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

Sung Jae Jun† Joris Pinkse ‡ Yuanyuan Wan§
Center for the Study of Auctions, Procurements, and Competition Policy
Department of Economics
The Pennsylvania State University

August 2009

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). Scaling the nonregular objective functions by a sample–size (n)–dependent input parameter α_n^2, we establish that the Laplace estimation procedure, given sufficient smoothness, 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 α_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 α_n translates to a faster convergence rate for our estimator $\hat{\boldsymbol{\theta}}$, albeit that the convergence rate of $\hat{\boldsymbol{\theta}}$ 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 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 Jeremy Fox, Bo Honoré, Joel Horowitz, Roger Koenker, Oliver Linton, Haiqing Xu, and Vicky Zinde–Walsh for helpful suggestions.
†sjun@psu.edu (corresponding author), 608 Kern Graduate Building, University Park, PA 16802
†joris@psu.edu
†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). Scaling the non-regular objective functions by a sample–size (n)–dependent input parameter α_n^2, we establish that the Laplace estimation procedure, given sufficient smoothness, 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 α_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 α_n translates to a faster convergence rate for our estimator $\hat{\boldsymbol{\theta}}$, albeit that the convergence rate of $\hat{\boldsymbol{\theta}}$ 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 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 θ_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) showed 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.[1] Chernozhukov and Hong (2003) derived 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

[1] 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 α_n^2 that scales the objective function.

If α_n diverges at a rate no slower than $\sqrt[3]{n}$ then our estimator $\hat{\boldsymbol{\theta}}$ is $\sqrt[3]{n}$–convergent. If α_n diverges faster than $\sqrt[3]{n}$ then $\hat{\boldsymbol{\theta}}$ 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 α_n increases at the rate of $\sqrt[3]{n}$, the limiting distribution of $\hat{\boldsymbol{\theta}}$ is characterized by the ratio of two integrals of a certain Gaussian process. The third type of limiting distribution, namely the normal, arises when α_n increases at a rate slower than $\sqrt[3]{n}$. Indeed, we show that subject to sufficient smoothness the convergence rate of $\hat{\boldsymbol{\theta}}$ can be as good as $\sqrt{n/\alpha_n}$; since α_n must increase to infinity with n, a convergence rate of $\sqrt{n}$ is not achievable. If α_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 α_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.[2] So inference is uniform in the choice of α_n; this result is at the heart of this paper.

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 α_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,

[2]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) derived 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 α_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 subtract an estimate of the bias and to use a special prior. Indeed, we show that using the equivalent of 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.[3]

For consistency, α_n must increase to infinity with the sample size; for consistency the rate at which α_n increases is immaterial. If one were to let α_n decrease to zero, however, then our estimator would converge 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 α_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, MSE) as a leading example. We emphasize the binary choice version of the MSE, albeit that the discussion applies equally to the ordered response and multinomial choice cases (Lee, 1992; Manski, 1975).

The 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 nonstandard limiting distribution; we are aware of only a few empirical uses of the MSE and its generalization, including Bajari, Fox, and Ryan (2008); Bajari and Fox (2009); Fox (2007, 2009).

Horowitz (1992) has shown that by replacing the maximum score objective function with a smoothed version thereof, a rate arbitrarily close to $\sqrt{n}$ is attainable with a limiting normal distribution, subject to sufficient smoothness. Subsequently de Jong and Woutersen (2007) and Kotlyarova and Zinde-Walsh (2009) have extended the Horowitz results. Kordas (2006) has extended the

[3]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).

smoothed maximum score estimator (Horowitz, 1992, SMS) to allow for quantiles other than the median and his procedure was implemented empirically by Belluzo (2004); using quantiles other than the median is also possible with the original MSE and the estimator proposed in this paper. Although the best attainable convergence rate of the SMS is the same as ours under similar conditions, our estimator is entirely different from Horowitz's, our estimator is not specific to the maximum score case, due to the nonconcavity of the objective function the SMS estimator is more difficult to compute than ours, and no uniform inference procedures are available for it.

One problem with the SMS is that the choice of bandwidth is determined by the unknown degree of smoothness. If the degree of smoothness used to determine the bandwidth differs from the true degree of smoothness then the convergence rate of the SMS is suboptimal. Indeed, Pollard (1993) has shown that $\sqrt[3]{n}$ is the best rate that can be achieved if only the smoothness conditions for the MSE are satisfied and that the SMS will then have a bias term which vanishes more slowly than $\sqrt[3]{n}$. The same problem arises for our estimator and the choice of α_n. Kotlyarova and Zinde-Walsh (2006, KZ) proposed an estimation procedure which automatically adapts to the unknown degree of smoothness. The advantage of our estimator over the SMS estimator, when both are combined with a KZ–like procedure, is the uniformity of our inference method across input parameter–values including the $\sqrt[3]{n}$–convergence case. For example, with a KZ–like estimation method, our inference procedure (with minor adaptations) could accommodate the possibility that only the smoothness conditions for the MSE are satisfied. It is difficult to see how — absent a uniform inference procedure — one could accomplish this for the SMS.

Horowitz (2002) established that the bootstrap (Efron and Tibshirani, 1997) offers an asymptotic refinement for the SMS estimator. In order for such a refinement to obtain, the rate at which the bandwidth tends to zero is different from the one resulting in the optimal convergence rate of the estimator.[4] We have made a preliminary investigation on the bootstrap in our context, but intend to establish rigorous results for a procedure providing asymptotic refinements in the future. On the basis of our preliminary work, we conclude that the bootstrap is inconsistent if α_n increases no slower than $\sqrt[3]{n}$ for much the same reasons that the bootstrap is inconsistent for the regular maximum score estimator (Abrevaya and Huang, 2005).[5] We further conclude that if α_n increases more slowly than $\sqrt[3]{n}$, then the bootstrap will be consistent. Moreover, asymptotic refinements will

[4]This is typical for nonparametric estimators, but an unattractive consequence is that the ratio of the width of the 'regular' confidence interval to that of the bootstrap confidence interval decreases to zero as the sample size tends to infinity.

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

obtain provided that there is sufficient smoothness that is not exploited to optimize the convergence rate of the estimator.

The bootstrap and the uniform inference procedure have different goals: with the bootstrap the actual coverage probability of the confidence interval converges faster to the nominal one if α_n increases sufficiently slowly and there is 'surplus' smoothness; the uniform inference procedure provides robustness to the choice of α_n. These two goals are mutually exclusive in the current context (and for KP estimators generally) since the bootstrap is inconsistent if α_n increases no slower than $\sqrt[3]{n}$. We focus on the uniform inference procedure here because, in view of the two discontinuities in the limiting distribution as a function of the rate of increase of α_n, we believe it is the more serious of the two problems these procedures address.

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 α_n resulting in less bias, but more variance. Further, the asymptotic bias can be corrected using Jeffreys' prior but if α_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 α_n.

There are issues of potential interest that are not studied in this paper in addition to those, like the bootstrap, that were discussed above. First, one could look at statistics other than the posterior mean, as do Chernozhukov and Hong (2003). Further, using our methodology for the least median of squares (LMS) estimator of Rousseeuw (1984) requires a nuisance parameter problem to be addressed and its breakdown point to be determined; see Zinde-Walsh (2002) for an adaptation of the SMS idea to the LMS case.

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.

2. Estimation Method

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

$$\boldsymbol{g_i}(\theta) = (2\boldsymbol{y_i} - 1)\big(I(\boldsymbol{a_i} \leq \boldsymbol{z_i^\intercal}\theta) - I(\boldsymbol{a_i} \leq \boldsymbol{z_i^\intercal}\theta_0)\big), \tag{1}$$

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

We consider Laplace–type estimators of the form

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

where $\{\alpha_n\}$ is some sequence for which $\alpha_n \to \infty$ as $n \to \infty$;[6] we will call π a *prior,* even though for most of our results we do not require it to be nonnegative. If the prior is nonnegative everywhere, then $\hat{\boldsymbol{\theta}}$ can be interpreted as the mean of a *posterior* distribution; we will use the term 'posterior' regardless of whether the prior is nonnegative. If $\boldsymbol{L_n}$ 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 θ_0 is a unique maximizer of $Q(\theta) = \mathbb{E}[\boldsymbol{g_i}(\theta)]$, 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 $\hat{\boldsymbol{\theta}}$. If α_n increases faster than $\sqrt[3]{n}$ then it turns out that $\hat{\boldsymbol{\theta}}$ is asymptotically equivalent to the estimator maximizing $\boldsymbol{L_n}$, for which Kim and Pollard (1990) derived limit results.[7] Although we emphasize the case in which α_n increases no faster than $\sqrt[3]{n}$, we provide results for α_n that increase faster, also. In subsequent sections we show that the best achievable convergence rate (given sufficient smoothness) 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 SMS estimator.

[6] In Chernozhukov and Hong (2003) $\alpha_n = \sqrt{n}$, but its choice does not affect the limiting distribution provided that α_n diverges to infinity.

[7] For the case in which $\boldsymbol{L_n}$ is the maximum score objective function, if α_n increases faster than $\sqrt{n}$ (not $\sqrt[3]{n}$), $\hat{\boldsymbol{\theta}}$ 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.

2.2. **Intuition for our results.** We now provide some intuition for our results when α_n increases no faster than $\sqrt[3]{n}$, but please note that our results also cover the case in which α_n increases faster than that. Let $\boldsymbol{S_n}(\theta) = n^{-1}\sum_{i=1}^{n}\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta)$ with $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta) = \boldsymbol{g_i}(\theta) - Q(\theta)$. Then (2) can be rewritten as

$$\hat{\boldsymbol{\theta}} = \frac{\int \theta\pi(\theta)\exp\big(\alpha_n^2\boldsymbol{S_n}(\theta) + \alpha_n^2 Q(\theta)\big)d\theta}{\int \pi(\theta)\exp\big(\alpha_n^2\boldsymbol{S_n}(\theta) + \alpha_n^2 Q(\theta)\big)d\theta}. \tag{3}$$

Let $\beta_n = \sqrt{\alpha_n^3/n}$. By applying the substitution $t = \alpha_n(\theta - \theta_0)$ to (3) we obtain

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

where $\pi_n(t) = \pi(\theta_0 + t/\alpha_n)$, $Q_n(t) = \alpha_n^2 Q(\theta_0 + t/\alpha_n)$ and $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) = \sqrt{n\alpha_n}\boldsymbol{S_n}(\theta_0 + t/\alpha_n)$.

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

$$\frac{1}{\beta_n}\frac{\int t\exp\big(\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) - t^{\mathsf{T}}Vt/2\big)dt}{\int \exp\big(\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) - t^{\mathsf{T}}Vt/2\big)dt} = \frac{1}{\beta_n}\frac{\int t\exp\big(\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big)\phi_V(t)dt}{\int \exp\big(\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big)\phi_V(t)dt}, \tag{5}$$

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

We establish in an appendix that $\tilde{\boldsymbol{S}}_{\boldsymbol{n}} \xrightarrow{w} \mathbb{G}$, where $\mathbb{G}$ 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

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

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

$$\int t\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\phi_V(t)dt,^{8}$$

which has a limiting normal distribution since $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ 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 $\mathscr{B}_n$. This bias term arises from the approximation of $Q_n(t)$ by $-t^{\mathsf{T}}Vt/2$. It is asymptotically negligble if α_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 $\sqrt{n}$.

[8] This approximation uses $\exp\big(\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big) \approx 1 + \beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)$ and the fact that $\int \phi_V(t)dt = 1$, $\int t\phi_V(t)dt = 0$.

So, letting α_n increase at a rate slower than $\sqrt[3]{n}$ implicitly smoothes out the discontinuities in $\boldsymbol{L_n}$, much like the SMS 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 case as an example to motivate the assumptions.

3.1. Assumptions.

The first of our assumptions is standard in the literature and is also found in Horowitz (1992).

Assumption A. *θ_0 is in the interior of some compact set Θ.*

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 $p(a, z) = \mathbb{P}[\boldsymbol{y} = 1 | \boldsymbol{a} = a, \boldsymbol{z} = z]$ and $f(a|z)$ denote the conditional density of $\boldsymbol{a}$ at a given $\boldsymbol{z} = z$. For his maximum score estimator Manski (1985) requires that the support of $\boldsymbol{x}$ is not contained in any proper linear subspace of $\Re^d$, that $0 < p(a, z) < 1$ for almost all a, z, and that for almost all z, $f(a|z) > 0$ for all $a \in \Re$, 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. *$\pi(\theta_0) \neq 0$, π is bounded in absolute value on Θ by $\bar{\pi}$, zero outside of Θ, 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 Θ. Further, for some $q \geq 0$, Q is $\Delta = q + 2$ times continuously differentiable at θ_0; π is q times continuously differentiable at θ_0.*[9] *$V = -Q''(\theta_0)$ is positive definite.*

[9] If $q = 0$, π is merely assumed continuous at θ_0.

Kim and Pollard (1990) derived the $\sqrt[3]{n}$–limiting distribution of the MSE assuming $\Delta = 2$. The degree of smoothness of Q depends on that of $p(\cdot, z)$ and $f(\cdot|z)$.

For the maximum score case, Horowitz (1992) obtained 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 $a = z^\intercal\theta_0$ 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 $a = z^\intercal\theta_0$, (iv) $\mathbb{E}\|\boldsymbol{z_i}\|^4 < \infty$, (v) V is positive definite.[10] To achieve the same convergence rate we need $\Delta = 3$ (see assumption H below), which implies the existence of three moments on $\boldsymbol{z_i}$, two partial derivatives of p with respect to a at $z^\intercal\theta_0$, and two derivatives of $f(\cdot|z)$ at $z^\intercal\theta_0$. So the conditions are different from but similar to those in Horowitz (1992); neither set of assumptions implies the other and neither estimator yields a better convergence rate under the conditions of the other. The comparison between the conditions necessary for the two estimators to obtain a certain convergence rate is similar under additional smoothness. Since we accommodate estimation problems other than maximum score, our conditions are less primitive. Finally, lemma G.1 shows that for the maximum score case $V = -2\mathbb{E}\left[\boldsymbol{z_i}\boldsymbol{z_i}^\intercal p_a(\boldsymbol{z_i}^\intercal\theta_0, \boldsymbol{z_i}) f(\boldsymbol{z_i}^\intercal\theta_0|\boldsymbol{z_i})\right]$, which by lemma G.2 is positive definite under weak conditions.

Assumption E. $\mathbb{E}\left[\sup_{\theta\in\Re^d} |\boldsymbol{g_i}(\theta)|\right] < \infty$ *and for some* $\iota, \iota^* > 0$ *with* $\iota + \iota^* > 1$, *some function* ν *for which* $\int \|t\|^{2+\iota}\nu(t)\phi_V(t)dt < \infty$, *and all* $t \in \Re^d$, $\lim_{\alpha\to\infty} \alpha^{\iota^*}\mathbb{E}\left|\tilde{\boldsymbol{g}}_{\boldsymbol{i}}(\theta_0 + t/\alpha)\right|^{2+\iota} \leq \nu(t)$.

Assumption E is trivially satisfied for the maximum score case 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, $H(t, s) = \lim_{\alpha\to\infty} \alpha\mathbb{E}[\boldsymbol{g_i}(\theta_0 + t/\alpha)\boldsymbol{g_i}(\theta_0 + s/\alpha)]$ *exists and is finite.*[11]

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

[10] Horowitz's conditions are phrased differently and use the conditional distribution function $F_{u|a,z}$ of 'latent variable equation' (Manski, 1975) errors $\boldsymbol{u_i}$ given regressors. Then, $p(a, z) = 1 - F_{\boldsymbol{u}|\boldsymbol{a},\boldsymbol{z}}(a - z^\intercal\theta_0|a, z)$ so that the degree of smoothness of $p(\cdot, z)$ corresponds to that of $F_{\boldsymbol{u}|\boldsymbol{a},\boldsymbol{z}}(\cdot|\cdot, z)$. Therefore, Δ–times differentiability of Q corresponds to $(\Delta - 1)$–times differentiability of both $f(\cdot|z)$ and $F_{\boldsymbol{u}|\boldsymbol{a},\boldsymbol{z}}(\cdot|\cdot, z)$. See also assumptions 8 and 9 in Horowitz (1992).

[11] Note that this implies that $\limsup_{\|t\|\to\infty} H(t, t)/\|t\| < \infty$.

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

(i) *For all ξ in the support Ξ of $\boldsymbol{\xi_i}$, $g(\xi,\theta)$ is right– (or left–) continuous at θ_0.*

(ii) *There exists an envelope function F_n such that for all $\xi \in \Xi$: $\sup_{t\in\mathscr{T}} \sqrt{\tilde{\alpha}_n}\left|g(\xi;\theta_0 + t/\tilde{\alpha}_n)\right| \leq F_n(\xi)$ and $\mathbb{E}[\boldsymbol{F_{ni}^2}] = O(1)$ where $\boldsymbol{F_{ni}} = F_n(\boldsymbol{\xi_i})$.*

(iii) *For any $\epsilon > 0$, $\mathbb{E}[\boldsymbol{F_{ni}^2} I(\boldsymbol{F_{ni}} > \epsilon\sqrt{n})] = o(1)$.*

(iv) *For any $\epsilon_n \downarrow 0$, $\sup_{\|t-s\|<\epsilon_n} \tilde{\alpha}_n \mathbb{E}\left[(\boldsymbol{g_i}(\theta_0 + t/\tilde{\alpha}_n) - \boldsymbol{g_i}(\theta_0 + s/\tilde{\alpha}_n))^2\right] = o(1)$.*

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

Assumption G is common, but is not always straightforward to verify. It ensures weak convergence of $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ to $\mathbb{G}$ and is satisfied in the maximum score case as theorems 1 and 2 show.

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

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 $\mathbb{E}\left[\sup_s f(s|\boldsymbol{z_i})\|\boldsymbol{z_i}\|\right] < \infty$.*

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

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} \frac{1}{c_\alpha^2} \frac{\int t \exp\left(c_\alpha^3 \mathbb{G}(t)\right)\phi_V(t)dt}{\int \exp\left(c_\alpha^3 \mathbb{G}(t)\right)\phi_V(t)dt}. \tag{6}$$

(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 $\tilde{\mathbb{G}}(t) = \mathbb{G}(t) - t^\intercal V t/2$,

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} \operatorname*{argmax}_t \tilde{\mathbb{G}}(t). \tag{7}$$

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 α_n increase at a rate no slower than $\sqrt[3]{n}$. In fact, part (ii) of theorem 3 demonstrates that if α_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 $c_\alpha \to \infty$ after $n \to \infty$, 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_α^4 is incorporated into $\boldsymbol{g_i}$ instead of into α_n^2.[12] Then

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} \frac{\int t \exp\{c_\alpha^4(\mathbb{G}(t) - t^\intercal V t/2)\}dt}{\int \exp\{c_\alpha^4(\mathbb{G}(t) - t^\intercal V t/2)\}dt}. \tag{8}$$

For large values of c_α, the right hand side of (8) is close to the value at which $\mathbb{G}(t) - t^\intercal V t/2$ 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 $\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0)$ converges to zero as $c_\alpha \downarrow 0$. More interestingly, noting that $\sqrt{n/\alpha_n} = \sqrt[3]{n}/c_\alpha$, (6) suggests that $\sqrt{n/\alpha_n}(\hat{\boldsymbol{\theta}} - \theta_0)$ has an approximate $\int t\mathbb{G}(t)\phi_V(t)dt$–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_α 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_α increases. But please note that these are asymptotic results; for small values of c_α 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 α_n increases more slowly than $\sqrt[3]{n}$. Let

$$\begin{cases} \mathscr{N}_{\boldsymbol{n}} &= \alpha_n^d \sqrt{n/\alpha_n} \int \pi(\theta)(\theta - \theta_0) \exp\big(\alpha_n^2 \boldsymbol{S_n}(\theta) + \alpha_n^2 Q(\theta)\big)d\theta, \\ \mathscr{D}_{\boldsymbol{n}} &= \alpha_n^d \int \pi(\theta) \exp\big(\alpha_n^2 \boldsymbol{S_n}(\theta) + \alpha_n^2 Q(\theta)\big)d\theta, \\ \mathscr{B}_n &= \alpha_n^d \sqrt{n/\alpha_n} \int \pi(\theta)(\theta - \theta_0) \exp\big(\alpha_n^2 Q(\theta)\big)d\theta. \end{cases} \tag{9}$$

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

$$\sqrt{\frac{n}{\alpha_n}}(\hat{\boldsymbol{\theta}} - \theta_0) - \frac{\mathscr{B}_n}{\mathscr{D}_{\boldsymbol{n}}} = \frac{\mathscr{N}_{\boldsymbol{n}} - \mathscr{B}_n}{\mathscr{D}_{\boldsymbol{n}}} \xrightarrow{d} N(0, \mathscr{V}), \tag{10}$$

where $\mathscr{V} = \iint ts^\intercal H(t,s)\phi_V(t)\phi_V(s)dtds$.

[12] Alternatively, one can carry out the substitution $\tilde{t} = t/c_\alpha^2$ and note that the process $\mathbb{G}^*$ with $\mathbb{G}^*(\tilde{t}) = \mathbb{G}(c_\alpha^2\tilde{t})/c_\alpha$ has the same statistical properties as $\mathbb{G}$.

In lemma G.4 we show that for the maximum score case in which the regressor vector includes a constant,[13]

$$\mathscr{V} = \frac{1}{2\sqrt{\pi}} V^{-1} \mathbb{E}\left[\frac{z_i z_i^\intercal}{\sqrt{z_i^\intercal V^{-1} z_i}} f(z_i^\intercal \theta_0 | z_i) \right] V^{-1}, \tag{11}$$

which suggests that using observation–specific weights can reduce (some scalar–valued function of) the asymptotic variance.[14] Some scale normalization of the weights should be imposed lest they assume the role of α_n. We do not pursue a weighting procedure in this paper.

The bias term $\mathscr{B}_n$ can affect the limiting distribution. As will become apparent below, if Q is sufficiently smooth then $\mathscr{B}_n$ 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 $n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0)$ 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 τ, $\mathscr{B}_n = \beta_n^{-1} \sum_{\tau=0}^{q} b_{q\tau}^* / \alpha_n^\tau + o\big(\alpha_n^{-q} \beta_n^{-1}\big)$.*

An expression for the values of $b_{q\tau}^*$ is provided in the proof of theorem 5. In particular, $b_{q1}^* = C_V \int \big(\pi_0 D_{Q3}(t) + D_{\pi 1}(t)\big) t \phi_V(t) dt$, where $C_V = 1/\phi_V(0)$, $\pi_0 = \pi(\theta_0)$, $D_{\pi 1}(t)$ is the first term in a Taylor expansion of $\pi(\theta_0 + t)$ about $\pi(\theta_0)$, i.e. $t^\intercal \pi_\theta(\theta_0)$, 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*

$$n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} N\left(\frac{\int \big(\pi_0 D_{Q3}(t) + D_{\pi 1}(t)\big) t \phi_V(t) dt}{\pi_0 c_\alpha^4}, c_\alpha^2 \mathscr{V} \right). \tag{12}$$

It is possible to minimize the asymptotic mean square error (AMSE, squared mean plus variance in (12)) by using an estimate of $c_{\alpha 1}^* = \sqrt[10]{4 \|b_{q1}^*\|^2 / C_V^2 \pi_0^2 \operatorname{tr}(\mathscr{V})}$ in lieu of c_α. However, because the asymptotic bias can be removed by the choice of a prior and since the asymptotic distribution using $c_{\alpha 1}^*$ would have nonzero mean — which is at odds with the uniform inference procedure at the heart of this paper — we do not include this result here; it is available from the authors' website.

[13] What is needed is in fact weaker: for some $j = 1, \ldots, d$: $\mathbb{P}[z_{ij} = 0] = 0$.

[14] 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 $\hat{\boldsymbol{\theta}}$ does not typically exist.

For the specific case of maximum score, the limit distribution of theorem 6 can be compared to that of the SMS 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 SMS, prior and c_α here) and the joint distribution of $(\boldsymbol{y_i}, \boldsymbol{a_i}, \boldsymbol{z_i})$. It can be shown that for the very special case in which $d = 1$, $\boldsymbol{z_i} = 1$ a.s., a normal kernel is used for the SMS estimator and a flat prior is used for ours, and bandwidth (SMS) and c_α 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 $(\boldsymbol{y_i}, \boldsymbol{a_i})$.[15] 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.

From (12) it follows that 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. As noted in theorem 5, $b^*_{q\tau} = 0$ for all even values of τ. Consequently, provided that an estimator $\hat{\boldsymbol{b}}^*_{\boldsymbol{q1}}$ of b^*_{q1} converges to b^*_{q1}, we obtain the result

$$n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0) - \frac{\hat{\boldsymbol{b}}^*_{\boldsymbol{11}}}{c_\alpha^4 \mathscr{D}_{\boldsymbol{n}}} \xrightarrow{d} N(0, c_\alpha^2 \mathscr{V}). \tag{13}$$

With $q = 3$ and $\alpha_n = c_\alpha^2 \sqrt[9]{n}$, (13) can be strengthened to

$$n^{4/9}(\hat{\boldsymbol{\theta}} - \theta_0) - \frac{n^{2/9}\hat{\boldsymbol{b}}^*_{\boldsymbol{31}}}{c_\alpha^4 \mathscr{D}_{\boldsymbol{n}}} - \frac{\hat{\boldsymbol{b}}^*_{\boldsymbol{33}}}{c_\alpha^8 \mathscr{D}_{\boldsymbol{n}}} \xrightarrow{d} N(0, c_\alpha^2 \mathscr{V}),$$

assuming $\hat{\boldsymbol{b}}^*_{\boldsymbol{31}}$ converges to b^*_{31} at a rate faster than $n^{2/9}$ which, in light of the smoothness condition $q = 3$, should not be problematic. This process can be repeated to obtain an estimator of θ_0 which converges at a rate arbitrarily close to $\sqrt{n}$, assuming sufficient smoothness. In particular, using $\alpha_n = c_\alpha^2 n^{1/(2q+3)}$ leads to

$$n^{\frac{q+1}{2q+3}}(\hat{\boldsymbol{\theta}} - \theta_0) - \frac{1}{\mathscr{D}_{\boldsymbol{n}}} \sum_{\tau=0}^{q} \frac{n^{\frac{q-\tau}{2q+3}} \hat{\boldsymbol{b}}^*_{\boldsymbol{q\tau}}}{c_\alpha^{2\tau+2}} \xrightarrow{d} N\big(0, c_\alpha^2 \mathscr{V}\big), \tag{14}$$

where $\hat{\boldsymbol{b}}^*_{\boldsymbol{q\tau}} = 0$ for even values of τ.

It turns out that removing the bias in theorem 6 can also be accomplished by choosing a prior that resembles Jeffreys' prior (Jeffreys, 1946), here $\pi(\theta) \propto \sqrt{\det\big(-Q''(\theta)\big)}$, near θ_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 θ_0. An example is $\pi_J(\theta) \propto \sqrt{\left|\det\big(-Q''(\theta)\big)\right|}$,

[15] The derivation is available upon request.

which can be estimated by $\hat{\boldsymbol{\pi}}_{\boldsymbol{J}}(\theta) \propto \sqrt{\left|\det\left(-\hat{\boldsymbol{Q}}''(\theta)\right)\right|}$ for some estimator $\hat{\boldsymbol{Q}}''$ which satisfies the following assumption. Let vec denote the familiar vec–operator, which stacks the columns of a matrix into a vector and let Q''' and $\hat{\boldsymbol{Q}}'''$ be defined as $\partial \operatorname{vec}^{\mathsf{T}} Q''/\partial\theta$ and $\partial \operatorname{vec}^{\mathsf{T}} \hat{\boldsymbol{Q}}''/\partial\theta$, respectively.

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

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

The conditions specified in assumption I are generic because the properties of Q'' depend on the nature of $\boldsymbol{g_i}$, 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 $n^{2/5}$ is sufficient, but when we require $r^* = 2$ more derivatives of Q'' will be available. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{J}}$ denote the estimator which is identical to $\hat{\boldsymbol{\theta}}$ except that it uses the prior $\hat{\boldsymbol{\pi}}_{\boldsymbol{J}}$ 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 $n^{2/5}(\hat{\boldsymbol{\theta}}_{\boldsymbol{J}} - \theta_0) \xrightarrow{d} N(0, c_\alpha^2 \mathscr{V})$.
(ii) If $\sqrt[9]{n} = O(\alpha_n)$, $q \geq 3$, and assumption I is satisfied with $r^* \geq 2$, then for some finite $\mathscr{B}_J$, $\sqrt{n/\alpha_n}(\hat{\boldsymbol{\theta}}_{\boldsymbol{J}} - \theta_0) - \mathscr{B}_J/\alpha_n^3\beta_n\mathscr{D}_{\boldsymbol{n}} \xrightarrow{d} N(0, c_\alpha^2\mathscr{V})$.

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 π at θ_0, such priors are straightforward to construct by using suitable polynomials in $\theta - \theta_0$ whose coefficients consist of estimated derivatives of Q'' at θ_0, albeit that this procedure would require an initial plugin estimator to use in lieu of θ_0, whereas such an initial plugin estimator is not required for $\hat{\boldsymbol{\pi}}$ 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 $\tilde{\pi}(\theta) = \sqrt{\det\left(-\tilde{Q}''(\theta)\right)}$, with

$$\begin{aligned}\tilde{Q}''(\theta) &= Q''(\theta_0)/2 \\ &+ \left(Q'''(\theta_0)\left(I \otimes (\theta-\theta_0)\right) + Q''(\theta_0)\right)^{\mathsf{T}} \left(Q''(\theta_0)\right)^{-1} \times \left(Q'''(\theta_0)\left(I \otimes (\theta-\theta_0)\right) + Q''(\theta_0)\right), \quad (15)\end{aligned}$$

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 maximum score case 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

$$\frac{\int \big(\pi_0 D_{\hat{Q}3}(t) + D_{\pi 1}(t)\big) t \phi_{\hat{V}}(t) dt}{\pi_0 c_\alpha^4}. \tag{16}$$

In particular, for $d = 1$ and a flat prior, (16) reduces to $\hat{\boldsymbol{Q}}'''(\hat{\boldsymbol{\theta}}_{\boldsymbol{p}})/2c_\alpha^4\big(\hat{\boldsymbol{Q}}''(\hat{\boldsymbol{\theta}}_{\boldsymbol{p}})\big)^2$, with $\hat{\boldsymbol{\theta}}_{\boldsymbol{p}}$ some preliminary estimator.

Since the form of an estimator of Q'' and its derivatives depends on the nature of $\boldsymbol{g_i}$, establishing general analytical results for the bias correction is not possible. It would be possible to do so for the maximum score case, but this would entail not much more than a rehash of results that are well–known and are available in the literature.

For the maximum score case, nonparametric estimators can be shown to be uniformly consistent for Q'' and Q''' in a neighborhood of θ_0 that decreases more slowly than $\hat{\boldsymbol{\theta}}_{\boldsymbol{p}}$. So consistency obtains and the bias up to order $n^{-2/5}$ is removed.

4. Uniform Inference

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

$$\boldsymbol{\Psi_n} = \frac{\int t \exp\big(\beta_{n1}\hat{\mathbb{G}}(t) - \beta_{n2} t^\intercal \hat{\boldsymbol{V}} t/2\big) dt}{\int \exp\big(\beta_{n1}\hat{\mathbb{G}}(t) - \beta_{n2} t^\intercal \hat{\boldsymbol{V}} t/2\big) dt}, \qquad \text{with} \begin{cases} \beta_{n1} = \min\big(\beta_n^{3/2}, \beta_n^{4/3}\big), \\ \beta_{n2} = \min\big(\beta_n^{2}, \beta_n^{4/3}\big), \end{cases} \tag{17}$$

have the same limiting distribution in all three cases.

Theorem 8. *If $\hat{\boldsymbol{V}} - V = o_p(1)$ and $\hat{\boldsymbol{H}}$ is a positive semidefinite covariance kernel*[16] *which is pointwise consistent for H and for some $0 < c_H < 2$ and all sample paths of $\hat{\boldsymbol{H}}$ satisfy $\limsup_{\|t\|\to\infty}\left(\hat{\boldsymbol{H}}(t,t)/\|t\|^{c_H}\right) < \infty$, then $\boldsymbol{\Psi_n}$ has the limiting distribution derived for $\boldsymbol{\chi_n}$ in the corresponding portion of theorems 3 and 4*[17] *if (i) $\beta_n = o(1)$, (ii) $1/\beta_n + \beta_n = O(1)$, (iii) $1/\beta_n = o(1)$.*

The requirement that $\hat{\boldsymbol{H}}$ be a positive semidefinite covariance kernel is both necessary and sufficient for there to exist a Gaussian process $\hat{\mathbb{G}}$ with that covariance kernel (Doob, 1953, theorem 3.1). Drawing random numbers using (17) is simple and is explained in appendix I.2.

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

The limiting distribution of Laplace–type estimators sometimes coincides with the limit of the posterior, i.e. $\lim_{n\to\infty} \exp\left(\alpha_n^2 \boldsymbol{L_n}(\theta_0 + t/\alpha_n)\right)$ up to a proportionality constant. Indeed, Chernozhukov and Hong (2003) derived conditions under which this is the case for objective functions that admit a quadratic expansion. In the case of coincidence of the true distribution and the limit of the posterior, the (Gibbs) draws used to obtain $\hat{\boldsymbol{\theta}}$ can then be used to construct confidence intervals. Here the limit of the posterior is a random object if α_n increases at the rate of $\sqrt[3]{n}$ (or faster) and is $N(0, V^{-1})$ if α_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 to choose input parameters π, α_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 α_n.

We consider a small number of designs, all in the context of the maximum score estimation problem. In all cases $\boldsymbol{y_i} = I\left(\boldsymbol{z_i}^\intercal \theta_0 - \boldsymbol{a_i} + \varsigma(\boldsymbol{a_i}, \boldsymbol{z_i})\boldsymbol{u_i} \geq 0\right)$, where $\boldsymbol{u_i}$ is a standard normal independent of $\boldsymbol{a_i}, \boldsymbol{z_i}$ which are also standard normal and have mutually independent elements. Unless otherwise stated $\theta_0 = [0, \vec{1}^\intercal]^\intercal$, where $\vec{1} \in \Re^{d-1}$ 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 $\varsigma^2 = (z_1^2 + 1)/2$. Figure 1 illustrates how (for a single data set) the normalized posterior varies with α_n; the value at θ_0 equals one. Larger values of α_n lead to a

[16] $\hat{\boldsymbol{H}}$ is a positive semidefinite covariance kernel if for any integer $0 < T^* < \infty$ and any values $t_1, t_2, \ldots, t_{T^*} \in \Re^d$, the matrix with (i,j)–element $\hat{\boldsymbol{H}}(t_i, t_j)$ is positive semidefinite.

[17] Please note that for part (i) of theorem 3, κ_n and $\sqrt[3]{n}$ can differ by a multiplicative constant.

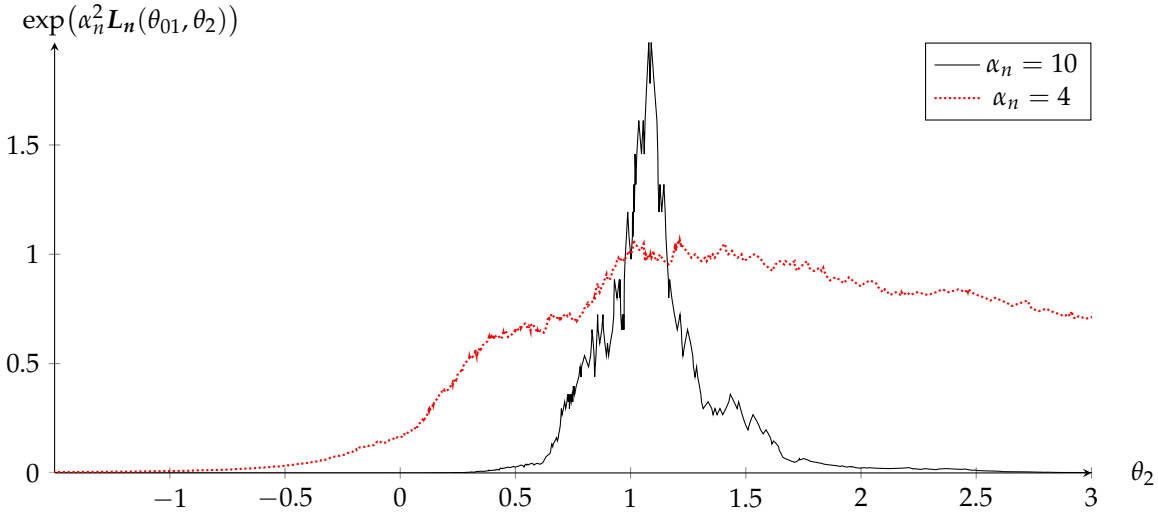

FIGURE 1. One draw of $\exp\left(\alpha_n^2 \boldsymbol{L_n}(0, \theta_2)\right)$; $n = 1,000$.

narrower posterior density, so $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ is an average over a comparatively small range of θ–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 θ–values whose range varies from one data set to another, $\hat{\boldsymbol{\theta}}$ for $\alpha_n = 10$ has a greater variance than for $\alpha_n = 4$. But the bias is less for greater values of α_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.

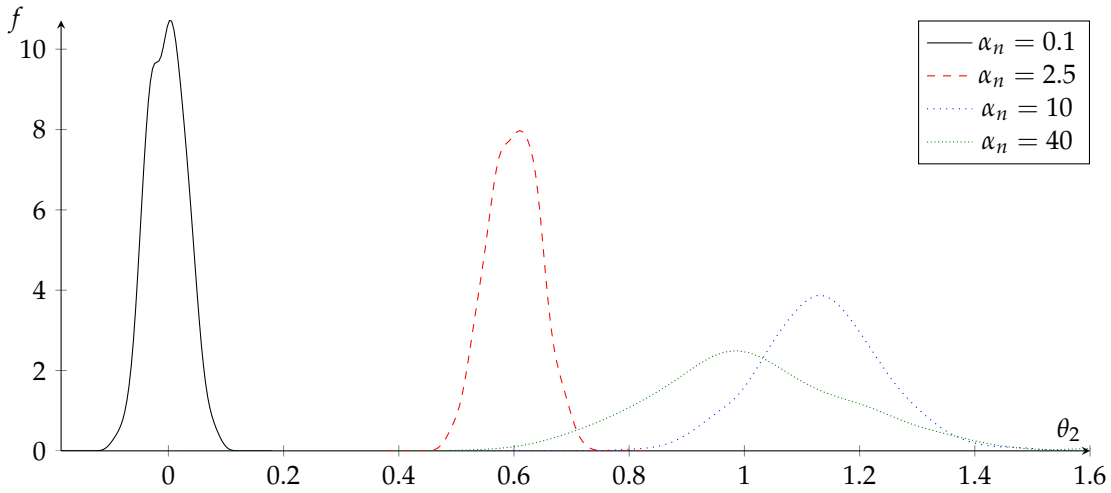

FIGURE 2. Density of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ for various choices of α_n; $d = 2$, $n = 1,000$

A secondary cause of the bias is that for smaller values of α_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 depends on the prior and its derivatives only at θ_0; the value of the prior and its derivatives at other values of θ are irrelevant for the asymptotic bias. For a flat prior the derivatives at θ_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 α_n, consider figure 2. For very small values of α_n the distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ is centered around the mean of the prior (zero). As α_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.

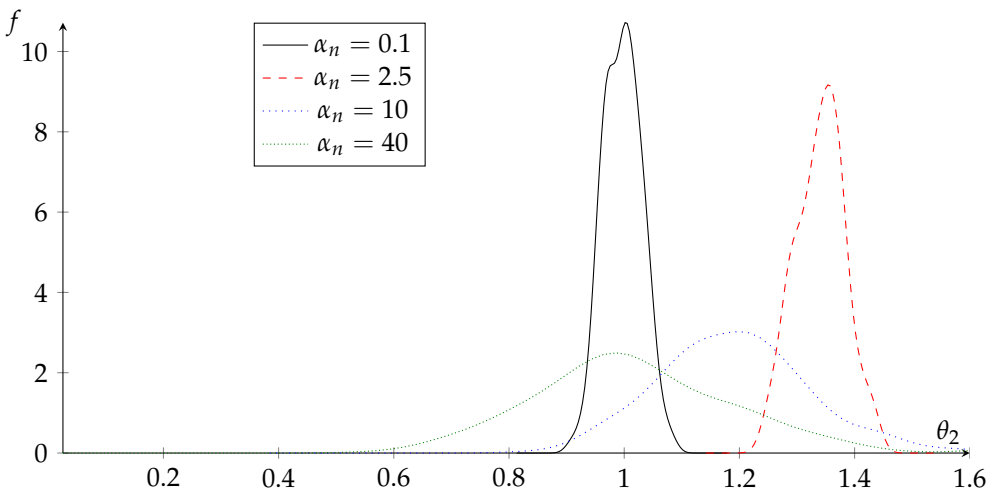

FIGURE 3. Density of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ for various choices of α_n; $d = 2$, $n = 1{,}000$, prior mean coincides with true value.

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 θ_{02}; both equal one. There is some bias for smaller values of α_n, but this bias is all in the same direction; positive.

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

objective function, e.g. one could construct a method for choosing α_n specific to the maximum score case.

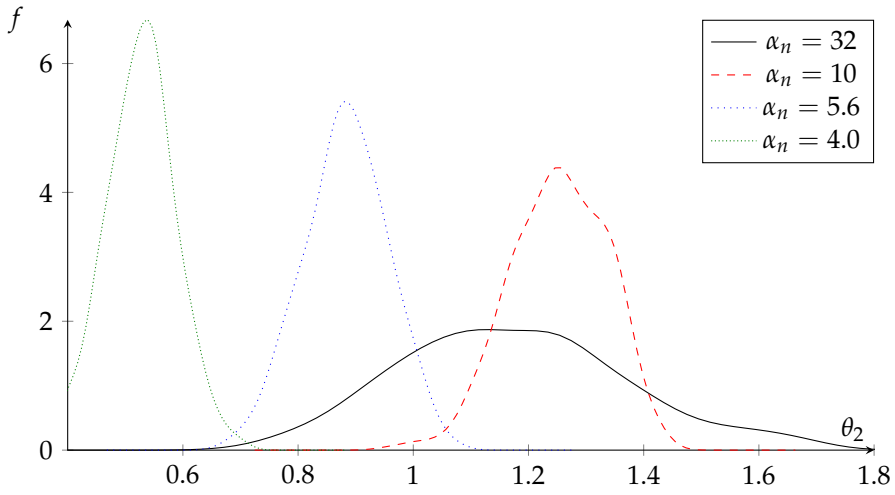

FIGURE 4. Density of $\hat{\boldsymbol{\theta}}_2$ for various values of α_n; $d = 10$, $n = 1,000$

The experiment depicted in figure 2 is repeated in figure 4 with $d = 10$. The results suggest that the value of α_n need not depend on d, but that the rewards for a good choice of α_n increase. Indeed, the estimator variance is substantial for $\alpha_n = 32$ and would be still greater for the maximum score estimator. In fact, for a $\sqrt[3]{n}$–consistent estimator to overcome the difference in variances between $\alpha_n = 10$ and $\alpha_n = 32$ over ten times as many observations would be needed.[18]

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. 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 here too the maximum score estimator[19] has greater variance than the Laplace–type estimators, which suggests that choosing $\alpha_n = \infty$ is suboptimal even without bias correction.

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 the

[18]The ratio of variances of $\hat{\boldsymbol{\theta}}$–values (for $\alpha_n = 32$ versus $\alpha_n = 10$) in the simulations is about 5.8; the number ten comes from the fact that $5.8^{3/2} = 14 > 10$.

[19]Computed using a grid search.

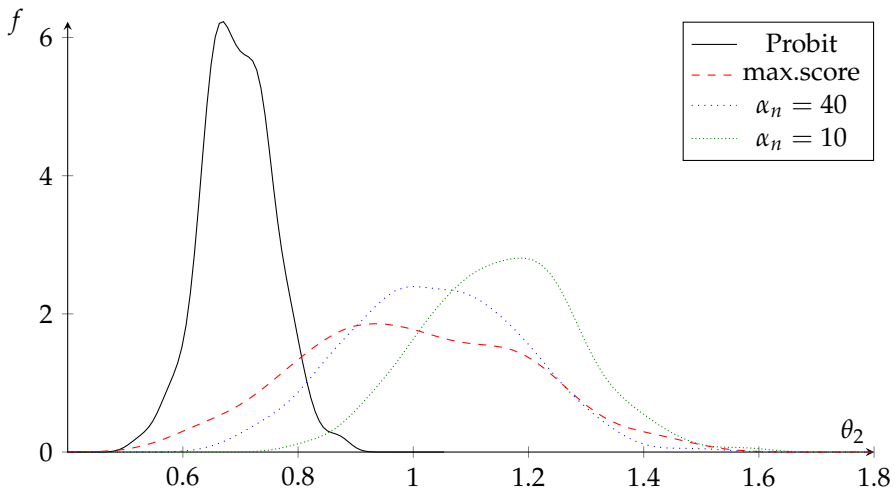

FIGURE 5. Density of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$; probit versus maximum score and Laplace; $d = 2$, $n = 1,000$

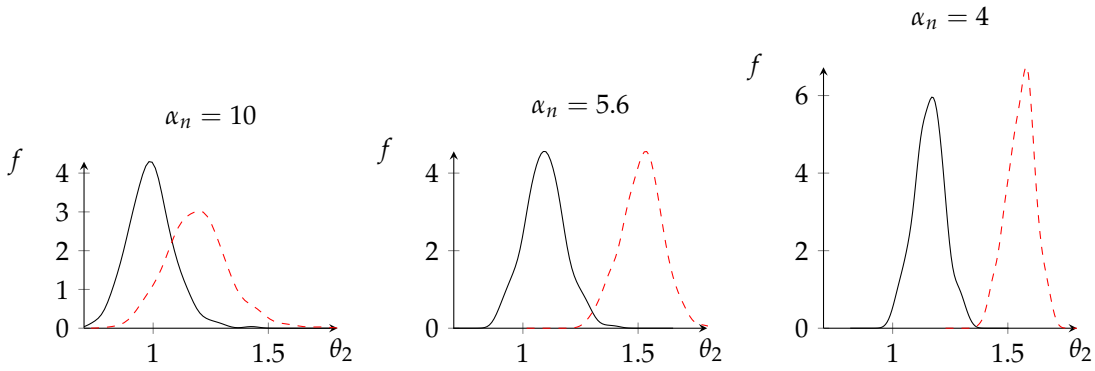

FIGURE 6. Density of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$; Jeffreys (solid) versus uniform (dashed) prior; $n = 1,000$, $d = 2$.

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 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 (15), 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 α_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.

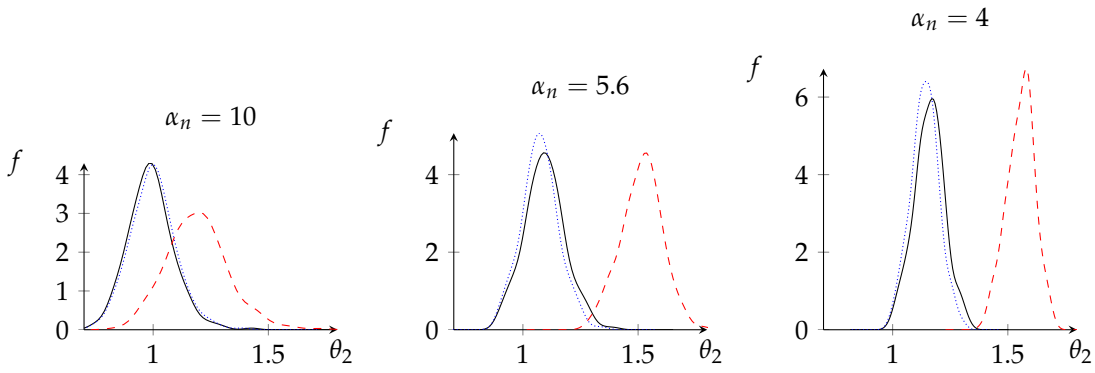

FIGURE 7. Density of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$; Jeffreys (solid), uniform (dashed), estimated Jeffreys (dotted) prior; $n = 1,000$, $d = 2$.

Figure 6 demonstrates that the bias decreases in α_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 α_n is small, which is due to the fact that for small values of α_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 descibed 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 α_n is chosen large.

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 α_n (and hence β_n) decreases. It does a good job approximating the finite sample distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$ except that for small values of α_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.[20]

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

[20] We do use Jeffreys' prior.

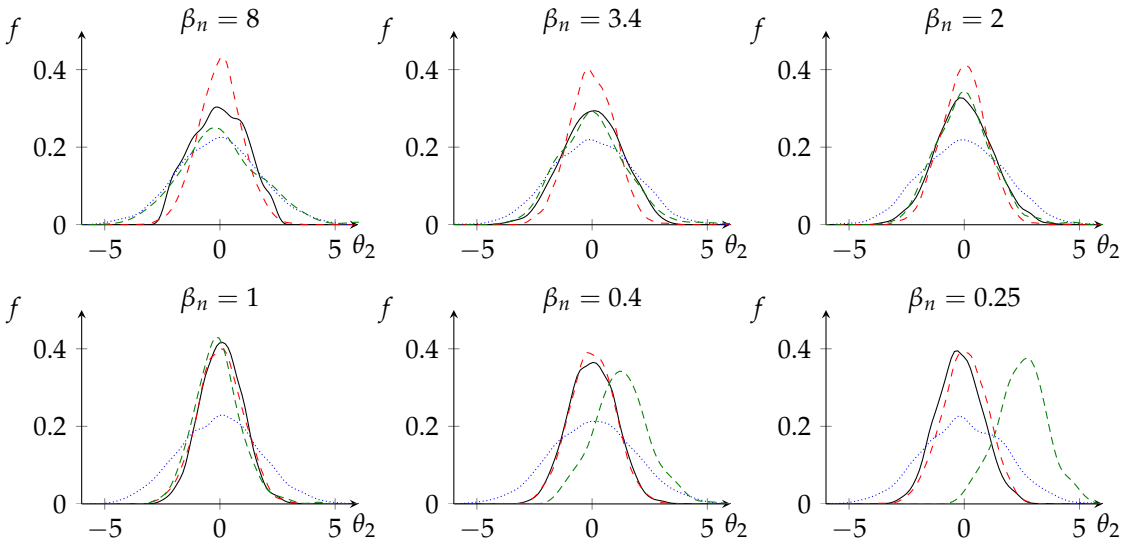

FIGURE 8. Density of $\kappa_n(\hat{\boldsymbol{\theta}}_{\mathbf{2}} - \theta_0)$ for various inference procedures: uniform (solid), normal (dashed), maximum score (dotted), finite sample (densely dashed); $n = 1,000$, $d = 2$, $R = 3,000$.

suggests that one should use our uniform inference procedure with a high value of β_n to simulate the limiting distribution of the maximum score estimator instead of trying to maximize $\hat{\mathbb{G}}(t) - t^{\intercal}\hat{\mathbb{V}}t/2$ 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.

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)*.

BELLUZO, W. (2004): "Semiparametric approaches to welfare evaluations in binary response models," *Journal of Business and Economic Statistics*, 22, 322–330.

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.

——— (2002): "Bootstrap critical values for tests based on the smoothed maximum score estimator," *Journal of econometrics*, 111(2), 141–167.

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. VOGELSANG (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.

KORDAS, G. (2006)): "Smoothed binary regression quantiles," *Journal of Applied Econometrics*, 21, 387–407.

KOSOROK, M. (2008): *Introduction to empirical processes and semiparametric inference*. Springer Verlag.

KOTLYAROVA, Y., AND V. ZINDE-WALSH (2006): "Non– and semi–parametric estimation in models with unknown smoothness," *Economics Letters*, 93(3), 379–386.

——— (2009): "Robust estimation in binary choice models," *Communications in Statistics (forthcoming)*.

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.

POLLARD, D. (1993): "The asymptotics of a binary choice model," Discussion paper, Yale University.

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.

ZINDE-WALSH, V. (2002): "Asymptotic theory for some high breakdown point estimators," *Econometric theory*, 18(05), 1172–1196.

APPENDIX A. PRELIMINARIES

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

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

Lemma A.1. *Let $\mathscr{T}_1 \subseteq \mathscr{T}_2 \cdots$ be a sequence of compact sets such that $0 \in \mathscr{T}_1$ and $\Re^d = \bigcup_{i=1}^\infty \mathscr{T}_i$. We then have $\tilde{\boldsymbol{S}}_{\boldsymbol{n}} \xrightarrow{w} \mathbb{G}$ in $\mathscr{L}^\infty(\mathscr{T}_1, \mathscr{T}_2, \ldots)$, where $\mathbb{G}(\cdot)$ 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 $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ in $\mathscr{L}^\infty(\mathscr{T})$ for an arbitrary compact set $\mathscr{T} \subset \Re^d$. Since $\mathscr{T}$ is dense, assumption G(i) ensures that for $j = 1, 2$,

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

is a pointwise measurable class, and hence $\mathscr{P}$–measurable for every $\mathscr{P}$; see van der Vaart and Wellner (1996, page 110). Note also that $\mathbb{E}\left[\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(s)\right] = \tilde{\alpha}_n E\left[\boldsymbol{g_i}(\theta_0 + t/\tilde{\alpha}_n)\boldsymbol{g_i}(\theta_0 + s/\tilde{\alpha}_n)\right] \to H(t,s)$ for every $t, s \in \mathscr{T}$ by assumption F. Therefore, the result follows from van der Vaart and Wellner (1996, theorem 2.11.22). □

Lemma A.2. *Let for $\tilde{\gamma} > 0$, $\tilde{\Gamma} = \{t \in \Re^d : \|t\| \leq \tilde{\gamma}\}$. Then (i) for any $0 < c < \infty$,*
$\lim_{\tilde{\gamma}\to\infty} \lim_{n\to\infty} \mathbb{P}\left[\sup_{t\in\tilde{\Gamma}^c} \left|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\right|/\|t\|^2 > c\right] = 0$ *and (ii)* $\mathbb{P}\left[\sup_{t\in\Gamma_n^c} \left|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\right|/\|t\|^2 > c\right] = o(1)$.

Proof. Let $\tilde{\tilde{\boldsymbol{S}}}_{\boldsymbol{n}}(t) = \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)/\|t\|$ and $\tilde{\boldsymbol{G}}(t) = \mathbb{G}(t)/\|t\|$. Suppose without loss of generality that $\alpha_1 = 1$. Then the equicontinuity of $\tilde{\tilde{\boldsymbol{S}}}_{\boldsymbol{n}}$ on Γ_1^c follows from that of $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$. Therefore on Γ_1^c, $\tilde{\tilde{\boldsymbol{S}}}_{\boldsymbol{n}} \xrightarrow{w} \tilde{\boldsymbol{G}}$, which is a uniformly bounded process in $\mathscr{L}^\infty(\Gamma_1^c)$. Hence, to establish (ii) ((i) is similar), note that $\sup_{t\in\Gamma_n^c}\left(\left|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\right|/\|t\|^2\right) \leq \sup_{t\in\Gamma_1^c}\left(\left|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\right|(\log\|t\|)/\|t\|^2\right)/\log\gamma_n = O_p(1)o(1) = o_p(1)$. □

Lemma A.3. *For any $c > 0$, $\sup_t\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - ct^\intercal Vt\} = O_p(1)$.*

Proof. Follows immediately from lemma A.1 and the properties of $\mathbb{G}$. □

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 θ_0. □

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 $Q_n(t) \leq -\min(\tilde{\alpha}_n^2 c_q, t^\intercal Vt/4)$ for all t for which $\theta_0 + t/\tilde{\alpha}_n \in \Theta$.*

Proof. Follows from the fact that (i) Θ is compact, (ii) Q has a unique maximum of zero at θ_0, and (iii) $Q_{\theta\theta}$ is continuous at θ_0 with value $-V$. □

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

Proof. We have
$\left|\exp(cb) - \sum_{s=0}^{j}(cb)^s/s!\right| \leq \left|\sum_{s=j+1}^{\infty}(cb)^s/s!\right| \leq |c|^{j+1}\sum_{s=j+1}^{\infty}|b|^s/s! \leq |c|^{j+1}\exp(|b|)$. □

A.4. **Further Auxiliary Results.** The results in this section presume that $0 \leq \lim_{n\to\infty}\beta_n = c_\beta < \infty$ 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 $\int_{\Gamma_n^c}\|P(t)\|\exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - c_3t^\intercal Vt\big)dt = o_p\big(\alpha_n^{-c_4}\big)$,
(ii) if $c_2 + c_3 > 0$ then $\int_{\Gamma_n^c}\|\pi_n(t)P(t)\|\exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| + c_2Q_n(t) - c_3t^\intercal Vt\big)dt = o_p\big(\alpha_n^{-c_4}\big)$.

Proof. First (i). By lemma A.2, the left hand side of (i) is for any c of smaller order than $\int_{\Gamma_n^c}\|P(t)\|\exp\big((cc_1c_\beta - c_3\lambda_-)\|t\|^2\big)dt$, which for sufficiently small c decreases exponentially in n. The left hand side in (ii) is by lemma A.5 of order no greater than

$$\begin{aligned}\int_{\Gamma_n^c}\|\pi_n(t)P(t)\|\exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - c_2\min(c_q\alpha_n^2, t^\intercal Vt/4) - c_3t^\intercal Vt\big)dt \\ \leq \int_{\Gamma_n^c}\|\pi_n(t)P(t)\|\exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - c_2c_q\alpha_n^2 - c_3t^\intercal Vt\big)dt \\ + \int_{\Gamma_n^c}\|\pi_n(t)P(t)\|\exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - (c_2/4+c_3)t^\intercal Vt\big)dt. \quad (18)\end{aligned}$$

The second right hand side term in (18) 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 $\sup_\theta |\boldsymbol{S_n}(\theta)| = o_p(1)$ by lemma A.3, it follows that for some $c^* > 0$, $\int_{\Gamma_n^c} \|\pi_n(t)P(t)\| \exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - c_2c_q\alpha_n^2\big)dt$ is of order no greater than $\exp\big(-c^*\alpha_n^2\big)$. □

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

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

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 $\int \|P(t)\| \exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| - c_3t^\intercal Vt\big)dt = O_p(1)$,
(ii) if $c_2 + c_3 > 0$ then $\int \|\pi_n(t)P(t)\| \exp\big(c_1\beta_n|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| + c_2Q_n(t) - c_3t^\intercal Vt\big)dt = O_p(1)$.

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. □

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

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

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

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

Proof. We have

$$\begin{aligned}
&\int \pi_n(t)t^j \exp\big(c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big)dt - C_V\pi_0 \int t^j \exp\big(c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big)\phi_V(t)dt \\
&= \int_{\Gamma_n^c} \pi_n(t)t^j \exp\big(c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big)dt + C_V \int_{\Gamma_n} \pi_n(t)t^j \exp\big(c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big)\Big(\exp\big(R_n(t)\big) - 1\Big)\phi_V(t)dt \\
&\qquad + C_V \int_{\Gamma_n} \big(\pi_n(t) - \pi_0\big)t^j \exp\big(c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big)\phi_V(t)dt - \pi_0 \int_{\Gamma_n^c} t^j \exp\big(c_\alpha^3\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) - t^\intercal Vt/2\big)dt. \quad (19)
\end{aligned}$$

The first and last right hand side terms in (19) are dealt with in lemma A.7. The second right hand side term in (19) is $o_p(1)$ by lemma B.1. Because π is continuous at θ_0, the third right hand side term in (19) is also $o_p(1)$ by lemma A.3. □

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

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

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

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

$$\sup_{t \in \tilde{\Gamma}} \big| Q_n(t) + t^\intercal V t/2 \big| \leq \tilde{\gamma}^2 \sup_{\|\theta - \theta_0\| \leq \tilde{\gamma}/\sqrt[3]{n}} \|Q_n''(\theta) + V\|/2 = o_p(1), \tag{20}$$

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

$$\int_{\tilde{\Gamma}} |\tilde{L}_n^*(t) - \tilde{\mathbb{G}}^*(t)|^2 dt = o(1). \tag{21}$$

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 $c > 0$, $T(\varkappa, c) = \{t \in \tilde{\Gamma} : |\varkappa(t) - \bar{\mathbb{G}}^*| \leq c\}$, where $\bar{\mathbb{G}}^* = \max_{t \in \tilde{\Gamma}} \tilde{\mathbb{G}}^*(t)$. We first establish that

$$d^*\big(T(\tilde{L}_n^*, c), T(\tilde{\mathbb{G}}^*, c)\big) = o(1). \tag{22}$$

Let $T_{1n}(c) = T(\tilde{L}_n^*, c) - T(\tilde{\mathbb{G}}^*, c)$ and $T_{2n}(c) = T(\tilde{\mathbb{G}}^*, c) - T(\tilde{L}_n^*, c)$. We show that $\mu\big(T_{2n}(c)\big) = o(1)$ where the same result for $T_{1n}(c)$ follows similarly. Let for arbitrary $c^* > 0$, $T_n^*(c^*) = \{t \in \tilde{\Gamma} : |\tilde{L}_n^*(t) - \tilde{\mathbb{G}}^*(t)| \leq c^*\}$. For the remainder of this lemma, define complements relative to $\tilde{\Gamma}$ (e.g. $T_n^{*c}(c^*) = \tilde{\Gamma} - T_n^*(c^*)$). Because $\mu\big(T_n^{*c}(c^*) \cap T_{2n}(c)\big) \leq \mu\big(T_n^{*c}(c^*)\big) = o(1)$ by (21), we only need to consider $\mu\big(T_n^*(c^*) \cap T_{2n}(c)\big)$.

Note first that $T_n^*(c^*) \cap T_{2n}(c) \subseteq T_n^{**}(c, c^*) = \{t \in \tilde{\Gamma} : c \leq |\tilde{L}_n^*(t) - \bar{\mathbb{G}}^*| \leq c + c^*\}$, such that by (21),

$$\begin{aligned} \lim_{c^* \downarrow 0} \lim_{n \to \infty} \mu\big(T_n^{**}(c, c^*)\big) &= \lim_{c^* \downarrow 0} \mu\big(\{t \in \tilde{\Gamma} : c \leq |\tilde{\mathbb{G}}^*(t) - \bar{\mathbb{G}}^*| \leq c + c^*\}\big) \\ &= \mu\big(\{t \in \tilde{\Gamma} : |\tilde{\mathbb{G}}^*(t) - \bar{\mathbb{G}}^*| = c\}\big) = 0, \end{aligned} \tag{23}$$

because $\tilde{\mathbb{G}}^*$ is continuous and nowhere differentiable. So (22) holds.

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

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

since the support of π_n is only increasing at a rate of α_n by assumption C. Thus,

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

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

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

$$\lim_{\tilde{\gamma} \to \infty} \lim_{n \to \infty} \mathbb{P}\left[\frac{\int_{\tilde{\Gamma}^c} \|t^j \pi_n(t)\| \exp\{\beta_n^{4/3} (\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t))\} dt}{\int \pi_n(t) \exp\{\beta_n^{4/3} (\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t))\} dt} > \epsilon \right] = o(1). \tag{24}$$

Proof. We first work on the numerator in (24). For $c > 0$, let $\boldsymbol{Z}_{\boldsymbol{n}}(\tilde{\gamma}, c) = \{t \in \tilde{\Gamma}^c : \pi_n(t) > 0 \wedge |\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| \leq c\|t\|^2\}$. For $t \in \boldsymbol{Z}_{\boldsymbol{n}}(\tilde{\gamma}, c)$ 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 $\exp\big(-\min(a,b)\big) \leq e^{-a} + e^{-b}$, by lemma A.5,

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

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

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

Finally the denominator in (24). If $\boldsymbol{Z}_{\boldsymbol{n}}^*(\bar{c}) = \{t : \pi_n(t) > 0 \wedge \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t) > -\bar{c}/2\}$ then

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

□

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

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

Proof. We have by lemma A.7,

$$\int \pi_n(t)\exp\big(Q_n(t)\big)dt - \pi_0 C_V = \int_{\Gamma_n} \pi_n(t)\exp\big(Q_n(t)\big)dt + o(1)$$
$$= C_V \int_{\Gamma_n} \pi_n(t)\Big(\exp\big(R_n(t)\big) - 1\Big)\phi_V(t)dt + C_V \int_{\Gamma_n} \big(\pi_n(t) - \pi_0\big)\phi_V(t)dt$$
$$- C_V \pi_0 \int_{\Gamma_n^c} \phi_V(t)dt + o(1). \quad (25)$$

The third right hand side term in (25) vanishes because γ_n increases to ∞, the first term is $o(1)$ by lemma A.8, and the second term is $o(1)$ by the continuity of π at θ_0. □

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

Proof. We have for ι^* defined in assumption E,

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

by assumption E. □

Lemma D.3. *Let $\mathscr{V}_N = \pi_0^2\iint ts^\intercal H(t,s)\phi_V(t)\phi_V(s)dtds$. Then $\int \pi_n(t)t\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\phi_V(t)dt \xrightarrow{d} N(0,\mathscr{V}_N)$.*

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

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

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

Lemma D.4. $\int \|\pi_n(t)t\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\|\Big(\exp\big(R_n(t)\big) - 1\Big)\phi_V(t)dt = o_p(1)$.

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. □

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

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

$$\int_{\Gamma_n} \|t\|^j|\pi_n(t)|\exp\big(|\beta_n\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)| + Q_n(t)\big)dt \le \exp(\beta_n\boldsymbol{S}_{\boldsymbol{n}}^*)\int_{\Gamma_n}\|t\|^j|\pi_n(t)|\exp\big(\beta_n t^\intercal Vt/4 + Q_n(t)\big)dt.$$

Since $\lim_{\|t\|\downarrow 0}\big(Q_n(t)/t^\intercal Vt\big) = -1/2$, it follows that for sufficiently large n, $Q_n(t) \le -t^\intercal Vt/4$ for all $t \in \Gamma_n$. Because $\beta_n = o(1)$ and $\boldsymbol{S}_{\boldsymbol{n}}^* = O_p(1)$ by lemma A.3, the stated result follows. □

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

Proof. Apply lemma A.6 with $c = \beta_n$ and $b = \tilde{\mathbf{S}}_{\boldsymbol{n}}(t)$, followed by lemma D.5. □

Lemma D.7. $\mathscr{N}_{\boldsymbol{n}} - \mathscr{B}_n \xrightarrow{d} N(0, C_V^2 \mathscr{V}_N)$.

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

$$\mathscr{N}_{\boldsymbol{n}} - \mathscr{B}_n = \frac{1}{\beta_n} \int \pi_n(t) t \{\exp(\beta_n \tilde{\mathbf{S}}_{\boldsymbol{n}}(t)) - 1\} \exp(Q_n(t)) dt =$$
$$\int \pi_n(t) t \tilde{\mathbf{S}}_{\boldsymbol{n}}(t) \exp(Q_n(t)) dt + \frac{1}{\beta_n} \int \pi_n(t) t \Big(\exp(\beta_n \tilde{\mathbf{S}}_{\boldsymbol{n}}(t)) - 1 - \beta_n \tilde{\mathbf{S}}_{\boldsymbol{n}}(t)\Big) \exp(Q_n(t)) dt. \quad (26)$$

The first right hand side term in (26) converges in distribution to the stated normal by lemmas D.3 and D.4. The last term in (26) is $O_p(\beta_n) = o_p(1)$ by lemma D.6. □

Lemma D.8. $\mathscr{D}_{\boldsymbol{n}} = \pi_0 C_V + o_p(1)$.

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

$$\mathscr{D}_{\boldsymbol{n}} = \int \pi_n(t) \exp(\beta_n \tilde{\mathbf{S}}_{\boldsymbol{n}}(t) + Q_n(t)) dt$$
$$= \int \pi_n(t) \exp(Q_n(t)) dt + \int \pi_n(t) \Big(\exp(\beta_n \tilde{\mathbf{S}}_{\boldsymbol{n}}(t)) - 1\Big) \exp(Q_n(t)) dt. \quad (27)$$

The first right hand side term in (27) is $\pi_0 C_V + o_p(1)$ by lemma D.1. The last term in (27) is $o_p(1)$ by lemma D.6. □

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 θ_0,

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

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

Lemma E.2.

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

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

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

Apply lemma E.1. □

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

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

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

E.2. **Bias Correction.** In this section the assumptions of theorem 7 are used; r^*, ρ_n are as defined in assumption I, take $\pi, \hat{\boldsymbol{\pi}}$ to equal $\pi_J, \hat{\boldsymbol{\pi}}_{\boldsymbol{J}}$, let $\hat{\boldsymbol{\pi}}_{\mathbf{0}} = \hat{\boldsymbol{\pi}}(\theta_0)$, and $\hat{\boldsymbol{\pi}}_{\boldsymbol{n}}(t) = \hat{\boldsymbol{\pi}}(\theta_0+t/\alpha_n)$. Let further $v(\theta) = \mathrm{vec}\big\{\big(-Q''(\theta)\big)^{-1}\big\}$ and $\hat{\boldsymbol{v}}(\theta) = \mathrm{vec}\big\{\big(-\hat{\boldsymbol{Q}}''(\theta)\big)^{-1}\big\}$.

Lemma E.4. *(i) $\hat{\boldsymbol{\pi}}_{\mathbf{0}} = \pi_0 + o_p(\alpha_n^{-r^*})$ and (ii) $\sup_{\|\theta-\theta_0\|\le\rho_n}\|\hat{\boldsymbol{\pi}}'(\theta) - \pi'(\theta)\| = o_p(\alpha_n^{-r^*})$.*

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

$$\pi'(\theta) = -\pi(\theta)Q'''(\theta)v(\theta)/2, \tag{28}$$

in a neighborhood of θ_0. Hence in such a neighborhood,

$$2\big(\hat{\boldsymbol{\pi}}'(\theta) - \pi'(\theta)\big) = -(\hat{\boldsymbol{\pi}}_{\mathbf{0}} - \pi_0)\hat{\boldsymbol{Q}}'''(\theta)\hat{\boldsymbol{v}}(\theta) - \pi_0\big\{\hat{\boldsymbol{Q}}'''(\theta)\hat{\boldsymbol{v}}(\theta) - Q'''(\theta)v(\theta)\big\}.$$

Apply part (i) and assumption I. □

Lemma E.5. $b_{q1}^* = \int\big(\pi_0 D_{Q3}(t) + D_{\pi1}(t)\big)t\phi_V(t)dt = 0$.

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

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

where the first equality follows from (28). □

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

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

$$\int tt^\intercal \hat{\boldsymbol{\pi}}'_{\boldsymbol{n}}(t)\phi_V(t)dt + \hat{\boldsymbol{\pi}}_{\mathbf{0}}\int D_{Q3}(t)t\phi_V(t)dt + \alpha_n^{-1}\int tt^\intercal \hat{\boldsymbol{\pi}}'_{\boldsymbol{n}}(t^*)D_{Q3}(t)\phi_V(t)dt,$$

which by lemma E.5 is equal to

$$\int tt^\intercal\big(\hat{\boldsymbol{\pi}}'_{\boldsymbol{n}}(t) - \pi'_0\big)\phi_V(t)dt + (\hat{\boldsymbol{\pi}}_{\mathbf{0}} - \pi_0)\int D_{Q3}(t)t\phi_V(t)dt$$
$$+ \alpha_n^{-1}\int tt^\intercal\big(\hat{\boldsymbol{\pi}}'_{\boldsymbol{n}}(t^*) - \pi'_0\big)D_{Q3}(t)\phi_V(t)dt + \alpha_n^{-1}\int D_{\pi 1}(t)D_{Q3}(t)t\phi_V(t)dt,$$

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. □

Appendix F. Uniform Inference

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

Lemma F.1. $\hat{\mathbb{G}} \xrightarrow{w} \mathbb{G}$ *in* $\mathscr{L}^\infty(\mathscr{T}_1, \mathscr{T}_2, \ldots)$.

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

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

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

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

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

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

Proof. Follows from lemma F.1 combined with the consistency of $\hat{\boldsymbol{V}}$. □

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

$$\int \psi(t) \exp\big(c_1\hat{\mathbb{G}}(t) - c_2 t^\intercal \hat{\boldsymbol{V}} t/2\big) dt \xrightarrow{d} \int \psi(t) \exp\big(c_1\mathbb{G}(t) - c_2 t^\intercal V t/2\big) dt.$$

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

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

$$\int t \exp\big(\beta_n^{4/3}(\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2)\big) dt \Big/ \int \exp\big(\beta_n^{4/3}(\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2)\big) dt \xrightarrow{d} \operatorname*{argmax}_{t \in \Re^d} \tilde{\mathbb{G}}(t).$$

Proof. Let $\hat{\boldsymbol{\psi}}_{\boldsymbol{j}}(t) = t^j \exp\big(\beta_n^{4/3}(\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2)\big)$ and $\psi_j(t) = t^j \exp\big(\beta_n^{4/3}(\mathbb{G}(t) - t^\intercal V t/2)\big)$. Then $\hat{\boldsymbol{\psi}}_{\mathbf{0}}, \hat{\boldsymbol{\psi}}_{\mathbf{1}}, \psi_0, \psi_1 \in \mathbb{L}(\Re^d, \mathbb{B}, \mu)$ and $\big(\hat{\boldsymbol{\psi}}_{\mathbf{0}}, \hat{\boldsymbol{\psi}}_{\mathbf{1}}\big) \xrightarrow{w} (\psi_0, \psi_1)$. Repeat the arguments of lemma C.1 following (20). □

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

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

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

$$\left\| \beta_n^{-1} \int t \exp\big(\beta_n^{3/2} \hat{\mathbb{G}}(t/\beta_n)\big) \phi_{\hat{V}}(t) dt - \sqrt{\beta_n} \int t \hat{\mathbb{G}}(t/\beta_n) \phi_{\hat{V}}(t) dt \right\| \\ \leq \beta_n C_V^{-1} \int \|t\| \exp\big(\sqrt{\beta_n} |\hat{\mathbb{G}}(t/\beta_n)| - t^\intercal \hat{V} t/2\big) dt,$$

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

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

$$\sqrt{\beta_n} \int t \hat{\mathbb{G}}(t/\beta_n) \phi_{\hat{V}}(t) dt \xrightarrow{d} N(0, \mathscr{V}). \tag{29}$$

Proof. The left hand side in (29) has the same distribution as $\int t \hat{\mathbb{G}}(t) \phi_{\hat{V}}(t) dt$, which by lemma F.1, the assumption that $\hat{\boldsymbol{V}} = V + o_p(1)$, and the conditions on $\hat{\boldsymbol{H}}$, converges in distribution to

$$\int t \mathbb{G}(t) \phi_V(t) dt, \tag{30}$$

Because $\tilde{\boldsymbol{S}}_{\boldsymbol{n}} \xrightarrow{w} \mathbb{G}$, (30) is also the limit of $\int t \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) \phi_V(t) dt$, which lemma D.3 established to have a limiting $N(0, \mathscr{V})$ distribution. □

Appendix G. Results Specific to the Maximum Score Case

G.1. $V, H, \mathscr{V}$.

Lemma G.1. $V = -2\mathbb{E}\big[\boldsymbol{z_i} \boldsymbol{z_i}^\intercal p_a(\boldsymbol{z_i}^\intercal \theta_0, \boldsymbol{z_i}) f(\boldsymbol{z_i}^\intercal \theta_0 | \boldsymbol{z_i})\big]$.

Proof. Let ∂_θ denote a partial derivative with respect to θ. In view of the definition of V, consider

$$\begin{aligned} \partial_{\theta\theta^\intercal} \mathbb{E} \boldsymbol{g_i}(\theta) = \partial_{\theta\theta^\intercal} \mathbb{E}\big[\big(2p(\boldsymbol{a_i}, \boldsymbol{z_i}) - 1\big) I(\boldsymbol{a_i} \leq \boldsymbol{z_i}^\intercal \theta)\big] &= \partial_{\theta\theta^\intercal} \mathbb{E}\left[\int_{-\infty}^{\boldsymbol{z_i}^\intercal \theta} \big(2p(a, \boldsymbol{z_i}) - 1\big) f(a|\boldsymbol{z_i}) da\right] \\ &= \partial_{\theta^\intercal} \mathbb{E}\big[\boldsymbol{z_i}\big(2p(\boldsymbol{z_i}^\intercal \theta, \boldsymbol{z_i}) - 1\big) f(\boldsymbol{z_i}^\intercal \theta | \boldsymbol{z_i})\big] \\ &= \mathbb{E}\big[\boldsymbol{z_i} \boldsymbol{z_i}^\intercal \big\{2p_a(\boldsymbol{z_i}^\intercal \theta, \boldsymbol{z_i}) f(\boldsymbol{z_i}^\intercal \theta | \boldsymbol{z_i}) + \big(2p(\boldsymbol{z_i}^\intercal \theta, \boldsymbol{z_i}) - 1\big) f'(\boldsymbol{z_i}^\intercal \theta | \boldsymbol{z_i})\big\}\big], \end{aligned}$$

which at θ_0 equals $2\mathbb{E}\big[\boldsymbol{z_i} \boldsymbol{z_i}^\intercal p_a(\boldsymbol{z_i}^\intercal \theta_0, \boldsymbol{z_i}) f(\boldsymbol{z_i}^\intercal \theta_0 | \boldsymbol{z_i})\big] = -V$. □

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

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

Lemma G.3. $H(t,s) = \mathbb{E}\big[\big|M(z_i^\intercal t, z_i^\intercal s, 0)\big| f(z_i^\intercal \theta_0 | z_i)\big]$.

Proof. Let

$$\mathscr{H}(t,s) = \lim_{\alpha\to\infty} \mathbb{E}\left[\alpha \int_{z_i^\intercal\theta_0}^{z_i^\intercal\theta_0 + \min(z_i^\intercal t, z_i^\intercal s)/\alpha} f(a|z_i)da\right] = \mathbb{E}\big[\min(z_i^\intercal t, z_i^\intercal s) f(z_i^\intercal\theta_0|z_i)\big].$$

by the dominated convergence theorem. Thus, noting that $(2y_i - 1)^2 = 1$, we have

$$\begin{aligned} H(t,s) &= \mathscr{H}(t,s) - \mathscr{H}(t,0) - \mathscr{H}(0,s) + \mathscr{H}(0,0) \\ &= \mathbb{E}\big[\{\min(z_i^\intercal t, z_i^\intercal s) - \min(z_i^\intercal t, 0) - \min(z_i^\intercal s, 0)\} f(z_i^\intercal\theta_0|z_i)\big] = \mathbb{E}\big[\big|M(z_i^\intercal t, z_i^\intercal s, 0)\big| f(z_i^\intercal\theta_0|z_i)\big]. \end{aligned}$$

□

Lemma G.4. *If for some* $\jmath = 1, \ldots, d$, $\mathbb{P}[z_{i\jmath} = 0] = 0$, *then* (11) *holds.*

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

$$\int_{t_1}^\infty \int s_j \phi_{\Sigma_{1j}^{-1}}(s_1, s_j) ds_j ds_1 = \sigma_{1j}\phi(t_1/\sigma_1)/\sigma_1, \tag{31}$$

$$\int_{t_1}^\infty s_1 \phi(s_1/\sigma_1) ds_1/\sigma_1 = \sigma_1 \phi(t_1/\sigma_1), \tag{32}$$

$$\iiint \int_{t_1}^\infty t_\ell t_1 s_j \phi_{\Sigma_{1j}^{-1}}(s_1, s_j)\phi_{\Sigma_{1\ell}^{-1}}(t_1, t_\ell) ds_1 ds_j dt_1 dt_\ell = \sigma_{1j}\sigma_{1\ell}/4\sigma_1\sqrt{\pi}, \tag{33}$$

$$\iint t_\ell s_j \min(t_1, s_1)\phi_{\Sigma^{-1}}(s)\phi_{\Sigma^{-1}}(t) ds dt = \sigma_{1j}\sigma_{1\ell}/2\sigma_1\sqrt{\pi}, \tag{34}$$

$$\iint t s^\intercal \min(t_1, s_1)\phi_{\Sigma^{-1}}(s)\phi_{\Sigma^{-1}}(t) ds dt = \Sigma e_1 e_1^\intercal \Sigma / 2\sqrt{\pi e_1^\intercal \Sigma e_1}, \tag{35}$$

where e_1 is the first unit vector. The left hand side in (31) 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 (32) is similar, but easier to establish, and the left hand side in (33) is by (31) 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 (34) then follows by applying (33) twice and (35) by repeatedly using (34).

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

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

Then by substitution of $\tilde{t} = \boldsymbol{A_i} t$ and $\tilde{s} = \boldsymbol{A_i} s$,

$$\begin{aligned}
\mathscr{V} = \iint ts^\intercal H(t,s)\phi_V(t)\phi_V(s)dtds &= \mathbb{E}\left[\boldsymbol{f_i} \iint ts^\intercal |M(\boldsymbol{z_i^\intercal} t, \boldsymbol{z_i^\intercal} s, 0)|\phi_V(t)\phi_V(s)dtd_s\right] \\
&= \mathbb{E}\left[\boldsymbol{f_i A_i^{-1}} \iint \tilde{t}\tilde{s}^\intercal |M(\tilde{t}_1, \tilde{s}_1, 0)|\phi_{\boldsymbol{V_i}}(t)\phi_{\boldsymbol{V_i}}(s)d\tilde{t}d\tilde{s}\big(\boldsymbol{A_i^\intercal}\big)^{-1}\right] \\
= \mathbb{E}\left[\boldsymbol{f_i A_i^{-1}} \iint \tilde{t}\tilde{s}^\intercal \min(\tilde{t}_1, \tilde{s}_1)\phi_{\boldsymbol{V_i}}(\tilde{t})\phi_{\boldsymbol{V_i}}(\tilde{s})d\tilde{t}d\tilde{s}\big(\boldsymbol{A_i^\intercal}\big)^{-1}\right] &= \frac{1}{2\sqrt{\pi}}\mathbb{E}\left[\boldsymbol{f_i}\frac{\boldsymbol{A_i^{-1}V_i^{-1}}e_1 e_1^\intercal \boldsymbol{V_i^{-1}}(\boldsymbol{A_i^\intercal})^{-1}}{\sqrt{e_1^\intercal \boldsymbol{V_i^{-1}} e_1}}\right] \\
&= \frac{1}{2\sqrt{\pi}} V^{-1}\mathbb{E}\left[\frac{\boldsymbol{z_i z_i^\intercal}}{\sqrt{\boldsymbol{z_i^\intercal} V^{-1} \boldsymbol{z_i}}} f(\boldsymbol{z_i^\intercal}\theta_0 | \boldsymbol{z_i})\right] V^{-1},
\end{aligned}$$

where the penultimate inequality follows from (35) and the last from the fact that $\boldsymbol{A_i^{-1}V_i^{-1}} = V^{-1}\boldsymbol{A_i^\intercal}$ and that $\boldsymbol{A_i^\intercal} e_1 = \boldsymbol{z_i}$. □

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

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

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

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

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

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

Lemma G.7. $\mathbb{E}\boldsymbol{F}_{ni}^2 = O(1)$.

Proof. By the law of iterated expectations,

$$\mathbb{E}\boldsymbol{F}_{ni}^2 = \tilde{\alpha}_n \mathbb{E}\Big[\mathbb{P}\big[\boldsymbol{z}_i^\intercal(\theta_0 + t_n(\boldsymbol{z}_i)/\tilde{\alpha}_n) < \boldsymbol{a}_i \leq \boldsymbol{z}_i^\intercal(\theta_0 + s_n(\boldsymbol{z}_i)/\tilde{\alpha}_n)\big|\boldsymbol{z}_i\big]\Big] \\ \leq \mathbb{E}\Big[\sup_s f_{\boldsymbol{a}|\boldsymbol{z}}(s|\boldsymbol{z}_i)\|\boldsymbol{z}_i\|\|t_n(\boldsymbol{z}_i) + s_n(\boldsymbol{z}_i)\|\Big].$$

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

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

Proof. Follows from $n/\tilde{\alpha}_n \to \infty$. □

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

Proof. Follows from

$$E\big[\tilde{\alpha}_n\big(\boldsymbol{g}_i(\theta_0 + t/\tilde{\alpha}_n) - \boldsymbol{g}_i(\theta_0 + s/\tilde{\alpha}_n)\big)^2\big] \leq \tilde{\alpha}_n \mathbb{E}\big[\mathbb{P}[\boldsymbol{z}_i^\intercal(\theta_0 + t/\tilde{\alpha}_n) < \boldsymbol{a}_i \leq \boldsymbol{z}_i^\intercal(\theta_0 + s/\tilde{\alpha}_n)|\boldsymbol{z}_i]\big] \\ + \tilde{\alpha}_n \mathbb{E}\big[\mathbb{P}[\boldsymbol{z}_i^\intercal(\theta_0 + s/\tilde{\alpha}_n) < \boldsymbol{a}_i \leq \boldsymbol{z}_i^\intercal(\theta_0 + t/\tilde{\alpha}_n)|\boldsymbol{z}_i]\big] \leq 2E\Big[\sup_s f_{\boldsymbol{a}|\boldsymbol{z}}(s|\boldsymbol{z}_i)\|\boldsymbol{z}_i\|\Big]\|t - s\|. \qquad \square$$

Lemma G.10. $\tilde{\boldsymbol{S}}_n \xrightarrow{w} \mathbb{G}$ *in* $\mathscr{L}^\infty(\mathscr{T}_1, \mathscr{T}_2, \ldots)$, *where* $\mathbb{G}$ *is a Gaussian process with covariance kernel* H *such that* $H(t,t) = O(\|t\|)$ *as* $\|t\| \to \infty$.

Proof. Recall that

$$\tilde{\alpha}_n E\big[\boldsymbol{g}_i(\theta_0 + t/\tilde{\alpha}_n)\boldsymbol{g}_i(\theta_0 + s/\tilde{\alpha}_n)\big] \to H(s,t) = \mathbb{E}\left[f(\boldsymbol{z}_i^\intercal\theta_0|\boldsymbol{a}_i)|M(\boldsymbol{z}_i^\intercal t, \boldsymbol{z}_i^\intercal s, 0)|\right].$$

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 $\mathscr{F}_n$ is a VČ class of which the VČ index is bounded by $2d + 5$. Therefore, for $0 < \varepsilon < 1$,

$$\mathscr{N}\left(\varepsilon\|\mathscr{F}\|_{\mathscr{Q},2}, \mathscr{F}_n, L_2(\mathscr{Q})\right) \leq C\varepsilon^{2-2(2d+5)}$$

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. □

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. □

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

$$\sqrt{\frac{n}{\alpha_n}}(\hat{\boldsymbol{\theta}} - \theta_0) = \sqrt{\frac{n}{\alpha_n}} \frac{\int \pi(\theta)(\theta - \theta_0) \exp\{\alpha_n^2 (\boldsymbol{S_n}(\theta) + Q(\theta))\} d\theta}{\int \pi(\theta) \exp\{\alpha_n^2 (\boldsymbol{S_n}(\theta) + Q(\theta))\} d\theta}$$
$$= \sqrt{\frac{n}{\alpha_n^3}} \frac{\int \pi_n(t) t \exp\big(c_\alpha^3 \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big) dt}{\int \pi_n(t) \exp\big(c_\alpha^3 \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + Q_n(t)\big) dt} = \frac{1}{c_\alpha^3} \frac{\int t \exp\big(c_\alpha^3 \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big) \phi_V(t) dt}{\int \exp\big(c_\alpha^3 \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\big) \phi_V(t) dt} + o_p(1),$$

by lemma B.2. Multiply both sides by c_α, then apply lemma A.1 and the continuous mapping theorem.

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

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

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

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

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, $\mathbb{N}_{\tilde{\Gamma}}/\mathbb{D}_{\tilde{\Gamma}} \xrightarrow{d} \operatorname{argmax}_{t \in \tilde{\Gamma}} \tilde{\mathbb{G}}(t)$ for any $\tilde{\gamma} > 0$. Let $\tilde{\gamma} \to \infty$. □

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

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

$$\mathscr{B}^*_n = \frac{C_V}{\beta_n} \sum_{p=0}^{q} \frac{1}{p!} \int \pi_n(t) t \big(R_n(t)\big)^p \phi_V(t) dt, \tag{36}$$

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

$$R_n(t) = \sum_{\delta=1}^{q} \frac{D_{Q,\delta+2}(t)}{\alpha_n^\delta} + \frac{D^*_{Q\Delta n}(t)}{\alpha_n^q}, \qquad \pi_n(t) = \sum_{\ell=0}^{q} \frac{D_{\pi\ell}(t)}{\alpha_n^\ell} + \frac{D^*_{\pi q n}(t)}{\alpha_n^q}. \tag{37}$$

Note that $\sup_{t \in \Gamma_n \setminus \{0\}} \big(|D^*_{Q\Delta n}(t)|/\|t\|^\Delta + |D^*_{\pi q n}(t)|/\|t\|^q\big) = o(1)$ by the continuity of the Δ–th derivative of Q and the q–th derivative of π at θ_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 (36) using (37) involving the remainder terms in the Taylor expansions of (37) are $o\big(\alpha_n^{-q} \beta_n^{-1}\big)$.

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

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

Using lemma E.3, we obtain

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

where

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

and $b_{q\tau}^* = C_V \sum_{r=0}^{\tau} b_{qr,\tau-r}$ Finally, $b_{q\tau}^* = 0$ for even–valued τ because $b_{qr\ell} = 0$ whenever r and ℓ 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. □

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

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

The first few lines of the proof of theorem 5 establish that the difference between $\mathscr{B}_n$ and $\mathscr{B}_n^*$ 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 π being nonrandom, so $\mathscr{B}_{\boldsymbol{Jn}} - \mathscr{B}_{\boldsymbol{Jn}}^* = o_p(\alpha_n^{-q}\beta_n^{-1})$.

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

$$\mathscr{B}_{\boldsymbol{Jn}}^* = \frac{C_V}{\beta_n} \sum_{p=0}^{q} \int \hat{\boldsymbol{\pi}}_{\boldsymbol{n}}(t) \Big(\sum_{\delta=1}^{q} \frac{D_{Q,\delta+2}(t)}{\alpha_n^\delta} \Big)^p t\phi_V(t)dt + o_p(\alpha_n^{-q}\beta_n^{-1}). \tag{38}$$

Part (i) of the theorem then follows by applying lemma E.6 to the first right hand side term in (38). For (ii) the first right hand side term in (38) equals

$$\frac{C_V}{\alpha_n^3\beta_n}\int \hat{\boldsymbol{\pi}}_{\boldsymbol{n}}(t)\big(D_{Q5}(t)+D_{Q3}(t)D_{Q4}(t)/2+D_{Q3}^3(t)/6\big)t\phi_V(t)dt \\ +\frac{C_V}{\alpha_n^2\beta_n}\int \hat{\boldsymbol{\pi}}_{\boldsymbol{n}}(t)\big(D_{Q4}(t)+D_{Q3}^2(t)\big)t\phi_V(t)dt. \quad (39)$$

By lemma E.4 the first term in (39) is

$$\frac{C_V\pi_0}{\alpha_n^3\beta_n}\int \big(D_{Q5}(t)+D_{Q3}(t)D_{Q4}(t)/2+D_{Q3}^3(t)/6\big)t\phi_V(t)dt+o_p(\alpha_n^{-3}\beta_n^{-1})=O_p(\alpha_n^{-3}\beta_n^{-1}). \quad (40)$$

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 (39) is

$$\frac{C_V\pi_0}{\alpha_n^2\beta_n}\int \big(D_{Q4}(t)+D_{Q3}^2(t)\big)t\phi_V(t)dt+\frac{C_V}{\alpha_n^3\beta_n}\int \big(D_{Q4}(t)+D_{Q3}^2(t)\big)\phi_V(t)tt^{\intercal}dt\ \pi'(\theta_0) \\ +o_p(\alpha_n^{-3}\beta_n). \quad (41)$$

The first term in (41) is zero because the normal distribution is even. The second term in (40) is $O(\alpha_n^{-3}\beta_n^{-1})$, so part (ii) of the theorem is satisfied with $\mathscr{B}_J$ equal to the sum of the first term in (40) and the second term in (41). □

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

$$\boldsymbol{\Psi}_{\boldsymbol{n}} = \int t\exp\big(c_\alpha^4\hat{\mathbb{G}}(t)-c_\alpha^4 t^{\intercal}\hat{\boldsymbol{V}}t/2\big)dt/\int \exp\big(c_\alpha^4\hat{\mathbb{G}}(t)-c_\alpha^4 t^{\intercal}\hat{\boldsymbol{V}}t/2\big)dt$$

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

$$c_\alpha\boldsymbol{\Psi}_{\boldsymbol{n}} \xrightarrow{d} c_\alpha\frac{\int t\exp\big(c_\alpha^{9/2}\mathbb{G}(t)-c_\alpha^6 t^{\intercal}Vt/2\big)dt}{\int \exp\big(c_\alpha^{9/2}\mathbb{G}(t)-c_\alpha^6 t^{\intercal}Vt/2\big)dt}=\frac{1}{c_\alpha^2}\frac{\int t\exp\big(c_\alpha^{9/2}\mathbb{G}(t/c_\alpha^3)\big)\phi_V(t)dt}{\int \exp\big(c_\alpha^{9/2}\mathbb{G}(t/c_\alpha^3)\big)\phi_V(t)dt}, \quad (42)$$

by substitution of $\tilde{t} = c_\alpha^3 t$ and replacing $\tilde{t}$ with t. The right hand side in (42) has the same distribution as the limit distribution in (6) since $c_\alpha^{9/2}\mathbb{G}(\cdot/c_\alpha^3)$ has the same properties as $c_\alpha^3\mathbb{G}(\cdot)$.

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

$$\boldsymbol{\Psi}_{\boldsymbol{n}} = \frac{1}{\beta_n}\frac{\int t\exp\big(\beta_n^{3/2}\hat{\mathbb{G}}(t/\beta_n)\big)\phi_{\hat{\boldsymbol{V}}}(t)dt}{\int \exp\big(\beta_n^{3/2}\hat{\mathbb{G}}(t/\beta_n)\big)\phi_{\hat{\boldsymbol{V}}}(t)dt}.$$

Apply lemma F.5 followed by lemma F.6. □

APPENDIX I. COMPUTATION

I.1. **Maximum Score.** 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 θ, z_i without its j–th element. Let $B_{ij} = B_{ij}(\theta_{-j}) = (a_i - z_{i,-j}^{\intercal}\theta_{-j})/z_{ij}$ if $z_{ij} \neq 0$ and be arbitrarily defined if $z_{ij} = 0$. Let further $S_{nj}(\theta_j) = \sum_{i=1}^{n}(2y_i - 1)\big(I(z_{ij} > 0)I(\theta_j \geq B_{ij}) + I(z_{ij} < 0)I(\theta_j \leq B_{ij})\big)/\sqrt[3]{n}$. Let $\mathscr{I}(\theta_j|\theta_{-j}) = \int_{-\infty}^{\theta_j} \pi_j(\theta_j|\theta_{-j}) \exp\big(S_{nj}(\theta_j)\big)d\theta_j$.

(1) Compute and renormalize probit estimates.
(2) Set $j = 1$.
(3) Compute B_{ij} for all i for which $z_{ij} \neq 0$ and sort all n^* of them in ascending order, i.e. $B_{1j} \leq B_{2j} \leq \ldots \leq B_{n^*j}$.
(4) Compute $S_{nj}(-\infty)$.
(5) Note that for all $\theta_j \leq B_{1j}$, $\mathscr{I}_{nj}(\theta_j) = \exp\big(S_{nj}(-\infty)\big)\Pi_j(\theta_j|\theta_{-j})$, where $\Pi_j(\theta_j|\theta_{-j})$ is the integrated conditional prior.
(6) For all $\theta_j \in (B_{1j}, B_{2j}]$, $\mathscr{I}_{nj}(\theta_j) = \mathscr{I}_{nj}(B_{1j}) + \exp\big\{S_{nj}\big((B_{1j} + B_{2j})/2\big)\big\}\big(\Pi_j(\theta_j|\theta_{-j}) - \Pi_j(B_{1j}|\theta_{-j})\big)$.
(7) Repeat this process until $\mathscr{I}_{nj}(\theta_j)$ is determined for all values of θ_j.
(8) Set $\mathscr{I}^*_{nj}(\theta_j) = \mathscr{I}_{nj}(\theta_j)/\mathscr{I}_{nj}(\infty)$, which is a distribution function.
(9) Draw a random number r from a uniform distribution.
(10) Compute $i^* = \max\{i : \mathscr{I}^*_{nj}(B_{ij}) \leq r\}$.
(11) Set $\theta_j = \theta_j(r) = \Pi^{-1}\big(\Pi(B_{i^*j}) + \mathscr{I}_{nj}(\infty)r + \mathscr{I}_{nj}(B_{ij})|\theta_{-j}\big)$.
(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 (17). Suppose that consistent estimates $\hat{V}$, $\hat{H}$ are available, where $\hat{H}$ is moreover a positive semidefinite covariance kernel.[21] One can then draw random numbers $t_1, \ldots, t_{T^*}$ from the multivariate normal $(\phi_{\beta_{n2}\hat{V}})$ and define

$$\hat{\zeta}_j = \frac{\sum_{s=1}^{T^*} t_s \exp(\beta_{n1} G_{js})}{\sum_{s=1}^{T^*} \exp(\beta_{n1} G_{js})}, \tag{43}$$

where $G_{j1}, \ldots, G_{jT^*}$ are drawn independently across j and have for all given $j = 1, \ldots, \mathbb{J}$ a joint normal distribution with $\mathrm{Cov}[G_{js}, G_{js^*}] = \hat{H}(t_s, t_{s^*})$, for all s, s^*. In the limit (n and T^*), each $\hat{\zeta}_j$ has

[21] In practice one only needs to ensure that the $T^* \times T^*$ matrix with (i,j)–element $\hat{H}(t_i, t_j)$ is positive semidefinite.

the same distribution as $\mathbf{\Psi_n}$ defined in (17). Note that the $\boldsymbol{t_s}$ draws are made only once. The draws $\hat{\boldsymbol{\zeta}}_{\mathbf{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\mathbb{J}}$ can then be used to construct confidence intervals.

I.3. **Derivatives of** Q**.** For the maximum score case, $Q(\theta) = \mathbb{E}\big[(2\boldsymbol{y_i} - 1)\big(I(\boldsymbol{a_i} \leq \boldsymbol{z_i}^\intercal\theta) - I(\boldsymbol{a_i} \leq \boldsymbol{z_i}^\intercal\theta_0)\big)\big]$. If derivatives of Q must be estimated, we recommend using the desired derivative of

$$\hat{\boldsymbol{Q}}'(\theta) = \frac{1}{n\tilde{b}} \sum_{i=1}^{n} \tilde{k}\left(\frac{\boldsymbol{z_i}^\intercal\theta - \boldsymbol{a_i}}{\tilde{b}}\right)(2\boldsymbol{y_i} - 1)\boldsymbol{z_i},$$

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