sup>13</sup>Like in other contexts (Cragg, 1992; Pinkse, 2006), a scalar-valued weight function which optimizes the asymptotic variance of all linear combinations of the coefficients of  $\hat{\theta}$  does not typically exist.

For the specific case of the maximum score estimator, the limit distribution of theorem 6 can be compared to that of the smoothed maximum score estimator of Horowitz (1992). Generally, which estimator has a smaller mean square error depends on the choice of input parameters (kernel and bandwidth in the case of smoothed maximum score, prior and  $c_\alpha$  here) and the joint distribution of  $(\mathbf{y}_i, \mathbf{a}_i, \mathbf{z}_i)$ . It can be shown that for the very special case in which  $d = 1$ ,  $\mathbf{z}_i = 1$  a.s., a normal kernel is used for the smoothed maximum score estimator and a flat prior is used for ours, and bandwidth (smoothed maximum score) and  $c_\alpha$  are chosen to minimize the asymptotic mean square error, the asymptotic mean square error of both estimators is identical regardless of the joint distribution of  $(\mathbf{y}_i, \mathbf{a}_i)$ .<sup>14</sup> A flat prior is however not a particularly good choice as theorem 7 demonstrates.

**3.4. Bias Correction.** There are many ways to correct the bias. Besides various resampling schemes, one could choose a prior to remove the bias or estimate the bias directly. We discuss both possibilities below.

It turns out that removing the bias in theorem 6 can be accomplished by choosing a prior that resembles Jeffreys' prior (Jeffreys, 1946), here  $\pi(\theta) \propto \sqrt{\det(-Q''(\theta))}$ , near  $\theta_0$ . However, since  $Q$  is not necessarily globally concave one should choose a prior which is both everywhere bounded and which resembles Jeffreys' prior in a neighborhood of  $\theta_0$ . An example is  $\pi_J(\theta) \propto \sqrt{|\det(-Q''(\theta))|}$ , which can be estimated by  $\hat{\pi}_J(\theta) \propto \sqrt{|\det(-\hat{Q}''(\theta))|}$  for some estimator  $\hat{Q}''$  which satisfies the following assumption. Let  $\text{vec}$  denote the familiar vec-operator, which stacks the columns of a matrix into a vector and let  $Q'''$  and  $\hat{Q}'''$  be defined as  $\partial \text{vec}^\top Q'' / \partial \theta$  and  $\partial \text{vec}^\top \hat{Q}'' / \partial \theta$ , respectively.

**Assumption I.** For  $r^* \geq 0$  and any  $\rho_n = o(1/\alpha_n)$ ,  $\hat{Q}''$  is bounded on  $\Theta$  and satisfies

$$\sup_{\|\theta - \theta_0\| \leq \rho_n} (\|\hat{Q}''(\theta) - Q''(\theta)\| + \|\hat{Q}'''(\theta) - Q'''(\theta)\|) = o_p(\alpha_n^{-r^*}).$$

The conditions specified in assumption I are generic because the properties of  $Q''$  depend on the nature of  $\mathbf{g}_i$ , on which we only imposed generic conditions. In the maximum score case,  $Q''$  and  $Q'''$  are *average derivatives* (Powell, Stock, and Stoker, 1989) and hence can typically be estimated at a rate much faster than the  $d + 1$ -dimensional nonparametric rate. For  $r^* = 0$ , no rate is required, for  $r^* = 2$  a rate of  $n^{-2/5}$ , but when we require  $r^* = 2$  more derivatives of  $Q''$  will be available. Let  $\hat{\theta}_J$  denote the estimator which is identical to  $\hat{\theta}$  except that it uses the prior  $\hat{\pi}_J$  defined above.

**Theorem 7.** Let assumptions A–H be satisfied.

(i) If  $\alpha_n = c_\alpha^2 \sqrt[5]{n}$  and assumption I is satisfied with  $r^* \geq 0$ , then  $n^{2/5}(\hat{\theta}_J - \theta_0) \xrightarrow{d} N(0, c_\alpha^2 \mathcal{V})$ .

<sup>14</sup>The derivation is available upon request.

(ii) If  $\sqrt[q]{n} = O(\alpha_n)$ ,  $q \geq 3$ , and assumption I is satisfied with  $r^* \geq 2$ , then for some finite  $\mathcal{B}_J$ ,  $\sqrt{n/\alpha_n}(\hat{\theta}_J - \theta_0) - \mathcal{B}_J/\alpha_n^3\beta_n\mathcal{D}_n \xrightarrow{d} N(0, c_\alpha^2\mathcal{V})$ .

So subject to the conditions of theorem 4, using (an estimate of) Jeffreys' prior yields a zero mean limiting normal distribution (part (i) of theorem 7), and with extra derivatives the  $n^{-4/9}$ -rate obtains.

Further bias reduction can be obtained by choosing a prior which knocks out the higher order bias terms, also. Because the bias only depends on derivatives of  $Q''$  and  $\pi$  at  $\theta_0$ , such priors are straightforward to construct by using suitable polynomials in  $\theta - \theta_0$  whose coefficients consist of estimated derivatives of  $Q''$  at  $\theta_0$ , albeit that this procedure would require an initial plugin estimator to use in lieu of  $\theta_0$ , whereas such an initial plugin estimator is not required for  $\hat{\pi}$  used in theorem 7. An example of such a polynomial prior which removes the  $\tau = 1$  term in the bias expansion of theorem 5 is  $\tilde{\pi}(\theta) = \sqrt{\det(-\tilde{Q}''(\theta))}$ , with

$$\begin{aligned} \tilde{Q}''(\theta) &= Q''(\theta_0)/2 \\ &+ \left( Q'''(\theta_0)(I \otimes (\theta - \theta_0)) + Q''(\theta_0) \right)^\top (Q''(\theta_0))^{-1} \times \left( Q'''(\theta_0)(I \otimes (\theta - \theta_0)) + Q''(\theta_0) \right), \end{aligned} \quad (13)$$

which is guaranteed to be negative definite since it is the sum of a negative definite and negative semidefinite matrix. Methods for estimating  $Q''$  and  $Q'''$  for the case of the maximum score estimator are discussed in appendix I.3.

An alternative to removing the bias by one's choice of prior is to estimate the bias directly. The bias term in theorem 6, for instance, can be estimated using

$$\frac{\int (\pi_0 D_{\hat{Q}_3}(t) + D_{\pi_1}(t)) t \phi_{\hat{V}}(t) dt}{\pi_0 c_\alpha^4}, \quad (14)$$

using e.g. nonparametric kernel estimators of  $Q$ . In particular, for  $d = 1$  and a flat prior, (14) reduces to  $\hat{Q}'''(\hat{\theta}_p)/2c_\alpha^4(\hat{Q}''(\hat{\theta}_p))^2$ , with  $\hat{\theta}_p$  some preliminary estimator.

Since the form of an estimator of  $Q''$  and its derivatives depends on the nature of  $g_i$ , establishing general analytical results for the bias correction is not possible. It would be possible to do so for the specific case of the maximum score estimator, but this would entail not much more than a rehash of results that are well-known and are available in the literature. So we only provide a sketch of what the theoretical justification would look like.

For the case of the maximum score estimator, nonparametric estimators can be shown to be uniformly consistent for  $Q''$  and  $Q'''$  in a neighborhood of  $\theta_0$  that decreases more slowly than  $\hat{\theta}_p$ . So consistency obtains and the bias up to order  $n^{-2/5}$  is removed.

If one is willing to assume the existence of five continuous derivatives of  $Q$  at  $\theta_0$  and three of  $\pi$  (i.e.  $\Delta = q + 2 = 5$ ), then to attain the  $n^{4/9}$ -rate for the bias corrected estimator,  $b_{33}^*$  must be estimated consistently and our estimator  $\hat{b}_{31}^*$  of  $b_{31}^*$  should satisfy  $\hat{b}_{31}^* - b_{31}^* = o(\alpha_n^{-2}) = o(n^{-2/9})$ . Because of the existence of additional derivatives, a nonparametric kernel estimator of  $b_{31}^*$  will be more efficient and beating the  $n^{2/9}$ -rate should not be problematic.

#### 4. UNIFORM INFERENCE

To conduct inference one can draw random numbers from the Kim and Pollard (1990) limit distribution if  $\alpha_n$  increases faster than  $\sqrt[3]{n}$ , from the limit distribution of theorem 3 if  $\alpha_n$  increases at the  $\sqrt[3]{n}$  rate, or use the normal of theorem 4 if  $\alpha_n$  increases more slowly than  $\sqrt[3]{n}$ . Since these are only rates for  $\alpha_n$ , in a sample of finite size it is generally not clear which of these three distributions should be used. Let  $\hat{\mathbf{G}}$  be  $\mathbf{G}$  with  $H$  replaced with some estimator  $\hat{H}$ . As it turns out, provided that the bias is asymptotically negligible, for  $\kappa_n$  defined in section 2.1 and  $\hat{V}$  some consistent estimator of  $V$ ,  $\chi_n = \kappa_n(\hat{\boldsymbol{\theta}} - \theta_0)$  and

$$\mathbf{\Psi}_n = \frac{\int t \exp(\beta_{n1} \hat{\mathbf{G}}(t) - \beta_{n2} t^\top \hat{V} t / 2) dt}{\int \exp(\beta_{n1} \hat{\mathbf{G}}(t) - \beta_{n2} t^\top \hat{V} t / 2) dt}, \quad \text{with } \begin{cases} \beta_{n1} = \min(\beta_n^{3/2}, \beta_n^{4/3}), \\ \beta_{n2} = \min(\beta_n^2, \beta_n^{4/3}), \end{cases} \quad (15)$$

have the same limiting distribution in all three cases.

**Theorem 8.** *If  $\hat{V} - V = o_p(1)$  and  $\hat{H}$  is a positive semidefinite covariance kernel<sup>15</sup> which is pointwise consistent for  $H$  and for some  $0 < c_H < 2$  and all sample paths of  $\hat{H}$  satisfies  $\limsup_{\|t\| \rightarrow \infty} (\hat{H}(t, t) / \|t\|^{c_H}) < \infty$ , then  $\mathbf{\Psi}_n$  has the limiting distribution derived for  $\chi_n$  in the corresponding portion of theorems 3 and 4<sup>16</sup> if (i)  $\beta_n = o(1)$ , (ii)  $1/\beta_n + \beta_n = O(1)$ , (iii)  $1/\beta_n = o(1)$ .*

The requirement that  $\hat{H}$  be a positive semidefinite covariance kernel is both necessary and sufficient for there to exist a Gaussian process  $\hat{\mathbf{G}}$  with that covariance kernel (Doob, 1953, theorem 3.1). Drawing random numbers using (15) is simple and is explained in appendix I.2.

So the proposed inference procedure is uniform in the divergence rate of  $\alpha_n$  despite the fact that the limiting distribution of  $\hat{\boldsymbol{\theta}}$  depends on the rate at which  $\alpha_n$  increases.

The limiting distribution of Laplace-type estimators sometimes coincides with the limit of the posterior, i.e.  $\lim_{n \rightarrow \infty} \exp(\alpha_n^2 \mathbf{L}_n(\theta_0 + t/\alpha_n))$  up to a proportionality constant. Indeed, Chernozhukov and Hong (2003) derive conditions under which this is the case for objective functions that admit a

<sup>15</sup> $\hat{H}$  is a positive semidefinite covariance kernel if for any integer  $0 < T^* < \infty$  and any values  $t_1, t_2, \dots, t_{T^*} \in \mathcal{R}^d$ , the matrix with  $(i, j)$ -element  $\hat{H}(t_i, t_j)$  is positive semidefinite.

<sup>16</sup>Please note that for part (i) of theorem 3,  $\kappa_n$  and  $\sqrt[3]{n}$  can differ by a multiplicative constant.

quadratic expansion. In the case of coincidence of the true distribution and the limit of the posterior, the (Gibbs) draws used to obtain  $\hat{\theta}$  can then be used to construct confidence intervals. Here the limit of the posterior is a random object if  $\alpha_n$  increases at the rate of  $\sqrt[3]{n}$  (or faster) and is  $N(0, V^{-1})$  if  $\alpha_n$  increases more slowly than  $\sqrt[3]{n}$ . So in neither case are the posterior and estimator limit distributions the same.

## 5. PERFORMANCE

As noted before, the main caveat of our estimator is the need for the choice of input parameters  $\pi, \alpha_n$ . The ability to choose a prior has been shown to be valuable in bias reduction, at least theoretically. Its practical value is examined later on in this section. We begin by analyzing the effect of one's choice of  $\alpha_n$ .

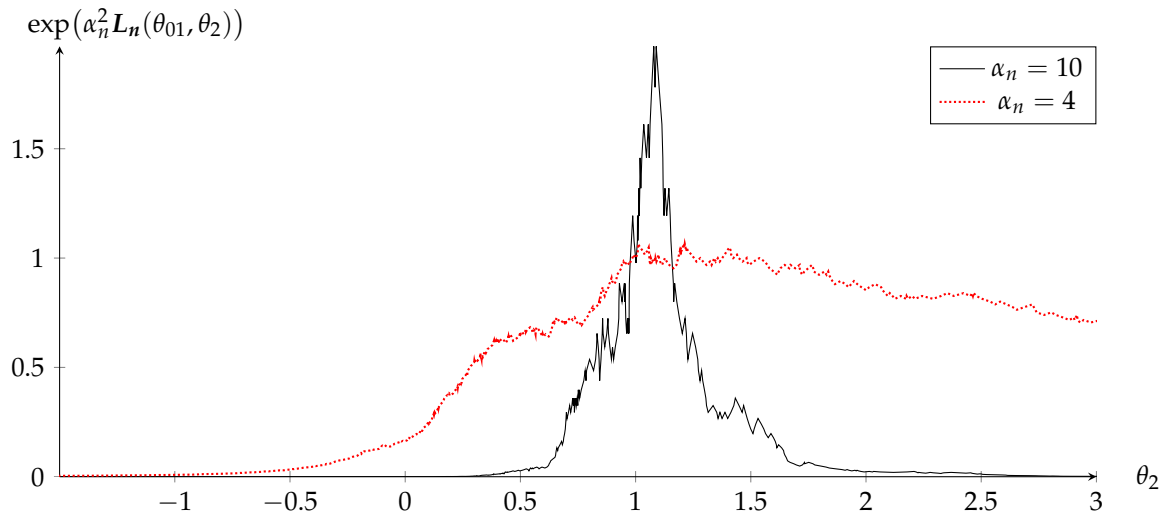


FIGURE 1. One draw of  $\exp(\alpha_n^2 L_n(0, \theta_2))$ ;  $n = 1,000$ .

We consider a small number of designs, all in the context of the maximum score estimation problem. In all cases  $\mathbf{y}_i = I(\mathbf{z}_i^\top \theta_0 - \mathbf{a}_i + \zeta(\mathbf{a}_i, \mathbf{z}_i) \mathbf{u}_i \geq 0)$ , where  $\mathbf{u}_i$  is a standard normal independent of  $\mathbf{a}_i, \mathbf{z}_i$  which are also standard normal and have mutually independent elements. Unless otherwise stated  $\theta_0 = [0, \bar{\mathbf{1}}^\top]^\top$ , where  $\bar{\mathbf{1}} \in \mathbb{R}^{d-1}$  is a vector of ones and the prior is uniform on a compact support, typically  $[-2, 2]$ .

For figure 1 we used  $d = 2$  and  $\zeta^2 = (z_1^2 + 1)/2$ . Figure 1 illustrates how (for a single data set) the normalized posterior varies with  $\alpha_n$ ; the value at  $\theta_0$  equals one. Larger values of  $\alpha_n$  lead to a narrower posterior density, so  $\hat{\theta}_2$  is an average over a comparatively small range of  $\theta$ -values. In this

particular instance the posterior density for  $\alpha_n = 10$  is close to the truth (one), but for other data sets it may not be. Because it is an average over fewer  $\theta$ -values whose range varies from one data set to another,  $\hat{\theta}$  for  $\alpha_n = 10$  has a greater variance than for  $\alpha_n = 4$ . But the bias is less for greater values of  $\alpha_n$ . One reason is that the posterior is asymmetric; in figure 1 the posterior for  $\alpha_n = 4$  tapers off much faster to the left than it does to the right, which is due to the particular design used.

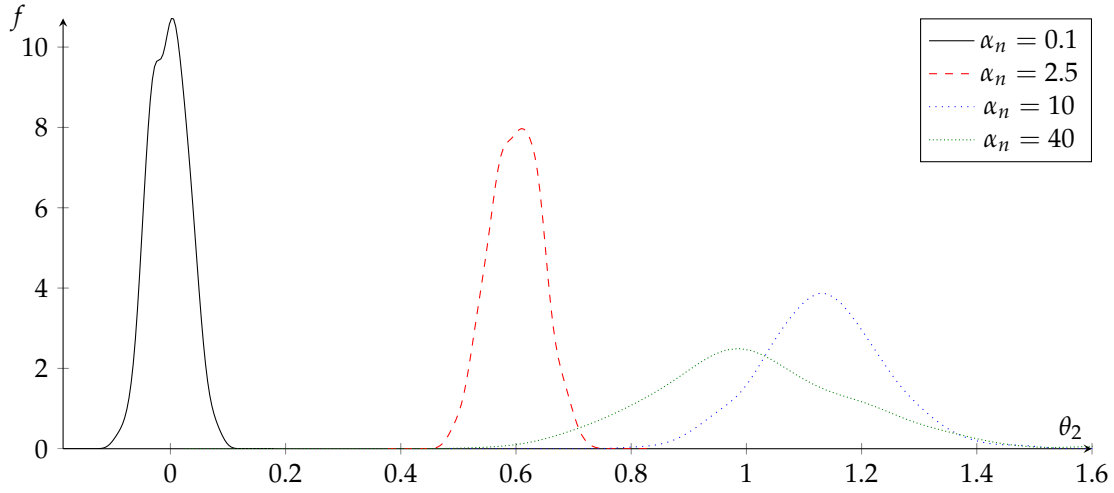


FIGURE 2. Density of  $\hat{\theta}_2$  for various choices of  $\alpha_n$ ;  $d = 2$ ,  $n = 1,000$

A secondary cause of the bias is that for smaller values of  $\alpha_n$ , the prior receives relatively more weight, thereby biasing one's estimates in the direction of the mean of the prior. Inspection of the formula for  $b_{q\tau}^*$  in the proof of theorem 5 demonstrates that the asymptotic bias only depends on the prior and its derivatives at  $\theta_0$ . For a flat prior the derivatives at  $\theta_0$  are all zero and the prior drops out of the bias formula. Therefore, the secondary bias is not correctible with the methods of section 3.4.

To see that this secondary bias can be substantial for small values of  $\alpha_n$ , consider figure 2. For very small values of  $\alpha_n$  the distribution of  $\hat{\theta}_2$  is centered around the mean of the prior (zero). As  $\alpha_n$  increases this secondary bias decreases until it becomes small compared to the primary (asymptotic) and correctible bias for  $\alpha_n = 10$ . Here the primary bias is positive.

This point is reinforced by figure 3, which represents the results of an experiment identical to the one for figure 2, except that the mean of the prior coincides with the value of  $\theta_{02}$ ; both equal one. There is some bias for smaller values of  $\alpha_n$ , but this bias is all in the same direction; positive.

A heuristic method for choosing  $\alpha_n$ , then, is to compute the value of  $\hat{\theta}$  for two flat priors with different means, say  $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}$ , and to choose the smallest value of  $\alpha_n$  for which the difference between  $\hat{\theta}_{(1)}$  and  $\hat{\theta}_{(2)}$  is small. Subsequently, the imputed value of  $\alpha_n$  can be used with a prior

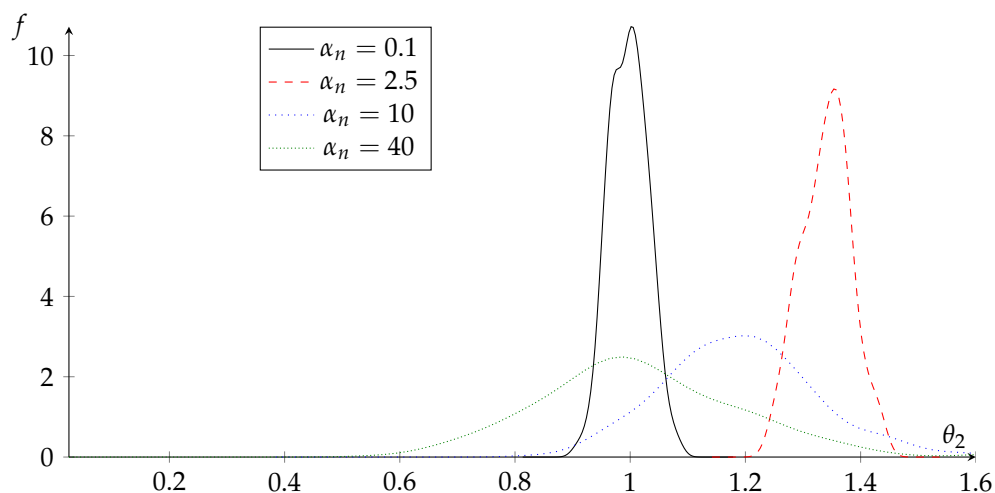


FIGURE 3. Density of  $\hat{\theta}_2$  for various choices of  $\alpha_n$ ;  $d = 2$ ,  $n = 1,000$ , prior mean coincides with true value.

of one's choosing. Alternatively, the procedure of choosing  $\alpha_n$  can be tailored to the nature of the objective function, e.g. one could construct a method for choosing  $\alpha_n$  specific to the maximum score case.

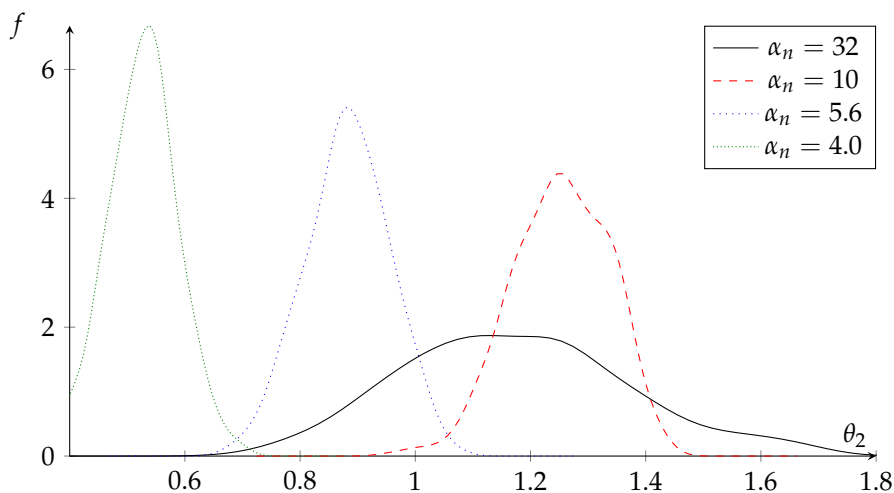


FIGURE 4. Density of  $\hat{\theta}_2$  for various values of  $\alpha_n$ ;  $d = 10$ ,  $n = 1,000$

The experiment depicted in figure 2 is repeated in figure 4 with  $d = 10$ . The results suggest that the value of  $\alpha_n$  need not depend on  $d$ .

Figure 5 underlines the value of heteroskedasticity-robust estimators for the binary choice model, such as the (smoothed) maximum score estimator and our Laplace-transform based estimator.

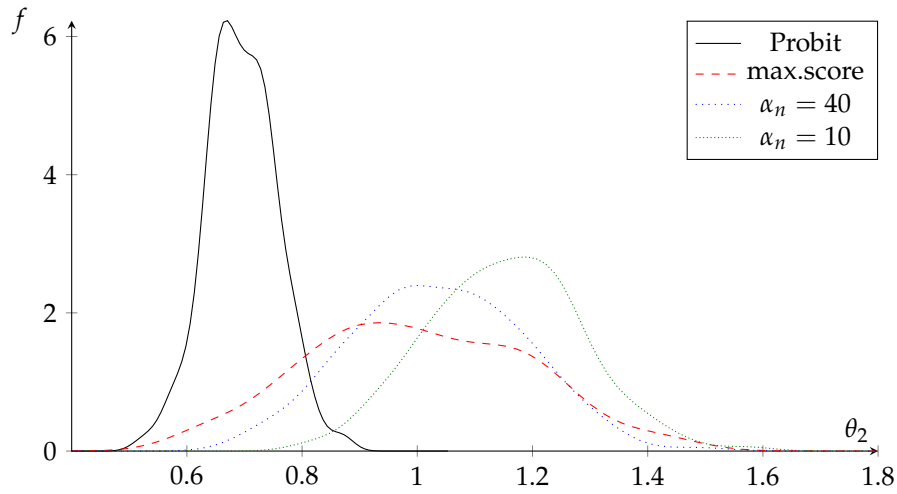


FIGURE 5. Density of  $\hat{\theta}_2$ ; probit versus maximum score and Laplace;  $d = 2$ ,  $n = 1,000$

Even though there is only a modest amount of heteroskedasticity and the error distribution is normal, the probit estimator has substantial bias. Naturally, under homoskedasticity with a normal error distribution the probit estimator will outperform the robust estimators since, as figure 5 indicates, its variance is considerably smaller than that of the other estimators. It is also clear that the maximum score estimator<sup>17</sup> has greater variance than the Laplace-type estimators, which suggests that choosing  $\alpha_n = \infty$  is suboptimal even without bias correction.

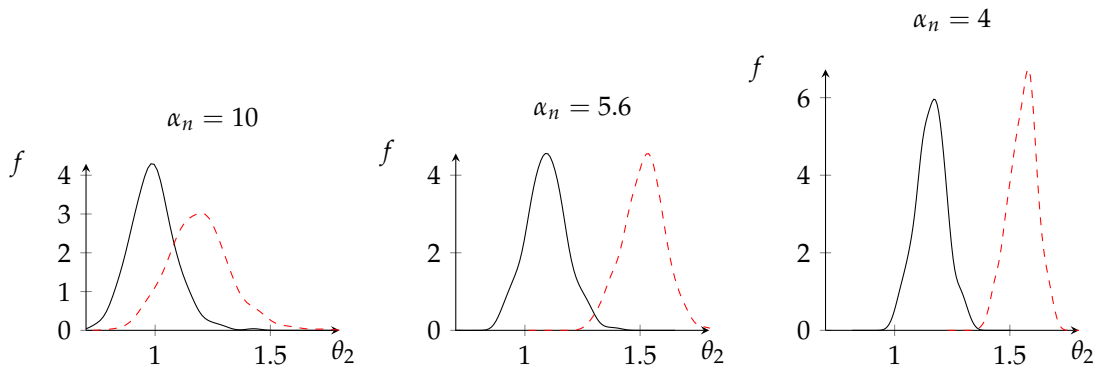


FIGURE 6. Density of  $\hat{\theta}_2$ ; Jeffreys (solid) versus uniform (dashed) prior;  $n = 1,000$ ,  $d = 2$ .

We now investigate the performance of one of the proposed bias-correction techniques, namely the use of Jeffreys' prior. It should be pointed out that using Jeffreys' prior takes more time than

<sup>17</sup>Computed using a grid search.

the flat prior unless the conditional prior of one element of the parameter vector given the others has a closed form solution. For our experiments, the cost of using the Jeffreys' prior was far from prohibitive. If computation time becomes a serious issue, one can instead opt for the more convenient, faster, and asymptotically equivalent form presented in (13), albeit that we expect the finite sample performance of such a procedure to be inferior.

Figure 6 depicts the bias for Jeffreys' prior and the correctly centered uniform prior for varying values of  $\alpha_n$ . The support of Jeffreys' prior was chosen such that the mean of Jeffreys' prior was also equal to one. The purpose of setting the mean of the prior equal to the truth is to assess the ability of Jeffreys' prior to correct the asymptotic bias without obfuscating the comparison by the asymptotically negligible bias caused by the discrepancy between the mean of the prior and the true parameter value.

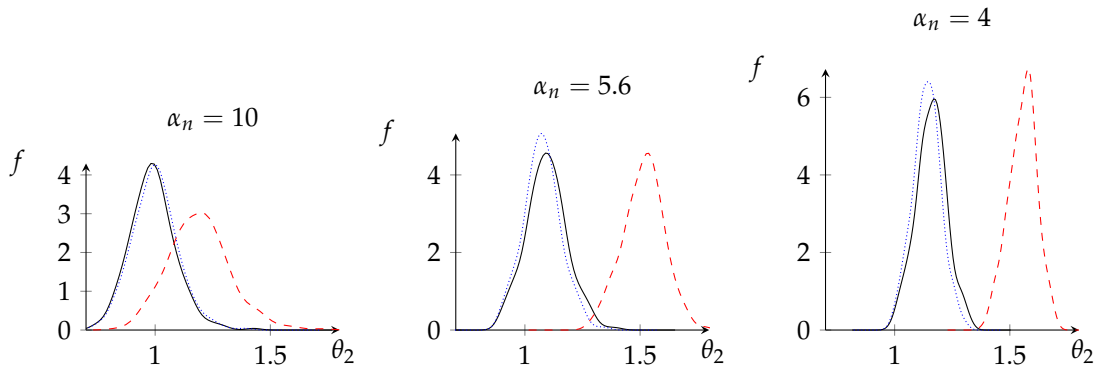


FIGURE 7. Density of  $\hat{\theta}_2$ ; Jeffreys (solid), uniform (dashed), estimated Jeffreys (dotted) prior;  $n = 1,000$ ,  $d = 2$ .

Figure 6 demonstrates that the bias decreases in  $\alpha_n$  for both priors, and that using Jeffreys' prior removes most of the bias of the uniform kernel, which is encouraging. It is apparent, however, that even Jeffreys' prior does not remove all asymptotic bias, especially when  $\alpha_n$  is small, which is due to the fact that for small values of  $\alpha_n$  the higher order terms in the bias expansion are of comparatively greater importance.

We repeated the same experiment, but now with the estimated version of Jeffreys' prior included. The results are depicted in figure 7.  $Q''$  was estimated using the method described in appendix I.3. Evidently, the error in estimating  $Q''$  is small relative to the bias reduction vis-à-vis the uniform prior. We conclude that an asymptotic bias correction using Jeffreys' prior is valuable unless  $\alpha_n$  is chosen large.

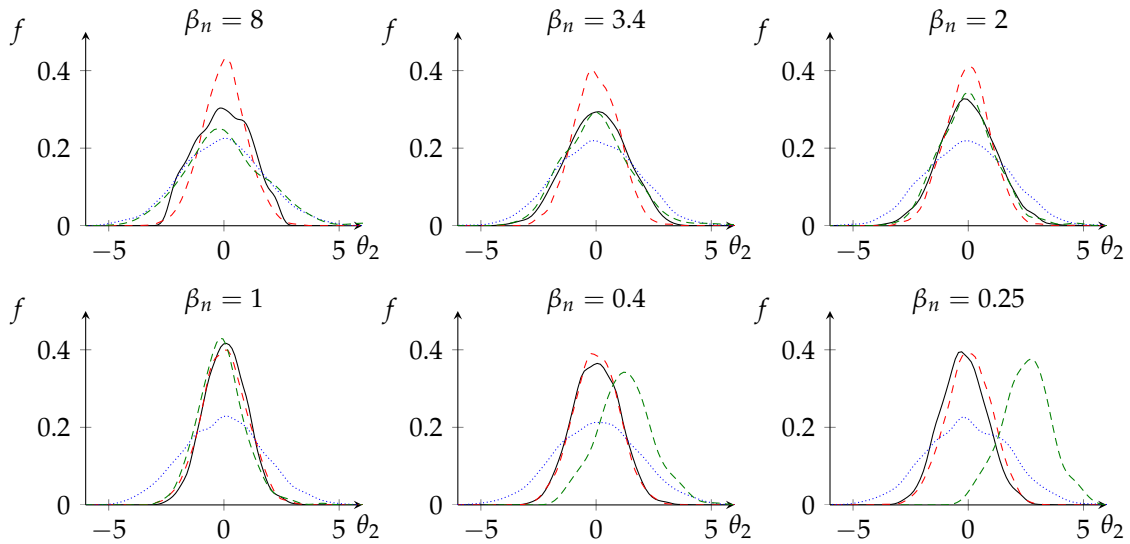


FIGURE 8. Density of  $\kappa_n(\hat{\theta}_2 - \theta_0)$  for various inference procedures: uniform (solid), normal (dashed), maximum score (dotted), finite sample (densely dashed);  $n = 1,000$ ,  $d = 2$ ,  $R = 3,000$ .

Finally, we evaluate the quality of our uniform inference procedure; the results are in figure 8. As figure 8 demonstrates, the uniform inference procedure of theorem 8 moves gradually from the maximum score limiting distribution to the normal limiting distribution as  $\alpha_n$  (and hence  $\beta_n$ ) decreases. It does a good job approximating the finite sample distribution of  $\hat{\theta}_2$  except that for small values of  $\alpha_n$  there is substantial bias, which is due to the fact that estimates used to generate the finite sample distribution are not higher order bias-corrected.<sup>18</sup>

The reason that the results in figure 8 use a design with  $d = 2$  only is that the limiting distribution of the maximum score estimator becomes very expensive to simulate for higher dimensions. This suggests that one should use our uniform inference procedure with a high value of  $\beta_n$  to simulate the limiting distribution of the maximum score estimator instead of trying to maximize  $\hat{G}(t) - t^\top \hat{V}t/2$  repeatedly. We found that the pattern of the remaining three densities for greater values of  $d$  (not shown) is similar to the one depicted in figure 8.

#### REFERENCES CITED

ABREVAYA, J., AND J. HUANG (2005): "On the bootstrap of the maximum score estimator," *Econometrica*, 73(4), 1175–1204.

<sup>18</sup>We do use Jeffreys' prior.

- ANDREWS, D., AND P. GUGGENBERGER (2008): "Hybrid and size-corrected subsampling methods," Discussion paper, Cowles Foundation.
- BAJARI, P., AND J. T. FOX (2009): "Measuring the efficiency of an FCC spectrum auction," Discussion paper, NBER.
- BAJARI, P., J. T. FOX, AND S. P. RYAN (2008): "Evaluating wireless carrier consolidation using semiparametric demand estimation," *Quantitative Marketing and Economics* (forthcoming).
- BILLINGSLEY, P. (1995): "Probability and Measure," .
- (1999): *Convergence of probability measures*. Wiley.
- CATTANEO, M. D., R. K. CRUMP, AND M. JANSSON (2008): "Small bandwidth asymptotics for density-weighted average derivatives," Discussion paper, UC Berkeley.
- CHERNOZHUKOV, V., AND H. HONG (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115(2), 293–346.
- CRAGG, J. G. (1992): "Quasi-Aitken estimation for heteroskedasticity of unknown form," *Journal of Econometrics*, 54(1-3), 179–201.
- DAVIDSON, J. (1994): *Stochastic limit theory*. Oxford University Press.
- DE JONG, R., AND T. WOUTERSEN (2007): "Dynamic time series binary choice," Discussion paper, Johns Hopkins University.
- DELGADO, M., J. RODRIGUEZ-POO, AND M. WOLF (2001): "Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator," *Economics Letters*, 73(2), 241–250.
- DOOB, J. L. (1953): *Stochastic processes*. Wiley.
- EFRON, B., AND R. TIBSHIRANI (1997): *An introduction to the bootstrap*. Chapman & Hall.
- FLORIOS, K., AND S. SKOURAS (2008): "Exact computation of max weighted score estimators," *Journal of Econometrics*, 146(1), 86–91.
- FOX, J. T. (2007): "Semiparametric estimation of multinomial discrete-choice models using a subset of choices," *Rand Journal of Economics*, 38(4), 1002–1019.
- (2009): "Estimating matching games with transfers," Discussion paper, University of Chicago.
- GEMAN, S., AND D. GEMAN (1984): "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- GUERRE, E., AND P. LAVERGNE (2005): "Data-driven rate-optimal specification testing in regression models," *Annals of Statistics*, pp. 840–870.

- HOROWITZ, J. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 60(3), 505–31.
- JEFFREYS, H. (1946): "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, pp. 453–461.
- KIEFER, N., AND T. VOGELANG (2005): "A new asymptotic theory for heteroskedasticity–autocorrelation robust tests," *Econometric Theory*, 21(06), 1130–1164.
- KIM, J., AND D. POLLARD (1990): "Cube root asymptotics," *Annals of Statistics*, 18(1), 191–219.
- KLEIN, R., AND R. SPADY (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61, 387–421.
- KOSOROK, M. (2008): *Introduction to empirical processes and semiparametric inference*. Springer Verlag.
- KOTLYAROVA, Y., AND V. ZINDE-WALSH (2003): "Improving the Efficiency and Robustness of the Smoothed Maximum Score Estimator," Discussion paper, McGill University.
- LEE, M. (1992): "Median regression for ordered discrete response," *Journal of Econometrics*, 51(1-2), 59–77.
- MANSKI, C. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3(3), 205–228.
- MANSKI, C. (1985): "Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313–333.
- MANSKI, C., AND T. THOMPSON (1986): "Operational characteristics of maximum score estimation," *Journal of Econometrics*, 32(1), 85–108.
- MIKUSHEVA, A. (2007): "Uniform inference in autoregressive models," *Econometrica*, 75(5), 1411–1452.
- MOON, H., AND F. SCHORFHEIDE (2009): "Bayesian and frequentist inference in partially identified models," Discussion paper, NBER.
- PINKSE, C. (1993): "On the computation of semiparametric estimates in limited dependent variable models," *Journal of Econometrics*, 58, 185–205.
- PINKSE, J. (2006): "Heteroskedasticity correction and dimension reduction," Discussion paper, Pennsylvania State University.
- POLITIS, D., J. ROMANO, AND M. WOLF (1999): *Subsampling*. Springer.
- POWELL, J., J. STOCK, AND T. STOKER (1989): "Semiparametric estimation of index coefficients," *Econometrica*, 57(6), 1403–1430.
- ROBERT, C., AND G. CASELLA (2004): *Monte Carlo statistical methods*. Springer.

ROUSSEEUW, P. (1984): “Least median of squares regression,” *Journal of the American Statistical Association*, pp. 871–880.

STAIGER, D., AND J. STOCK (1997): “Instrumental variables regression with weak instruments,” *Econometrica*, 65, 557–586.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak convergence and empirical processes*. Springer.

## APPENDIX A. PRELIMINARIES

**A.1. Notation.** Let  $\gamma_n = \alpha_n^{1/(4q+4)}$ ,  $\Gamma_n = \{t \in \mathfrak{R}^d : \|t\| \leq \gamma_n\}$ , and  $\Gamma_n^c = \mathfrak{R}^d - \Gamma_n$ . Let further for vector-valued  $t$ ,  $t^j$  denote 1 if  $j = 0$  and  $t$  if  $j = 1$ . Define  $C_V = \int \exp(-t^\top V t/2) dt = 1/\phi_V(0) = (2\pi)^{d/2}/\sqrt{\det(V)}$  and let  $\lambda_-$  be the smallest eigenvalue of  $V$ . Finally, let  $R_n(t) = Q_n(t) + t^\top V t/2$ .

**A.2. Weak Convergence.** The results in this section presume that assumptions F and G are satisfied.

**Lemma A.1.** *Let  $\mathcal{T}_1 \subseteq \mathcal{T}_2 \cdots$  be a sequence of compact sets such that  $0 \in \mathcal{T}_1$  and  $\mathfrak{R}^d = \bigcup_{i=1}^\infty \mathcal{T}_i$ . We then have  $\tilde{\mathbf{S}}_n \xrightarrow{w} \mathbf{G}$  in  $\mathcal{L}^\infty(\mathcal{T}_1, \mathcal{T}_2, \dots)$ , where  $\mathbf{G}(\cdot)$  is a Gaussian process with covariance kernel  $H$ .*

*Proof.* By van der Vaart and Wellner (1996, theorem 1.6.1), it suffices to establish the weak convergence of  $\tilde{\mathbf{S}}_n$  in  $\mathcal{L}^\infty(\mathcal{T})$  for an arbitrary compact set  $\mathcal{T} \subset \mathfrak{R}^d$ . Since  $\mathcal{T}$  is dense, assumption G(i) ensures that for  $j = 1, 2$ ,

$$\mathcal{F}_{n,\epsilon}^j = \{\tilde{\alpha}_n^{j/2} (g(\cdot; \theta_0 + t/\tilde{\alpha}_n) - g(\cdot; \theta_0 + t/\tilde{\alpha}_n))^j\}_{\|t-s\| < \epsilon}$$

is a pointwise measurable class, and hence  $\mathcal{P}$ -measurable for every  $\mathcal{P}$ ; see van der Vaart and Wellner (1996, page 110). Note also that  $\mathbb{E}[\tilde{\mathbf{S}}_n(t)\tilde{\mathbf{S}}_n(s)] = \tilde{\alpha}_n \mathbb{E}[\mathbf{g}_i(\theta_0 + t/\tilde{\alpha}_n)\mathbf{g}_i(\theta_0 + s/\tilde{\alpha}_n)] \rightarrow H(t, s)$  for every  $t, s \in \mathcal{T}$  by assumption F. Therefore, the result follows from van der Vaart and Wellner (1996, theorem 2.11.22).  $\square$

**Lemma A.2.** *Let for  $\tilde{\gamma} > 0$ ,  $\tilde{\Gamma} = \{t \in \mathfrak{R}^d : \|t\| \leq \tilde{\gamma}\}$ . Then (i) for any  $0 < c < \infty$ ,*

$$\lim_{\tilde{\gamma} \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{P}[\sup_{t \in \tilde{\Gamma}^c} |\tilde{\mathbf{S}}_n(t)|/\|t\|^2 > c] = 0 \text{ and (ii) } \mathbb{P}[\sup_{t \in \Gamma_n^c} |\tilde{\mathbf{S}}_n(t)|/\|t\|^2 > c] = o(1).$$

*Proof.* Let  $\tilde{\tilde{\mathbf{S}}}_n(t) = \tilde{\mathbf{S}}_n(t)/\|t\|$  and  $\tilde{\mathbf{G}}(t) = \mathbf{G}(t)/\|t\|$ . Suppose without loss of generality that  $\alpha_1 = 1$ . Then the equicontinuity of  $\tilde{\tilde{\mathbf{S}}}_n$  on  $\Gamma_1^c$  follows from that of  $\tilde{\mathbf{S}}_n$ . Therefore on  $\Gamma_1^c$ ,  $\tilde{\tilde{\mathbf{S}}}_n \xrightarrow{w} \tilde{\mathbf{G}}$ , which is a uniformly bounded process in  $\mathcal{L}^\infty(\Gamma_1^c)$ . Hence, to establish (ii) ((i) is similar), note that  $\sup_{t \in \Gamma_n^c} (|\tilde{\mathbf{S}}_n(t)|/\|t\|^2) \leq \sup_{t \in \Gamma_1^c} (|\tilde{\mathbf{S}}_n(t)|(\log \|t\|)/\|t\|^2) / \log \gamma_n = O_p(1)o(1) = o_p(1)$ .  $\square$

**Lemma A.3.** *For any  $c > 0$ ,  $\sup_t \{|\tilde{\mathbf{S}}_n(t)| - ct^\top V t\} = O_p(1)$ .*

*Proof.* Follows immediately from lemma A.1 and the properties of  $\mathbf{G}$ .  $\square$

### A.3. Auxiliary Results.

**Lemma A.4.**  $\lim_{\epsilon \downarrow 0} \sup_{\|\theta - \theta_0\| \leq \epsilon} \|Q''(\theta) + V\| = 0.$

*Proof.* Follows from the continuity of  $Q''$  at  $\theta_0$ .  $\square$

**Lemma A.5.** For some  $0 < c_q < \infty$  and all  $\theta \in \Theta$ ,  $Q(\theta) \leq -\min(c_q, (\theta - \theta_0)'V(\theta - \theta_0)/4)$  and  $Q_n(t) \leq -\min(\tilde{\alpha}_n^2 c_q, t^\top V t/4)$  for all  $t$  for which  $\theta_0 + t/\tilde{\alpha}_n \in \Theta$ .

*Proof.* Follows from the fact that (i)  $\Theta$  is compact, (ii)  $Q$  has a unique maximum of zero at  $\theta_0$ , and (iii)  $Q_{\theta\theta}$  is continuous at  $\theta_0$  with value  $-V$ .  $\square$

**Lemma A.6.** For any  $|c| \leq 1$ , any  $b$ , and any nonnegative integer  $j$ ,  $|\exp(cb) - \sum_{s=0}^j (cb)^s/s!| \leq |c|^{j+1} \exp(|b|).$

*Proof.* We have

$$|\exp(cb) - \sum_{s=0}^j (cb)^s/s!| \leq |\sum_{s=j+1}^{\infty} (cb)^s/s!| \leq |c|^{j+1} \sum_{s=j+1}^{\infty} |b|^s/s! \leq |c|^{j+1} \exp(|b|). \quad \square$$

**A.4. Further Auxiliary Results.** The results in this section presume that  $0 \leq \lim_{n \rightarrow \infty} \beta_n = c_\beta < \infty$  and that  $\tilde{\alpha}_n = \alpha_n$ .

**Lemma A.7.** For all nonnegative and finite  $c_1, c_2, c_3, c_4$  and any polynomial  $P$ ,

(i) if  $c_3 > 0$  then  $\int_{\Gamma_n^c} \|P(t)\| \exp(c_1 \beta_n |\tilde{\mathbf{S}}_n(t)| - c_3 t^\top V t) dt = o_p(\alpha_n^{-c_4}),$

(ii) if  $c_2 + c_3 > 0$  then  $\int_{\Gamma_n^c} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\tilde{\mathbf{S}}_n(t)| + c_2 Q_n(t) - c_3 t^\top V t) dt = o_p(\alpha_n^{-c_4}).$

*Proof.* First (i). By lemma A.2, the left hand side of (i) is for any  $c$  of smaller order than

$$\int_{\Gamma_n^c} \|P(t)\| \exp((c c_1 c_\beta - c_3 \lambda_-) \|t\|^2) dt, \text{ which for sufficiently small } c \text{ decreases exponentially in } n.$$

The left hand side in (ii) is by lemma A.5 of order no greater than

$$\begin{aligned} & \int_{\Gamma_n^c} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\tilde{\mathbf{S}}_n(t)| - c_2 \min(c_q \alpha_n^2, t^\top V t/4) - c_3 t^\top V t) dt \\ & \leq \int_{\Gamma_n^c} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\tilde{\mathbf{S}}_n(t)| - c_2 c_q \alpha_n^2 - c_3 t^\top V t) dt \\ & \quad + \int_{\Gamma_n^c} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\tilde{\mathbf{S}}_n(t)| - (c_2/4 + c_3) t^\top V t) dt. \quad (16) \end{aligned}$$

The second right hand side term in (16) was dealt with in (i). If  $c_3 > 0$ , the first right hand side term is bounded by the same expression with  $c_2 = 0$ , which was dealt with in (i), also. Finally, the first right hand side term if  $c_3 = 0$ . Since  $\sup_\theta |\mathbf{S}_n(\theta)| = o_p(1)$  by lemma A.3, it follows that for some  $c^* > 0$ ,  $\int_{\Gamma_n^c} \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\tilde{\mathbf{S}}_n(t)| - c_2 c_q \alpha_n^2) dt$  is of order no greater than  $\exp(-c^* \alpha_n^2)$ .  $\square$

**Lemma A.8.** For any  $c > 0$ , (i)  $\mathbb{P}[\sup_{t \in \Gamma_n} I(|R_n(t)| > ct^\top Vt/4) \neq 0] = o(1)$ .

*Proof.* Follows from lemma A.4,  $\gamma_n = o(\alpha_n)$ , and the fact that  $R_n(t) = t^\top (Q''(\theta_0 + t/\alpha_n) + V)t/2$ .  $\square$

**Lemma A.9.** For all nonnegative and finite  $c_1, c_2, c_3, c_4$  and any polynomial  $P$ ,

(i) if  $c_3 > 0$  then  $\int \|P(t)\| \exp(c_1 \beta_n |\tilde{\mathcal{S}}_n(t)| - c_3 t^\top Vt) dt = O_p(1)$ ,

(ii) if  $c_2 + c_3 > 0$  then  $\int \|\pi_n(t)P(t)\| \exp(c_1 \beta_n |\tilde{\mathcal{S}}_n(t)| + c_2 Q_n(t) - c_3 t^\top Vt) dt = O_p(1)$ .

*Proof.* By lemma A.7, we only need to show that the integrals over  $t \in \Gamma_n$  are  $O_p(1)$ . For (i) the stated result follows from lemma A.1. For (ii), it follows from lemmas A.1 and A.8.  $\square$

#### APPENDIX B. $\alpha_n = c_\alpha^2 \sqrt[3]{n}$

**Lemma B.1.** For  $j = 0, 1$ ,  $\int_{\Gamma_n} \pi_n(t) t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t)) (\exp(R_n(t)) - 1) \phi_V(t) dt = o_p(1)$ .

*Proof.* By lemma A.6 for  $b = R_n(t)/c$ , the left hand side is bounded in absolute value by  $c \int_{\Gamma_n} \|\pi_n(t) t^j\| \exp(c_\alpha^3 |\tilde{\mathcal{S}}_n(t)| + |R_n(t)/c| - t^\top Vt/2) dt$ , which with probability approaching one is bounded above by  $c \int_{\Gamma_n} \|\pi_n(t) t^j\| \exp(c_\alpha^3 |\tilde{\mathcal{S}}_n(t)| - t^\top Vt/4) dt \xrightarrow{d} c\mathbf{Y}$  by lemma A.1 for some  $\mathbf{Y}$  independent of  $c$ . Let  $c \downarrow 0$ .  $\square$

**Lemma B.2.** For  $j = 0, 1$ ,  $\int \pi_n(t) t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t) + Q_n(t)) dt = C_V \pi_0 \int t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t)) \phi_V(t) dt + o_p(1)$ .

*Proof.* We have

$$\begin{aligned} & \int \pi_n(t) t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t) + Q_n(t)) dt - C_V \pi_0 \int t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t)) \phi_V(t) dt \\ &= \int_{\Gamma_n^c} \pi_n(t) t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t) + Q_n(t)) dt + C_V \int_{\Gamma_n} \pi_n(t) t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t)) (\exp(R_n(t)) - 1) \phi_V(t) dt \\ & \quad + C_V \int_{\Gamma_n} (\pi_n(t) - \pi_0) t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t)) \phi_V(t) dt - \pi_0 \int_{\Gamma_n^c} t^j \exp(c_\alpha^3 \tilde{\mathcal{S}}_n(t) - t^\top Vt/2) dt. \quad (17) \end{aligned}$$

The first and last right hand side terms in (17) are dealt with in lemma A.7. The second right hand side term in (17) is  $o_p(1)$  by lemma B.1. Because  $\pi$  is continuous at  $\theta_0$ , the third right hand side term in (17) is also  $o_p(1)$  by lemma A.3.  $\square$

APPENDIX C.  $\sqrt[3]{n} = o(\alpha_n)$ 

**Lemma C.1.** *Let  $\tilde{\gamma}, \tilde{\Gamma}$  be as in lemma A.2. Let further  $\tilde{L}_n(t) = \tilde{S}_n(t) + Q_n(t)$  Then*

$$\frac{\int_{\tilde{\Gamma}} \pi_n(t) t \exp(\beta_n^{4/3} \tilde{L}_n(t)) dt}{\int_{\tilde{\Gamma}} \pi_n(t) \exp(\beta_n^{4/3} \tilde{L}_n(t)) dt} \xrightarrow{d} \operatorname{argmax}_{t \in \tilde{\Gamma}} \tilde{G}(t).$$

*Proof.* Note that  $\mathbb{L}^2(\tilde{\Gamma}, \mathbb{B}, \mu)$  is separable where  $\mathbb{B}$  is the Borel sigma algebra. Since  $\tilde{G}(t) = \mathbb{G}(t) - t^\top V t / 2$ ,  $\tilde{L}_n, \tilde{G}$  are both in  $\mathbb{L}^2(\tilde{\Gamma}, \mathbb{B}, \mu)$  and  $\tilde{L}_n \xrightarrow{w} \tilde{G}$  by lemma A.1 and the fact that

$$\sup_{t \in \tilde{\Gamma}} |Q_n(t) + t^\top V t / 2| \leq \tilde{\gamma}^2 \sup_{\|\theta - \theta_0\| \leq \tilde{\gamma} / \sqrt[3]{n}} \|Q_n''(\theta) + V\| / 2 = o_p(1), \quad (18)$$

by lemma A.4. Hence, by the Skorokhod representation theorem (Billingsley, 1999, theorem 6.7), there exist  $\tilde{L}_n^*, \tilde{G}^*$  with the same distributions as  $\tilde{L}_n, \tilde{G}$ , such that for an arbitrary sample path  $\tilde{L}_n^* = \tilde{L}_n^*(\cdot; \omega)$  of  $\tilde{L}_n^*$  and corresponding sample path  $\tilde{G}^* = \tilde{G}^*(\cdot; \omega)$  of  $\tilde{G}^*$ ,

$$\int_{\tilde{\Gamma}} |\tilde{L}_n^*(t) - \tilde{G}^*(t)|^2 dt = o(1). \quad (19)$$

Let for arbitrary sets  $T_1, T_2 \subset \tilde{\Gamma}$ ,  $d^*(T_1, T_2) = \mu(T_1 - T_2) + \mu(T_2 - T_1)$ . Let further for arbitrary  $c > 0$ ,  $T(\varkappa, c) = \{t \in \tilde{\Gamma} : |\varkappa(t) - \tilde{G}^*| \leq c\}$ , where  $\tilde{G}^* = \max_{t \in \tilde{\Gamma}} \tilde{G}^*(t)$ . We first establish that

$$d^*(T(\tilde{L}_n^*, c), T(\tilde{G}^*, c)) = o(1). \quad (20)$$

Let  $T_{1n}(c) = T(\tilde{L}_n^*, c) - T(\tilde{G}^*, c)$  and  $T_{2n}(c) = T(\tilde{G}^*, c) - T(\tilde{L}_n^*, c)$ . We show that  $\mu(T_{2n}(c)) = o(1)$  where the same result for  $T_{1n}(c)$  follows similarly. Let for arbitrary  $c^* > 0$ ,  $T_n^*(c^*) = \{t \in \tilde{\Gamma} : |\tilde{L}_n^*(t) - \tilde{G}^*(t)| \leq c^*\}$ . For the remainder of this lemma, define complements relative to  $\tilde{\Gamma}$  (e.g.  $T_n^{*c}(c^*) = \tilde{\Gamma} - T_n^*(c^*)$ ). Because  $\mu(T_n^{*c}(c^*) \cap T_{2n}(c)) \leq \mu(T_n^{*c}(c^*)) = o(1)$  by (19), we only need to consider  $\mu(T_n^*(c^*) \cap T_{2n}(c))$ .

Note first that  $T_n^*(c^*) \cap T_{2n}(c) \subseteq T_n^{**}(c, c^*) = \{t \in \tilde{\Gamma} : c \leq |\tilde{L}_n^*(t) - \tilde{G}^*| \leq c + c^*\}$ , such that by (19),

$$\begin{aligned} \lim_{c^* \downarrow 0} \lim_{n \rightarrow \infty} \mu(T_n^{**}(c, c^*)) &= \lim_{c^* \downarrow 0} \mu(\{t \in \tilde{\Gamma} : c \leq |\tilde{G}^*(t) - \tilde{G}^*| \leq c + c^*\}) \\ &= \mu(\{t \in \tilde{\Gamma} : |\tilde{G}^*(t) - \tilde{G}^*| = c\}) = 0, \end{aligned} \quad (21)$$

because  $\tilde{G}^*$  is continuous and nowhere differentiable. So (20) holds.

Finally, note that for  $j = 0, 1$  and some finite constant  $C$ ,

$$\int_{T^c(\tilde{L}_n^*, c)} \|t\|^j |\pi_n(t)| \exp\{\beta_n^{4/3} (\tilde{L}_n^*(t) - \tilde{G}^*)\} dt \leq C \exp(-\beta_n^{4/3} c) \int \|t\|^j |\pi_n(t)| dt = o(1),$$

since the support of  $\pi_n$  is only increasing at a rate of  $\alpha_n$  by assumption C. Thus,

$$\frac{\int_{\tilde{\Gamma}} \pi_n(t) t \exp(\beta_n^{4/3} \tilde{L}_n^*(t)) dt}{\int_{\tilde{\Gamma}} \pi_n(t) \exp(\beta_n^{4/3} \tilde{L}_n^*(t)) dt} \leq \text{ess sup } T(\tilde{L}_n^*, c) + o(1) = \text{ess sup } T(\tilde{\mathbf{G}}^*, c) + o(1).$$

Repeat the arguments to get a lower bound equal to  $\text{ess inf } T(\tilde{\mathbf{G}}^*, c) + o(1)$ . The stated result then follows from the fact that  $\lim_{c \downarrow 0} \text{ess inf } T(\tilde{\mathbf{G}}^*, c) = \lim_{c \downarrow 0} \text{ess sup } T(\tilde{\mathbf{G}}^*, c) = \text{argmax}_{t \in \tilde{\Gamma}} \tilde{\mathbf{G}}^*(t)$  for almost all sample paths by lemma 2.6 of Kim and Pollard (1990).  $\square$

**Lemma C.2.** For  $\tilde{\Gamma}$  as defined in lemma A.2 and any  $\epsilon > 0$ ,

$$\lim_{\tilde{\gamma} \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{P} \left[ \frac{\int_{\tilde{\Gamma}^c} \|t^j \pi_n(t)\| \exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} dt}{\int \pi_n(t) \exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} dt} > \epsilon \right] = o(1). \quad (22)$$

*Proof.* We first work on the numerator in (22). For  $c > 0$ , let  $\mathbf{Z}_n(\tilde{\gamma}, c) = \{t \in \tilde{\Gamma}^c : \pi_n(t) > 0 \wedge |\tilde{\mathbf{S}}_n(t)| \leq c \|t\|^2\}$ . For  $t \in \mathbf{Z}_n(\tilde{\gamma}, c)$  we have by assumption C that for some fixed  $c^* > 0$ ,  $\sup_{t: \pi_n(t) > 0} \|t\| \leq c^* \sqrt[3]{n}$  such that using  $\exp(-\min(a, b)) \leq e^{-a} + e^{-b}$ , by lemma A.5,

$$\exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} \leq \exp\{\beta_n^{4/3}(cc^{*2} - c_q)n^{2/3}\} + \exp\{\beta_n^{4/3}(c - \lambda^-/4)\|t\|^2\},$$

which for sufficiently small  $c$ , some  $c^{**} > 0$  independent of  $\tilde{\gamma}, n$ , and sufficiently large  $n$  is bounded by  $2 \exp(-\beta_n^{4/3} c^{**} \tilde{\gamma}^2)$ . So for any  $0 < \bar{c} < \infty$ , some finite  $C$ , and any  $\epsilon > 0$  (using  $\mathbb{P}[\mathcal{E}_1] \leq \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] + \mathbb{P}[\mathcal{E}_2^c]$ ),

$$\begin{aligned} & \lim_{\tilde{\gamma} \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{P} \left[ \exp(\bar{c} \beta_n^{4/3}) \int_{\tilde{\Gamma}^c} \|t^j \pi_n(t)\| \exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} dt > \epsilon \right] \\ & \leq \lim_{\tilde{\gamma} \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{P} \left[ 2C \alpha_n^j \exp\{(\bar{c} - c^{**} \tilde{\gamma}^2) \beta_n^{4/3}\} > \epsilon \right] + \lim_{\tilde{\gamma} \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{t \in \tilde{\Gamma}^c} |\tilde{\mathbf{S}}_n(t)| / \|t\|^2 > c \right] = 0. \end{aligned}$$

Finally the denominator in (22). If  $\mathbf{Z}_n^*(\bar{c}) = \{t : \pi_n(t) > 0 \wedge \tilde{\mathbf{S}}_n(t) + Q_n(t) > -\bar{c}/2\}$  then

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{P} \left[ \int \pi_n(t) \exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} dt \geq \exp(-\beta_n^{4/3} \bar{c}) \right] \\ & \geq \lim_{n \rightarrow \infty} \mathbb{P} \left[ \underline{\pi} \int_{\mathbf{Z}_n^*(\bar{c})} \exp(-\beta_n^{4/3} \bar{c}/2) dt \geq \exp(-\beta_n^{4/3} \bar{c}) \right] \\ & \geq \lim_{n \rightarrow \infty} \mathbb{P} \left[ \underline{\pi} \exp(-\bar{c} \beta_n^{4/3}/2) \mu(\mathbf{Z}_n^*(\bar{c})) \geq \exp(-\beta_n^{4/3} \bar{c}) \right] = 1. \quad \square \end{aligned}$$

APPENDIX D.  $\alpha_n = o(\sqrt[3]{n})$

**Lemma D.1.**  $\int \pi_n(t) \exp(Q_n(t)) dt = \pi_0 C_V + o(1)$ .

*Proof.* We have by lemma A.7,

$$\begin{aligned} \int \pi_n(t) \exp(Q_n(t)) dt - \pi_0 C_V &= \int_{\Gamma_n} \pi_n(t) \exp(Q_n(t)) dt + o(1) \\ &= C_V \int_{\Gamma_n} \pi_n(t) \left( \exp(R_n(t)) - 1 \right) \phi_V(t) dt + C_V \int_{\Gamma_n} (\pi_n(t) - \pi_0) \phi_V(t) dt \\ &\quad - C_V \pi_0 \int_{\Gamma_n^c} \phi_V(t) dt + o(1). \end{aligned} \quad (23)$$

The third right hand side term in (23) vanishes because  $\gamma_n$  increases to  $\infty$ , the first term is  $o(1)$  by lemma A.8, and the second term is  $o(1)$  by the continuity of  $\pi$  at  $\theta_0$ .  $\square$

**Lemma D.2.** Let  $\mu_{in} = \sqrt{\alpha_n/n} \int t \tilde{g}_i(\theta_0 + t/\alpha_n) \phi_V(t) dt$ , such that  $\{\mu_{in}\}$  is an independent mean zero array. Then  $\sum_{i=1}^n \mathbb{E} \|\mu_{in}\|^{2+\iota} = o(1)$  for  $\iota$  defined in assumption E.

*Proof.* We have for  $\iota^*$  defined in assumption E,

$$\sum_{i=1}^n \mathbb{E} \|\mu_{in}\|^{2+\iota} \leq \alpha_n^{1-\iota-\iota^*} \left( \frac{\alpha_n^3}{n} \right)^{\iota/2} \int \|t\|^{2+\iota} \alpha_n^{\iota^*} \mathbb{E} |\tilde{g}_i(\theta_0 + t/\alpha_n)|^{2+\iota} \phi_V(t) dt = o(1) O(1) O(1) = o(1),$$

by assumption E.  $\square$

**Lemma D.3.** Let  $\mathcal{V}_N = \pi_0^2 \iint ts^\top H(t,s) \phi_V(t) \phi_V(s) dt ds$ . Then  $\int \pi_n(t) t \tilde{S}_n(t) \phi_V(t) dt \xrightarrow{d} N(0, \mathcal{V}_N)$ .

*Proof.* By lemma A.1,  $\int (\pi_n(t) - \pi_0) t \tilde{S}_n(t) \phi_V(t) dt = o_p(1)$ . Further,

$$\pi_0 \int t \tilde{S}_n(t) \phi_V(t) dt = n^{-1/2} \sum_{i=1}^n \pi_0 \sqrt{\alpha_n} \int t \tilde{g}_i(\theta_0 + t/\alpha_n) \phi_V(t) dt \xrightarrow{d} N(0, \mathcal{V}_N),$$

by Lindeberg's theorem (see e.g. theorem 23.6 of Davidson, 1994); the Lindeberg condition is satisfied by lemma D.2.  $\square$

**Lemma D.4.**  $\int \|\pi_n(t) t \tilde{S}_n(t)\| \left( \exp(R_n(t)) - 1 \right) \phi_V(t) dt = o_p(1)$ .

*Proof.* For  $\int_{\Gamma_n^c}$ , use lemma A.7 and for  $\int_{\Gamma_n}$  follow the same steps as in the proof of lemma B.1.  $\square$

**Lemma D.5.** For any  $j = 0, 1$ ,  $\int \|t\|^j |\pi_n(t)| \exp(|\beta_n \tilde{S}_n(t)| + Q_n(t)) dt = O_p(1)$ .

*Proof.* Define  $S_n^* = \sup_{t \in \Gamma_n} (|\tilde{S}_n(t)| - t^\top V t / 4)$ . For  $\int_{\Gamma_n^c}$  the stated result follows from lemma A.7. So we only need to deal with

$$\int_{\Gamma_n} \|t\|^j |\pi_n(t)| \exp(|\beta_n \tilde{S}_n(t)| + Q_n(t)) dt \leq \exp(\beta_n S_n^*) \int_{\Gamma_n} \|t\|^j |\pi_n(t)| \exp(\beta_n t^\top V t / 4 + Q_n(t)) dt.$$

Since  $\lim_{\|t\| \downarrow 0} (Q_n(t) / t^\top V t) = -1/2$ , it follows that for sufficiently large  $n$ ,  $Q_n(t) \leq -t^\top V t / 4$  for all  $t \in \Gamma_n$ . Because  $\beta_n = o(1)$  and  $S_n^* = O_p(1)$  by lemma A.3, the stated result follows.  $\square$

**Lemma D.6.** For  $j = 0, 1$ ,  $\int \pi_n(t) t^j \{ \exp(\beta_n \tilde{\mathfrak{S}}_n(t)) - \sum_{s=0}^j (\beta_n \tilde{\mathfrak{S}}_n(t))^s \} \exp(Q_n(t)) dt = O_p(\beta_n^{j+1})$ .

*Proof.* Apply lemma A.6 with  $c = \beta_n$  and  $b = \tilde{\mathfrak{S}}_n(t)$ , followed by lemma D.5.  $\square$

**Lemma D.7.**  $\mathcal{N}_n - \mathcal{B}_n \xrightarrow{d} N(0, C_V^2 \mathcal{V}_N)$ .

*Proof.* Use  $t = \alpha_n(\theta - \theta_0)$  to obtain

$$\begin{aligned} \mathcal{N}_n - \mathcal{B}_n &= \frac{1}{\beta_n} \int \pi_n(t) t \{ \exp(\beta_n \tilde{\mathfrak{S}}_n(t)) - 1 \} \exp(Q_n(t)) dt = \\ &\int \pi_n(t) t \tilde{\mathfrak{S}}_n(t) \exp(Q_n(t)) dt + \frac{1}{\beta_n} \int \pi_n(t) t \left( \exp(\beta_n \tilde{\mathfrak{S}}_n(t)) - 1 - \beta_n \tilde{\mathfrak{S}}_n(t) \right) \exp(Q_n(t)) dt. \end{aligned} \quad (24)$$

The first right hand side term in (24) converges in distribution to the stated normal by lemmas D.3 and D.4. The last term in (24) is  $O_p(\beta_n) = o_p(1)$  by lemma D.6.  $\square$

**Lemma D.8.**  $\mathcal{D}_n = \pi_0 C_V + o_p(1)$ .

*Proof.* Use  $t = \alpha_n(\theta - \theta_0)$  to obtain

$$\begin{aligned} \mathcal{D}_n &= \int \pi_n(t) \exp(\beta_n \tilde{\mathfrak{S}}_n(t) + Q_n(t)) dt \\ &= \int \pi_n(t) \exp(Q_n(t)) dt + \int \pi_n(t) \left( \exp(\beta_n \tilde{\mathfrak{S}}_n(t)) - 1 \right) \exp(Q_n(t)) dt. \end{aligned} \quad (25)$$

The first right hand side term in (25) is  $\pi_0 C_V + o_p(1)$  by lemma D.1. The last term in (25) is  $o_p(1)$  by lemma D.6.  $\square$

## APPENDIX E. BIAS

**E.1. Bias Expansion.** In this subsection, the assumptions of theorem 5 are used.

**Lemma E.1.** For some finite  $C_R$ ,  $\sup_{t \in \Gamma_n} |R_n(t)| \leq C_R \gamma_n^3 / \alpha_n = o(1)$ .

*Proof.* Because  $Q$  is three times continuously differentiable at  $\theta_0$ ,

$$\sup_{t \in \Gamma_n} |R_n(t)| \leq \gamma_n^2 \sup_{\|\theta - \theta_0\| \leq \gamma_n / \alpha_n} \|Q''(\theta) + V\| / 2 \leq C_R \gamma_n^3 / \alpha_n,$$

which is  $o(1)$  by the definition of  $\gamma_n$ .

**Lemma E.2.**

$$\int_{\Gamma_n} \pi_n(t) t \left( \exp(R_n(t)) - \sum_{p=0}^q \frac{(R_n(t))^p}{p!} \right) \phi_V(t) dt = O((\gamma_n^3 / \alpha_n)^{q+1}).$$

*Proof.* The length of the left hand side is by lemma A.6 for  $s_n^* = \sup_{t \in \Gamma_n} |R_n(t)|$  equal to

$$\left\| \sum_{p=q+1}^{\infty} \int_{\Gamma_n} \pi_n(t) t \frac{(R_n(t))^p}{p!} \phi_V(t) dt \right\| \leq (s_n^*)^{q+1} \exp(s_n^*) \int_{\Gamma_n} \|\pi_n(t)t\| \phi_V(t) dt = O((s_n^*)^{q+1}).$$

Apply lemma E.1. □

Let  $\mathcal{M}_{pq}$  be the collection of  $q$ -vectors  $m$  of nonnegative integers for which  $\sum_{\ell=1}^q m_\ell = p$ . Let further  $\mathcal{M}_{pqr}^* = \{m \in \mathcal{M}_{pq} : \sum_{\ell=1}^q \ell m_\ell = r\}$ , let  $D_{Q\delta}(t)$  be the  $\delta$ -th term in a Taylor expansion of  $Q(\theta_0 + t)$  about  $Q(\theta_0)$  and let  $D_{\pi\delta}(t)$  be likewise for  $\pi$ , e.g.  $D_{Q2}(t) = t^\top Q''(\theta_0)t/2$ . Finally, let  $D_{Q\delta_n}^*(t)$  be  $\alpha_n^\delta$  times the remainder term in a  $\delta$ -th order Taylor expansion of  $Q(\theta_0 + t/\alpha_n)$  about  $Q(\theta_0)$  and let  $D_{\pi\delta_n}^*(t)$  be likewise for  $\pi$ .

**Lemma E.3.** For any vector  $a \in \mathfrak{R}^q$ ,  $(\sum_{\ell=1}^q a_\ell)^p / p! = \sum_{m \in \mathcal{M}_{pq}} \prod_{\ell=1}^q a_\ell^{m_\ell} / m_\ell!$ .

*Proof.* This is a restatement of the multinomial theorem. □

**E.2. Bias Correction.** In this section the assumptions of theorem 7 are used;  $r^*, \rho_n$  are as defined in assumption I, take  $\pi, \hat{\pi}$  to equal  $\pi_J, \hat{\pi}_J$ , let  $\hat{\pi}_0 = \hat{\pi}(\theta_0)$ , and  $\hat{\pi}_n(t) = \hat{\pi}(\theta_0 + t/\alpha_n)$ . Let further  $v(\theta) = \text{vec}\{(-Q''(\theta))^{-1}\}$  and  $\hat{v}(\theta) = \text{vec}\{(-\hat{Q}''(\theta))^{-1}\}$ .

**Lemma E.4.** (i)  $\hat{\pi}_0 = \pi_0 + o_p(\alpha_n^{-r^*})$  and (ii)  $\sup_{\|\theta - \theta_0\| \leq \rho_n} \|\hat{\pi}'(\theta) - \pi'(\theta)\| = o_p(\alpha_n^{-r^*})$ .

*Proof.* Part (i) is implied by assumptions I and D ( $\det(V) > 0$ ). For (ii), note that

$$\pi'(\theta) = -\pi(\theta)Q'''(\theta)v(\theta)/2, \tag{26}$$

in a neighborhood of  $\theta_0$ . Hence in such a neighborhood,

$$2(\hat{\pi}'(\theta) - \pi'(\theta)) = -(\hat{\pi}_0 - \pi_0)\hat{Q}'''(\theta)\hat{v}(\theta) - \pi_0\{\hat{Q}'''(\theta)\hat{v}(\theta) - Q'''(\theta)v(\theta)\}.$$

Apply part (i) and assumption I. □

**Lemma E.5.**  $b_{q1}^* = \int (\pi_0 D_{Q3}(t) + D_{\pi 1}(t)) t \phi_V(t) dt = 0$ .

*Proof.* Let  $V^{j\ell}$  denote the  $(j, \ell)$  element of  $V^{-1}$  and let  $\mathcal{K}_{ps\ell} = \partial_{\theta_p \theta_s \theta_\ell} Q(\theta_0)$ . Then the  $j$ -th element of  $b_{q1}^*$  is equal to

$$\begin{aligned} & \frac{\pi_0}{6} \sum_{p,s,\ell=1}^d \mathcal{K}_{ps\ell} \int t_p t_s t_\ell t_j \phi_V(t) dt + \sum_{\ell=1}^d \partial_{\theta_\ell} \pi(\theta_0) \int t_\ell t_j \phi_V(t) dt \\ &= \frac{\pi_0}{6} \left( \sum_{p,s,\ell=1}^d \mathcal{K}_{ps\ell} (V^{ps} V^{j\ell} + V^{p\ell} V^{sj} + V^{pj} V^{s\ell}) - 3 \sum_{p,s,\ell=1}^d \mathcal{K}_{ps\ell} V^{ps} V^{j\ell} \right) = 0, \end{aligned}$$

where the first equality follows from (26).  $\square$

**Lemma E.6.**  $\int (\alpha_n + D_{Q3}(t)) \hat{\pi}_n(t) t \phi_V(t) dt = o_p(\alpha_n^{-r^*})$ .

*Proof.* The left hand side in the lemma statement is by the mean value theorem for some  $t^*$  between zero and  $t$  equal to

$$\int t t^\top \hat{\pi}'_n(t) \phi_V(t) dt + \hat{\pi}_0 \int D_{Q3}(t) t \phi_V(t) dt + \alpha_n^{-1} \int t t^\top \hat{\pi}'_n(t^*) D_{Q3}(t) \phi_V(t) dt,$$

which by lemma E.5 is equal to

$$\begin{aligned} & \int t t^\top (\hat{\pi}'_n(t) - \pi'_0) \phi_V(t) dt + (\hat{\pi}_0 - \pi_0) \int D_{Q3}(t) t \phi_V(t) dt \\ &+ \alpha_n^{-1} \int t t^\top (\hat{\pi}'_n(t^*) - \pi'_0) D_{Q3}(t) \phi_V(t) dt + \alpha_n^{-1} \int D_{\pi 1}(t) D_{Q3}(t) t \phi_V(t) dt, \end{aligned}$$

whose first three terms are  $o_p(\alpha_n^{-r^*})$  by lemma E.4, dominated convergence, and assumption C, and whose last term is zero because of the symmetry of the normal distribution.  $\square$

## APPENDIX F. UNIFORM INFERENCE

The results in appendix F presume that the assumptions of theorem 8 are satisfied.

**Lemma F.1.**  $\hat{\mathbf{G}} \xrightarrow{w} \mathbf{G}$  in  $\mathcal{L}^\infty(\mathcal{T}_1, \mathcal{T}_2, \dots)$ .

*Proof.* For any fixed and distinct  $t_1, \dots, t_j$ , we can write

$$\begin{bmatrix} \hat{\mathbf{G}}(t_1) \\ \vdots \\ \hat{\mathbf{G}}(t_j) \end{bmatrix} = \hat{\mathbf{\Omega}}^{1/2}(t_1, \dots, t_j) \boldsymbol{\eta},$$

with  $\boldsymbol{\eta} \sim N(0, I_j)$  and  $\hat{\boldsymbol{\Omega}}(t_1, \dots, t_j) \in \mathfrak{R}^{j \times j}$  with  $(\ell, p)$ -element  $\hat{H}(t_\ell, t_p)$ . Let  $\Omega$  be as  $\hat{\boldsymbol{\Omega}}$  with  $\hat{H}$  replaced by  $H$ . Since  $\hat{H}$  is a consistent estimator of  $H$ ,

$$\hat{\boldsymbol{\Omega}}^{1/2}(t_1, \dots, t_j)\boldsymbol{\eta} = \Omega^{1/2}(t_1, \dots, t_j)\boldsymbol{\eta} + o_p(1) \sim N(0, \Omega(t_1, \dots, t_j)) + o_p(1),$$

which is the joint distribution of  $\mathbf{G}(t_1), \dots, \mathbf{G}(t_j)$ . The stated result then follows from the fact that  $\hat{\mathbf{G}}$  is tight.  $\square$

**Lemma F.2.** For any  $c = [c_1, c_2]^\top$  with  $0 < c_1, c_2 < \infty$ , Let  $\hat{\mathbf{G}}(t; c) = c_1 \hat{\mathbf{G}}(t) - c_2 t^\top \hat{\mathbf{V}}t/2$  and  $\tilde{\mathbf{G}}(t; c) = c_1 \mathbf{G}(t) - c_2 t^\top \mathbf{V}t/2$ . Then  $\hat{\mathbf{G}}(\cdot; c) \xrightarrow{w} \tilde{\mathbf{G}}(\cdot; c)$  in  $\mathcal{L}^\infty(\mathcal{T}_1, \mathcal{T}_2, \dots)$ .

*Proof.* Follows from lemma F.1 combined with the consistency of  $\hat{\mathbf{V}}$ .  $\square$

**Lemma F.3.** For  $c$  as in lemma F.2 and any vector-valued function  $\psi$  for which for some  $c_\psi < \infty$ ,  $\sup_t \|\psi(t)\| / (\|t\|^{c_\psi} + 1) < \infty$ ,

$$\int \psi(t) \exp(c_1 \hat{\mathbf{G}}(t) - c_2 t^\top \hat{\mathbf{V}}t/2) dt \xrightarrow{d} \int \psi(t) \exp(c_1 \mathbf{G}(t) - c_2 t^\top \mathbf{V}t/2) dt.$$

*Proof.* By lemma F.2 and the Skorokhod representation theorem (Billingsley, 1999, theorem 6.7), there exist  $\hat{\mathbf{G}}^*, \tilde{\mathbf{G}}^*$ , such that  $\hat{\mathbf{G}}^*, \tilde{\mathbf{G}}^*$  have the same properties as  $\hat{\mathbf{G}}, \tilde{\mathbf{G}}$  and such that for all  $\omega$ :  $\hat{\mathbf{G}}^*(\cdot; \omega) \rightarrow \tilde{\mathbf{G}}^{**}(\cdot; \omega)$ . Since  $\|\psi(t)\| \exp(c_1 \hat{\mathbf{G}}(t) - c_2 t^\top \hat{\mathbf{V}}t/2)$  is a.s. integrable by the assumptions on  $\hat{H}, \psi$ , the stated result then follows from the dominated convergence theorem (Billingsley, 1995, theorem 16.4).  $\square$

**Lemma F.4.** If  $1 = o(\beta_n)$  then

$$\int t \exp(\beta_n^{4/3}(\hat{\mathbf{G}}(t) - t^\top \hat{\mathbf{V}}t/2)) dt / \int \exp(\beta_n^{4/3}(\hat{\mathbf{G}}(t) - t^\top \hat{\mathbf{V}}t/2)) dt \xrightarrow{d} \operatorname{argmax}_{t \in \mathfrak{R}^d} \tilde{\mathbf{G}}(t).$$

*Proof.* Let  $\hat{\boldsymbol{\psi}}_j(t) = t^j \exp(\beta_n^{4/3}(\hat{\mathbf{G}}(t) - t^\top \hat{\mathbf{V}}t/2))$  and  $\boldsymbol{\psi}_j(t) = t^j \exp(\beta_n^{4/3}(\mathbf{G}(t) - t^\top \mathbf{V}t/2))$ . Then  $\hat{\boldsymbol{\psi}}_0, \hat{\boldsymbol{\psi}}_1, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1 \in \mathbb{L}(\mathfrak{R}^d, \mathbb{B}, \mu)$  and  $(\hat{\boldsymbol{\psi}}_0, \hat{\boldsymbol{\psi}}_1) \xrightarrow{w} (\boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ . Repeat the arguments of lemma C.1 following (18).  $\square$

**Lemma F.5.** If  $\beta_n = o(1)$  then

$$\frac{1}{\beta_n^j} \int t^j \exp(\beta_n^{3/2} \hat{\mathbf{G}}(t/\beta_n)) \phi_{\hat{\mathbf{V}}}(t) dt = o_p(1) + \begin{cases} 1, & j = 0, \\ \sqrt{\beta_n} \int t \hat{\mathbf{G}}(t/\beta_n) \phi_{\hat{\mathbf{V}}}(t) dt, & j = 1. \end{cases}$$

*Proof.* We show the stated result for  $j = 1$ ; the case  $j = 0$  is similar. By lemma A.6, we have

$$\begin{aligned} & \left\| \beta_n^{-1} \int t \exp(\beta_n^{3/2} \hat{\mathbf{G}}(t/\beta_n)) \phi_{\hat{\mathbf{V}}}(t) dt - \sqrt{\beta_n} \int t \hat{\mathbf{G}}(t/\beta_n) \phi_{\hat{\mathbf{V}}}(t) dt \right\| \\ & \leq \beta_n C_V^{-1} \int \|t\| \exp(\sqrt{\beta_n} |\hat{\mathbf{G}}(t/\beta_n)| - t^\top \hat{\mathbf{V}} t/2) dt, \end{aligned}$$

which has the same distribution as  $\beta_n C_V^{-1} \int \|t\| \exp(|\hat{\mathbf{G}}(t)| - t^\top \hat{\mathbf{V}} t/2) dt = O_p(\beta_n) = o_p(1)$ .  $\square$

**Lemma F.6.** *If  $\beta_n = o(1)$  then*

$$\sqrt{\beta_n} \int t \hat{\mathbf{G}}(t/\beta_n) \phi_{\hat{\mathbf{V}}}(t) dt \xrightarrow{d} N(0, \mathcal{V}). \quad (27)$$

*Proof.* The left hand side in (27) has the same distribution as  $\int t \hat{\mathbf{G}}(t) \phi_{\hat{\mathbf{V}}}(t) dt$ , which by lemma F.1, the assumption that  $\hat{\mathbf{V}} = V + o_p(1)$ , and the conditions on  $\hat{\mathbf{H}}$ , converges in distribution to

$$\int t \mathbf{G}(t) \phi_V(t) dt, \quad (28)$$

Because  $\tilde{\mathbf{S}}_n \xrightarrow{w} \mathbf{G}$ , (28) is also the limit of  $\int t \tilde{\mathbf{S}}_n(t) \phi_V(t) dt$ , which lemma D.3 established to have a limiting  $N(0, \mathcal{V})$  distribution.  $\square$

## APPENDIX G. RESULTS SPECIFIC TO THE MAXIMUM SCORE ESTIMATOR

### G.1. $V, H, \mathcal{V}$ .

**Lemma G.1.**  $V = -2\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top p_a(\mathbf{z}_i^\top \theta_0, \mathbf{z}_i) f(\mathbf{z}_i^\top \theta_0 | \mathbf{z}_i)]$ .

*Proof.* Let  $\partial_\theta$  denote a partial derivative with respect to  $\theta$ . In view of the definition of  $V$ , consider

$$\begin{aligned} \partial_{\theta\theta^\top} \mathbb{E} \mathbf{g}_i(\theta) &= \partial_{\theta\theta^\top} \mathbb{E} [(2p(\mathbf{a}_i, \mathbf{z}_i) - 1) I(\mathbf{a}_i \leq \mathbf{z}_i^\top \theta)] = \partial_{\theta\theta^\top} \mathbb{E} \left[ \int_{-\infty}^{\mathbf{z}_i^\top \theta} (2p(a, \mathbf{z}_i) - 1) f(a | \mathbf{z}_i) da \right] \\ &= \partial_{\theta^\top} \mathbb{E} [\mathbf{z}_i (2p(\mathbf{z}_i^\top \theta, \mathbf{z}_i) - 1) f(\mathbf{z}_i^\top \theta | \mathbf{z}_i)] \\ &= \mathbb{E} [\mathbf{z}_i \mathbf{z}_i^\top \{2p_a(\mathbf{z}_i^\top \theta, \mathbf{z}_i) f(\mathbf{z}_i^\top \theta | \mathbf{z}_i) + (2p(\mathbf{z}_i^\top \theta, \mathbf{z}_i) - 1) f'(\mathbf{z}_i^\top \theta | \mathbf{z}_i)\}], \end{aligned}$$

which at  $\theta_0$  equals  $2\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top p_a(\mathbf{z}_i^\top \theta_0, \mathbf{z}_i) f(\mathbf{z}_i^\top \theta_0 | \mathbf{z}_i)] = -V$ .  $\square$

**Lemma G.2.** *If  $\mathbf{u}_i$  is the error term in the latent variable equation of Manski (1975) and Manski's conditional median assumption is satisfied, then  $V = 2\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top f_{u|z}(0, \mathbf{z}_i^\top \theta_0 | \mathbf{z}_i)]$ .*

*Proof.* Note first that  $p(a, z) = 1 - F_{u|az}(a - z^\top \theta_0 | a, z)$ , whose partial derivative with respect to  $a$  for  $a = z^\top \theta_0, z = z$  is  $f_{u|az}(0 | z^\top \theta_0, z)$  since  $F_{u|az}(0 | a, z) = 1/2$  for all values of  $a, z$  by the conditional median assumption.  $\square$

**Lemma G.3.**  $H(t, s) = \mathbb{E} [ |M(\mathbf{z}_i^\top t, \mathbf{z}_i^\top s, 0)| f(\mathbf{z}_i^\top \theta_0 | \mathbf{z}_i) ]$ .

*Proof.* Let

$$\mathcal{H}(t, s) = \lim_{\alpha \rightarrow \infty} \mathbb{E} \left[ \alpha \int_{\mathbf{z}_i^\top \theta_0}^{\mathbf{z}_i^\top \theta_0 + \min(\mathbf{z}_i^\top t, \mathbf{z}_i^\top s) / \alpha} f(a | \mathbf{z}_i) da \right] = \mathbb{E} [\min(\mathbf{z}_i^\top t, \mathbf{z}_i^\top s) f(\mathbf{z}_i^\top \theta_0 | \mathbf{z}_i)].$$

by the dominated convergence theorem. Thus, noting that  $(2\mathbf{y}_i - 1)^2 = 1$ , we have

$$\begin{aligned} H(t, s) &= \mathcal{H}(t, s) - \mathcal{H}(t, 0) - \mathcal{H}(0, s) + \mathcal{H}(0, 0) \\ &= \mathbb{E} [\{ \min(\mathbf{z}_i^\top t, \mathbf{z}_i^\top s) - \min(\mathbf{z}_i^\top t, 0) - \min(\mathbf{z}_i^\top s, 0) \} f(\mathbf{z}_i^\top \theta_0 | \mathbf{z}_i)] = \mathbb{E} [ |M(\mathbf{z}_i^\top t, \mathbf{z}_i^\top s, 0)| f(\mathbf{z}_i^\top \theta_0 | \mathbf{z}_i) ]. \end{aligned}$$

□

**Lemma G.4.** *If for some  $j = 1, \dots, d$ ,  $\mathbb{P}[\mathbf{z}_{ij} = 0] = 0$ , then (11) holds.*

*Proof.* We first establish the following five results for a generic variance matrix  $\Sigma \in \mathbb{R}^{d \times d}$ , where  $\Sigma_{\ell j}$  denotes the  $2 \times 2$  submatrix of  $\Sigma$  containing  $\sigma_{\ell\ell}, \sigma_{\ell j}, \sigma_{j\ell}, \sigma_{jj}$  with  $\sigma_{\ell j}$  the  $(\ell, j)$  element of  $\Sigma$ .

$$\int_{t_1}^{\infty} \int s_j \phi_{\Sigma_{1j}^{-1}}(s_1, s_j) ds_j ds_1 = \sigma_{1j} \phi(t_1 / \sigma_1) / \sigma_1, \quad (29)$$

$$\int_{t_1}^{\infty} s_1 \phi(s_1 / \sigma_1) ds_1 / \sigma_1 = \sigma_1 \phi(t_1 / \sigma_1), \quad (30)$$

$$\iiint \int_{t_1}^{\infty} t_\ell t_1 s_j \phi_{\Sigma_{1j}^{-1}}(s_1, s_j) \phi_{\Sigma_{1\ell}^{-1}}(t_1, t_\ell) ds_1 ds_j dt_1 dt_\ell = \sigma_{1j} \sigma_{1\ell} / 4\sigma_1 \sqrt{\pi}, \quad (31)$$

$$\iint t_\ell s_j \min(t_1, s_1) \phi_{\Sigma^{-1}}(s) \phi_{\Sigma^{-1}}(t) ds dt = \sigma_{1j} \sigma_{1\ell} / 2\sigma_1 \sqrt{\pi}, \quad (32)$$

$$\iint t s^\top \min(t_1, s_1) \phi_{\Sigma^{-1}}(s) \phi_{\Sigma^{-1}}(t) ds dt = \Sigma e_1 e_1^\top \Sigma / 2 \sqrt{\pi e_1^\top \Sigma e_1}, \quad (33)$$

where  $e_1$  is the first unit vector. The left hand side in (29) is

$$\frac{\sigma_{1j}}{\sigma_1^3} \int_{t_1}^{\infty} s_1 \phi(s_1 / \sigma_1) ds_1 = -\frac{\sigma_{1j}}{\sigma_1} \int_{t_1 / \sigma_1}^{\infty} \phi'(s_1) ds_1 = \frac{\sigma_{1j}}{\sigma_1} \phi(t_1 / \sigma_1).$$

Equality (30) is similar, but easier to establish, and the left hand side in (31) is by (29) equal to

$$\begin{aligned} \frac{\sigma_{1j}}{\sigma_1} \iint t_\ell t_1 \phi(t_1 / \sigma_1) \phi_{\Sigma_{1\ell}^{-1}}(t_1, t_\ell) dt_\ell dt_1 &= \frac{\sigma_{1j} \sigma_{1\ell}}{\sigma_1^4} \int t_1^2 \phi^2(t_1 / \sigma_1) dt_1 = \frac{\sigma_{1j} \sigma_{1\ell}}{\sqrt{2^3} \sigma_1} \int t_1^2 \phi^2(t_1 / \sqrt{2}) dt_1 \\ &= \frac{\sigma_{1j} \sigma_{1\ell}}{4\sigma_1 \sqrt{\pi}}. \end{aligned}$$

Equality (32) then follows by applying (31) twice and (33) by repeatedly using (32).

Now suppose without loss of generality that  $j = 1$  and denote the remaining elements by  $\tilde{z}_i$ . Let further

$$\mathbf{A}_i = \begin{bmatrix} z_{i1} & \tilde{z}_i^\top \\ 0 & I_{d-1} \end{bmatrix}, \quad \mathbf{V}_i = (\mathbf{A}_i^\top)^{-1} V \mathbf{A}_i^{-1}, \quad f_i = f(z_i^\top \theta_0 | z_i).$$

Then by substitution of  $\tilde{t} = \mathbf{A}_i t$  and  $\tilde{s} = \mathbf{A}_i s$ ,

$$\begin{aligned} \mathcal{V} &= \iint ts^\top H(t, s) \phi_V(t) \phi_V(s) dt ds = \mathbb{E} \left[ f_i \iint ts^\top |M(z_i^\top t, z_i^\top s, 0)| \phi_V(t) \phi_V(s) dt ds \right] \\ &= \mathbb{E} \left[ f_i \mathbf{A}_i^{-1} \iint \tilde{t} \tilde{s}^\top |M(\tilde{t}_1, \tilde{s}_1, 0)| \phi_{V_i}(t) \phi_{V_i}(s) d\tilde{t} d\tilde{s} (\mathbf{A}_i^\top)^{-1} \right] \\ &= \mathbb{E} \left[ f_i \mathbf{A}_i^{-1} \iint \tilde{t} \tilde{s}^\top \min(\tilde{t}_1, \tilde{s}_1) \phi_{V_i}(\tilde{t}) \phi_{V_i}(\tilde{s}) d\tilde{t} d\tilde{s} (\mathbf{A}_i^\top)^{-1} \right] = \frac{1}{2\sqrt{\pi}} \mathbb{E} \left[ f_i \frac{\mathbf{A}_i^{-1} \mathbf{V}_i^{-1} e_1 e_1^\top \mathbf{V}_i^{-1} (\mathbf{A}_i^\top)^{-1}}{\sqrt{e_1^\top \mathbf{V}_i^{-1} e_1}} \right] \\ &= \frac{1}{2\sqrt{\pi}} V^{-1} \mathbb{E} \left[ \frac{z_i z_i^\top}{\sqrt{z_i^\top V^{-1} z_i}} f(z_i^\top \theta_0 | z_i) \right] V^{-1}, \end{aligned}$$

where the penultimate inequality follows from (33) and the last from the fact that  $\mathbf{A}_i^{-1} \mathbf{V}_i^{-1} = V^{-1} \mathbf{A}_i^\top$  and that  $\mathbf{A}_i^\top e_1 = z_i$ .  $\square$

**G.2. Weak Convergence.** In this section, we show that the maximum score estimator satisfies assumption G. In view of equation (1), part i is satisfied. Let  $\mathcal{I}(\mathcal{A})$  be the Vapnik–Černovenkis (VČ) index of a function class  $\mathcal{A}$ . Let  $\mathcal{T} \subset \mathfrak{R}^d$  be an arbitrary compact set.

**Lemma G.5.** *Let  $\tilde{v}_n(\xi, t) = I(z^\top(\theta_0 + t/\tilde{\alpha}_n) \geq a)$  and  $\tilde{\mathcal{F}}_n = \{\tilde{v}_n(\cdot; t)\}_{t \in \mathcal{T}}$  and every  $n$ ,  $\mathcal{I}(\tilde{\mathcal{F}}_n) \leq d + 3$ .*

*Proof.* Since  $\tilde{\mathcal{F}}_n$  is a collection of indicator functions, its VČ index is equal to the VČ index of the collection of sets  $\{(z, a) : z^\top(\theta_0 + t/\tilde{\alpha}_n) \geq a\}$  with  $t$  ranging over  $\mathcal{T}$ . For every  $n$ , this collection is a subcollection of the sets  $\{(z, a) : z^\top t + as \geq 0\}$  with  $(t, s)$  ranging over  $\mathfrak{R}^{d+1}$ . The VČ index of the latter collection of sets is equal to  $d + 3$  by Kosorok (2008, lemma 9.12).  $\square$

**Lemma G.6.** *For all  $n$ ,  $\mathcal{I}(\mathcal{F}_n) \leq 2d + 5$ .*

*Proof.* Note that for every element  $v \in \mathcal{F}_n$  there is an element  $\tilde{v}_n \in \tilde{\mathcal{F}}_n$  such that  $v_n(\xi) = \sqrt{\tilde{\alpha}_n}(2y - 1)(\tilde{v}_n(\xi) - I(z^\top \theta_0 \geq a))$ . Therefore,  $\mathcal{I}(\mathcal{F}_n)$  is bounded by  $2\mathcal{I}(\tilde{\mathcal{F}}_n) - 1$  by Kosorok (2008, lemma 9.9). The conclusion then follows from lemma G.5.  $\square$

Let  $t_n(z) = \operatorname{arginf}_{t \in \mathcal{T}} z^\top(\theta_0 + t/\tilde{\alpha}_n)$  and  $s_n(z) = \operatorname{argsup}_{t \in \mathcal{T}} z^\top(\theta_0 + t/\tilde{\alpha}_n)$  and note that  $F_n$  defined in assumption G is  $F_n(\xi) = \sqrt{\tilde{\alpha}_n} I\{z^\top(\theta_0 + t_n(z)/\tilde{\alpha}_n) < a \leq z^\top(\theta_0 + s_n(z)/\tilde{\alpha}_n)\}$ .

**Lemma G.7.**  $\mathbb{E}F_{ni}^2 = O(1)$ .

*Proof.* By the law of iterated expectations,

$$\begin{aligned} \mathbb{E}F_{ni}^2 &= \tilde{\alpha}_n \mathbb{E} \left[ \mathbb{P} \left[ \mathbf{z}_i^\top (\theta_0 + t_n(\mathbf{z}_i) / \tilde{\alpha}_n) < \mathbf{a}_i \leq \mathbf{z}_i^\top (\theta_0 + s_n(\mathbf{z}_i) / \tilde{\alpha}_n) \mid \mathbf{z}_i \right] \right] \\ &\leq \mathbb{E} \left[ \sup_s f_{a|z}(s | \mathbf{z}_i) \|\mathbf{z}_i\| \|t_n(\mathbf{z}_i) + s_n(\mathbf{z}_i)\| \right]. \end{aligned}$$

The stated result then follows from  $t_n(\mathbf{z}_i), s_n(\mathbf{z}_i) \in \mathcal{T}$  and  $\mathbb{E}[\sup_s f_{a|z}(s | \mathbf{z}_i) \|\mathbf{z}_i\|] < \infty$ .  $\square$

**Lemma G.8.** For all  $\epsilon > 0$ ,  $\mathbb{E}[F_{ni}^2 I(\mathbf{F}_{ni} > \epsilon \sqrt{n})] = o(1)$ .

*Proof.* Follows from  $n/\tilde{\alpha}_n \rightarrow \infty$ .  $\square$

**Lemma G.9.** For every  $\epsilon_n \downarrow 0$ ,  $\sup_{\|s-t\| < \epsilon_n} \mathbb{E}[\tilde{\alpha}_n (\mathbf{g}_i(\theta_0 + t/\tilde{\alpha}_n) - \mathbf{g}_i(\theta_0 + s/\tilde{\alpha}_n))^2] = o(1)$ .

*Proof.* Follows from

$$\begin{aligned} \mathbb{E}[\tilde{\alpha}_n (\mathbf{g}_i(\theta_0 + t/\tilde{\alpha}_n) - \mathbf{g}_i(\theta_0 + s/\tilde{\alpha}_n))^2] &\leq \tilde{\alpha}_n \mathbb{E}[\mathbb{P}[\mathbf{z}_i^\top (\theta_0 + t/\tilde{\alpha}_n) < \mathbf{a}_i \leq \mathbf{z}_i^\top (\theta_0 + s/\tilde{\alpha}_n) \mid \mathbf{z}_i]] \\ &\quad + \tilde{\alpha}_n \mathbb{E}[\mathbb{P}[\mathbf{z}_i^\top (\theta_0 + s/\tilde{\alpha}_n) < \mathbf{a}_i \leq \mathbf{z}_i^\top (\theta_0 + t/\tilde{\alpha}_n) \mid \mathbf{z}_i]] \leq 2\mathbb{E} \left[ \sup_s f_{a|z}(s | \mathbf{z}_i) \|\mathbf{z}_i\| \right] \|t - s\|. \end{aligned} \quad \square$$

**Lemma G.10.**  $\tilde{\Sigma}_n \xrightarrow{w} \mathbf{G}$  in  $\mathcal{L}^\infty(\mathcal{T}_1, \mathcal{T}_2, \dots)$ , where  $\mathbf{G}$  is a Gaussian process with covariance kernel  $H$  such that  $H(t, t) = O(\|t\|)$  as  $\|t\| \rightarrow \infty$ .

*Proof.* Recall that

$$\tilde{\alpha}_n \mathbb{E}[\mathbf{g}_i(\theta_0 + t/\tilde{\alpha}_n) \mathbf{g}_i(\theta_0 + s/\tilde{\alpha}_n)] \rightarrow H(s, t) = \mathbb{E} [f(\mathbf{z}_i^\top \theta_0 | \mathbf{a}_i) | M(\mathbf{z}_i^\top t, \mathbf{z}_i^\top s, 0) |].$$

By lemmas G.8–G.9 and theorem A.1, showing that part v holds will complete the proof. Note however that by lemma G.6 each  $\mathcal{F}_n$  is a VČ class of which the VČ index is bounded by  $2d + 5$ .

Therefore, for  $0 < \epsilon < 1$ ,

$$\mathcal{N}(\epsilon \|\mathcal{F}\|_{\mathcal{Q}, 2, \mathcal{F}_n, L_2(\mathcal{Q})}) \leq C \epsilon^{2-2(2d+5)}$$

for some constant  $C$  that only depends on  $d$ ; see e.g. van der Vaart and Wellner (1996, theorem 2.6.7).

Hence, part v of assumption G is satisfied, and the conclusion follows.  $\square$

## APPENDIX H. PROOFS OF THEOREMS

Please note that the proofs use some additional notation that was introduced in appendix A.1.

*Proof of Theorem 1.* See lemma A.1.  $\square$

*Proof of Theorem 2.* See lemma G.10. □

*Proof of Theorem 3.* First (i). Note that by substitution of  $t = \alpha_n(\theta - \theta_0)$  we get

$$\begin{aligned} \sqrt{\frac{n}{\alpha_n}}(\hat{\theta} - \theta_0) &= \sqrt{\frac{n}{\alpha_n}} \frac{\int \pi(\theta)(\theta - \theta_0) \exp\{\alpha_n^2(\mathbf{S}_n(\theta) + Q(\theta))\} d\theta}{\int \pi(\theta) \exp\{\alpha_n^2(\mathbf{S}_n(\theta) + Q(\theta))\} d\theta} \\ &= \sqrt{\frac{n}{\alpha_n^3}} \frac{\int \pi_n(t)t \exp(c_\alpha^3 \tilde{\mathbf{S}}_n(t) + Q_n(t)) dt}{\int \pi_n(t) \exp(c_\alpha^3 \tilde{\mathbf{S}}_n(t) + Q_n(t)) dt} = \frac{1}{c_\alpha^3} \frac{\int t \exp(c_\alpha^3 \tilde{\mathbf{S}}_n(t)) \phi_V(t) dt}{\int \exp(c_\alpha^3 \tilde{\mathbf{S}}_n(t)) \phi_V(t) dt} + o_p(1), \end{aligned}$$

by lemma B.2. Multiply both sides by  $c_\alpha$ , then apply lemma A.1 and the continuous mapping theorem.

Now (ii). By substitution of  $t = \sqrt[3]{n}(\theta - \theta_0)$  we get

$$\sqrt[3]{n}(\hat{\theta} - \theta_0) = \frac{\int \pi_n(t)t \exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} dt}{\int \pi_n(t) \exp\{\beta_n^{4/3}(\tilde{\mathbf{S}}_n(t) + Q_n(t))\} dt} = \frac{\mathbb{N}}{\mathbb{D}},$$

Now, for  $\tilde{\Gamma} = \{t : \|t\| \leq \tilde{\gamma}\}$  for finite positive  $\tilde{\gamma}$ . Let further  $\mathbb{N}_{\tilde{\Gamma}} = \int_{\tilde{\Gamma}} \pi_n(t)t \exp\{\dots\} dt$ ,  $\mathbb{N}_{\tilde{\Gamma}^c} = \int_{\tilde{\Gamma}^c} \pi_n(t)t \exp\{\dots\} dt$ ,  $\mathbb{D}_{\tilde{\Gamma}} = \int_{\tilde{\Gamma}} \pi_n(t) \exp\{\dots\} dt$ ,  $\mathbb{D}_{\tilde{\Gamma}^c} = \int_{\tilde{\Gamma}^c} \pi_n(t) \exp\{\dots\} dt$ . Simple manipulations show that

$$\frac{\mathbb{N}}{\mathbb{D}} = \frac{\mathbb{N}_{\tilde{\Gamma}}}{\mathbb{D}_{\tilde{\Gamma}}} \left(1 - \frac{\mathbb{D}_{\tilde{\Gamma}^c}}{\mathbb{D}}\right) + \frac{\mathbb{N}_{\tilde{\Gamma}^c}}{\mathbb{D}}.$$

By lemma C.2 both  $\mathbb{D}_{\tilde{\Gamma}^c}/\mathbb{D}$  and  $\mathbb{N}_{\tilde{\Gamma}^c}/\mathbb{D}$  are  $o_p(1)$  and by lemma C.1,  $\mathbb{N}_{\tilde{\Gamma}}/\mathbb{D}_{\tilde{\Gamma}} \xrightarrow{d} \operatorname{argmax}_{t \in \tilde{\Gamma}} \tilde{\mathbf{G}}(t)$  for any  $\tilde{\gamma} > 0$ . Let  $\tilde{\gamma} \rightarrow \infty$ . □

*Proof of Theorem 4.* Combine lemmas D.7 and D.8. □

*Proof of Theorem 5.* For  $\mathcal{M}_{pqr}^*$  defined as in appendix E. By lemmas A.7 and E.2, the difference between  $\mathcal{B}_n$  and

$$\mathcal{B}_n^* = \frac{C_V}{\beta_n} \sum_{p=0}^q \frac{1}{p!} \int \pi_n(t)t (R_n(t))^p \phi_V(t) dt, \quad (34)$$

is  $o(\alpha_n^{-q} \beta_n^{-1})$ . We employ the Taylor expansions

$$R_n(t) = \sum_{\delta=1}^q \frac{D_{Q,\delta+2}(t)}{\alpha_n^\delta} + \frac{D_{Q\Delta n}^*(t)}{\alpha_n^q}, \quad \pi_n(t) = \sum_{\ell=0}^q \frac{D_{\pi\ell}(t)}{\alpha_n^\ell} + \frac{D_{\pi q n}^*(t)}{\alpha_n^q}. \quad (35)$$

Note that  $\sup_{t \in \Gamma_n \setminus \{0\}} (|D_{Q\Delta n}^*(t)|/\|t\|^\Delta + |D_{\pi q n}^*(t)|/\|t\|^q) = o(1)$  by the continuity of the  $\Delta$ -th derivative of  $Q$  and the  $q$ -th derivative of  $\pi$  at  $\theta_0$ , and the fact that  $\gamma_n = o(\alpha_n)$ . Hence, by lemmas A.7 and A.9 it follows that any expressions in an expansion of (34) using (35) involving the remainder terms in the Taylor expansions of (35) are  $o(\alpha_n^{-q} \beta_n^{-1})$ .

Omitting the remainder terms in the definition of  $\mathcal{B}_n^*$  yields

$$\mathcal{B}_n^{**} = \frac{C_V}{\beta_n} \sum_{p=0}^q \frac{1}{p!} \int \left( \sum_{\ell=0}^q \frac{D_{\pi\ell}(t)}{\alpha_n^\ell} \right) \left( \sum_{\delta=1}^q \frac{D_{Q,\delta+2}(t)}{\alpha_n^\delta} \right)^p t\phi_V(t) dt.$$

Using lemma E.3, we obtain

$$\begin{aligned} \mathcal{B}_n^{**} &= \frac{C_V}{\beta_n} \sum_{p=0}^q \int \left( \sum_{\ell=0}^q \frac{D_{\pi\ell}(t)}{\alpha_n^\ell} \right) \sum_{\mathcal{M}_{pq}} \prod_{\delta=1}^q \left\{ \left( \frac{D_{Q,\delta+2}(t)}{\alpha_n^\delta} \right)^{m_\delta} \frac{1}{m_\delta!} \right\} t\phi_V(t) dt \\ &= \frac{C_V}{\beta_n} \sum_{p=0}^q \sum_{\mathcal{M}_{pq}} \sum_{\ell=0}^q \frac{1}{\alpha_n^{\ell+\sum_{\delta=1}^q m_\delta}} \int D_{\pi\ell}(t) \left( \prod_{\delta=1}^q \frac{D_{Q,\delta+2}^{m_\delta}(t)}{m_\delta!} \right) t\phi_V(t) dt \\ &= \frac{C_V}{\beta_n} \sum_{\ell=0}^q \sum_{r=0}^{q-\ell} \frac{b_{qr\ell}}{\alpha_n^{r+\ell}} + O(\alpha_n^{-q-1}) = \frac{1}{\beta_n} \sum_{\tau=0}^q \frac{b_{q\tau}^*}{\alpha_n^\tau} + o(\alpha_n^{-q}\beta_n^{-1}), \end{aligned}$$

where

$$b_{qr\ell} = \sum_{p=0}^q \sum_{\mathcal{M}_{pqr}^*} \int D_{\pi\ell}(t) \left( \prod_{\delta=1}^q \frac{D_{Q,\delta+2}^{m_\delta}(t)}{m_\delta!} \right) t\phi_V(t) dt,$$

and  $b_{q\tau}^* = C_V \sum_{r=0}^\tau b_{qr,\tau-r}$ . Finally,  $b_{q\tau}^* = 0$  for even-valued  $\tau$  because  $b_{qr\ell} = 0$  whenever  $r$  and  $\ell$  add up to an even number. Indeed, then there is at least one  $t_s$  taken to an odd power in (H), whose integral with respect to the normal density equals zero.  $\square$

**Proof of Theorem 6.** Given theorems 4 and 5, we only need to show that (i)  $b_{q0}^* = 0$  and (ii)  $b_{q1}^* = C_V \int (\pi_0 D_{Q3}(t) + D_{\pi 1}(t)) t\phi_V(t) dt$ . Both results follow from the definition of  $b_{q\tau}^*$  in the proof of theorem 5.  $\square$

**Proof of Theorem 7.** Let  $\mathcal{N}_n, \mathcal{D}_{Jn}, \mathcal{B}_{Jn}, \mathcal{B}_{Jn}^*$  be defined as  $\mathcal{N}_n, \mathcal{D}_n, \mathcal{B}_n, \mathcal{B}_n^*$  with  $\hat{\pi} = \hat{\pi}_J$  and  $\hat{\pi}_n = \hat{\pi}(\theta_0 + t/\alpha_n)$  replacing  $\pi, \pi_n$ . By lemma E.4 we have  $\mathcal{N}_n - \mathcal{B}_{Jn} \xrightarrow{d} N(0, C_V^2 \mathcal{V}_N)$ ,  $\mathcal{D}_{Jn} = \pi_0 C_V + o_p(1)$ , following the same steps as in lemmas D.7 and D.8. It therefore suffices to consider  $\mathcal{B}_{Jn}$ .

The first few lines of the proof of theorem 5 establish that the difference between  $\mathcal{B}_n$  and  $\mathcal{B}_n^*$  is  $o(\alpha_n^{-q}\beta_n^{-1})$ . The argument there and those in lemmas A.7 and E.2 used therein do not depend on  $\pi$  being nonrandom, so  $\mathcal{B}_{Jn} - \mathcal{B}_{Jn}^* = o_p(\alpha_n^{-q}\beta_n^{-1})$ .

Expanding  $R_n(t)$  and ignoring the remainder term like in the proof of theorem 5, we get

$$\mathcal{B}_{Jn}^* = \frac{C_V}{\beta_n} \sum_{p=0}^q \int \hat{\pi}_n(t) \left( \sum_{\delta=1}^q \frac{D_{Q,\delta+2}(t)}{\alpha_n^\delta} \right)^p t\phi_V(t) dt + o_p(\alpha_n^{-q}\beta_n^{-1}). \quad (36)$$

Part (i) of the theorem then follows by applying lemma E.6 to the first right hand side term in (36). For (ii) the first right hand side term in (36) equals

$$\begin{aligned} \frac{C_V}{\alpha_n^3 \beta_n} \int \hat{\pi}_n(t) (D_{Q5}(t) + D_{Q3}(t)D_{Q4}(t)/2 + D_{Q3}^3(t)/6) t \phi_V(t) dt \\ + \frac{C_V}{\alpha_n^2 \beta_n} \int \hat{\pi}_n(t) (D_{Q4}(t) + D_{Q3}^2(t)) t \phi_V(t) dt. \end{aligned} \quad (37)$$

By lemma E.4 the first term in (37) is

$$\frac{C_V \pi_0}{\alpha_n^3 \beta_n} \int (D_{Q5}(t) + D_{Q3}(t)D_{Q4}(t)/2 + D_{Q3}^3(t)/6) t \phi_V(t) dt + o_p(\alpha_n^{-3} \beta_n^{-1}) = O_p(\alpha_n^{-3} \beta_n^{-1}). \quad (38)$$

Using the same arguments as in the proof of theorem 5 about the remainder terms in an expansion of the prior, the second term in (37) is

$$\begin{aligned} \frac{C_V \pi_0}{\alpha_n^2 \beta_n} \int (D_{Q4}(t) + D_{Q3}^2(t)) t \phi_V(t) dt + \frac{C_V}{\alpha_n^3 \beta_n} \int (D_{Q4}(t) + D_{Q3}^2(t)) \phi_V(t) t t^\top dt \pi'(\theta_0) \\ + o_p(\alpha_n^{-3} \beta_n). \end{aligned} \quad (39)$$

The first term in (39) is zero because the normal distribution is even. The second term in (38) is  $O(\alpha_n^{-3} \beta_n^{-1})$ , so part (ii) of the theorem is satisfied with  $\mathcal{B}_J$  equal to the sum of the first term in (38) and the second term in (39).  $\square$

**Proof of Theorem 8.** First part (ii). If  $c_\alpha \geq 1$ , then  $\kappa_n = \sqrt[3]{n}$  and by lemma F.3,

$$\mathbf{\Psi}_n = \int t \exp(c_\alpha^4 \hat{\mathbf{G}}(t) - c_\alpha^4 t^\top \hat{\mathbf{V}} t / 2) dt / \int \exp(c_\alpha^4 \hat{\mathbf{G}}(t) - c_\alpha^4 t^\top \hat{\mathbf{V}} t / 2) dt$$

has the limiting distribution given in (8). For  $c_\alpha < 1$ , we have likewise  $\kappa_n = \sqrt[3]{n}/c_\alpha$  and

$$c_\alpha \mathbf{\Psi}_n \xrightarrow{d} c_\alpha \frac{\int t \exp(c_\alpha^{9/2} \mathbf{G}(t) - c_\alpha^6 t^\top \mathbf{V} t / 2) dt}{\int \exp(c_\alpha^{9/2} \mathbf{G}(t) - c_\alpha^6 t^\top \mathbf{V} t / 2) dt} = \frac{1}{c_\alpha^2} \frac{\int t \exp(c_\alpha^{9/2} \mathbf{G}(t/c_\alpha^3)) \phi_V(t) dt}{\int \exp(c_\alpha^{9/2} \mathbf{G}(t/c_\alpha^3)) \phi_V(t) dt}, \quad (40)$$

by substitution of  $\tilde{t} = c_\alpha^3 t$  and replacing  $\tilde{t}$  with  $t$ . The right hand side in (40) has the same distribution as the limit distribution in (6) since  $c_\alpha^{9/2} \mathbf{G}(\cdot/c_\alpha^3)$  has the same properties as  $c_\alpha^3 \mathbf{G}(\cdot)$ .

Part (iii) is established in lemma F.4.

Finally, part (i). By substitution of  $\tilde{t} = \beta_n t$  and replacing  $\tilde{t}$  by  $t$  we obtain

$$\mathbf{\Psi}_n = \frac{1}{\beta_n} \frac{\int t \exp(\beta_n^{3/2} \hat{\mathbf{G}}(t/\beta_n)) \phi_{\hat{\mathbf{V}}}(t) dt}{\int \exp(\beta_n^{3/2} \hat{\mathbf{G}}(t/\beta_n)) \phi_{\hat{\mathbf{V}}}(t) dt}.$$

Apply lemma F.5 followed by lemma F.6.  $\square$

## APPENDIX I. COMPUTATION

**I.1. Maximum Score Estimates.** We now describe the algorithm used to compute our estimates for the maximum score case. A computer program written in C is available upon request.

Let  $\theta_{-j}, z_{i,-j}$  be respectively  $\theta, z_i$  without its  $j$ -th element. Let  $\mathbf{B}_{ij} = \mathbf{B}_{ij}(\theta_{-j}) = (\mathbf{a}_i - \mathbf{z}_{i,-j}^\top \theta_{-j}) / z_{ij}$  if  $z_{ij} \neq 0$  and be arbitrarily defined if  $z_{ij} = 0$ . Let further  $\mathcal{S}_{nj}(\theta_j) = \sum_{i=1}^n (2\mathbf{y}_i - 1) (I(z_{ij} > 0)I(\theta_j \geq \mathbf{B}_{ij}) + I(z_{ij} < 0)I(\theta_j \leq \mathbf{B}_{ij})) / \sqrt[3]{n}$ . Let  $\mathcal{A}(\theta_j | \theta_{-j}) = \int_{-\infty}^{\theta_j} \pi_j(\theta_j | \theta_{-j}) \exp(\mathcal{S}_{nj}(\theta_j)) d\theta_j$ .

- (1) Compute and renormalize probit estimates.
- (2) Set  $j = 1$ .
- (3) Compute  $\mathbf{B}_{ij}$  for all  $i$  for which  $z_{ij} \neq 0$  and sort all  $n^*$  of them in ascending order, i.e.  $\mathbf{B}_{1j} \leq \mathbf{B}_{2j} \leq \dots \leq \mathbf{B}_{n^*j}$ .
- (4) Compute  $\mathcal{S}_{nj}(-\infty)$ .
- (5) Note that for all  $\theta_j \leq \mathbf{B}_{1j}$ ,  $\mathcal{I}_{nj}(\theta_j) = \exp(\mathcal{S}_{nj}(-\infty))\Pi_j(\theta_j | \theta_{-j})$ , where  $\Pi_j(\theta_j | \theta_{-j})$  is the integrated conditional prior.
- (6) For all  $\theta_j \in (\mathbf{B}_{1j}, \mathbf{B}_{2j}]$ ,  $\mathcal{I}_{nj}(\theta_j) = \mathcal{I}_{nj}(\mathbf{B}_{1j}) + \exp\{\mathcal{S}_{nj}((\mathbf{B}_{1j} + \mathbf{B}_{2j})/2)\} (\Pi_j(\theta_j | \theta_{-j}) - \Pi_j(\mathbf{B}_{1j} | \theta_{-j}))$ .
- (7) Repeat this process until  $\mathcal{I}_{nj}(\theta_j)$  is determined for all values of  $\theta_j$ .
- (8) Set  $\mathcal{I}_{nj}^*(\theta_j) = \mathcal{I}_{nj}(\theta_j) / \mathcal{I}_{nj}(\infty)$ , which is a distribution function.
- (9) Draw a random number  $r$  from a uniform distribution.
- (10) Compute  $i^* = \max\{i : \mathcal{I}_{nj}^*(\mathbf{B}_{ij}) \leq r\}$ .
- (11) Set  $\theta_j = \theta_j(r) = \Pi^{-1}(\Pi(\mathbf{B}_{i^*j}) + \mathcal{I}_{nj}(\infty)r + \mathcal{I}_{nj}(\mathbf{B}_{i^*j}) | \theta_{-j})$ .
- (12) Increase  $j$ . If  $j > d$  set  $j = 1$ . Go to step 3.
- (13) Repeat for a large number of times to burn in.
- (14) Then start using them as actual draws.

**I.2. Inference.** We now describe how one can draw  $\mathbb{J}$  random numbers with the same distribution as (15). Suppose that consistent estimates  $\hat{\mathbf{V}}, \hat{\mathbf{H}}$  are available, where  $\hat{\mathbf{H}}$  is moreover a positive semi-definite covariance kernel.<sup>19</sup> One can then draw random numbers  $\mathbf{t}_1, \dots, \mathbf{t}_{T^*}$  from the multivariate normal  $(\phi_{\beta_{n2}\hat{\mathbf{V}}})$  and define

$$\hat{\boldsymbol{\zeta}}_j = \frac{\sum_{s=1}^{T^*} \mathbf{t}_s \exp(\beta_{n1} \mathbf{G}_{js})}{\sum_{s=1}^{T^*} \exp(\beta_{n1} \mathbf{G}_{js})}, \quad (41)$$

where  $\mathbf{G}_{j1}, \dots, \mathbf{G}_{jT^*}$  are drawn independently across  $j$  and have for all given  $j = 1, \dots, \mathbb{J}$  a joint normal distribution with  $\text{Cov}[\mathbf{G}_{js}, \mathbf{G}_{js^*}] = \hat{\mathbf{H}}(\mathbf{t}_s, \mathbf{t}_{s^*})$ , for all  $s, s^*$ . In the limit ( $n$  and  $T^*$ ), each  $\hat{\boldsymbol{\zeta}}_j$  has

<sup>19</sup>In practice one only needs to ensure that the  $T^* \times T^*$  matrix with  $(i, j)$ -element  $\hat{\mathbf{H}}(\mathbf{t}_i, \mathbf{t}_j)$  is positive semidefinite.

the same distribution as  $\Psi_n$  defined in (15). Note that the  $t_s$  draws are made only once. The draws  $\hat{\zeta}_1, \dots, \hat{\zeta}_J$  can then be used to construct confidence intervals.

**I.3. Derivatives of  $Q$ .** For the specific case of the maximum score estimator,  $Q(\theta) = \mathbb{E}[(2y_i - 1)(I(\mathbf{a}_i \leq \mathbf{z}_i^\top \theta) - I(\mathbf{a}_i \leq \mathbf{z}_i^\top \theta_0))]$ . If derivatives of  $Q$  must be estimated, we recommend using the desired derivative of

$$\hat{Q}'(\theta) = \frac{1}{n\tilde{b}} \sum_{i=1}^n \tilde{k}\left(\frac{\mathbf{z}_i^\top \theta - \mathbf{a}_i}{\tilde{b}}\right) (2y_i - 1) \mathbf{z}_i,$$

where  $\tilde{k}$  is a kernel and  $\tilde{b}$  a bandwidth.