Integrated Score Estimation

Sung Jae Jun[1] Joris Pinkse[2] Yuanyuan Wan[3]
CAPCP, Department of Economics Department of Economics
The Pennsylvania State University The University of Toronto

We study the properties of the integrated score estimator (ISE), which is the Laplace version of Manski's maximum score estimator (MMSE). The ISE belongs to a class of estimators whose basic asymptotic properties were studied in Jun, Pinkse, and Wan (2015). Here, we establish that the MMSE, or more precisely $\sqrt[3]{n}|\hat{\boldsymbol{\theta}}_M - \theta_0|$, (locally first order) stochastically dominates the ISE under the conditions necessary for the MMSE to attain its $\sqrt[3]{n}$ convergence rate and that the ISE has the same convergence rate as Horowitz's smoothed maximum score estimator (SMSE) under somewhat weaker conditions. An implication of the stochastic dominance result is that the confidence intervals of the MMSE are for any given coverage rate wider than those of the ISE, provided that the input parameter α_n is not chosen too large. Further, we introduce an inference procedure that is not only rate adaptive as established in Jun, Pinkse, and Wan (2015), but also uniform in the choice of α_n. We propose three different first order bias elimination procedures and we discuss the choice of input parameters. We develop a computational algorithm for the ISE based on the Gibbs sampler and we examine implementational issues in detail. We argue in favor of normalizing the norm of the parameter vector as opposed to fixing one of the coefficients. Finally, we evaluate the computational efficiency of the ISE and the performance of the ISE and the proposed inference procedure in an extensive Monte Carlo study.

Key Words: Binary Choice Model, Semiparametric Maximum Score, Laplace Estimation, Efficiency, Uniform Inference.

[1](corresponding author) sjun@psu.edu; Department of Economics, The Pennsylvania State University, 303 Kern Graduate Building, University Park 16802; This paper is based on research supported by NSF grant SES–0922127. We thank the Human Capital Foundation (www.hcfoundation.ru), especially Andrey P. Vavilov, for their support of CAPCP (http://capcp.psu.edu) at the Pennsylvania State University. We thank Don Andrews, Miguel Delgado, Jeremy Fox, Bo Honoré, Joel Horowitz, Roger Koenker, Arthur Lewbel, Runze Li, Oliver Linton, Peter Robinson, Neil Wallace, Haiqing Xu, Vicky Zinde–Walsh, six anonymous referees, and numerous departmental seminar and conference participants for helpful suggestions.

[2]joris@psu.edu
[3]yuanyuan.wan@utoronto.ca

1 Motivation

The MMSE proposed in Manski (1975, M75) is an intuitive and appealing estimator for the standard single equation binary choice model, which has been extended to multinomial choice and fixed effects panel data models (Manski, 1987). Its principal attraction is that, unlike the probit estimator, it neither requires the error in the latent variable equation to have a known distribution nor it to be independent of the regressors: heteroskedasticity of unknown form is permitted. This added level of generality comes at a cost, however: the MMSE converges at a $\sqrt[3]{n}$ rate, is set–valued,[4] has a Chernoff rather than a normal limit distribution, is difficult to compute (Pinkse, 1993; Florios and Skouras, 2008), and cannot be bootstrapped in its standard form (Abrevaya and Huang, 2005).[5]

Horowitz (1992, H92) has shown that if the MMSE objective function, which is a step function, is smoothed out then under additional smoothness conditions on the distributions of the model variables the convergence rate of the SMSE can exceed that of the MMSE and the SMSE will have a normal limit distribution. Readers are referred to Chen and Zhang (2014) for the local polynomial analog and Hong, Mahajan, and Nekipelov (2010) for an alternative based on numerical derivatives. However, as Chamberlain (1986) has shown, the (regular) parametric $\sqrt{n}$ rate is not attainable unless additional restrictions are imposed like the independence of errors and regressors (Powell, Stock, and Stoker, 1989; Klein and Spady, 1993; Ichimura, 1993) or the availability of a 'special regressor' (Lewbel, 1998, 2000). Moreover, Pollard (1993) has shown that $\sqrt[3]{n}$ is the best rate achievable under the smoothness conditions required for the MMSE: the SMSE converges more slowly than the MMSE if the additional conditions needed for the SMSE are not satisfied.

The ISE uses the idea of Chernozhukov and Hong (2003, CH03), *pretends* that the MMSE objective function (times an input parameter α_n^2/n) is a loglikelihood function, and computes a *quasi*–Bayesian posterior mean. Jun, Pinkse, and Wan (2015, JPW15) provide generic results for $\sqrt[3]{n}$–consistent estimators that we here adapt to establish that, depending on the pseudo prior π and the rate of α_n, the ISE has the same convergence rate as the MMSE or SMSE under similar conditions. Indeed, depending on the choice of α_n and the degree of smoothness (e.g. of the limit objective function), the limit distribution can be *(a)* Chernoff, *(b)* a ratio of integrals over Gaussian processes, or *(c)* normal. This result thus bridges the gap between the Chernoff limit distribution of the MMSE and the normal limit distribution of the SMSE, which is related with Hong, Mahajan, and Nekipelov (2010)'s finding that maximizing the MMSE objective function by numerical derivative methods can also lead to a limit distribution that is a hybrid of a normal and a Chernoff. After our discussion of asymptotics for fixed choices of input parameters, we then conduct an in–depth analysis of various issues that have not been studied before, including uniform inference and efficiency.

We propose an inference procedure for the ISE which (as mentioned before) is not only rate–adaptive as established in JPW15[6] but also *uniformly valid* with respect to the choice of the input parameter α_n (provided

[4] The MMSE objective function is not smooth: it consists of polytope–shaped plateaus. Since the parameter of interest is point identified, all maximizers are *asymptotically* equivalent and the theory does not provide guidance about which one to use. However, in finite samples, the set of maximizers is a union of polytopes.

[5] Although there are bootstrap procedures modified for the MMSE (e.g. Lee and Pun, 2006; Patra, Seijo, and Sen, 2011), their practical applicability is limited because of the computational challenges of the MMSE.

[6] Readers are also referred to an early working paper version Jun, Pinkse, and Wan (2009).

that it does not diverge too slowly), which is not available for other methods. This is an alternative to first choosing between MMSE and SMSE and if one chooses the SMSE then being careful to choose a bandwidth that is compatible with the limiting normal distribution (assuming that the additional smoothness conditions are satisfied). Indeed, we show that if one uses the SMSE asymptotic distribution for a bandwidth that tends to zero for a fixed sample size n then both the size and the power of (SMSE) t tests goes to zero. So with the ISE one has to guard against excessive smoothing whereas with the SMSE one has to guard both against excessive and insufficient smoothing to get reliable inference results. Our procedure, however, does not adapt to the degree of smoothness q of the limit objective function.[7]

The above discussion does not dictate a particular choice of input parameters α_n, π. In fact, as we will argue later, there does not exist a theoretically optimal pair (α_n, π) even if the degree of smoothness q is known. This is analogous to the problem with the kernel regression estimator that if the unknown regression function is twice continuously differentiable then the optimal bandwidth can be determined for a given second order kernel, but for higher order kernels no optimal bandwidth exists. This issue arises equally for the MMSE and SMSE. Fortunately, performance of the ISE appears to be fairly robust over a wide range of input parameter choices and, unlike with the SMSE, the choice of α_n does not depend on the scaling of the regressors. We therefore make a simple specific recommendation that is straightforward to implement: choose $\alpha_n = 1.5\sqrt[3]{n}$ and choose $\pi(\theta) \propto \left\{1 + \|\theta\|^2\right\}^{-(1+d)/2}$, where $d+1$ is the dimension of the regressor vector, and correct for the asymptotic bias in the inference procedure.

We further show that the ISE is stochastically dominated by the MMSE under the conditions that Kim and Pollard (1990, KP90) spelled out to obtain the $\sqrt[3]{n}$–consistency of the MMSE. Under the KP90 conditions $\sqrt[3]{n}$ is the best attainable *rate*, but we show here that the Chernoff limit distribution of the MMSE is then not the best attainable limit *distribution*. This result complements the result established in H92 (and shown here to be shared with the ISE) that the SMSE (and the ISE) converge faster under *additional* conditions.

We compute our estimator using a simple Gibbs sampling procedure. Since the ISE is the quasi posterior mean of a density function that is proportional to the product of a (chosen) pseudo prior and the exponential of a step function, a draw from the conditional quasi posterior distribution of one coefficient given the remaining coefficients is simple and relatively inexpensive. Indeed, computing the SMSE by simulated annealing was considerably slower in every scenario than computing the ISE using the Gibbs algorithm.[8] The uniform inference procedure is also based on simulations and entails little more than taking draws from a multivariate normal and summing.

We provide simulation results to highlight a number of features. In section 3 we complement our theoretical results by simulations that illustrate graphically (see figure 1) and powerfully how the choice of input parameter affects asymptotic efficiency (in terms of a comparison of the limit distribution functions) in the case in which only the MMSE conditions are known to hold.

[7] Kotlyarova and Zinde-Walsh (2006) propose a smoothness–adaptive estimation procedure for a different estimation problem, which we discuss later.

[8] We should point out that the simulated annealing routine that we use to compute the SMSE may not be optimal, but the same is true for our Gibbs sampling routine to compute the ISE. All computing times increase roughly linearly both in the sample size and in the number of unknowns. See table 3 for more details.

The remaining numerical results are contained in section 7. There we study the behavior of our estimators in three different designs: standard probit, probit with heteroskedasticity, and a homoskedastic binary choice model in which the error term follows a Laplace (symmetric exponential) distribution, i.e. one design in which the probit estimator is consistent, one in which theory suggests that the SMSE or the ISE with slowly increasing α_n converge fastest, and one in which the SMSE conditions (and the ISE conditions needed for faster convergence) are violated. We conclude, as noted above, that performance is fairly stable over a large range of α_n values and that choosing $\alpha_n = \infty$ (or equivalently using the MMSE) is indeed inefficient in every scenario.

We analyze how the choice of input parameters α_n, π affects performance (see e.g. figure 3) and compare a measure of estimation error across estimators for the three designs mentioned above. We consider three different sample sizes, five and nine regressors, and two choices of priors. Our simulation results reported in table 1 demonstrate that the ISE with a t–based prior performs better than the ISE with a uniform prior, whose performance compares favorably to that of the SMSE. But, it should be pointed out that performance of the SMSE could likely be improved with a different choice of kernel, or indeed by using the alternatives proposed in Chen and Zhang (2014) and Hong, Mahajan, and Nekipelov (2010).[9]

Finally, we document the behavior of our uniform inference procedure, which appears to perform well for the t–based prior, as evidenced by the size and power plots in figures 4 and 5, and described in detail in section 7.

The remainder of the paper is organized as follows. In section 2 we present the binary choice model and establish the asymptotic properties of the ISE, noting that most of the results are implied by or follow quickly from those in JPW15. Section 3 documents the inefficiency of the MMSE, compared with the ISE, under the conditions needed for the MMSE to be $\sqrt[3]{n}$–consistent. Section 4 discusses the choice of input parameters. Section 5 proposes our uniform inference procedure and establishes its uniformity properties analytically. Finally, section 6 documents our computation method and section 7 contains the results of our extensive simulation study.

2 Asymptotics

Consider the binary choice model

$$\boldsymbol{y_i} = \mathbb{1}\big(\theta_0^{\mathsf{T}} \boldsymbol{z_i} - \boldsymbol{a_i} + \boldsymbol{u_i} \geq 0\big), \quad i = 1, \ldots, n,$$

where $\mathbb{1}$ denotes the indicator function, $\boldsymbol{x_i} = [\boldsymbol{a_i} \ \boldsymbol{z_i^{\mathsf{T}}}]^{\mathsf{T}}$ is a vector of regressors, $\boldsymbol{u_i}$ an unobservable error term, and $\theta_0 \in \mathbb{R}^d$ the parameter vector of interest. The coefficient on $\boldsymbol{a_i}$ is assumed to equal minus one in lieu of normalizing the norm of the coefficients of $\boldsymbol{z_i}$ to be one. The usual scale normalization is equivalent to setting the absolute value of the coefficient of $\boldsymbol{a_i}$ to be one, but we focus on the case where it equals minus one

[9] We did not include further comparisons since the computer time demands of our current experiments were already substantial and most of these experiments predate our awareness of Chen and Zhang (2014).

for the sake of presentational simplicity.[10] The objective is to estimate θ_0 using an i.i.d. sample $\{(\boldsymbol{y_i}, \boldsymbol{x_i})\}$.

As in Manski (1985) we assume that

$$\text{Med}(\boldsymbol{u_1}|\boldsymbol{x_1}) = 0 \text{ a.s.}, \tag{1}$$

and that the conditional distribution of $\boldsymbol{a_1}$ given $\boldsymbol{z_1} = z$ is at almost all z absolutely continuous with respect to the Lebesgue measure with density function $f(\cdot|z)$. These assumptions will be stated formally further down this section in a somewhat different guise. We let $p(a, z) = \mathbb{E}\big(\boldsymbol{y_1}|\boldsymbol{x_1} = [a, z^\intercal]^\intercal\big)$ such that (1) implies that $2p\big(\theta_0^\intercal z, z\big) = 1$ at almost all z.

The regressor $\boldsymbol{a_1}$ resembles the special regressor of e.g. Lewbel (2000). However, Lewbel (2000) makes the additional assumption that $\boldsymbol{a_1}$ is independent of $\boldsymbol{u_1}$ given $\boldsymbol{z_1}$. If $\boldsymbol{a_1}$ is known to be independent of $\boldsymbol{u_1}$ given $\boldsymbol{z_1}$, then the special regressor method can be preferable to ours; see Khan and Tamer (2010) for general results on convergence rates and efficiency with the additional conditional independence assumption.

The maximum score estimator (MMSE) proposed in M75 maximizes the objective function[11]

$$\boldsymbol{L_n}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(2\boldsymbol{y_i} - 1)\big\{\mathbb{1}\big(\boldsymbol{a_i} \leq \theta^\intercal \boldsymbol{z_i}\big) - \mathbb{1}\big(\boldsymbol{a_i} \leq \theta_0^\intercal \boldsymbol{z_i}\big)\big\} \tag{2}$$

H92 proposed replacing the indicator function in (2) with an integrated kernel to obtain the smoothed maximum score estimator (SMSE) which features a faster convergence rate and asymptotic normality under additional smoothness conditions. Instead of smoothing, we follow CH03 and use a Laplace–type estimator, namely

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

where α_n, π are input parameters. In CH03 α_n was implicitly chosen to equal $\sqrt{n}$ under the assumption that the objective function $\boldsymbol{L_n}$ allows for a stochastic quadratic expansion. However, their assumption is not satisfied in our case and consequently the choices of α_n and π affect the first order asymptotic properties of our estimator.

Assumption.

A. θ_0 *is in the interior of a compact set* Θ*;*

B. $2p(\theta_0^\intercal z, z) = 1$ *for almost all* z, $\mathbb{P}(v^\intercal \boldsymbol{x_1} = 0) < 1$ *for any* $v \neq 0$, $0 < p(a, z) < 1$ *for almost all* a, z, *and* $f(z^\intercal \theta_0|z) > 0$ *for almost all* z*;*

C. for some $q \geq 0$, p *is* $q + 1$ *times continuously differentiable with respect to* a *at* $(\theta_0^\intercal z, z)$ *for almost all* z*;*

D. $f(\cdot|z)$ *is* q *times continuously differentiable at* $\theta_0^\intercal z$ *for almost all* z*;*

E. $\mathbb{E}\{\sup_a f(a|\boldsymbol{z_1})\|\boldsymbol{z_1}\|\} < \infty$*;*

[10] Since the sign of the coefficient can be estimated super–efficiently, taking the sign to be known is innocuous.

[11] We are subtracting out a constant for convenience.

F. *for some* $\mu \geq 2$, $\mathbb{E}\|\boldsymbol{z_1}\|^{\mu} < \infty$;

G. $0 < V = -2\mathbb{E}\{\boldsymbol{z_1}\boldsymbol{z_1}^{\intercal}\partial_a p(\theta_0^{\intercal}\boldsymbol{z_1}, \boldsymbol{z_1}) f(\theta_0^{\intercal}\boldsymbol{z_1}|\boldsymbol{z_1})\}$;

H. π *is* q *times continuously differentiable at* θ_0;

I. $\pi(\theta) > 0$ *at all* θ *in the interior of* Θ *and* $\pi(\theta) = 0$ *at all* $\theta \notin \Theta$.

We now compare our conditions to what is needed to establish a limit distribution for the MMSE and SMSE. The MMSE conditions should be compared to ours for the case $q = 0$ and the SMSE for $q = 1$.

Manski (1985) normalized the parameter space to be a unit circle so that assumption A was automatic. Assumption A was also used in H92.

Assumption B is equivalent to assumption 2 in Manski (1985). Assumptions C, D, F and G with $q = 0$ are equivalent to KP90, condition (iv). Assumption E is used to establish L_2–continuity (see A.5) and is implied by condition (vii) in KP90.[12] Hong, Mahajan, and Nekipelov (2010) assume the differentiability of the expectation of (2), which roughly corresponds to assumptions C and D. Assumptions H and I are input parameters for our estimator and hence play no role in the comparison. Thus, our assumptions are equivalent to those necessary to obtain a limit distribution of the MMSE.

Compared to H92, assumptions A, B and G also appear in H92, assumptions C to E are implied by assumptions 8 and 9 of H92, and assumption F is weaker than assumption 5.[13] Again, assumptions H and I are conditions on input parameters which can be satisfied by their choice and are hence irrelevant for the comparison of assumptions. Thus, our assumptions are weaker than those in H92. To obtain normality, Hong, Mahajan, and Nekipelov (2010) make the same assumptions as H92.

We are now in a position to state the asymptotics of the Laplace version of the MMSE under various assumptions. Let $M(\cdot)$ denote the ('sample') median of its arguments,

$$H(t,s) = \mathbb{E}\{f(\theta_0^{\intercal}\boldsymbol{z_1}|\boldsymbol{z_1})\big|M(t^{\intercal}\boldsymbol{z_1}, s^{\intercal}\boldsymbol{z_1}, 0)\big|\}, \tag{3}$$

and let $\mathbb{G}$ be a zero mean Gaussian process with covariance kernel H.

Theorem 1 (Asymptotic distribution)**.**

(i) *If assumptions A to I are satisfied for* $q = 0$ *and* $\mu = 2$, *and moreover* $\alpha_n \succ \sqrt[3]{n}$ *then*

$$\sqrt[3]{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} \mathbb{C},$$

where $\mathbb{C}$ *is the Chernoff distribution, i.e. the distribution of* $\operatorname{argmax}_t\{\mathbb{G}(t) - t^{\intercal}Vt/2\}$;

(ii) *If assumptions A to I are satisfied for* $q = 0$ *and* $\mu = 2$, *and* $0 < \lim_{n\to\infty} \alpha_n / \sqrt[3]{n} = c_\alpha^2 < \infty$ *then*

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

[12] Letting $\boldsymbol{g_1}(\theta) = (2\boldsymbol{y_1} - 1)\{\mathbb{1}(\boldsymbol{a_1} \leq \theta^{\intercal}\boldsymbol{z_1}) - \mathbb{1}(\boldsymbol{a_1} \leq \theta_0^{\intercal}\boldsymbol{z_1})\}$, note that $\{\boldsymbol{g_1}(\theta) - \boldsymbol{g_1}(\tilde{\theta})\}^2 = |\boldsymbol{g_1}(\theta) - \boldsymbol{g_1}(\tilde{\theta})|$.

[13] Denoting Horowitz's p and F on page 510 by p_H and F_H, we have $p_H(s|z) = f(\theta_0' z - s|z)$, $F_H(-s|s,z) = 1 - p(\theta_0' z - s, z)$.

where ϕ_V is the density of $N(0, V^{-1})$.

(iii) If assumptions A to I are satisfied for $q = 1$ and $\mu = 3$, and $\lim_{n\to\infty} \alpha_n / \sqrt[5]{n} = c_\alpha^2 < \infty$ then

$$n^{2/5}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} N(c_\alpha^{-4}\mathscr{B}, c_\alpha^2 \mathscr{V}),$$

with $\mathscr{V} = \iint ts^\intercal H(t,s)\phi_V(t)\phi_V(s)dtds$ and $\mathscr{B} = \int\{D_{\pi 1}(t) + \pi_0 D_{Q3}(t)\}t\phi_V(t)dt/\pi_0$, where D_{Q3}, $D_{\pi 1}$ are the third order term in a Taylor expansion of $Q(\theta_0 + t)$ around θ_0 and the first order term in a Taylor expansion of $\pi(\theta_0 + t)$ around θ_0,[14] respectively, with $\pi_0 = \pi(\theta_0)$ and Q the expectation of $\boldsymbol{L_n}$. □

Theorem 1 establishes a trichotomy in the limit distribution, where the limit distribution and the convergence rate depend both on the rate of α_n and on the degree of smoothness. The limit distribution is one of a Chernoff distribution, a ratio of Gaussian integrals, or a normal distribution.[15] The $n^{2/5}$ rate in the third part of the theorem is the fastest achievable rate under the stated conditions; if α_n goes to zero more slowly then the bias will dominate unless more smoothness is assumed.

The hard work for theorem 1 was already done in JPW15, which contains generic results for (originally) $\sqrt[3]{n}$–consistent estimators. What is new here is the specific application to the maximum score case.

In theorem 1 we normalize the coefficient of the last regressor to be -1, which is similar to H92. When a different normalization is used, the limit distributions described in theorem 1 need to be adjusted accordingly. For example, under Manski's normalization (i.e. the norm of the $d + 1$ dimensional parameter vector equals one), the Delta method shows that the limit distribution of the first d elements of the normalized estimator is characterized by

$$\frac{\left(\|\theta_0\|^2 + 1\right)I - \theta_0\theta_0^\intercal}{\left(\|\theta_0\|^2 + 1\right)^{3/2}}$$

times the limit distributions indicated in theorem 1.

The bias in the normality case needs to be dealt with when conducting inference. JPW15 discussed two possibilities: either directly estimating the bias or using a bias–eliminating prior such as in a neighborhood of θ_0

$$\pi^*(\theta) \propto \sqrt{-\det\{\partial_{\theta\theta^\intercal} Q(\theta)\}},$$

which we call the Jeffreys prior for its resemblance with the Jeffreys prior in the Bayesian literature.[16] In this paper we propose another simulation–based approach, which we believe is preferable to the first two methods because computation is easier than if one uses the Jeffreys prior and performance is better than if one

[14] For scalar–valued θ, $D_{Q3}(t/\alpha_n) = Q'''(\theta_0)t^3/6\alpha_n^3$.

[15] Since the SMSE with a small bandwidth behaves like the MMSE, the SMSE's asymptotic behavior can possibly be analyzed like we analyze the behavior of the ISE in theorem 1. However, since our focus is on the ISE, we do not pursue this question here.

[16] For issues of using an estimated prior, see JPW15.

subtracts out the bias; these issues are discussed in greater detail in section 6 and the recommended choice of prior in section 4.1.

The new idea is related with the fact that we do not choose a particular limiting distribution to conduct inference but we simulate some random variables $\hat{\boldsymbol{\Psi}}$ such that the limit distribution of $\hat{\boldsymbol{\Psi}}$ automatically adapts to the rate of α_n. In fact, in section 5 we show that this approach is not only rate adaptive but also *uniformly* valid within the class of *all* input parameters that satisfy a certain rate condition. We will show there how to simulate $\hat{\boldsymbol{\Psi}}$ such that the bias in the normality case is automatically incorporated into $\hat{\boldsymbol{\Psi}}$. For details, see section 5.

3 Efficiency

Theorem 1 demonstrates that the limit distribution of our estimator depends both on the choice of input parameters and on the smoothness of f, p. Under the weaker set of assumptions ($q = 0$ and $\mu = 2$) our estimator, under the same assumptions as KP90 for the MMSE, has a $\sqrt[3]{n}$ convergence rate, which is known to be the best rate attainable (Pollard, 1993). If $\alpha_n \succ \sqrt[3]{n}$ then the limit distributions of our estimator and the MMSE coincide. If $\alpha_n \sim \sqrt[3]{n}$ rate, however, the limit distributions differ. The main content of this section is an efficiency comparison in the two cases, i.e. $\alpha_n \succ \sqrt[3]{n}$ (or, indeed the MMSE) and $\alpha_n \sim \sqrt[3]{n}$.

Indeed, the following theorems illustrate that the MMSE is generally suboptimal. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{c_\alpha}}$ denote $\hat{\boldsymbol{\theta}}$ using $\alpha_n = c_\alpha^2 \sqrt[3]{n}$ with $\hat{\boldsymbol{\theta}}_{\boldsymbol{\infty}}$ as the special case with the same limit distribution as the MMSE. We now compare the limit distributions of $\hat{\boldsymbol{\theta}}_{\boldsymbol{c_\alpha}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\infty}}$. We denote the distribution function of the limit distribution of $\left| \sqrt[3]{n} \iota^\intercal (\hat{\boldsymbol{\theta}}_{\boldsymbol{c_\alpha}} - \theta_0) \right|$ by F_{c_α}, where ι is an arbitrary vector in $\mathbb{R}^d$ with $\|\iota\| = 1$.

Theorem 2 (Tail probabilities)**.** *For all* $0 < K, \Gamma < \infty$ *there exists a* $c_\alpha^* > 0$ *such that*

$$\inf_{0 < c_\alpha < c_\alpha^*} F_{c_\alpha}(K) / F_\infty(K) > \Gamma. \tag{4}$$

Further, F_∞ *locally first order stochastic dominates (LFOSD)* F_{c_α} *in following sense: for any* $\bar{K} > 0$*, there exists a* $c_\alpha^* > 0$ *such that for any* $0 \leq K \leq \bar{K}$*,*

$$\inf_{0 < c_\alpha < c_\alpha^*} \left\{ F_{c_\alpha}(K) - F_\infty(K) \right\} \geq 0,$$

where the inequality is strict for some $0 < K < \bar{K}$*.*

Note first that being stochastically dominated is a good thing here since we want small values for $\sqrt[3]{n} \|\hat{\boldsymbol{\theta}}_{\boldsymbol{c_\alpha}} - \theta_0\|$. Further, LFOSD is a weaker concept than FOSD. Indeed, the proof of theorem 2 does not generalize to FOSD, although it does not rule it out either. In fact, the simulation results reported later in this section do suggest the presence of a FOSD relationship, but we have failed to prove it. We can establish second order stochastic dominance (SOSD), however, as is asserted in theorem 3 below.

Theorem 2 has implications for the width of confidence intervals. Indeed, it can be interpreted as saying that for any given level of confidence, c_α can be chosen small enough to ensure that the ISE produces a

narrower confidence interval than MMSE.

Theorem 3 (Stochastic dominance)**.** *There exists a* $c_\alpha^* > 0$ *such that for any* $K^* \geq 0$*,*

$$\inf_{0<c_\alpha<c_\alpha^*} \int_0^{K^*} \{F_{c_\alpha}(K) - F_\infty(K)\} dK \geq 0,$$

where the inequality is strict for all $0 < K^* < \bar{K}$ *for some* $\bar{K} > 0$*.*

The key point of theorem 3 is uniformity, i.e. c_α^* does not depend on K^*. Therefore, we can choose a sufficiently small c_α such that $\hat{\boldsymbol{\theta}}_\infty$ (and hence the MMSE) is less efficient in the SOSD sense than $\hat{\boldsymbol{\theta}}_{c_\alpha}$ for finite c_α.

Note that, although it was established in H92 that the SMSE converges faster under *additional smoothness* conditions, Pollard (1993) showed that the SMSE converges more slowly than the MMSE if $q = 0$. So the results in H92 that establish that the SMSE converges faster under additional conditions neither imply nor contradict theorem 3.

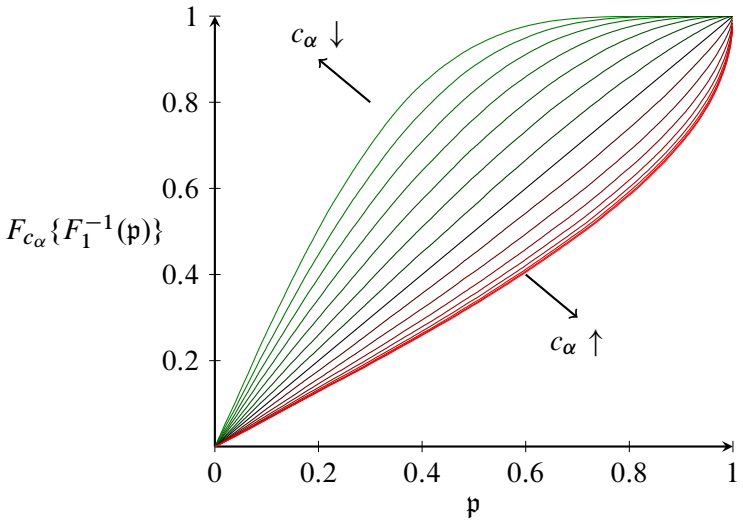

Figure 1: Efficiency of our estimator; $d = 1$

Figure 1 provides further support to our claim that the MMSE is inefficient, even if $q = 0$. Indeed, figure 1 depicts how F_{c_α} behaves as c_α changes. We computed

$$\frac{\sum_{j=1}^{\mathfrak{J}} t_j (t_j^2 + 1) \exp\{c_\alpha^2 \mathbb{G}(t_j) - c_\alpha^2 t_j^2/2 - c_\alpha^2 \mathbb{G}(\boldsymbol{M}) + c_\alpha^2 \boldsymbol{M}^2/2\}}{\sum_{j=1}^{\mathfrak{J}} (t_j^2 + 1) \exp\{c_\alpha^2 \mathbb{G}(t_j) - c_\alpha^2 t_j^2/2 - c_\alpha^2 \mathbb{G}(\boldsymbol{M}) + c_\alpha^2 \boldsymbol{M}^2/2\}},$$

with $\mathfrak{J} = 10^6$ (a million), $t_j = \tan\{\pi(2j - 1)/4\mathfrak{J}\}$, $\boldsymbol{M} = \text{argmax}_{j=1,\ldots,\mathfrak{J}}\{\mathbb{G}(t_j) - t_j^2/2\}$ 167,400 times for sixteen different values of c_α, ranging from $\sqrt{100}$ down by factors of $\sqrt[16]{100}$.[17] Here, $\mathbb{G}(t_1), \ldots, \mathbb{G}(t_{\mathfrak{J}})$ are

[17] We subtract $c_\alpha^2\{\mathbb{G}(\boldsymbol{M}) - \boldsymbol{M}^2\}$ to reduce rounding error and use the specified nodes t_j based on (the quantile function of) the Cauchy distribution to ensure that the tails receive sufficient weight. We ran a job for 23 hours on a cluster and 167,400 is the number of times these computations could be made within that time span.

jointly normal random draws consistent with a two–sided Brownian motion.

On the vertical axis is the value of $F_{c_\alpha}\left\{F_1^{-1}(\mathfrak{p})\right\}$ with $\mathfrak{p}$ the value on the horizontal axis. The 45 degree line then corresponds to $c_\alpha = 1$ and curves above the 45 degree line correspond to smaller values of c_α and curves below the 45 degree line to larger values of c_α. The efficiency differences across limit distributions is striking with the MMSE distribution being the least efficient of all. For instance, the median of F_1 corresponds to approximately the first quartile of F_∞.

Figure 1 suggests that F_{c_α} changes monotonically in c_α, which implies that F_∞ first order stochastically dominates F_{c_α}, recalling that the estimator with the dominated limit distribution has greater asymptotic efficiency than the estimator with the dominating limit distribution. This would be a far stronger result than theorem 2, but one for which (as noted earlier) we have no proof.

The above discussion may give the false impression that picking a small value of c_α (and hence of α_n) is necessarily better. However, note that the above comparison analyzes the limit distribution F_{c_α}, which is obtained by taking $n \to \infty$ first with a fixed c_α. If one lets $c_\alpha \to 0$ first then $\hat{\boldsymbol{\theta}}$ converges to the mean of the prior for any fixed sample size n. An informal way of describing what is happening is that choosing a finite c_α introduces bias in the finite sample distribution of $\hat{\boldsymbol{\theta}}$ which vanishes as $n \to \infty$.

It is of course possible to make c_α depend on n and let $c_\alpha \to 0$ as $n \to \infty$. This would be in line with theorem 1(iii) of our estimator. However, as Pollard (1993) has shown for the SMSE — and as is also the case here — if $q = 0$ then the MMSE has a $\sqrt[3]{n}$ convergence rate but the SMSE may have a convergence rate that is slower than $\sqrt[3]{n}$ and with a degenerate limit distribution. This happens because oversmoothing causes the bias term to dominate.

With additional smoothness assumptions, of course, both our estimator for $\alpha_n = c_\alpha^2 n^{2/5}$ (see theorem 1(iii)) and the SMSE converge faster than the MMSE and both have a more convenient normal limit distribution.

A comparison of the efficiency of the SMSE and our estimator is not fruitful. As already noted, under the stated conditions the asymptotic bias can be made to vanish by picking a suitable prior π for our estimator and a higher order kernel for the SMSE. The asymptotic mean square error can then be made arbitrarily small by picking a very small c_α (or a large bandwidth for the SMSE). Since the asymptotic mean square error can be made arbitrarily close to zero for both estimators but the convergence rate cannot be improved without making stronger assumptions, no meaningful efficiency comparison can be made between the two estimators.

4 Input Parameters

4.1 Prior

The choice of prior we recommend is a prior based on the d–variate Cauchy distribution, i.e. to choose (on a bounded set)

$$\pi(\theta) \propto \frac{1}{(1 + \|\theta\|^2)^{(1+d)/2}}, \tag{5}$$

which is shown in theorem 4 to be equivalent to imposing a uniform prior on the unit half–sphere in $\mathbb{R}^{d+1}$, i.e. an uninformative prior when the only information imposed is the sign of the last coefficient. Since the Cauchy distribution is a t distribution with one degree of freedom and the conditional distribution of one element given the remaining elements is again a t distribution we will call this choice of prior the *t–based prior.* Theorem 4 establishes this result; similar results can be found in Phillips (1984, 1989).

Theorem 4. *Choosing π equal to the d–variate Cauchy density function (5) is equivalent to imposing a uniform distribution on the unit half–sphere in $\mathbb{R}^{d+1}$. Further, if $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d)$ is a draw from a d–variate Cauchy then $\boldsymbol{\xi_1} \Big/ \sqrt{\sum_{i=2}^{d} \boldsymbol{\xi}_i^2 + 1}$ is (conditional on $\boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_d$) a draw from a t–distribution with d degrees of freedom.*

Normalizing the coefficient of the last regressor $\boldsymbol{a}_i$ to equal -1 is equivalent to restricting the parameter space to a half–sphere. Imposing a flat prior on the half–sphere is more reasonable than imposing a flat prior on θ_0 because the t–based prior treats all elements of the parameter vector symmetrically while the flat prior on a subset of $\mathbb{R}^d$ penalizes different deviations from θ_0 differently. For instance, if for $d = 1$ the true parameter vector equals $[1, -1]^\mathsf{T}$ then with the simple uniform prior $[0, -1]^\mathsf{T}$ and $[2, -1]^\mathsf{T}$ are equally far from the 'truth.' So an estimate that suggests that the first coefficient is twice as large as the second (in absolute value) is equally bad as an estimate that suggests that the second coefficient is infinitely many times as large as the first. This does not seem reasonable.

The above choice of prior does not eliminate bias, and certainly not higher order bias. However, based on our experience with the simulations reported in section 7, the efficiency improvements from substantial amounts of smoothing (paired with complicated bias corrections) are unlikely to be realized in practice unless sample sizes are unusually large.

4.2 Smoothing

The "optimal" choice of α_n depends on the degree of smoothness q and the choice of prior. In practice the degree of smoothness is unknown. In a different context Kotlyarova and Zinde-Walsh (2006) proposed a procedure which automatically adapts to the degree of smoothness, but found that the desirable theoretical properties were not reflected in their simulation results.

Alternatively, one can assume a degree of smoothness, choose a prior, and then choose the value of α_n that minimizes e.g. the asymptotic mean square error.[18] This is the route followed by H92, who (if one assumes $q = 1$) minimizes the sum of the squared norm of the bias and the trace of the variance. The analogous choice in our case would be

$$\alpha_n = \left(\frac{4 \|\mathscr{B}\|^2 n}{\mathrm{tr}(\mathscr{V})} \right)^{1/5}, \tag{6}$$

or indeed an estimator thereof.

[18] For $q = 0$ the limit distribution is not normal so other loss functions may be preferable.

However, if it is known that $q = 1$ then one can choose a bias–eliminating prior (or higher order kernel for the SMSE) and there is then no (asymptotically) optimal choice of c_α in $\alpha_n = c_\alpha^2 \sqrt[5]{n}$ because the optimal convergence rate is $n^{2/5}$ which requires that $\alpha_n \sim \sqrt[5]{n}$ but the asymptotic mean square error is increasing in c_α; the same issue arises with the SMSE or indeed nonparametric kernel estimation more generally.[19]

Prediction optimality via (e.g. k–fold) cross validation is another possibility. The idea of maximum score is to estimate the parameter vector by optimizing the median–based prediction. This suggests choosing the input parameter value that performs the best in multi–fold cross validation. To be more specific, recall that the model in (1) implies that

$$\text{Med}(\boldsymbol{y_1}|\boldsymbol{x_1}) = \mathbb{1}(\boldsymbol{z_1}^\mathsf{T}\theta_0 - \boldsymbol{a_1} \geq 0).$$

Therefore, for a candidate value α_n, we can compute predictions

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}}(\alpha_n) = \mathbb{1}\{\boldsymbol{z_i}^\mathsf{T}\hat{\boldsymbol{\theta}}(\alpha_n) - \boldsymbol{a_i} \geq 0\},$$

which can be cross–validated using e.g. the mean absolute deviation criterion. Since α_n is a scalar, it is easy to minimize the prediction error criterion. Further, although the values of α_n that minimize the prediction error criterion are not necessarily unique, it is reasonable to choose the largest minimizer. This approach is computationally the most expensive but it requires no prior information about degrees of smoothness and other unknown model primitives.

In view of the discussion above and in view of the fact that our inference procedure proposed in the following section is adaptive to the rate of α_n, we do not pursue an automatic (or optimal) choice of input parameter α_n. Our procedure does, however, have one fortuitous feature: α_n multiplies the objective function but does not scale the regressors, as it does with the SMSE.

Our simulation results suggest that $\alpha_n = 1.5\sqrt[3]{n}$ is a reasonable choice for all designs considered and yields $\sqrt[3]{n}$–consistent estimators regardless of the actual degree of smoothness q. More slowly increasing choices of α_n work better in theory if it is known that $q \geq 1$ and worse if $q = 0$, but the potential efficiency gain promised by asymptotic theory does not appear to be substantial, certainly not in comparison with the efficiency gain and computational simplicity of our estimator relative to the MMSE. Indeed, under the SMSE assumptions ($q = 1$) our estimator using $\alpha_n \sim \sqrt[5]{n}$ converges at a faster rate, and is hence infinitely more efficient, than our estimator using the suggested choice of α_n, but such efficiency gains are not borne out by the simulations.

5 Uniform Inference

We now present theoretical results supporting a simulation–based inference method that does not require assumptions on the choice (including the rate) of the input parameter sequence other than it lying between

[19] For the SMSE the asymptotic bias is zero if a higher order kernel is used. Hence the asymptotic mean square error is decreasing in the choice of bandwidth.

two diverging bounds. Unlike JPW15 the proposed method is shown to be not only *rate–adaptive* but also uniformly valid for *any* such sequence of the input parameter. Unlike JPW15 we do not require that the bias be removed in the estimator itself. This is an advantage because using the Jeffreys prior can be expensive in terms of computer time, because implementing the Gibbs sampler to draw from the quasi–posterior requires inverting the distribution function corresponding to the Jeffreys prior. See section 6 for more details. Instead of removing the bias from the estimator we incorporate the bias correction into the simulation of the limit distribution.

For the purpose of inference we propose drawing random numbers by

$$\hat{\boldsymbol{\Psi}} = \frac{1}{\sqrt[3]{n}} \frac{\int t\lambda\big(\hat{\boldsymbol{\theta}} + t/\sqrt[3]{n}\big) \exp\big[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2\}\big] dt}{\int \lambda\big(\hat{\boldsymbol{\theta}} + t/\sqrt[3]{n}\big) \exp\big[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2\}\big] dt}, \tag{7}$$

to mimic the behavior of $\hat{\boldsymbol{\theta}} - \theta_0$, where $\beta_n = \sqrt{\alpha_n^3/n}$, $\hat{\boldsymbol{V}}$ is a consistent estimator of V, and the prior λ need not be the same as the prior π used for estimation. Indeed, the choice of λ can be used to address the bias issue in the faster convergence case. We will assume that $\hat{\mathbb{G}}$ is constructed by using an estimated H that satisfies e.g. assumptions O, P, Q in JPW15. Possible choices for $\hat{\mathbb{G}}, \hat{\boldsymbol{V}}, \lambda$ are discussed in section 6.2.

The discussion below focuses on the case $q = 1$, but other values of q can be accommodated analogously. However, the analysis below presumes that a minimum value of q is known, so the procedure does not adapt when the degree of (minimal) smoothness is unknown.

We make the following assumptions.

Assumption J. *α_n satisfies $c_\beta^{2/3}\sqrt[5]{n} \le \alpha_n \le C_\beta^2\sqrt[3]{n}$ for $0 < c_\beta \le C_\beta < \infty$ which implies that $c_\beta n^{-1/5} \le \beta_n \le C_\beta$.*

Assumption J restricts the sequences of input parameters that are allowed. Note that assumption J does not fix any specific rate of α_n (as in theorem 1). Indeed, it allows for *any* sequence whose elements fall within the specified bounds: e.g. $\beta_n = 2 + \sin n$ is allowed, as is $\beta_n = n^{-1/5}\mathbb{1}(n \text{ odd}) + \mathbb{1}(n \text{ even})$. So the results in this section are *uniform* within the class of input parameters satisfying assumption J.

Assumption J covers the case $\alpha_n \sim \sqrt[5]{n}$ (i.e. case (iii) in theorem 1) and the case $\alpha_n \sim \sqrt[3]{n}$ (i.e. case (ii) in theorem 1). The case of $\alpha_n \succ \sqrt[3]{n}$ (i.e. case (i) in theorem 1) is omitted. Note that there is no discontinuity between cases (i) and (ii) in theorem 1 in that as $c_\alpha \to \infty$, the limit distribution of case (ii) converges to that of case (i). Discontinuity between cases (ii) and (iii) is the inferential difficulty we address.

Assumption K. *The function λ used in (7) satisfies $\partial_\theta \log\lambda(\theta_0) = \partial_\theta \log\pi(\theta_0) - \partial_\theta \log\pi^*(\theta_0)$, with $\lambda_0 = \lambda(\theta_0), \pi_0 = \pi(\theta_0), \pi_0^* = \pi^*(\theta_0)$, where π^* is a first order bias–eliminating prior (e.g. the Jeffreys prior) bounded away from zero on Θ. Further, there exists a $\rho > 1$ such that*

$$\lim_{\delta\downarrow 0} \sup_{\|\theta-\theta_0\|<\delta} \frac{\big|\lambda(\theta) - \lambda_0 - \partial_{\theta^\intercal}\lambda(\theta_0)(\theta-\theta_0)\big|}{\delta^\rho} = 0. \tag{8}$$

The second part of assumption K is a smoothness condition on λ, which is implied by twice differentiability. The first part of assumption K ensures that $\hat{\boldsymbol{\Psi}}$ has the same bias as the estimator, even if $\alpha_n \sim \sqrt[5]{n}$. An

example of a function λ that satisfies the above conditions is any function λ for which

$$\lambda(\theta) = \frac{\pi(\theta)}{\pi^*(\theta)}, \tag{9}$$

for θ near θ_0.

To see how the proposed method works, consider

$$\begin{aligned}\sqrt{n/\alpha_n}\hat{\boldsymbol{\Psi}} &\simeq \frac{1}{\beta_n}\frac{\int t\{\lambda_0 + t^\intercal \partial_\theta \lambda(\theta_0)/\alpha_n\}\exp\{\beta_n \hat{\mathbb{G}}(t)\}\phi_V(t)dt}{\lambda_0 \int \exp\{\beta_n \hat{\mathbb{G}}(t)\}\phi_V(t)dt} \\ &\simeq \frac{m_1(\hat{\mathbb{G}}, \beta_n) + (\alpha_n\beta_n)^{-1}\int t\{D_{\lambda 1}(t)/\lambda_0\}\phi_V(t)dt}{1 + \beta_n m_3(\hat{\mathbb{G}}, \beta_n)},\end{aligned} \tag{10}$$

where $D_{\lambda 1}$ is the natural analog to $D_{\pi 1}$, and where m_1 and m_3 are continuous functionals of $\mathbb{G}$ using e.g. a sup norm on compacta such that they are continuous in β_n, also. The precise definitions of the m–functions can be found in (19) in appendix B; appendix B also contains a rigorous justification for this expansion. The expansion in (10) should be compared with the asymptotic expansion of the estimator. Indeed, it is shown in appendix B that (as a byproduct of theorem 1) under assumptions J and K we have the expansion

$$\sqrt{n/\alpha_n}(\hat{\boldsymbol{\theta}} - \theta_0) \simeq \frac{m_1(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n) + (\alpha_n\beta_n)^{-1}\int t\{D_{\pi 1}(t)/\pi_0 + D_{Q3}(t)\}\phi_V(t)dt}{1 + \beta_n m_3(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)}, \tag{11}$$

where $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ is a stochastic process defined in (21) (in an appendix) that converges weakly to $\mathbb{G}$.

The stochastic terms in (10) and (11) determine the shape of the (limit) distributions: they depend on the rate of α_n. The nonstochastic terms in the numerators of (10) and (11) account for the bias if $\alpha_n \sim \sqrt[5]{n}$ and they are asymptotically negligible if $\alpha_n \succ \sqrt[5]{n}$. The choice of λ by assumption K ensures that the two nonstochastic terms coincide. Indeed, recall that π^* satisfies $\int t\{D_{\pi^* 1}(t)/\pi_0^* + D_{Q3}(t)\}\phi_V(t)dt = 0$. Therefore, by assumption K,

$$\begin{aligned}\int t\{D_{\lambda 1}(t)/\lambda_0\}\phi_V(t)dt &= \int t\{D_{\pi 1}(t)/\pi_0 - D_{\pi^* 1}(t)/\pi_0^*\}\phi_V(t)dt \\ &= \int t\{D_{\pi 1}(t)/\pi_0 + D_{Q3}(t)\}\phi_V(t)dt,\end{aligned}$$

which is exactly the bias term in (11).

Since assumption J is all that is needed for the expansions in (10) and (11), the intuitive arguments above suggest that the inference based on $\hat{\boldsymbol{\Psi}}$ is uniformly valid among the class of all α_n sequences that satisfy assumption J. We now formalize this result. Let $\mathcal{A}$ be the collection of all sequences $\{\alpha_n\}$ that satisfy assumption J.

Theorem 5. *Suppose that the assumptions of theorem 1 and assumptions J and K are satisfied and that* $q = 1$.

For any $x \in \mathbb{R}$ *and* $w \in \mathbb{R}^d$,

$$\sup_{\{\alpha_n\} \in \mathcal{A}} \left| \mathbb{P}(w^\intercal \sqrt{n/\alpha_n} \hat{\boldsymbol{\Psi}} \leq x) - \mathbb{P}\{w^\intercal \sqrt{n/\alpha_n}(\hat{\boldsymbol{\theta}} - \theta_0) \leq x\} \right| \to 0.$$

The trichotomy of theorem 1 suggests that inferential uncertainty is a practical issue, because what is chosen in practice is the value of α_n, not its rate. Theorem 5 shows that this problem can be resolved by simulating quantiles of $\hat{\boldsymbol{\Psi}}$ for the purpose of inference. If α_n has a specific rate, then theorem 5 shows that inference based on $\hat{\boldsymbol{\Psi}}$ will be automatically adaptive. However, theorem 5 is a stronger result than *rate–adaptive* inference, because theorem 5 does not presume the convergence of the distribution functions. For instance, assumption J allows for the possibility that α_n oscillates between $\sqrt[3]{n}$ and $\sqrt[5]{n}$. Therefore, simulating quantiles of $\hat{\boldsymbol{\Psi}}$ for the purpose of inference is not only *rate–adaptive* but also *uniform* within the class of all input parameters that satisfy assumption J.

Finally, note that there are at least two alternative methods for making the bias in the limit distribution reflect that in the estimator to the one proposed above. Indeed, one can replace (7) with either one of

$$\begin{cases} \dfrac{1}{\sqrt[3]{n}} \dfrac{\int t\{\pi(\hat{\boldsymbol{\theta}} + t/\sqrt[3]{n}) - \pi(\hat{\boldsymbol{\theta}}) \partial_{\theta^\intercal} \log \pi^*(\hat{\boldsymbol{\theta}}) t/\sqrt[3]{n}\} \exp\left[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2\}\right] dt}{\int \{\pi(\hat{\boldsymbol{\theta}} + t/\sqrt[3]{n}) - \pi(\hat{\boldsymbol{\theta}}) \partial_{\theta^\intercal} \log \pi^*(\hat{\boldsymbol{\theta}}) t/\sqrt[3]{n}\} \exp\left[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2\}\right] dt}, & (12) \\ \dfrac{1}{\sqrt[3]{n}} \dfrac{\int t\pi(\hat{\boldsymbol{\theta}} + t/\sqrt[3]{n}) \exp\left[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2\}\right] dt}{\int \pi(\hat{\boldsymbol{\theta}} + t/\sqrt[3]{n}) \exp\left[\beta_n^{4/3}\{\hat{\mathbb{G}}(t) - t^\intercal \hat{\boldsymbol{V}} t/2\}\right] dt} - \dfrac{1}{\alpha_n^2} \hat{\boldsymbol{V}}^{-1} \partial_\theta \log \pi^*(\hat{\boldsymbol{\theta}}). & (13) \end{cases}$$

The advantage of both (12) and (13) over (9) is computational simplicity and that they obviate the need to deal with the problem of figuring out where to apply truncation to prevent small π^* values. Note that (13) is equivalent to estimating and subtracting the bias from the estimator itself and not applying a bias correction in the limit experiment. Our simulation results (not tabulated in this paper) suggest that (12) performs somewhat better than (9) and (13) in practice.

6 Computation

6.1 Estimates

We now describe the Gibbs sampling scheme used to obtain our estimates: a more general Metropolis–Hasting algorithm is possible but we focus on the Gibbs sampler here. With the Gibbs sampler from a given starting value one repeatedly draws θ_j from the conditional posterior of θ_j conditional on the values of the remaining coefficients θ_{-j}, iterating over j, until convergence. Below we explain how to obtain a random draw $\boldsymbol{\theta}_j^*$ from the conditional (pseudo) posterior of θ_j given θ_{-j}. As will become apparent, this procedure is straightforward to implement.

The conditional (pseudo) posterior of θ_j given θ_{-j} is

$$\tilde{r}(\theta_j|\theta_{-j}) = \frac{\pi(\theta_j|\theta_{-j}) \exp\{\alpha_n^2 \boldsymbol{L_n}(\theta)\}}{\int \pi(\theta_j|\theta_{-j}) \exp\{\alpha_n^2 \boldsymbol{L_n}(\theta)\} d\theta_j}.$$

Therefore, we can obtain a random draw from $\tilde{r}(\theta_j|\theta_{-j})$ by computing

$$\boldsymbol{\theta_j^*} = \tilde{R}^{-1}(\boldsymbol{\zeta}|\theta_{-j}), \tag{14}$$

where $\boldsymbol{\zeta} \sim U(0,1)$, $\tilde{R}(\theta_j^*|\theta_{-j}) = \int_{-\infty}^{\theta_j^*} \tilde{r}(\theta_j|\theta_{-j})d\theta_j$ is the distribution function of the conditional posterior, and $\tilde{R}^{-1}$ is the (generalized) inverse of $\tilde{R}$. We now address the question how best to do this.

The key insight is that $\boldsymbol{L_n}$ is a simple step function in each dimension. Indeed, let $\boldsymbol{z_{ij}}$ denote the j–th element of $\boldsymbol{z_i}$ and $\boldsymbol{z_{i,-j}}$ the vector $\boldsymbol{z_i}$ without its j–th element. For $\boldsymbol{z_{ij}} \neq 0$ define

$$\boldsymbol{B_i} = \frac{\boldsymbol{a_i} - \theta_{-j}^{\mathsf{T}} \boldsymbol{z_{i,-j}}}{\boldsymbol{z_{ij}}},$$

ignoring observation i if $\boldsymbol{z_{ij}} = 0$.[20] Assume that the $\boldsymbol{B_i}$'s are sorted in ascending order, that the $\boldsymbol{B_i}$'s outside the support of the conditional prior $\pi(\cdot|\theta_{-j})$ are omitted,[21] and that the observations are indexed $0, \ldots, n-1$ instead of $1, \ldots, n$. Set $\boldsymbol{B_{-1}} = -\infty$ and $\boldsymbol{B_n} = \infty$.

Then

$$\boldsymbol{L_n}(\theta) = \frac{1}{n}\sum_{i=0}^{n-1}(2\boldsymbol{y_i} - 1)\mathbb{1}(\boldsymbol{z_i}^{\mathsf{T}}\theta \geq \boldsymbol{a_i}) = \\ \frac{1}{n}\sum_{i=0}^{n-1}(2\boldsymbol{y_i} - 1)\operatorname{sgn}(\boldsymbol{z_{ij}})\mathbb{1}(\theta_j \geq \boldsymbol{B_i}) + \frac{1}{n}\sum_{i=0}^{n-1}(2\boldsymbol{y_i} - 1)\mathbb{1}(\boldsymbol{z_{ij}} < 0), \tag{15}$$

where $\operatorname{sgn}(s) = \mathbb{1}(s > 0) - \mathbb{1}(s < 0)$. Note that the second right hand side term in (15) does not depend on θ. Define

$$\boldsymbol{T_i} = \exp\left\{\frac{\alpha_n^2}{n}\sum_{s=0}^{i}(2\boldsymbol{y_s} - 1)\operatorname{sgn}(\boldsymbol{z_{sj}})\right\}, \qquad \boldsymbol{R_i} = \sum_{s=1}^{i}\boldsymbol{T_{s-1}}(\boldsymbol{\Pi_s} - \boldsymbol{\Pi_{s-1}}),$$

where $\boldsymbol{\Pi_i} = \Pi(\boldsymbol{B_{i-1}}|\theta_{-j})$ is the conditional distribution function of the prior.

For given θ_j^*, let i be such that $\boldsymbol{B_{i-1}} \leq \theta_j^* < \boldsymbol{B_i}$. Then

$$\begin{aligned}\tilde{R}(\theta_j^*|\theta_{-j}) &\propto \int^{\theta_j^*} \pi(\theta_j|\theta_{-j})\exp\{\alpha_n^2 \boldsymbol{L_n}(\theta)\}d\theta_j \\ &= \sum_{s=0}^{i-1}\int_{\boldsymbol{B_{s-1}}}^{\boldsymbol{B_s}} \pi(\theta_j|\theta_{-j})\exp\{\alpha_n^2 \boldsymbol{L_n}(\theta)\}d\theta_j + \int_{\boldsymbol{B_{i-1}}}^{\theta_j^*} \pi(\theta_j|\theta_{-j})\exp\{\alpha_n^2 \boldsymbol{L_n}(\theta)\}d\theta_j \\ &\propto \sum_{s=1}^{i}(\boldsymbol{\Pi_s} - \boldsymbol{\Pi_{s-1}})\boldsymbol{T_{s-1}} + \{\Pi(\theta_j^*|\theta_{-j}) - \boldsymbol{\Pi_i}\}\boldsymbol{T_i} = \boldsymbol{R_i} + \{\Pi(\theta_j^*|\theta_{-j}) - \boldsymbol{\Pi_i}\}\boldsymbol{T_i},\end{aligned}$$

[20] We use $\boldsymbol{B_i}$ instead of $\boldsymbol{B_{ij}}$ to simplify notation; there is little scope for ambiguity.

[21] We do not account for such omissions in the notation and continue to use n to denote the number of observations (remaining).

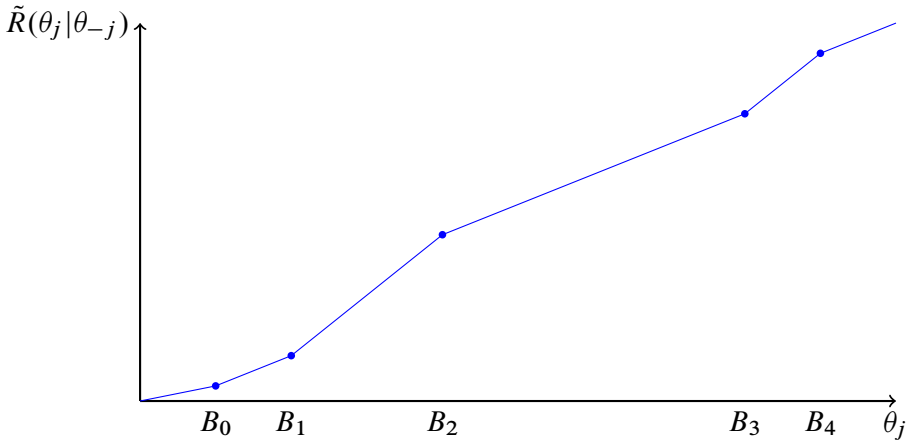

Figure 2: The conditional posterior of θ_j given θ_{-j} for $n = 5$ and a constant prior on compact support.

where the constant of proportionality in the $\propto$ relations equals $\boldsymbol{R_{n+1}}$ and hence does not depend on θ_j^*. Thus,

$$\tilde{R}(\theta_j^*|\theta_{-j}) = \frac{\boldsymbol{R_i} + \{\Pi(\theta_j^*|\theta_{-j}) - \boldsymbol{\Pi_i}\}\boldsymbol{T_i}}{\boldsymbol{R_{n+1}}}, \tag{16}$$

which for a uniform prior is drawn in figure 2.

Let $\boldsymbol{\zeta}$ be as in (14) and let $\boldsymbol{i^*}$ be the largest integer for which $\boldsymbol{R_{i^*}} \leq \boldsymbol{\zeta R_{n+1}}$. Then it follows from (14) and (16) that

$$\boldsymbol{\theta_j^*} = \tilde{R}^{-1}(\boldsymbol{\zeta}|\theta_{-j}) = \Pi^{-1}\left(\frac{\boldsymbol{R_{n+1}\zeta - R_i}}{\boldsymbol{T_i}} + \boldsymbol{\Pi_{i^*}}|\theta_{-j}\right),$$

is a draw from the desired conditional posterior.

6.2 Limit distribution

We now describe the method we used to obtain draws from the limit distribution in section 7. The method described here is likely inefficient and better methods can be found in various sources, e.g. Stroud (1971).

For some large T, draw i.i.d. $\boldsymbol{t_1}, \ldots, \boldsymbol{t_T}$ from a multivariate normal with mean zero and variance $2\hat{\boldsymbol{V}}^{-1}$. Then compute weights $\boldsymbol{\chi_1}, \ldots, \boldsymbol{\chi_T}$ with $\chi_j = \exp(-\boldsymbol{t_j^\intercal \hat{V} t_j}/4)$. Then compute $\hat{\boldsymbol{H}}$ with j, ℓ element $n^{-1}\sum_{i=1}^n \left|M(\boldsymbol{t_j^\intercal z_i}, \boldsymbol{t_\ell^\intercal z_i}, 0)\right| \hat{\boldsymbol{f}}(\hat{\boldsymbol{\theta}}^\intercal \boldsymbol{z_i}|\boldsymbol{z_i})$. Examples of $\hat{\boldsymbol{V}}$, $\hat{\boldsymbol{f}}$ are the Hessian of the SMS objective function and a kernel density estimator, respectively, albeit that some care should be taken if $\boldsymbol{z_i}$ has unbounded support.

Now, for each replication, draw $\hat{\mathbb{G}}_1, \ldots, \hat{\mathbb{G}}_T$ from a multivariate normal with mean zero and variance $\hat{\boldsymbol{H}}$ and compute e.g.

$$\frac{\sum_{j=1}^T \boldsymbol{t_j}\left[\{\pi(\hat{\boldsymbol{\theta}} + \boldsymbol{t_j}/\alpha_n) - \pi(\hat{\boldsymbol{\theta}})\boldsymbol{t_j^\intercal}\hat{\boldsymbol{\gamma}}/\alpha_n\}\exp(\beta_n \hat{\mathbb{G}}_j)\boldsymbol{\chi_j}\right]}{\sum_{j=1}^T\left[\{\pi(\hat{\boldsymbol{\theta}} + \boldsymbol{t_j}/\alpha_n) - \pi(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\gamma}}/\alpha_n\}\exp(\beta_n \hat{\mathbb{G}}_j)\boldsymbol{\chi_j}\right]},$$

if (12) is desired and similarly for (9) and (13), where $\hat{\boldsymbol{\gamma}} = \partial_\theta \log \pi^*(\hat{\boldsymbol{\theta}})$.

7 Simulations

7.1 Design

We have implemented our methodology using three designs. Each of the designs features a constant term and a number of mutually independent standard normal regressors. The first design is a standard probit model, and the second design has $\boldsymbol{u_i} = \{|\boldsymbol{u_i^*}| - \Phi^{-1}(0.75)\}\boldsymbol{x_{i2}}$, where $\boldsymbol{u_i^*}$ is standard normal and independent of $\boldsymbol{x_i}$, and $\boldsymbol{x_{i2}}$ is the first slope regressor. The third design is like a standard probit model albeit that the errors are drawn from a Laplace distribution instead of a normal distribution. We will refer to these designs as *probit*, *hetero*, and *laplace*, respectively. In all three designs the errors have zero median conditional on the regressors. The second distribution features both an asymmetric error distribution and heteroskedasticity and the third nondifferentiability of the error distribution at zero.

We computed SMSE estimates using an adapted version of a simulated annealing algorithm kindly provided to us by Yulia Kotlyarova and using a normal kernel and a range of bandwidths.[22] For the Laplace estimators we used the Gibbs sampling scheme described in section 6 with a single chain using a burn–in period of 10,000 draws and an average taken over 5,000 subsequent draws. These numbers are arbitrary and suboptimal in multiple ways: it is probably best to run multiple chains simultaneously and to make the length of the burn–in period depend on certain convergence criteria; see e.g. Cowles and Carlin (1996). Since we were doing these computations for three designs, using two different priors, for two sample sizes, with five and nine regressors, with sixteen different smoothing parameters, using 1,000 replications, budgetary constraints precluded a deeper investigation into the optimal chain design or convergence criteria.[23] In all cases probit estimates were taken as the starting values.

The SMSE search was restricted to $[-50, 50]^d$. The two priors that we used for the Laplace estimator are a constant prior on $[-50, 50]^d$ and the t–based prior proposed in section 4.1 on the same support.[24] Because of computing time limitations we did not use the bias–eliminating Jeffreys prior. We used sample sizes of both 1,000 and 2,000 and five and nine regressors (including an intercept and the regressor whose coefficient is normalized). We used a large number of input parameters differing by a factor of 1.25 each (for all three estimators) in order to study the effect of the choice of the input parameter on performance. In all cases $\theta_0 = [1, \ldots, 1]^\intercal$. We sometimes include the normalized coefficient in the parameter vector, in which case we refer to the *extended parameter vector*.

If α_n is chosen too small then the Laplace estimator will be close to the mean of the prior, i.e. zero. Likewise, if the SMSE bandwidth is chosen too large then the SMSE is approximately equal to a large constant times $\mathbb{E}\{(2\boldsymbol{y_1} - 1)\boldsymbol{z_1}\}$ (if nonzero).[25] In *probit* the SMSE extended parameter vector estimates (once

[22] We used sixteen different bandwidths, separated by a factor of 2, and centered at $n^{-1/5}$.

[23] It would perhaps be possible to conduct a more expansive simulation study with fewer observations and only a couple of regressors, but we aim to appeal to empiricists, for whom large numbers of observations and regressors are often the norm.

[24] Therefore, the t–based prior has mean zero.

[25] Note that the SMSE maximizes $n^{-1}\sum_{i=1}^n (2\boldsymbol{y_i} - 1)K\{(\boldsymbol{z_i}^\intercal\theta - \boldsymbol{a_i})/h\} \approx K(0)\mathbb{E}(2\boldsymbol{y_i} - 1) + k(0)\mathbb{E}\{(2\boldsymbol{y_i} - 1)(\boldsymbol{z_i}^\intercal\theta - \boldsymbol{a_i})\}/h$,

normalized to have norm one) these come out to be approximately $[0.527, 0.493, \ldots, 0.493, -0.000]^\mathsf{T}$, and $[0.365, 0.352, \ldots, 0.352, -0.000]^\mathsf{T}$,[26] for five and nine regressors, respectively, which is closer to $[\theta_0^\mathsf{T}, -1]^\mathsf{T}/\sqrt{d+1}$ than the prior used for the Laplace estimator. As will become apparent, it does not seem to matter much in practice: if one oversmooths, performance is poor for both estimators as expected and improves as α_n increases or h decreases. We have therefore not developed designs that do not favor the SMSE in this way.

7.2 Dependence on the input parameter

To investigate the dependence on the input parameter α_n of the Laplace estimator, we have graphed the quantity

$$\frac{1}{R}\sum_{r=1}^{R} \|\hat{\boldsymbol{\theta}}_r^* - \theta_0^*\|, \tag{17}$$

where $\hat{\boldsymbol{\theta}}_r^* = [\hat{\boldsymbol{\theta}}_r^\mathsf{T}, -1]^\mathsf{T}/(\|\hat{\boldsymbol{\theta}}_r\|^2+1)^{1/2}$ and $\theta_0^* = [\theta_0^\mathsf{T}, -1]^\mathsf{T}/(\|\theta_0\|^2+1)^{1/2}$ with $\hat{\boldsymbol{\theta}}_r$ the Laplace estimator in replication r, as a function of α_n for several designs and input parameter choices.[27] Figure 3 contains the results which we only depict for the t–based prior since the pictures for the uniform prior are qualitatively the same.

Each row corresponds to a single design with the right graph providing an amplified detail of the left graph with the smaller values of α_n omitted. There are four curves in each graph corresponding to a sample size, number of regressors pair where dots indicate the value at which the minimum is achieved.

Larger samples and fewer regressors result in less estimation error, which is not surprising. The fact that the minimizing value of α_n does not always vary with n can be attributed in part to the coarseness of the grid of input parameters used. Going from $n = 1,000$ to $n = 2,000$ doubles the sample size, which (since we do not use a bias–eliminating prior) means that the optimal α_n increases by a factor of $2^{1/5} = 1.15$ in the first two graphs and a factor $2^{1/3} = 1.26$ in the last graph where the grid points are a factor 1.25 apart.

As expected, the estimation error (17) for small values of α_n is large since the mean of the prior (zero) is different from θ_0. As α_n increases the estimation error drops rapidly and then increases more slowly in all three cases, again as our theoretical results would indicate. There is a small fluctuation in the middle in all three designs. In an early working paper (Jun, Pinkse, and Wan, 2009) we found a similar pattern in a different design where the prior bias (due to the fact that $\theta_0 \neq 0$) is partially offset by the asymptotic bias, before the asymptotic bias disappears also. In other designs the asymptotic bias may amplify the prior bias.

The gain from using our estimator over the MMSE ($\alpha_n = \infty$) appears to be especially large in *probit* because the graphs in the other two cases are flatter for large values of α_n. This may in part be attributable to

which is maximized at the largest value of θ proportional to $\mathbb{E}\{(2\boldsymbol{y_i}-1)\boldsymbol{z_i}\}$ in the parameter space if one abstracts away from parameter space shape issues. In other words, the SMSE converges in probability to $[\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_2, -1]^\mathsf{T}$ where $\bar{\theta}_1 = C\mathbb{E}(2\boldsymbol{y_i}-1)$ and $\bar{\theta}_2 = C\mathbb{E}\{(2\boldsymbol{y_i}-1)\boldsymbol{z_{i2}}\}$ for some large C whose value depends on the size of the parameter space. After normalization, the probability limit becomes $[\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_2, -1/C]^\mathsf{T}/\{\bar{\theta}_1^2 + (d-2)\bar{\theta}_2^2 + 1/C^2\}^{1/2}$.

[26] The last elements are negative numbers that are small in absolute value. See footnote 25.

[27] We impose the renormalization to treat all regression coefficients equally: without the normalization, one would be giving infinite weight to the coefficient that is normalized to equal -1.

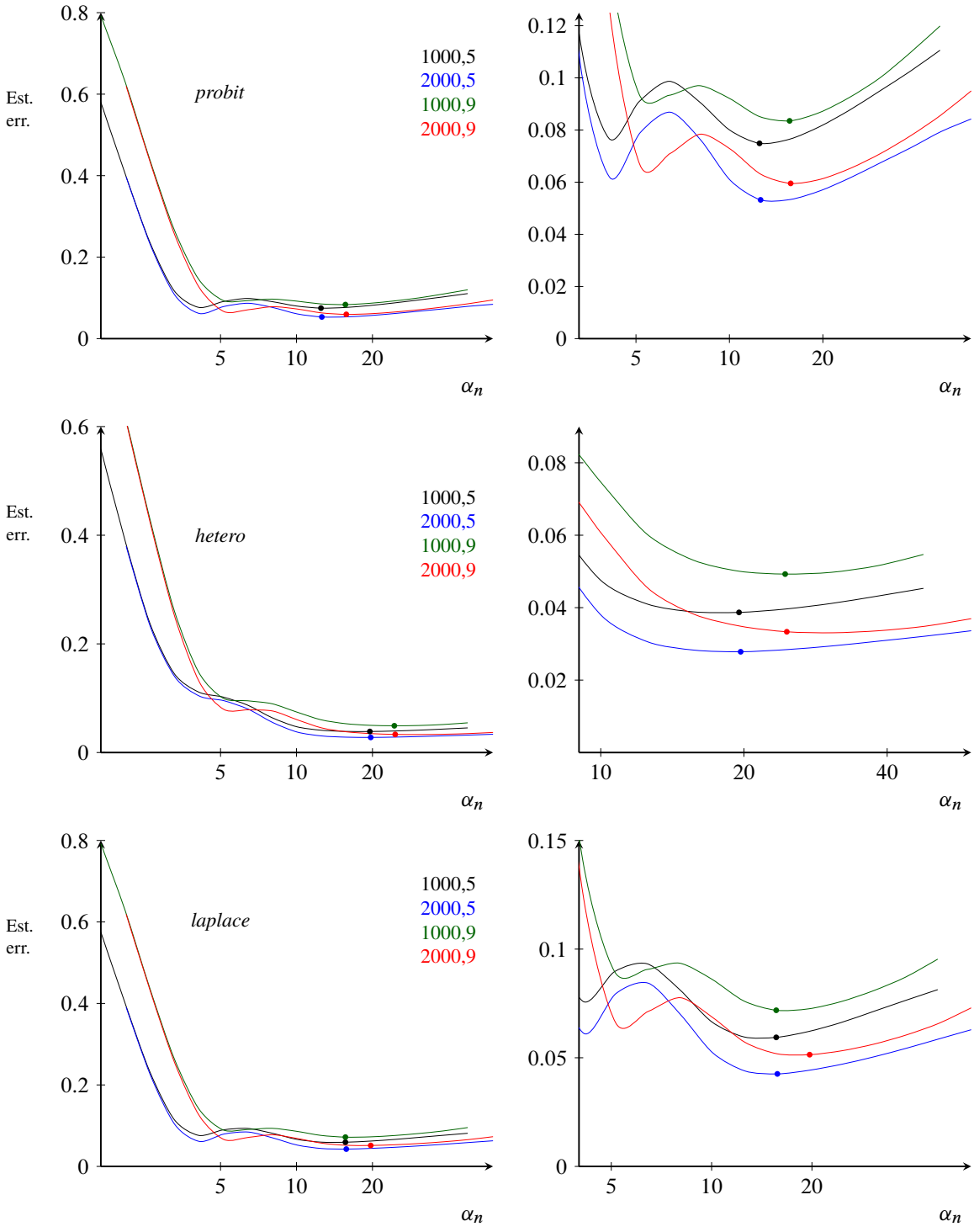

Figure 3: How (17) varies with the choice of α_n, sample size, and number of regressors, for the t–distribution–based prior and design 0,2,4. The right graphs provide amplified details of the left. The legend numbers show the sample size and the number of regressors, respectively.

the difficulty of computing the MMSE, i.e. we may not achieve the true maximum using a single chain and the chain length used in our simulations.

Perhaps the main conclusion that can be drawn from these graphs is that while it is clearly suboptimal to choose $\alpha_n = \infty$ one should be careful not to pick α_n too small; $\alpha_n = 1.5\sqrt[3]{n}$ appears to be a reasonable choice for the estimator with the t prior, as mentioned in section 4.2. These choices are 15 and 18.9 for $n = 1,000$ and $n = 2,000$ respectively, whereas the choices of α_n suggested by (6) in the probit case are 11.5, 13.2, 12.3, and 14.2 for the (1000, 5), (2000, 5), (1000, 9), and (2000, 9) cases, respectively;[28] see table 2 for more information. These numbers are all in the range of near–optimal performance.

7.3 Estimator performance

The quantity defined in (17) can be compared across both designs and estimators. For each estimator/design combination we picked the value of the input parameter that minimized (17) as is illustrated in figure 3; the results are displayed in table 1 and the actual minimizing input parameter values are in table 2. As the results of section 7.2 indicate, estimation error is fairly flat over a large range and although the entries in the table would vary a bit, the choice of only presenting the minimized values is immaterial for the qualitative conclusions.

	probit					*hetero*				*laplace*			
	1000		2000		4000	1000		2000		1000		2000	
Estimator	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=8$
SMSE	950	1162	664	843	511	462	569	336	382	701	913	504	632
uniform	969	1118	698	822	531	403	519	294	344	718	887	525	632
t	749	835	532	595	398	387	493	278	331	594	719	426	514

Table 1: Estimation error (17) $\times 10,000$ across estimators

As one would expect estimation error is less in larger samples and greater if there are more regressors. The estimation error decreases by approximately a factor 1.4 if one goes from 1,000 to 2,000 observations. Since $1.4 \approx \sqrt{2}$ this constitutes a more sizable improvement than first order asymptotics would suggest (a factor $2^{2/5} = 1.32$ for *probit* and *hetero* and $2^{1/3} = 1.26$ for *laplace*). We expect the rate of improvement of the estimation error to level off to that suggested by asymptotic theory as the sample size increases and this is indeed borne out by the results going from 2,000 to 4,000 observations for *probit*.

Throughout, it appears that the t–based prior does a bit better than the uniform prior and the SMSE. This may be due to the form of the loss function (17) since both the SMSE and the Laplace estimator with a uniform prior penalize deviations in some directions less than others, but first order asymptotics suggest that the choice of loss function would not affect the ranking of estimators. Alternatively, and more plausibly, the t–based prior may result in less bias than the other two estimators.[29]

[28] In the heteroskedastic case, $Q''(\theta_0)$ is not finite, which violates our assumptions. In the Laplace estimator case, $\sqrt[3]{n}$ is the best achievable rate.

[29] As we discussed in section 4.1, the t–based prior treats all coefficients symmetrically, whereas the uniform prior places infinite

Whatever the explanation, it should be noted that the choice of prior is asymptotically immaterial in *laplace* because $\sqrt[3]{n}$ is the best achievable rate. In the other two cases, one can in principle bring up the convergence rate very close to $\sqrt{n}$ by eliminating higher order bias and choosing a small α_n or large h. *In theory* once one fixes the choice of α_n and prior (or h and kernel) for one method one can improve over the asymptotic properties of that estimator by choosing the input parameters of another estimator.

Nevertheless, the results in table 1 suggest that using a Laplace estimator with a t–based prior beats using SMSE with a normal kernel.

	probit					*hetero*				*laplace*			
	1000		2000		4000	1000		2000		1000		2000	
Estimator	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=8$
SMSE	288	309	250	269	218	147	158	103	138	230	309	200	269
uniform	244	305	246	384	310	244	305	246	384	244	305	246	384
t	125	156	126	157	159	195	244	197	308	156	156	157	197
SMSE	*262*	*278*	*228*	*242*	*199*	} *asymptotically optimal values*							
uniform	*241*	*340*	*277*	*391*	*318*								
t	*115*	*123*	*132*	*142*	*152*								

Table 2: Input parameters corresponding to table 1 in the paper, $\times 1,000$ for SMSE, $\times 10$ for the others and in italics below the line the corresponding asymptotically optimal values.

The (minimizing) input parameters used in table 1 are in table 2, along with the (infeasible) asymptotically optimal input parameter choices for the probit case.[30] The differences between the input parameter values that minimize (17) and the corresponding asymptotically optimal ones seem reasonable, especially when taking into account the coarseness of the grid used in the simulations, the distinction between finite sample and asymptotic behavior, the fact that the asymptotically optimal values are based on a first order mean square error criterion and the finite sample ones on (17), and the flatness of (17) over a range of input parameter values.

	probit					*hetero*				*laplace*			
	1000		2000		4000	1000		2000		1000		2000	
Estimator	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=8$	$d=4$	$d=8$
SMSE	47.8	94.8	96.7	200.0	189.0	39.7	84.8	73.3	167.1	45.2	96.6	92.0	195.9
uniform	9.0	16.5	19.5	35.9	42.6	9.1	17.6	19.5	37.6	9.0	16.9	19.9	37.0
t	14.9	28.8	31.7	59.9	67.2	15.0	30.0	31.4	62.1	14.9	29.1	32.2	61.5

Table 3: Time in seconds to compute an estimate

We now turn our discussion to the issue of computation times, which are reported in table 3 for the estimates of table 1. The programs are written in C. To provide an idea of the magnitudes reported here, to

weight on the last element.

[30] We have no procedure for determining the optimal input parameters for hetero and laplace, since the conditions for $n^{2/5}$ convergence are not met. See footnote 28.

compute the SMSE for the probit case with 2,000 observations and nine regressors using sixteen different bandwidth values in 1,000 replications takes approximately $200 \times 16 \times 1000 = 3,200,000$ seconds (37 days) of CPU time. Thankfully, parallel processing and a large cluster made this feasible.

Computation times appear to be approximately linear in the number of unknown coefficients and the number of observations, albeit that our use of the same parameters for the routines across designs, sample sizes, and number of coefficients, is unlikely to be optimal.

That said, in our simulations the Laplace estimator with a uniform prior was on average about five times faster than the SMSE and significantly faster than the Laplace estimator with a t–based prior. Using the t prior is slower because inverting the distribution function of a t distribution is more time–consuming than inverting the distribution function of a uniform distribution.

7.4 Inference

We now turn to an evaluation of our uniform inference procedure of section 5. In our evaluations we use the limit distribution described in H92 to obtain critical values for the SMSE and use the bias expansion–based uniform inference procedure described in equation (12) of section 5 to produce ones for ours. Because of computing time feasibility constraints we had to use the true rather than the estimated limit distribution in all cases, which is less than ideal. Nevertheless, the simulations provide a clear picture of some important features.

First consider figures 4 and 5, both of which correspond to the *probit* design with five regressors. In each figure the top two graphs correspond to $n = 1,000$ and the bottom two graphs to $n = 2,000$ with the left graphs depicting size and the right graphs depicting power for hypotheses about the first slope coefficient. The difference between figures 4 and 5 is in the choice of input parameter. What constitutes a 'moderate amount' of smoothing (i.e. smaller values of α_n and larger values of bandwidth) or 'little smoothing' (i.e. larger values of α_n and smaller values of bandwidth) is somewhat arbitrary and the curves are hence not directly comparable, but it does not matter for the overall conclusions.[31]

The size of all estimators exceeds nominal size (indicated by the dashed 45 degree line) for $n = 1,000$, which appears to be mostly due to bias for the uniform prior case except for the SMSE and little smoothing for reasons that will become apparent below. For 2,000 observations the size is noticeably better in the moderate smoothing case. With little smoothing actual size still exceeds nominal size for the Laplace estimators which we attribute to the fact that with little smoothing one is essentially trying to compute the MMSE. Since the objective function has a less regular shape, it would take more and longer chains for the Gibbs sampler to achieve actual convergence. Note that there is no method that can confirm convergence of the chains. The opposite happens for the SMSE, which is natural since its rejection probability tends to zero as the bandwidth goes to zero for *fixed* n.[32] So if one would further reduce the bandwidth then the size and power curves for the SMSE (for fixed n) would eventually get arbitrarily close to the horizontal axis. As anticipated the influence

[31] In figure 4 the bandwidths used for the SMSE were 0.449 and 0.391 and the values for α_n were 12.5 and 15.75. For figure 5, they were 0.118 and 0.102 versus 38.15 and 48.06.

[32] Inference for the SMSE is based on an approximate normal distribution with mean proportional to h^2 and variance proportional to $1/nh$. Since for $\xi \sim N(0,1)$ and any finite $b, C > 0$, $\mathbb{P}(\xi/\sqrt{nh} - h^2 b > C) \to 0$ as $h \to 0$, both size and power tend to zero.

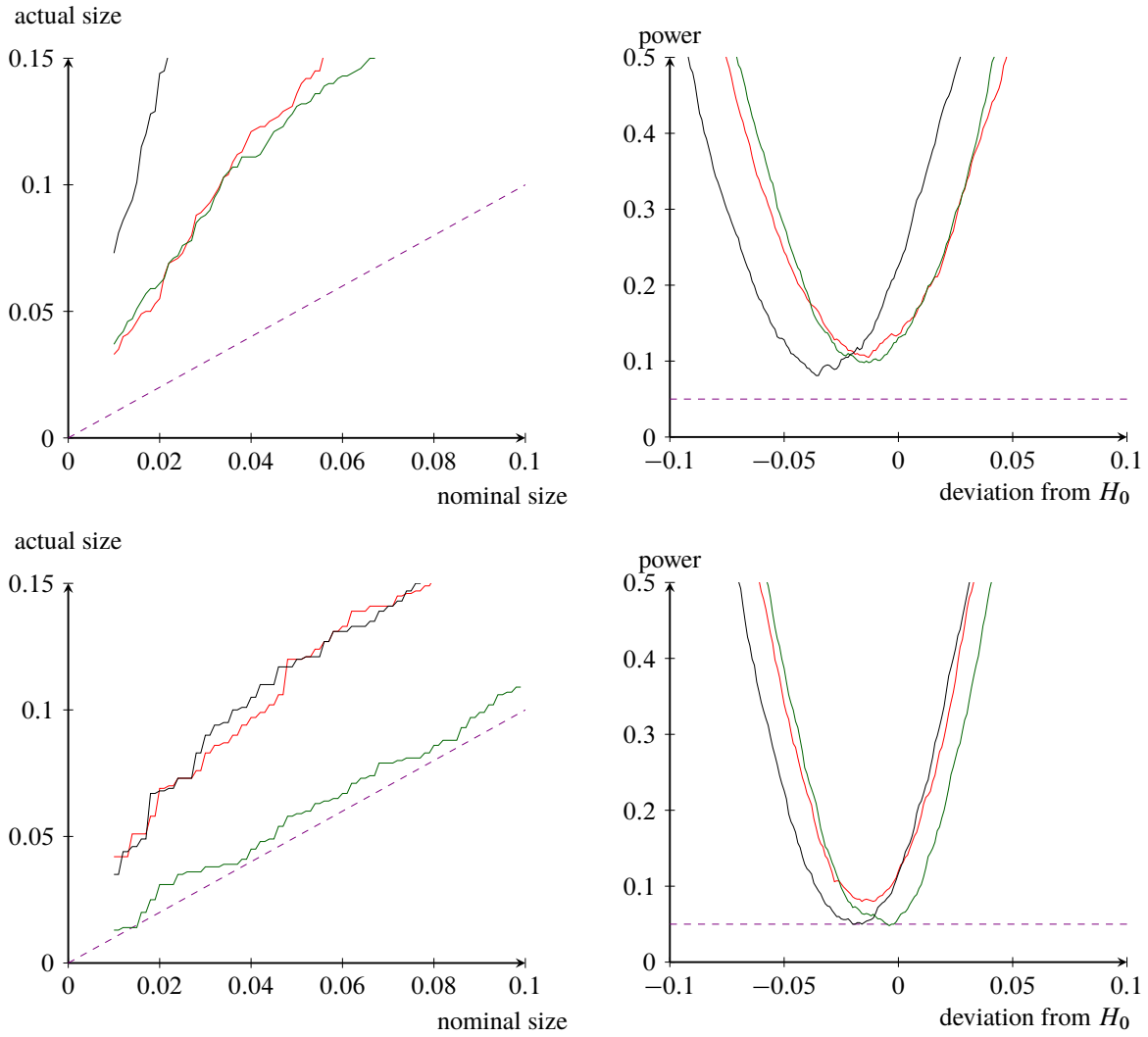

Figure 4: Size (left) and power (right) plots for $n = 1,000$ (top) and $n = 2,000$ (bottom) for *probit* with five regressors and a moderate amount of smoothing for the SMSE (red) and the Laplace estimator with uniform prior (black) and t prior (green)

of the prior diminishes as α_n increases, which is borne out by the fact that the size and powers curves of the Laplace estimators for large α_n almost coincide.

The way to interpret the power graphs is as follows. Zero on the horizontal axis corresponds to the null hypothesis where we would hope to see a vertical axis value equal to the nominal rejection probability 0.05. With little smoothing the graph is approximately symmetric. When one introduces smoothing bias becomes an issue, especially for the uniform prior estimator.

The power improves as the sample size increases, which can be seen by comparing the power values for a given deviation from H_0 in the top and bottom graphs.

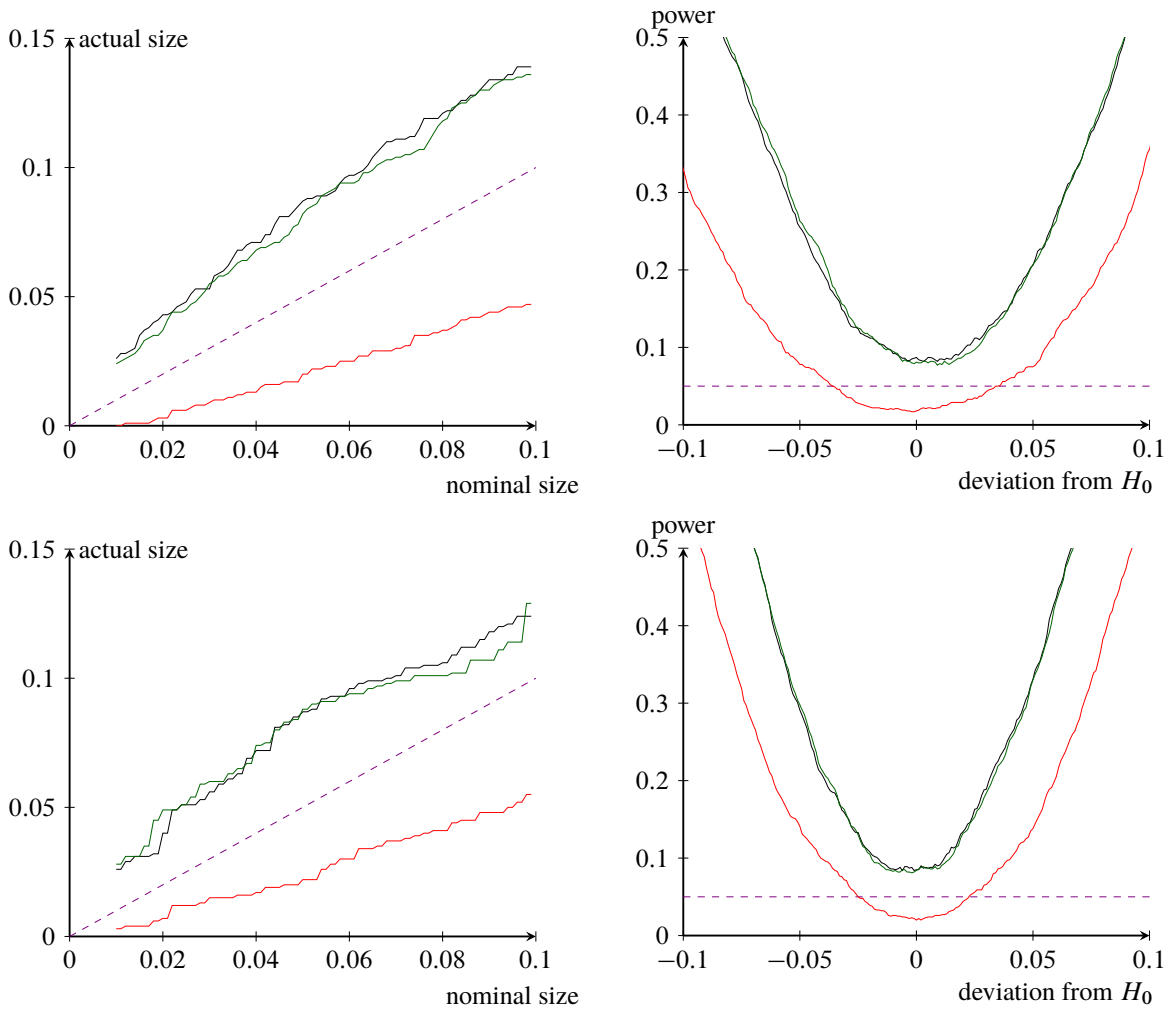

Figure 5: Size (left) and power (right) plots for $n = 1,000$ (top) and $n = 2,000$ (bottom) for *probit* with five regressors and little smoothing for the SMSE (red) and the Laplace estimator with uniform prior (black) and t prior (green)

8 Concluding remarks

We have proposed a Laplace estimator for a semiparametric binary choice model which, depending on the choice of input parameters and smoothness conditions, has one of three limit distributions and is either $\sqrt[3]{n}$–convergent or converges faster than that. We show that our estimator is more efficient than Manski's maximum score estimator under the same conditions. Further, when extra smoothness conditions are satisfied, our estimator and Horowitz's smoothed maximum score estimator have the same convergence rate, but the estimators cannot be ranked in terms of asymptotic efficiency. Our estimator is easier to compute than both the original and the smoothed maximum score estimators; computation of our estimator is based on the Gibbs sampler. We provide an inference procedure that is uniform in the choice of input parameter: no analogous procedure is available for maximum score–type estimators. Finally, our estimator appears to work well in a simulation study.

References

Abrevaya, J., and J. Huang (2005): "On the bootstrap of the maximum score estimator," *Econometrica*, 73(4), 1175–1204.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), 189–218.

Chen, S., and H. Zhang (2014): "Binary quantile regression with local polynomial smoothing," Discussion paper, HKUST discussion paper.

Chernozhukov, V., and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115(2), 293–346.

Cowles, M. K., and B. P. Carlin (1996): "Markov chain Monte Carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association*, 91(434), 883–904.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford university press.

Florios, K., and S. Skouras (2008): "Exact computation of max weighted score estimators," *Journal of Econometrics*, 146(1), 86–91.

Hong, H., A. Mahajan, and D. Nekipelov (2010): "Extremum estimation and numerical derivatives," Discussion paper, The University of California at Berkeley.

Horowitz, J. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, 60(3), 505–31.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58(1), 71–120.

Jun, S. J., J. Pinkse, and Y. Wan (2009): "Cube root N and faster convergence, Laplace estimators, and uniform inference," Discussion paper, Pennsylvania State University discussion paper.

— (2015): "Classical Laplace estimation for-consistent estimators: Improved convergence rates and rate-adaptive inference," *Journal of Econometrics*, 187(1), 201–216.

Khan, S., and E. Tamer (2010): "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, 78(6), 2021–2042.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *Annals of Statistics*, 18(1), 191–219.

Klein, R., and R. Spady (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61, 387–421.

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

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

Lee, S. M. S., and M. Pun (2006): "On m out of n bootstrapping for nonstandard M-estimation with nuisance parameters," *Journal of the American Statistical Association*, 101(475), 1185–1197.

Lewbel, A. (1998): "Semiparametric latent variable model estimation with endogenous or mismeasured regressors," *Econometrica*, 66(1), 105–121.

— (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97(1), 145–177.

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. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, pp. 357–362.

Müller, M. E. (1959): "A note on a method for generating points uniformly on n-dimensional spheres," *Communications of the ACM*, 2(4), 19–20.

Patra, R. K., E. Seijo, and B. Sen (2011): "A consistent bootstrap procedure for the maximum score estimator," *arXiv preprint*.

Phillips, P. C. B. (1984): "The exact distribution of LIML: I," *International Economic Review*, 25(1), 249–261.

— (1989): "Partially identified econometric models," *Econometric Theory*, 5(02), 181–240.

Pinkse, C. (1993): "On the computation of semiparametric estimates in limited dependent variable models," *Journal of Econometrics*, 58, 185–205.

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.

Stroud, A. H. (1971): *Approximate calculation of multiple integrals*. Prentice-Hall.

van der Vaart, A., and J. Wellner (1996): *Weak convergence and empirical processes*. Springer.

Appendices

A Technical Lemmas

Let $\boldsymbol{\xi_i} = [\boldsymbol{y_i}, \boldsymbol{x_i^\intercal}]^\intercal$. Let $g(\boldsymbol{\xi_i}, \theta) = \boldsymbol{g_i}(\theta) = (2\boldsymbol{y_i} - 1)\{\mathbb{1}(\boldsymbol{a_i} \le \theta^\intercal \boldsymbol{z_i}) - \mathbb{1}(\boldsymbol{a_i} \le \theta_0^\intercal \boldsymbol{z_i})\}$ such that θ_0 is the maximizer of $\mathbb{E}\boldsymbol{g_i}(\theta)$. Let further $\mathscr{I}(\mathscr{A})$ be the Vapnik–Černovenkis (VČ) index of a given function class $\mathscr{A}$. Further, for some sequence $\tilde{\alpha}_n$ with $1/\tilde{\alpha}_n = o(1)$ and $\tilde{\alpha}_n = o(n)$, let $\mathscr{F}_n = \left\{\sqrt{\tilde{\alpha}_n} g(\cdot, \theta_0 + t/\tilde{\alpha}_n)/c_t\right\}_{t \in \mathbb{R}^d}$, where $c_t = 1 + \|t\|$. Note that $\mathscr{F}_n$ has an envelope function $F_n(\xi) = \sqrt{\tilde{\alpha}_n}\|z\|/(\|z\| + \tilde{\alpha}_n |a - z^\intercal \theta_0|)$. We will write $\boldsymbol{F_{ni}}$ for $F_n(\boldsymbol{\xi_i})$.

Lemma A.1. $\partial_{\theta\theta^\intercal} \mathbb{E}\boldsymbol{g_i}(\theta) = 2\mathbb{E}\left[\boldsymbol{z_i z_i^\intercal} \partial_a p(\boldsymbol{z_i^\intercal}\theta_0, \boldsymbol{z_i}) f(\boldsymbol{z_i^\intercal}\theta_0 | \boldsymbol{z_i})\right] = -V < 0.$

Proof. The left hand side of the lemma statement equals

$$\partial_{\theta\theta^\intercal} \mathbb{E}\left\{\int_{-\infty}^{\boldsymbol{z_i^\intercal}\theta} \big(2p(a, \boldsymbol{z_i}) - 1\big) f(a|\boldsymbol{z_i}) da\right\} = \partial_{\theta^\intercal} \mathbb{E}\left\{\boldsymbol{z_i}\big(2p(\boldsymbol{z_i^\intercal}\theta, \boldsymbol{z_i}) - 1\big) f(\boldsymbol{z_i^\intercal}\theta | \boldsymbol{z_i})\right\}$$
$$= 2\mathbb{E}\left\{\boldsymbol{z_i z_i^\intercal} \partial_a p(\boldsymbol{z_i^\intercal}\theta, \boldsymbol{z_i}) f(\boldsymbol{z_i^\intercal}\theta | \boldsymbol{z_i})\right\} + \mathbb{E}\left\{\boldsymbol{z_i z_i^\intercal}\big(2p(\boldsymbol{z_i^\intercal}\theta, \boldsymbol{z_i}) - 1\big) f'(\boldsymbol{z_i^\intercal}\theta | \boldsymbol{z_i})\right\}. \quad (18)$$

The second right hand side term in (18) equals zero at θ_0. To see that $V > 0$, recall that $p(a, z) = \mathbb{P}(\boldsymbol{y_i} = 1 | \boldsymbol{a_i} = a, \boldsymbol{z_i} = z) = \mathbb{P}(\boldsymbol{u_i} \ge a - \theta_0^\intercal z | \boldsymbol{a_i} = a, \boldsymbol{z_i} = z)$, which equals 0.5 if $a = \theta_0^\intercal z$. □

Lemma A.2. $\lim_{\alpha\to\infty} \alpha \mathbb{E}\{\boldsymbol{g_1}(\theta_0 + t/\alpha)\boldsymbol{g_1}(\theta_0 + s/\alpha)\} = \mathbb{E}\left\{\left|M(\boldsymbol{z_i^\intercal}t, \boldsymbol{z_i^\intercal}s, 0)\right| f(\boldsymbol{z_i^\intercal}\theta_0 | \boldsymbol{z_i})\right\} = H(t, s).$

Proof. Let

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

where the second equality holds by the dominated convergence theorem. Thus, noting that $(2\boldsymbol{y_i} - 1)^2 = 1$, the left hand side of the lemma statement equals

$$\mathscr{H}(t, s) - \mathscr{H}(t, 0) - \mathscr{H}(0, s) + \mathscr{H}(0, 0) =$$
$$\mathbb{E}\left[\left\{\min(\boldsymbol{z_i^\intercal}t, \boldsymbol{z_i^\intercal}s) - \min(\boldsymbol{z_i^\intercal}t, 0) - \min(\boldsymbol{z_i^\intercal}s, 0)\right\} f(\boldsymbol{z_i^\intercal}\theta_0 | \boldsymbol{z_i})\right] = \mathbb{E}\left[\left|M(\boldsymbol{z_i^\intercal}t, \boldsymbol{z_i^\intercal}s, 0)\right| f(\boldsymbol{z_i^\intercal}\theta_0 | \boldsymbol{z_i})\right]. \quad \square$$

Lemma A.3. $\mathbb{E}\boldsymbol{F_{ni}^2} = O(1).$

Proof. Take n large enough to ensure that $\tilde{\alpha}_n \ge 1$. Then for $z \ne 0$,

$$\mathbb{E}\big(\boldsymbol{F_{ni}^2} | \boldsymbol{z_i} = z\big) = \tilde{\alpha}_n \int \frac{f(a|z)}{\left(1 + \tilde{\alpha}_n \frac{|a - z^\intercal\theta_0|}{\|z\|}\right)^2} da \le \sup_a f(a|z)\|z\| \int \frac{db}{(1 + |b|)^2} \le 2 \sup_a f(a|z)\|z\|,$$

noting that the left and right hand side expressions are equal if $z = 0$. Apply assumption E. □

Lemma A.4. *For all* $\epsilon > 0$, $\mathbb{E}\{\boldsymbol{F}_{\boldsymbol{ni}}^{\boldsymbol{2}} \mathbb{1}(\boldsymbol{F}_{\boldsymbol{ni}} > \epsilon\sqrt{n})\} = o(1)$.

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

Lemma A.5. *For every* $\epsilon_n \downarrow 0$, $\sup_{\|s-t\|<\epsilon_n} \mathbb{E}\big[\tilde{\alpha}_n\{\boldsymbol{g_i}(\theta_0 + t/\tilde{\alpha}_n)/c_t - \boldsymbol{g_i}(\theta_0 + s/\tilde{\alpha}_n)/c_s\}^2\big] = o(1)$.

Proof. Note that

$$\begin{aligned}
\mathbb{E}\bigg[\tilde{\alpha}_n\bigg\{\boldsymbol{g_i}\Big(\theta_0 + \frac{t}{\tilde{\alpha}_n}\Big)/c_t - \boldsymbol{g_i}\Big(\theta_0 + \frac{s}{\tilde{\alpha}_n}\Big)/c_s\bigg\}^2\bigg] \leq \\
\frac{2\tilde{\alpha}_n}{c_t^2}\mathbb{E}\bigg\{\boldsymbol{g_i}\Big(\theta_0 + \frac{t}{\tilde{\alpha}_n}\Big) - \boldsymbol{g_i}\Big(\theta_0 + \frac{s}{\tilde{\alpha}_n}\Big)\bigg\}^2 + 2\tilde{\alpha}_n\Big(\frac{1}{c_t} - \frac{1}{c_s}\Big)^2\mathbb{E}\boldsymbol{g_i^2}\Big(\theta_0 + \frac{s}{\tilde{\alpha}_n}\Big) \\
\leq 2\|t-s\|\mathbb{E}\Big\{\sup_a f(a|\boldsymbol{z_i})\|z_i\|\Big\} + \frac{2\|s\|(c_s-c_t)^2}{c_s^2c_t^2}\mathbb{E}\Big\{\sup_a f(a|\boldsymbol{z_i})\|z_i\|\Big\} \\
\leq \{2\|t-s\|(1+\|t-s\|)\}\mathbb{E}\Big\{\sup_a f(a|\boldsymbol{z_i})\|z_i\|\Big\},
\end{aligned}$$

which tends to zero as $\|t-s\|$ tends to zero. □

Let $\mathscr{C}$ be a class of sets and let A be an arbitrary set. Let $\mathcal{G} = \{\mathbb{1}(\cdot \in C) - \mathbb{1}(\cdot \in A) : C \in \mathscr{C}\}$ and $\tilde{\mathcal{G}} = \{\kappa\{\mathbb{1}(\cdot \in C) - \mathbb{1}(\cdot \in A)\} : C \in \mathscr{C}, \kappa > 0\}$.

Lemma A.6. $\mathscr{I}(\mathcal{G}) = \mathscr{I}(\tilde{\mathcal{G}})$.

Proof. For any (fixed) $\kappa > 0$ and $C \in \mathscr{C}$ and any $(\check{x}, \check{y})$,

$$\begin{cases} 0 \leq \check{y} \leq \kappa\{\mathbb{1}(\check{x} \in C) - \mathbb{1}(\check{x} \in A)\} \iff 0 \leq \check{y} \leq \mathbb{1}(\check{x} \in C) - \mathbb{1}(\check{x} \in A), \\ \kappa\{\mathbb{1}(\check{x} \in C) - \mathbb{1}(\check{x} \in A)\} \leq \check{y} \leq 0 \iff \mathbb{1}(\check{x} \in C) - \mathbb{1}(\check{x} \in A) \leq \check{y} \leq 0, \end{cases}$$

because $\mathbb{1}(\check{x} \in C) - \mathbb{1}(\check{x} \in A) \in \{-1, 0, 1\}$. Therefore, $\{(\check{x}_1, \check{y}_1), \ldots, (\check{x}_n, \check{y}_n)\}$ is shattered by the between–graphs of $\mathscr{G}$ if and only if it is shattered by those of $\tilde{\mathscr{G}}$. □

Lemma A.7. $\mathscr{I}(\mathscr{F}_n) \leq 2d + 5$.

Proof. Let $\tilde{\mathscr{F}}_n = \big\{\{\upsilon_n(\cdot;t) - \upsilon_n(\cdot;0)\}/c_t\big\}_{t\in\mathbb{R}^d}$, where $\upsilon_n(\xi,t) = \mathbb{1}\big(z^\intercal(\theta_0 + t/\tilde{\alpha}_n) \geq a\big)$. Further let $\bar{\mathscr{F}} = \{\psi(\cdot;t,s)\}_{(t,s)\in\mathbb{R}^{d+1}}$, where $\psi(\xi;t,s) = \mathbb{1}(z^\intercal t + as \geq 0)$. Since every element of $\mathscr{F}_n$ has the form $\sqrt{\tilde{\alpha}_n}(2y-1)\{\upsilon_n(\cdot;t) - \upsilon_n(\cdot;0)\}/c_t$, it follows from Kosorok (2008, lemma 9.12) and A.6 that

$$\mathscr{I}(\mathscr{F}_n) \leq 2\mathscr{I}(\tilde{\mathscr{F}}_n) - 1 \leq 2\mathscr{I}(\bar{\mathscr{F}}) - 1.$$

Therefore, the lemma assertion follows from the fact that $\mathscr{I}(\bar{\mathscr{F}}) \leq d + 3$ by Kosorok (2008, lemma 9.12). □

Lemma A.8. *The pair* $(\mathscr{F}_n, F_n)$ *satisfies*

$$\limsup_n \sup_{\mathscr{Q}} \int_0^1 \sqrt{\log \mathscr{N}\{\varepsilon \|F_n\|_{\mathscr{Q},2}, \mathscr{F}_n, L_2(\mathscr{Q})\}} d\varepsilon < \infty,$$

where the supremum is taken over all (finitely) discrete probability measures $\mathscr{Q}$ *with* $\|F_n\|_{\mathscr{Q},2} > 0$.

Proof. By A.7, $\mathscr{F}_n$ has a finite VČ index that does not depend on n. Therefore, the lemma follows from Kosorok (2008, lemma 11.21). □

Lemma A.9.

$$\mathbb{E}\left[\frac{\int \|t\mathbb{G}(t)\| \exp\{|\mathbb{G}(t)|\}\phi_V(t)dt}{\int \exp\{-|\mathbb{G}(t)|\}\phi_V(t)dt}\right] < \infty.$$

Proof. By the Jensen inequality,

$$\frac{1}{\int \exp\{-|\mathbb{G}(t)|\}\phi_V(t)dt} \leq \int \exp\{|\mathbb{G}(t)|\}\phi_V(t)dt.$$

So it suffices to show that

$$\mathbb{E}\left[\int \|t\mathbb{G}(t)\| \exp\{|\mathbb{G}(t)| + |\mathbb{G}(s)|\}\phi_V(t)\phi_V(s)dtds\right] < \infty.$$

Repeated application of the Schwarz inequality reduces the problem to showing that for any fixed and finite C,

$$\int \mathbb{E} \exp\{C|\mathbb{G}(t)|\}\phi_V(t)dt < \infty.$$

But by the properties of the lognormal distribution,

$$\int \mathbb{E} \exp\{C|\mathbb{G}(t)|\}\phi_V(t)dt \leq 2\int \mathbb{E} \exp\{C\mathbb{G}(t)\}\phi_V(t)dt = 2\int \exp\{C^2 H(t,t)\}\phi_V(t)dt < \infty,$$

as desired. □

B Lemmas for Uniform Inference

We will be using the mapping

$$m_j(s,\beta) = \begin{cases} \beta^{-1} M_j\{\exp(\beta s) - 1\}, & \beta > 0, \\ M_j s, & \beta = 0, \end{cases} \qquad j = 1, 2, \tag{19}$$

where

$$M_j a = \int \eta_j(t) a(t) \phi_V(t) dt,$$

with $\eta_1(t) = t$, $\eta_2(t) = tD_{\pi 1}(t)/\pi_0$, and $\eta_3(t) = 1$. Note that

$$\partial_\beta m_j(s,\beta) = \begin{cases} \beta^{-1} M_j[s\{\exp(\beta s) - 1\}] - \beta^{-2} M_j\{\exp(\beta s) - 1 - \beta s\}, & \beta > 0, \\ M_j s^2/2, & \beta = 0. \end{cases} \tag{20}$$

Below we establish the expansions in (10) and (11) under, among others, assumptions J and K. Letting $\tilde{\alpha}_n = \alpha_n$, recall that $\boldsymbol{g_i}(\theta) = (2\boldsymbol{y_i} - 1)\{\mathbb{1}(\boldsymbol{a_i} \le \theta^\intercal \boldsymbol{z_i}) - \mathbb{1}(\boldsymbol{a_i} \le \theta_0^\intercal \boldsymbol{z_i})\}$, and define[33]

$$\begin{cases} \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) = \sqrt{n\tilde{\alpha}_n}\{\boldsymbol{L_n}(\theta_0 + t/\tilde{\alpha}_n) - Q(\theta_0 + t/\tilde{\alpha}_n)\}, \\ Q_n(t) = \tilde{\alpha}_n^2 Q(\theta_0 + t/\tilde{\alpha}_n), \\ \pi_n(t) = \pi(\theta_0 + t/\tilde{\alpha}_n). \end{cases} \tag{21}$$

We first show that $\sqrt{n/\alpha_n}(\hat{\boldsymbol{\theta}} - \theta_0)$ is for $\beta_n = \sqrt{\alpha_n^3/n}$ equal to

$$\begin{aligned} &\frac{1}{\beta_n} \frac{\int t\pi_n(t) \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + \alpha_n^2 Q(\theta_0 + t/\alpha_n)\} dt}{\int \pi_n(t) \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t) + \alpha_n^2 Q(\theta_0 + t/\alpha_n)\} dt} \\ &\qquad \simeq \frac{1}{\beta_n} \frac{\int t\{1 + D_{\pi 1}(t)/\pi_0\alpha_n\}\{1 + D_{Q3}(t)/\alpha_n\} \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} \phi_V(t) dt}{\int \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} \phi_V(t) dt} \\ &\qquad \simeq \frac{m_1(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n) + \alpha_n^{-1} m_2(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n) + \alpha_n^{-1}\beta_n^{-1} \int \{tD_{\pi 1}(t)/\pi_0 + D_{Q3}(t)\} \phi_V(t) dt}{1 + \beta_n m_3(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)} \\ &\qquad \simeq \frac{m_1(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n) + \alpha_n^{-1}\beta_n^{-1} \int \{tD_{\pi 1}(t)/\pi_0 + D_{Q3}(t)\} \phi_V(t) dt}{1 + \beta_n m_3(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)}. \end{aligned} \tag{22}$$

We then show that when we simulate $\hat{\boldsymbol{\Psi}}$ as in (7) we get exactly the same object as (22) with $\hat{\mathbb{G}}$ in lieu of $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$. In particular, after substitution of $t \leftarrow \beta_n^{2/3} t$ in (7), $\sqrt{n/\alpha_n}\hat{\boldsymbol{\Psi}}$ is equal to

$$\frac{1}{\beta_n} \frac{\int t\lambda\big(\hat{\boldsymbol{\theta}} + t/\alpha_n\big) \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt}{\int \lambda\big(\hat{\boldsymbol{\theta}} + t/\alpha_n\big) \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt},$$

which will be shown to be approximated by

$$\begin{aligned} &\frac{1}{\beta_n} \frac{\int t\big\{1 + \alpha_n^{-1} t^\intercal \partial_\theta \log \lambda(\theta_0)\big\} \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt}{\int \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt} \\ &\qquad \simeq \frac{m_1(\hat{\mathbb{G}}, \beta_n) + (\alpha_n\beta_n)^{-1} \int t\{D_{\lambda 1}(t)/\lambda_0\} \phi_V(t) dt}{1 + m_3(\hat{\mathbb{G}}, \beta_n)}. \end{aligned}$$

Lemma B.1. *For $j = 0, 1$ and $R_n(t) = \pi_n(t) \exp\big[\alpha_n^2\{Q_n(t) + t^\intercal V t/2\}\big] - \{\pi_0 + D_{\pi 1}(t)/\alpha_n\}\{1 +$*

[33] These definitions are consistent with those in JPW15.

$D_{Q3}(t)/\alpha_n\}$,

$$\left\| \int t^j \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} R_n(t) \exp(-t^\intercal V t/2)\big] dt \right\| = o_p(\beta_n).$$

Proof. It suffices to show that

$$\begin{cases} \left\| \int t^j [\exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} - 1] R_n(t) \exp(-t^\intercal V t/2) dt \right\| = o_p(\beta_n), & (23) \\ \left\| \int t^j R_n(t) \exp(-t^\intercal V t/2) dt \right\| = o_p(\beta_n). & (24) \end{cases}$$

Since (24) follows from the fact that π satisfies (8), we focus on (23). Because for an arbitrary polynomial P, $\left\| P(t)\big[\exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} - 1\big] \right\| \le \beta_n \| P(t) \| \exp\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\}$, it suffices to show that

$$\int \| P(t) \| \exp\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\} \| R_n(t) \| \exp(-t^\intercal V t/2) dt = o_p(1),$$

which follows from JPW15, lemma B12, because $\exp\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\} \le \exp\{\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} + \exp\{-\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\}$. □

Lemma B.2. $(\alpha_n^2 \beta_n)^{-1} \int t D_{\pi 1}(t) D_{Q3}(t) \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} \phi_V(t) dt = o_p(1)$.

Proof. Since by JPW15, lemma B3,

$$(\alpha_n^2 \beta_n)^{-1} \int \| t D_{\pi 1}(t) D_{Q3}(t) \| \big| \exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} - 1 \big| \phi_V(t) dt$$
$$\le \alpha_n^{-2} \int \| t D_{\pi 1}(t) D_{Q3}(t) \| \exp\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\} \phi_V(t) dt$$

the lemma statement follows from JPW15, lemma B6. □

Lemma B.3. $(\alpha_n \beta_n)^{-1} \int t\{D_{\pi 1}(t) + D_{Q3}(t)\}[\exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} - 1] \phi_V(t) dt = o_p(1)$.

Proof. Letting $P(t) = t\{D_{\pi 1}(t) + D_{Q3}(t)\}$, note that by JPW15, lemma B3,

$$\left\| P(t)[\exp\{\beta_n \tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\} - 1] \right\| \le \beta_n \| P(t) \| \exp\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\}.$$

Therefore, it suffices to show that $\int \| P(t) \| \exp\{|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\} \phi_V(t) dt = O_p(1)$, which follows from JPW15, lemma B6. □

Lemma B.4. *For* $j = 0, 1$,

$$\left\| \int t^j \{\lambda(\hat{\boldsymbol{\theta}} + t/\alpha_n) - \lambda_0 - D_{\lambda 1}(t)/\alpha_n\} \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_{\hat{\boldsymbol{V}}}(t) dt \right\| = o_p(\beta_n).$$

Proof. It suffices to show that

$$\begin{cases} \left\| \int t^j \{\lambda(\hat{\boldsymbol{\theta}} + t/\alpha_n) - \lambda_0 - D_{\lambda 1}(t)/\alpha_n\} \big[\exp\{\beta_n \hat{\mathbb{G}}(t)\} - 1\big] \phi_{\hat{\boldsymbol{V}}}(t) dt \right\| = o_p(\beta_n), & (25) \\ \left\| \int t^j \{\lambda(\hat{\boldsymbol{\theta}} + t/\alpha_n) - \lambda_0 - D_{\lambda 1}(t)/\alpha_n\} \phi_{\hat{\boldsymbol{V}}}(t) dt \right\| = o_p(\beta_n). & (26) \end{cases}$$

Because (26) follows from assumption K we focus on (25). Since for an arbitrary polynomial P, $\left\| P(t)\big[\exp\{\beta_n \hat{\mathbb{G}}(t)\} - 1\big] \right\| \leq \beta_n \|P(t)\| \exp\{|\hat{\mathbb{G}}(t)|\}$, it suffices to show that

$$\int \|t^j \{\lambda(\hat{\boldsymbol{\theta}} + t/\alpha_n) - \lambda_0 - D_{\lambda 1}(t)/\alpha_n\} \exp\{|\hat{\mathbb{G}}(t)|\} \phi_{\hat{\boldsymbol{V}}}(t) dt = o_p(1). \quad (27)$$

By JPW15, lemma H4,

$$\int \|P(t)\| \exp\{|\hat{\mathbb{G}}(t)|\} \phi_{\hat{\boldsymbol{V}}}(t) dt = O_p(1), \quad (28)$$

which implies that

$$\int t^j \lambda(\hat{\boldsymbol{\theta}} + t/\alpha_n) \exp\{|\hat{\mathbb{G}}(t)|\} \phi_{\hat{\boldsymbol{V}}}(t) dt = \int t^j \{\lambda(\hat{\boldsymbol{\theta}}) + t^\intercal \partial_\theta \lambda(\hat{\boldsymbol{\theta}})/\alpha_n\} \exp\{|\hat{\mathbb{G}}(t)|\} \phi_{\hat{\boldsymbol{V}}}(t) dt + o_p(1). \quad (29)$$

Finally, (27) follows from (28) and (29), the continuity of λ and $\partial_\theta \lambda$, and the consistency of $\hat{\boldsymbol{\theta}}$. □

Lemma B.5. *For any polynomial P,*

$$\int P(t) \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_{\hat{\boldsymbol{V}}}(t) dt = \int P(t) \exp\{\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt + o_p(1).$$

Proof. Note first that

$$\int \|P(t)\| \exp\{\beta_n \hat{\mathbb{G}}(t)\} \big|\phi_{\hat{\boldsymbol{V}}}(t) - \phi_V(t)\big| dt \leq \sqrt{\int \|P(t)\|^2 \exp\{2\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt \int \big\|\phi_{\hat{\boldsymbol{V}}}(t)/\phi_V(t) - 1\big\|^2 \phi_V(t) dt},$$

where $\int \|P(t)\|^2 \exp\{2\beta_n \hat{\mathbb{G}}(t)\} \phi_V(t) dt = O_p(1)$ by JPW15, lemma H2. Moreover, $\int \big|\phi_{\hat{\boldsymbol{V}}}(t)/\phi_V(t)^{-1} - 1\big|^2 \phi_V(t) dt = o_p(1)$ by the dominated convergence theorem. □

Lemma B.6. *For any polynomial P,*

$$\int P(t)\big[\exp\{\beta_n \hat{\mathbb{G}}(t)\} - 1\big] \phi_{\hat{\boldsymbol{V}}}(t) dt = \int P(t)\big[\exp\{\beta_n \hat{\mathbb{G}}(t)\} - 1\big] \phi_V(t) dt + o_p(\beta_n).$$

Proof. Note that $\left\| P(t)\big[\exp\{\beta_n \hat{\mathbb{G}}(t)\} - 1\big] \right\| \leq \beta_n \|P(t)\| \exp\{\hat{\mathbb{G}}(t)\}$ and follow the same logic as B.5.

Lemma B.7.

$$\sqrt{n/\alpha_n}\hat{\boldsymbol{\Psi}} = \frac{m_1(\hat{\mathbb{G}}, \beta_n) + (\alpha_n\beta_n)^{-1}\int t\{D_{\lambda 1}(t)/\lambda_0\}\phi_V(t)dt}{1 + m_3(\hat{\mathbb{G}}, \beta_n)} + o_p(1).$$

Proof. By substitution of $t \leftarrow \beta_n^{2/3}t$ and by B.4 to B.6,

$$\begin{aligned}\sqrt{n/\alpha_n}\hat{\boldsymbol{\Psi}} &= \frac{1}{\beta_n}\frac{\int t\{1 + D_{\lambda 1}(t)/\lambda_0\alpha_n\}\exp\{\beta_n\hat{\mathbb{G}}(t)\}\phi_V(t)dt}{\int \exp\{\beta_n\hat{\mathbb{G}}(t)\}\phi_V(t)dt} + o_p(1) \\ &= \frac{m_1(\hat{\mathbb{G}}, \beta_n) + (\alpha_n\beta_n)^{-1}\int t\{D_{\lambda 1}(t)/\lambda_0\}\phi_V(t)dt}{1 + m_3(\hat{\mathbb{G}}, \beta_n)} + o_p(1).\end{aligned}$$

□

Below let $B = [0, C_\beta]$ and let $m(s, \beta) = \left[m_1(s, \beta), m_2(s, \beta), \beta m_3(s, \beta)\right]^\intercal$.

Lemma B.8. *$\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \cdot)\}$ and $\{m(\hat{\mathbb{G}}, \cdot)\}$ are stochastically equicontinuous.*

Proof. Noting that by (20) and JPW15, lemma B2, $\max_{\beta\in B}\left\|\partial_\beta m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta)\right\| = O_p(1)$, the stated result follows from theorem 21.10 in Davidson (1994). The case with $\hat{\mathbb{G}}$ in lieu of $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ can be similarly dealt with by using JPW15, lemma H2. □

Lemma B.9. *For any continuous function ω with $|\omega| \le 1/2$ and for $\mathfrak{S}$ being either $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ or $\hat{\mathbb{G}}$,*

$$\lim_{\delta\downarrow 0}\mathbb{E}\sup_{|\beta-\tilde{\beta}|<\delta}\left|\omega\{m(\mathfrak{S}, \beta)\} - \omega\{m(\mathfrak{S}, \tilde{\beta})\}\right| = 0, \tag{30}$$

where $\beta, \tilde{\beta}$ are implicitly assumed to belong to B.

Proof. Consider $\mathfrak{S} = \tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ first. Choose $\epsilon > 0$. For $C, \nu > 0$, define the events

$$A_C = \left\{\max_{\beta\in B}\left\|m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta)\right\| \le C\right\}, \qquad B_\nu = \left\{\sup_{|\beta-\tilde{\beta}|<\delta}\left\|m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta) - m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \tilde{\beta})\right\| \le \nu\right\}.$$

By Boole's inequality the left hand side in (30) is bounded above by

$$\mathbb{P}\left(A_C^c\right) + \mathbb{P}\left(B_\nu^c\right) + \mathbb{E}\left[\mathbb{1}\left(A_C\right)\mathbb{1}\left(B_\nu\right)\sup_{|\beta-\tilde{\beta}|<\delta}\left|\omega\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta)\} - \omega\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \tilde{\beta})\}\right|\right]. \tag{31}$$

Since ω is continuous, it is uniformly continuous on $\{m : \|m\| \le C\}$. So there exists a $\nu > 0$ for which the expectation in (31) is bounded by ϵ. Further, by B.8, $\mathbb{P}\left(B_\nu^c\right)$ can be made less than ϵ by choosing δ sufficiently small. Finally, $\mathbb{P}\left(A_C^c\right)$ can be made smaller than ϵ by choosing a sufficiently large C, because

$$\max_{\beta\in B}\left\|m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta)\right\| \le \int\left\|\eta(t)\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)\right\|\exp\{C_\beta|\tilde{\boldsymbol{S}}_{\boldsymbol{n}}(t)|\}\phi_V(t)dt = O_p(1)$$

by JPW15, lemma B2. Hence (31) is bounded by 3ϵ for a sufficiently small δ. The case of $\mathfrak{S} = \hat{\mathbb{G}}$ can be similarly dealt with by using JPW15, lemma H2. □

Lemma B.10. *For any continuous bounded function ω and for $\mathfrak{S}$ being either $\tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ or $\hat{\mathbb{G}}$,*

$$\max_{\beta \in \boldsymbol{B}} \left| \mathbb{E}\omega\{m(\mathfrak{S}, \beta)\} - \mathbb{E}\omega\{m(\mathbb{G}, \beta)\} \right| = o(1). \tag{32}$$

Proof. Consider $\mathfrak{S} = \tilde{\boldsymbol{S}}_{\boldsymbol{n}}$ first. Let $\omega_n^*(\beta) = \mathbb{E}\omega\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta)\}$ and $\omega^*(\beta) = \mathbb{E}\omega\{m(\mathbb{G}, \beta)\}$. Divide B up into T intervals of length $\delta = C_\beta / T$ and let $\beta_{(t)}$ denote an element of interval t. By the triangle inequality, the left hand side in (32) is bounded by

$$\sup_{|\beta - \tilde{\beta}| < \delta} \left| \omega_n^*(\beta) - \omega_n^*(\tilde{\beta}) \right| + \sup_{|\beta - \tilde{\beta}| < \delta} \left| \omega^*(\beta) - \omega^*(\tilde{\beta}) \right| + \max_{t = 1, \ldots, T} \left| \omega_n^*(\beta_{(t)}) - \omega^*(\beta_{(t)}) \right|. \tag{33}$$

The last term in (33) is $o(1)$ by JPW15, lemma B2. Further, the second term in (33) is arbitrarily small when δ is sufficiently small, because ω^* is continuous on a compact set. The first term in (33) is $o(1)$ by B.9. The case of $\mathfrak{S} = \hat{\mathbb{G}}$ can be similarly dealt with by using JPW15, lemma H2, instead of JPW15, lemma B2. □

Lemma B.11. *$m(\mathbb{G}, \beta)$ has a density with respect to the Lebesgue measure that is continuous in β.*

Proof. By the Karhunen–Loève theorem $\mathbb{G}(t)$ can be written as $\sum_{i=1}^{\infty} \boldsymbol{z_i} \varphi_i(t)$ (on compacta), where $\{\boldsymbol{z_i}\}$ is i.i.d. $N(0, 1)$. This convergence is in $\mathbb{L}^2$ and uniform in t, and therefore $m(\mathbb{G}, \beta)$ and $m\{\sum_{i=1}^{\infty} \boldsymbol{z_i} \varphi, \beta\}$ are distributionally equivalent. Consider $\boldsymbol{m_N}(\beta) = m\{\sum_{i=1}^{N} \boldsymbol{z_i} \varphi_i, \beta\}$. Then $\boldsymbol{m_N}(\beta)$ has a density (of bounded variation) $\mathfrak{f}_N(\cdot; \beta)$ which is continuous in both arguments; $\mathfrak{f}_N$ can be deduced from the density of $\boldsymbol{z_1}, \ldots, \boldsymbol{z_N}$. By Helly's selection theorem, there exists a subsequence $\{\mathfrak{f}_{N_k}(\cdot; \beta)\}$ that converges a.e. to some $\mathfrak{f}(\cdot; \beta)$. Note that $m\{\sum_{i=1}^{\infty} \boldsymbol{z_i} \varphi, \beta\}$ is a continuous random variable, because the derivative of $m\{\sum_{i=1}^{\infty} \boldsymbol{z_i} \varphi, \beta\}$ with respect to any $\boldsymbol{z_i}$ is nonzero. Therefore, the distribution function of $\boldsymbol{m_N}(\beta)$ converges uniformly to that of $m(\mathbb{G}, \beta)$, which leads to

$$\begin{aligned} \mathbb{P}\{m(\mathbb{G}, \beta) \le a\} - \mathbb{P}\{m(\mathbb{G}, \beta) \le b\} &= \lim_{k \to \infty} \left[\mathbb{P}\{\boldsymbol{m_{N_k}}(\beta) \le a\} - \mathbb{P}\{\boldsymbol{m_{N_k}}(\beta) \le b\} \right] \\ &= \lim_{k \to \infty} \int_b^a \mathfrak{f}_{N_k}(m; \beta) dm = \int_b^a \mathfrak{f}(m; \beta) dm. \end{aligned}$$

Hence $\mathfrak{f}(\cdot; \beta)$ is the density of $m(\mathbb{G}, \beta)$. The continuity of $\mathfrak{f}(m; \beta)$ in β for every m follows from the convergence and continuity of $\mathfrak{f}_{N_k}$. □

Lemma B.12. *Let $h : \mathbb{R}^3 \to \mathbb{R}$ be a continuous function such that $h\{m(\mathbb{G}, \beta)\}$ is a continuous random variable. Further, let $\mathcal{B}$ be the collection of sequences $\{\beta_n\}$ with $\beta_n \in B$. Then, for any t,*

$$\sup_{\{\beta_n\} \in \mathcal{B}} \left| \mathbb{P}[h\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)\} \le t] - \mathbb{P}[h\{m(\hat{\mathbb{G}}, \beta_n)\} \le t] \right| = o(1).$$

Proof. It suffices to show that

$$\begin{cases} \sup_{\beta_n \in B} \left| \mathbb{P}[h\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)\} \leq t] - \mathbb{P}[h\{m(\mathbb{G}, \beta_n)\} \leq t] \right| = o(1), & (34) \\ \sup_{\beta_n \in B} \left| \mathbb{P}[h\{m(\hat{\mathbb{G}}, \beta_n)\} \leq t] - \mathbb{P}[h\{m(\mathbb{G}, \beta_n)\} \leq t] \right| = o(1). & (35) \end{cases}$$

Since (34) and (35) are similar, we only show (34) here. Fix $t \in \mathbb{R}$ and $\epsilon > 0$. We will use B.10 choosing the convenient continuous and bounded functions

$$\bar{\omega}_{\epsilon t}(r) = \begin{cases} 1, & r \leq t, \\ 1 - (r-t)/\epsilon, & t < m \leq t + \epsilon, \\ 0, & r > t + \epsilon. \end{cases} \qquad \tilde{\omega}_{\epsilon t}(r) = \begin{cases} 1, & r \leq t - \epsilon, \\ (t-r)/\epsilon, & t - \epsilon < r \leq t, \\ 0, & r > t. \end{cases}$$

Then, letting $\bar{\omega}^*_{\epsilon t} = \bar{\omega}_{\epsilon t} \circ h$ and $\tilde{\omega}^*_{\epsilon t} = \tilde{\omega}_{\epsilon t} \circ h$,

$$\begin{aligned} \mathbb{E}\tilde{\omega}^*_{\epsilon t}\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)\} - \mathbb{E}\bar{\omega}^*_{\epsilon t}\{m(\mathbb{G}, \beta_n)\} &\leq \mathbb{P}[h\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)\} \leq t] - \mathbb{P}[h\{m(\mathbb{G}, \beta_n)\} \leq t] \\ &\leq \mathbb{E}\bar{\omega}^*_{\epsilon t}\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta_n)\} - \mathbb{E}\tilde{\omega}^*_{\epsilon t}\{m(\mathbb{G}, \beta_n)\}. \quad (36) \end{aligned}$$

The majorant side in (36) is for $G_{\epsilon t} = \bar{\omega}^*_{\epsilon t} - \tilde{\omega}^*_{\epsilon t}$ bounded by

$$\sup_{\beta \in B} \left| \mathbb{E}\bar{\omega}^*_{\epsilon t}\{m(\tilde{\boldsymbol{S}}_{\boldsymbol{n}}, \beta)\} - \mathbb{E}\bar{\omega}^*_{\epsilon t}\{m(\mathbb{G}, \beta)\} \right| + \sup_{\beta \in B} \mathbb{E}G_{\epsilon t}[h\{m(\mathbb{G}, \beta)\}], \quad (37)$$

The first term in (37) converges to zero by B.10. We now show that $\lim_{\epsilon \downarrow 0} \sup_{\beta \in B} \mathbb{E}G_{\epsilon t}[h\{m(\mathbb{G}, \beta)\}] = 0$. Note that by B.11 $h\{m(\mathbb{G}, \beta)\}$ has density (with respect to the Lebesgue measure) $\mathfrak{f}_\beta$, which is continuous in β. Therefore, $\bar{\mathfrak{f}}(x) = \max_{\beta \in B} \mathfrak{f}_\beta(x)$ is a real–valued function, and we have

$$\sup_{\beta} \mathbb{E}G_{\epsilon t}[h\{m(\mathbb{G}, \beta)\}] \leq \sup_{\beta} \mathbb{P}[|h\{m(\mathbb{G}, \beta)\} - t| \leq \epsilon] \leq \int_{t-\epsilon}^{t+\epsilon} \bar{\mathfrak{f}}(x)dx,$$

which converges to zero as $\epsilon \downarrow 0$. □

C Lemmas for the Efficiency Results

Lemma C.1. *For all $c_\alpha > 0$, F_{c_α} has a density that is continuous in c_α and that is positive at $K = 0$.*

Proof. Recall that F_{c_α} is the distribution function of the absolute value of

$$\frac{\int \iota^\intercal t \exp\{c_\alpha^2 \mathbb{G}(t) - c_\alpha^2 t^\intercal V t/2\} dt}{\int \exp\{c_\alpha^2 \mathbb{G}(t) - c_\alpha^2 t^\intercal V t/2\} dt}. \quad (38)$$

We will work with (38) since its density at zero is half of $F'_{c_\alpha}(0)$. First, (38) can for implicitly defined

$\mathbb{N}, \mathbb{N}^*, \mathbb{D}$, and $\mathbb{D}^*$ be written as

$$\frac{\mathbb{N}-\mathbb{N}^*}{\mathbb{D}+\mathbb{D}^*} = \frac{\int_0^\infty \iota^\intercal t \exp\{c_\alpha^2 \mathbb{G}(t) - c_\alpha^2 t^\intercal V t/2\} dt - \int_0^\infty \iota^\intercal t \exp\{c_\alpha^2 \mathbb{G}^*(t) - c_\alpha^2 t^\intercal V t/2\} dt}{\int_0^\infty \exp\{c_\alpha^2 \mathbb{G}(t) - c_\alpha^2 t^\intercal V t/2\} dt + \int_0^\infty \exp\{c_\alpha^2 \mathbb{G}^*(t) - c_\alpha^2 t^\intercal V t/2\} dt}, \tag{39}$$

where $\mathbb{G}^*$ is an independent copy of $\mathbb{G}$ which implies that $(\mathbb{N}^*, \mathbb{D}^*)$ is an independent copy of $(\mathbb{N}, \mathbb{D})$. Note that $\mathbb{N}, \mathbb{D}, \mathbb{N}^*, \mathbb{D}^*$ are all positive (nonzero) with probability one and have density functions. By a change–of–variables, the density of (39) is hence equal to

$$2\int_0^\infty \mathbb{E}(\mathbb{D}|\mathbb{N}=t) f_{\mathbb{N}}^2(t) dt, \tag{40}$$

where $f_{\mathbb{N}}$ is the density of $\mathbb{N}$. Now, (40) is positive unless $\mathbb{E}(\mathbb{D}|\mathbb{N}=t)$ is zero for almost all t, which will occur only if the distribution of $\mathbb{D}$ given $\mathbb{N}=t$ is degenerate at zero. But $\mathbb{D}$ is a.s. positive. □

Lemma C.2. *For any $c_\alpha > 0$, the object in (38) has infinitely many finite moments.*

Proof. Let $\boldsymbol{N}, \boldsymbol{D}$ be the numerator and denominator in (38), respectively. Let $\tau(t) = \exp\bigl(-c_\alpha^2 t^\intercal V t/2\bigr)$. Then, by the Jensen inequality, Tonelli's theorem, properties of the lognormal distribution, and the fact that $H(t,t)$ is linear in t (see (3)),

$$\frac{\mathbb{E}|\boldsymbol{N}|^\delta}{\{\int \tau(t)dt\}^{\delta-1}} \le \mathbb{E}\left[\int \|t\|^\delta \exp\{\delta c_\alpha^2 \mathbb{G}(t)\}\tau(t)dt\right] =$$
$$\int \|t\|^\delta \mathbb{E}\exp\{\delta c_\alpha^2 \mathbb{G}(t)\}\tau(t)dt = \int \|t\|^\delta \exp\{\delta^2 c_\alpha^4 H(t,t)/2\}\tau(t)dt < \infty.$$

Therefore, it suffices to show that for any $\delta > 1$, $\mathbb{E}\boldsymbol{D}^{1-\delta} < \infty$, for which by integration by parts it is sufficient to establish that

$$\lim_{K \downarrow 0} \frac{\mathbb{P}(\boldsymbol{D} \le K)}{K^\delta} = 0.$$

Now, since $\boldsymbol{D} \ge \int_{\|t\|\le 1} \exp\{c_\alpha^2 \mathbb{G}(t)\}\tau(t)dt$, it follows that for $C = 1/\int_{\|t\|\le 1}\tau(t)dt > 0$,

$$\mathbb{P}(\boldsymbol{D} \le K) \le \mathbb{P}\Bigl[\min_{\|t\|\le 1} \exp\{c_\alpha^2 \mathbb{G}(t)\} \le CK\Bigr] = \mathbb{P}\left\{\min_{\|t\|\le 1} \mathbb{G}(t) \le \frac{\log(CK)}{c_\alpha^2}\right\}$$
$$= \mathbb{P}\left\{\max_{\|t\|\le 1} \mathbb{G}(t) \ge -\frac{\log(CK)}{c_\alpha^2}\right\}, \tag{41}$$

by symmetry. The right hand side in (41) is for $K < 1/C$ bounded by

$$\mathbb{P}\left\{\max_{\|t\|\le 1} |\mathbb{G}(t)| \ge -\frac{\log(CK)}{c_\alpha^2}\right\} \le 2\exp\left\{-\frac{\log^2(CK)}{8c_\alpha^4 \mathbb{E}\max_{\|t\|\le 1}\mathbb{G}^2(t)}\right\} \tag{42}$$

by Borell's inequality,[34] where $\mathbb{E}\max_{\|t\|\leq 1}\mathbb{G}^2(t) < \infty$ by van der Vaart and Wellner (1996, prop.A.2.3) because $\sup_{\|t\|\leq 1} H(t,t) < \infty$. Finally, substituting $\tilde{K}$ for $-\log(CK)$ and taking $\tilde{K}\to\infty$ shows that the right hand side in (42) goes to zero faster than any power of K. □

D Proofs of Theorems

In this appendix we will use $\bar{F}$ to denote the survivor function corresponding to F, i.e. $\bar{F} = 1 - F$, and likewise for other symbols.

Proof of Theorem 1. By A.1 to A.5 and A.8, Assumptions A through G in JPW15 are satisfied. Therefore, the assertions follow from Theorem 1 of JPW15 by letting $\tilde{\alpha}_n = \alpha_n$ in cases ii and iii, and letting $\tilde{\alpha}_n = \sqrt[3]{n}$ in case i. □

Proof of Theorem 2. Fix $0 < K < \infty$ and define

$$\Upsilon = \frac{\int \|t\mathbb{G}(t)\| \exp\{|\mathbb{G}(t)|\}\phi_V(t)dt}{\int \exp\{-|\mathbb{G}(t)|\}\phi_V(t)dt}.$$

The limit of the numerator in (4) does not depend on c_α and is nonzero. So we show that the denominator in (4) can be made arbitrarily small. We will pick $c_\alpha \leq 1$; the upper bound is immaterial as long as it is fixed and finite. Now, by substitution of $t \leftarrow c_\alpha t$ and a simple application of the mean value theorem,[35]

$$\lim_{n\to\infty}\mathbb{P}\{\sqrt[3]{n}\|\hat{\boldsymbol{\theta}}_{\boldsymbol{c_\alpha}} - \theta_0\| > K\} = \mathbb{P}\left[\left\|\frac{\int t\exp\{c_\alpha^2\mathbb{G}(t) - c_\alpha^2 t^\intercal V t/2\}dt}{\int \exp\{c_\alpha^2\mathbb{G}(t) - c_\alpha^2 t^\intercal V t/2\}dt}\right\| > K\right] =$$
$$\mathbb{P}\left[\left\|\frac{\int t\exp\{c_\alpha^{3/2}\mathbb{G}(t)\}\phi_V(t)dt}{c_\alpha\int \exp\{c_\alpha^{3/2}\mathbb{G}(t)\}\phi_V(t)dt}\right\| > K\right] \leq \mathbb{P}\left[\frac{\int \|t\mathbb{G}(t)\|\exp\{c_\alpha^{3/2}|\mathbb{G}(t)|\}\phi_V(t)dt}{\int \exp\{-c_\alpha^{3/2}|\mathbb{G}(t)|\}\phi_V(t)dt} > \frac{K}{\sqrt{c_\alpha}}\right]$$
$$\leq \mathbb{P}\left(\Upsilon > K/\sqrt{c_\alpha}\right) \leq \mathbb{P}\left(\Upsilon > K/\sqrt{c_\alpha^*}\right).$$

Now, pick c_α^* small enough to satisfy (4).

The above argument also establishes the second half of the theorem with the exception of an area near $K = 0$. However, because

$$\lim_{K\downarrow 0}\frac{\bar{F}_\infty(K) - \bar{F}_{c_\alpha}(K)}{K} = F'_{c_\alpha}(0) - F'_\infty(0)$$

and $F'_\infty(0) < \infty$, it follows from C.1 that $F'_{c_\alpha}(0) > 0$ and $F'_{c_\alpha}(0) > F'_\infty(0)$ for all sufficiently small c_α. □

[34] Christer Borell, not Émile Borel; see van der Vaart and Wellner (1996, p.438).

[35] $\int t\exp\{c_\alpha^{3/2}\mathbb{G}(t)\}\phi_V(t)dt = \int t\exp\{0\mathbb{G}(t)\}\phi_V(t)dt + c_\alpha^{3/2}\int t\mathbb{G}(t)\exp\{\dot{c}^{3/2}\mathbb{G}(t)\}\phi_V(t)dt$ for some $\dot{c}$ between 0 and c_α. The first right hand side equals zero. Taking norms in appropriate places does the rest.

Proof of Theorem 3. Since by C.2 (38) has a finite mean, $\int_0^\infty \bar{F}_{c_\alpha}(K)dK < \infty$. By theorem 2 there exist $\check{K}, \epsilon > 0$ such that for some $c_\alpha^* = c_\alpha^*(\check{K}, \epsilon)$

$$\begin{cases} \sup_{c_\alpha \in [0,1]} \int_{\check{K}}^\infty \bar{F}_{c_\alpha}(K)dK & < \epsilon, \\ \inf_{0<c_\alpha<c_\alpha^*} \int_0^{\check{K}} \{\bar{F}_\infty(K) - \bar{F}_{c_\alpha}(K)\}dK & > \epsilon. \end{cases}$$

We now show that for any $K^* \geq 0$,

$$\inf_{0<c_\alpha<c_\alpha^*} \int_0^{K^*} \{\bar{F}_\infty(K) - \bar{F}_{c_\alpha}(K)\}dK \geq 0.$$

For $K^* \leq \check{K}$ this result is implied by theorem 2. For $K^* > \check{K}$,

$$\begin{aligned} &\inf_{0<c_\alpha<c_\alpha^*} \int_0^{K^*} \{\bar{F}_\infty(K) - \bar{F}_{c_\alpha}(K)\}dK \\ &\qquad \geq \inf_{0<c_\alpha<c_\alpha^*} \int_0^{\check{K}} \{\bar{F}_\infty(K) - \bar{F}_{c_\alpha}(K)\}dK - \sup_{c_\alpha \in [0,1]} \int_0^{\check{K}} \bar{F}_{c_\alpha}(K)dK \geq \epsilon - \epsilon \geq 0. \end{aligned}$$

□

Proof of Theorem 4. By Müller (1959) if $\boldsymbol{\zeta} \sim N(0, I_{d+1})$ then $\boldsymbol{\zeta}/\|\boldsymbol{\zeta}\|$ has a uniform distribution on the sphere. The current normalization, however, normalizes the last element to equal one so the corresponding quantity is $\boldsymbol{\zeta}^* = \tilde{\boldsymbol{\zeta}}/|\zeta_{d+1}|$ where $\boldsymbol{\zeta} = [\tilde{\boldsymbol{\zeta}}^\intercal, \zeta_{d+1}]^\intercal$. Let $\Lambda = \sqrt{\|\zeta^*\|^2 + 1}$. Then a simple change of variables argument reveals the density of $\boldsymbol{\zeta}^*$ at ζ^* to be

$$\frac{2}{(2\pi)^{(d+1)/2}} \int_0^\infty t^d \exp\left(-t^2\Lambda^2/2\right)dt \propto \Lambda^{-d-1},$$

which is the density of a d–variate Cauchy distribution. To obtain the conditional density in the theorem statement observe that for $\Lambda^* = \sqrt{\sum_{j=2}^d \zeta_j^{*2} + 1}$ the conditional density is proportional to

$$\frac{(\Lambda^*)^d}{\Lambda^{d+1}} = \frac{1}{\Lambda^*\{(t/\Lambda^*)^2 + 1\}^{(d+1)/2}},$$

which after a scale adjustment to remove Λ^* indeed leads to a t distribution with d degrees of freedom. □

Proof of Theorem 5. It follows from B.12. □