$\sqrt{n}$–Consistent Robust Integration–Based Estimation*

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

April 2010

We propose a new robust estimator of the regression coefficients in a linear regression model. The proposed estimator is the only robust estimator based on integration rather than optimization. It allows for dependence between errors and regressors, is $\sqrt{n}$–consistent, and asymptotically normal. It moreover has the best–achievable *breakdown point* of regression–invariant estimators, has bounded *gross error sensitivity*, is both *affine invariant* and *regression–invariant*, and the number of operations required for its computation is linear in n. An extension would result in bounded *local shift sensitivity*, also.

*This paper is based on research supported by NSF grant SES–0922127. We thank the Human Capital Foundation for their support of CAPCP. We thank Don Andrews, Roger Koenker, Peter Robinson, Neil Wallace and Haiqing Xu for helpful suggestions.

†sjun@psu.edu (corresponding author), 303 Kern Graduate Building, University Park, PA 16802
‡joris@psu.edu
†yxw162@psu.edu

1. INTRODUCTION

We propose a new estimator for the regression coefficients in a linear regression model, which is robust to 'contamination.' Our estimator is inspired by the *least median of squares* (LMS) estimator of Rousseeuw (1984) and the *Laplace estimator* of Chernozhukov and Hong (2003); see also Jun, Pinkse, and Wan (2009). Like Laplace estimators, our estimator is defined as the ratio of two integrals involving an exponential transform of (in our case) the LMS objective function, but this is where the similarity ends.

Suppose that the parameter vector of interest θ_0 is the unique minimizer of a population objective function Ω over a compact parameter space Θ. Laplace estimators then employ the fact that θ_0 satisfies

$$\theta_0 = \lim_{n\to\infty} \frac{\int \theta \varpi(\theta) \exp\{-\alpha_n \Omega(\theta)\} d\theta}{\int \varpi(\theta) \exp\{-\alpha_n \Omega(\theta)\} d\theta}, \tag{1}$$

where ϖ is a pseudo–prior defined on Θ and $\{\alpha_n\}$ is a scalar–valued deterministic sequence diverging to infinity with the sample size n. Note here that the density $\varpi(\theta) \exp\{-\alpha_n \Omega(\theta)\} \,/\, \int \varpi(\theta) \exp\{-\alpha_n \Omega(\theta)\} d\theta$ becomes more concentrated around θ_0 as α_n increases. Replacing Ω in (1) with its sample analog $\hat{\mathbf{\Omega}}$[1] results in a Laplace estimator. If a quadratic expansion of $\hat{\mathbf{\Omega}}$ is available then the Laplace estimator is generally $\sqrt{n}$–consistent (Chernozhukov and Hong, 2003) and the divergence rate of α_n immaterial. Absent such a quadratic expansion, as in the case of the LMS estimator, the resulting estimator is not $\sqrt{n}$–consistent, and the divergence rate of α_n partly determines the convergence rate of the Laplace estimator (Jun, Pinkse, and Wan, 2009).

We, instead, use the fact that in our case Ω is symmetric around θ_0, which implies that

$$\theta_0 = \frac{\int \theta \exp\{-\Omega(\theta)\} d\theta}{\int \exp\{-\Omega(\theta)\} d\theta}, \tag{2}$$

where the integrals are taken over the entire Euclidean space. There are four fundamental differences between (1) and (2): in (2) there is no limit, there is no α_n, there is no compact parameter space requirement, and there is no ϖ. Because there is no limit in (2), α_n is not needed anymore. Since the symmetry of Ω around θ_0 is used, the parameter space should not be artificially restricted and no prior can be used. Our estimator $\hat{\boldsymbol{\theta}}$ is obtained by replacing Ω in (2) with $\hat{\mathbf{\Omega}}$.

In this paper we focus our attention on the case in which $\hat{\mathbf{\Omega}}$ is the LMS objective function, or a close relative thereof. We show that, subject to assumptions outlined in subsequent sections, $\hat{\boldsymbol{\theta}}$ is

[1] We use bold face for random variables.

$\sqrt{n}$–consistent and asymptotically normal with many robustness properties, which will be further explained below.

Instead of basing an estimator on (2), as we do in this paper, one could alternatively consider $\hat{\boldsymbol{\theta}}_L$, the Laplace estimator using $\hat{\boldsymbol{\Omega}}$. However, because the LMS objective function does not allow for a quadratic expansion (Kim and Pollard, 1990), $\hat{\boldsymbol{\theta}}_L$ will not be $\sqrt{n}$–consistent. Indeed, this scenario is similar to the one studied in Jun, Pinkse, and Wan (2009) for the objective functions of other $\sqrt[3]{n}$–consistent estimators.

The pioneering work of Huber (1973) has spawned an abundance of papers proposing estimators with ever more desirable robustness properties. The main differences between the estimators are their robustness properties, their asymptotic behavior absent contamination, their equivariance properties, and their degree of computational complexity. These properties are summarized in table 1. Our estimator is attractive in all four respects, as the exposition below will make apparent.

One notion of robustness is the finite sample *breakdown point* (Donoho and Huber, 1982),[2] which is the fraction of the sample that must be changed to push the value of an estimator arbitrarily far. The breakdown point of the least squares estimator equals $1/n$ and the breakdown point of the least absolute deviations estimator (Koenker and Bassett, 1978) depends on the regressor distribution and can be arbitrarily close to zero in large samples (Hampel, Ronchetti, Rousseeuw, and Stahel, 1986, p.328). Most estimators, however, have a finite sample breakdown point close to 0.5 if the regressors are in *general position* (Rousseeuw, 1984). Notable exceptions are Huber (1973); Krasker (1980); Mallows (1983). Our estimator has the best achievable breakdown point of regression invariant estimators, determined in Rousseeuw (1984).

Because the requirement that regressors be in general position is strong, we provide results that are more general than that. Specifically, it can be preferable (from a breakdown point perspective) to use a quantile q other than the median. Details can be found in section 3.

Other commonly used notions of robustness are the *gross error sensitivity* (GES) and the *local shift sensitivity* (LSS), both due to Hampel (1968, 1974). The GES of an estimator is finite if its *influence function* (Hampel, 1968, 1974) is bounded. Many, but not all, robust estimators have a bounded influence function, including ours.

[2] An asymptotic version can be found in Hampel (1971) and a different breakdown point concept in Sakata and White (1995, 1998).

Estimation Method	acronym	BDP =0.5	GES finite	LSS finite	$\sqrt{n}$ rate	Normal	Comp. # oper.	Equivariance Scale	Affine	Regr.
Huber (1973)	HUB				✓	✓	?		✓	✓
Koenker and Bassett (1978)	LAD				✓	✓	n [a]	✓	✓	✓
Krasker (1980)	HK		✓		✓	✓	?		✓	✓
Siegel (1982)[b]	RM	✓	✓		✓	✓	n^d	✓		✓
Mallows (1983)	MAL		✓	✓	✓	✓	?		✓	✓
Rousseeuw (1984)	LMS	✓	✓[c]				n^d [d]	✓	✓	✓
Rousseeuw (1984)	LTS	✓	✓		✓	✓	$n \log n$	✓	✓	✓
Rousseeuw and Yohai (1984)	SEST	✓			✓	✓	$n^2 \log n$	✓	✓	✓
Yohai (1987)	MM	✓			✓	✓	?		✓	✓
Yohai and Zamar (1988)	TAU	✓			✓	✓	?		✓	✓
Croux, Rousseeuw, and Hossjer (1994)	GS	✓			✓	✓	$n^2 \log n$ [e]		✓	✓
Hossjer (1994)	LTA	✓	✓		✓	✓	$n \log n$	✓	✓	✓
Hadi and Luceno (1997)	MTLE	✓	✓		✓	✓	?	✓	✓	✓
Chang, McKean, Naranjo, and Sheather (1999)	HBRR	✓	✓[f]	?	✓	✓	?	✓	✓	✓
Zinde-Walsh (2002)	SLMS	✓				✓	?		✓	✓
Cizek (2008)	GTE	✓	✓		✓	✓	?	✓	✓	✓
New		✓	✓	[g]	✓	✓	n		✓	✓

TABLE 1. Comparison of robust estimators of the coefficients in a linear regression model.

[a] With preprocessing; see Portnoy and Koenker (1997).

[b] Asymptotics are due to Hossjer (1994).

[c] If the constant is not varied, infinite if varied; see Davies (1993).

[d] See Croux, Rousseeuw, and Hossjer (1994).

[e] See Croux, Rousseeuw, and Hossjer (1994).

[f] If the constant is not varied.

[g] Can be modified to have a finite LSS.

The LSS is finite if the partial derivative of the influence function with respect to regressor and regressand–values is bounded.[3] We know of only one estimator, namely Mallows (1983), which is known to have a finite LSS. The proposed estimator does not have a finite LSS if the tails of the error distribution are thin. We do, however, describe a modification of our estimator which can achieve a finite LSS.

Virtually all existing robust estimators, and ours, are $\sqrt{n}$–consistent and asymptotically normal. The two exceptions are the two other LMS–based estimators, Rousseeuw (1984); Zinde-Walsh (2002). The original LMS estimator has been shown to be $\sqrt[3]{n}$–consistent and has a complicated limit distribution (see Kim and Pollard, 1990). Zinde-Walsh (2002) smoothes out the LMS objective function to obtain a better convergence rate and a limiting normal distribution, but her estimator does not achieve the desired $\sqrt{n}$–rate and its GES is infinite.

Like Zinde-Walsh (2002), but unlike most of the other estimators mentioned here, we do allow for dependence between errors and regressors. There are many examples in economics in which e.g. heteroskedasticity is important. Unlike Zinde-Walsh (2002), however, we do not allow for time series dependence, but follow the rest of the literature and assume independent and identically distributed (i.i.d.) data.

Almost all existing estimators, and ours, are both *affine invariant* and *regression invariant*. About half are also *scale invariant*, meaning that if the regressand is scaled, the vector of regression coefficient estimates is scaled by the same amount. Our estimator is not scale invariant, and scaling does have a material impact on its performance. Issues pertaining to scaling are discussed in detail in section 6.

Finally, there is great variation in the computational complexity of estimators, both in terms of computation time and the difficulty of writing a program. Ours is the only estimator for which the number of operations required for its computation is linear in n, albeit that the constant multiplying n can be large and increases with the number of regressors d. Because our estimator is the ratio of two integrals, it can be computed using any of a number of numerical integration techniques. For low–dimensional (small d) problems Gaussian quadrature works well. For many regressors, (quasi) Monte Carlo techniques can be used. For the numbers produced in this paper, we use Gibbs sampling (Geman and Geman, 1984). A simple Gibbs sampling procedure is described in appendix F; a C program using a faster algorithm is available from the authors upon request.

[3]The definition of the LSS is more general in that it allows for left– and right–derivatives to be different.

The remainder of our paper is organized as follows. In section 2 we define our estimator. Its breakdown point properties are established in section 3. Section 4 contains the asymptotic results absent contamination and section 5 a discussion of its asymptotic robustness properties (GES and LSS). Finally, section 6 addresses the effects of scaling of observables and section 7 the computation of the proposed estimator.

2. ESTIMATOR

For some $0 < q < 1$ to be chosen, let $N = \lfloor qn \rfloor + 1$, where $\lfloor \cdot \rfloor$ denotes the largest integer no greater than its argument. Let further $\mathscr{Q}^*(\boldsymbol{\xi}; q^*)$ denote the q^*–quantile of the distribution of $\boldsymbol{\xi}$, $\mathscr{Q}(\boldsymbol{\xi}) = \mathscr{Q}^*(\boldsymbol{\xi}; q)$, and let $\hat{\mathscr{Q}}(\boldsymbol{\xi}_i)$ be the N–th order statistic of $\boldsymbol{\xi}_1, \ldots \boldsymbol{\xi}_n$ for arbitrary $\boldsymbol{\xi}$'s. In case a quantile is not unique, in the sense that there are multiple values m that satisfy $\mathbb{P}(\boldsymbol{\xi} < m) \leq q \leq 1 - \mathbb{P}(\boldsymbol{\xi} > m)$, $\mathscr{Q}$ is taken to be any such value.[4]

Let $\{(\boldsymbol{x_i}, \boldsymbol{y_i})\}$ be an i.i.d. sample of size n where $\boldsymbol{x_i} \in \mathbb{R}^d$. The object of interest is the vector of regression coefficients in the linear regression model

$$\boldsymbol{y_i} = \boldsymbol{x_i}^\intercal \theta_0 + \boldsymbol{u_i}, \quad i = 1, \ldots, n.$$

Under conditions to be developed in section 4, θ_0 is unique and given by

$$\theta_0 = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathscr{Q}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal \theta|^2). \tag{3}$$

Our estimator of θ_0 is

$$\hat{\boldsymbol{\theta}} = \frac{\int \theta \exp\{-\hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal \theta|^2)\} d\theta}{\int \exp\{-\hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal \theta|^2)\} d\theta}. \tag{4}$$

The estimator $\hat{\boldsymbol{\theta}}$ resembles a Laplace estimator (Chernozhukov and Hong, 2003; Jun, Pinkse, and Wan, 2009), albeit that (as mentioned in the introduction) there is no sample–size–dependent input parameter scaling the objective function and no pseudo–prior. Indeed, in Chernozhukov and Hong (2003) and Jun, Pinkse, and Wan (2009) the objective function must be multiplied by a parameter which tends to infinity with the sample size to ensure consistency; in Chernozhukov and Hong (2003) the parameter is set to n, in Jun, Pinkse, and Wan (2009) it is chosen by the practitioner. This is not needed here because $m(t) = \mathscr{Q}\{|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal (\theta_0 + t)|\}$ happens to be symmetric in t.

[4]For instance, the median of a binary random variable $\boldsymbol{\xi}$ with $\mathbb{P}(\boldsymbol{\xi} = 1) = 1/2$ is any value between 0 and 1, inclusive.

By substitution of $t = \theta - \theta_0$ in (4) it follows that for $\hat{m}_n(t) = \hat{\mathscr{Q}}\{|y_i - x_i^\intercal(\theta_0 + t)|\}$,

$$\hat{\theta} - \theta_0 = \frac{\int t \exp\{-\hat{m}_n^2(t)\}dt}{\int \exp\{-\hat{m}_n^2(t)\}dt}. \tag{5}$$

The representation (5) will be frequently used in the remainder of the paper, especially in the proofs.

3. Breakdown Point

We now establish a general result concerning the breakdown properties of our estimator, which implies that the breakdown properties of our estimator are no worse than those of Rousseeuw (1984).

Let $\hat{\mathscr{D}}_n = \sup_{\|t\|=1} n^{-1} \sum_{i=1}^n I(|x_i^\intercal t| = 0)$, $\hat{\gamma} = n\hat{\mathscr{D}}_n$ and $\mathscr{D} = \sup_{\|t\|=1} \mathbb{P}(|x_i^\intercal t| = 0)$. The numbers $\hat{\mathscr{D}}_n, \hat{\gamma}$ represent the degree of noncollinearity in the sample and $\mathscr{D}$ that in the population. The best breakdown point obtains when observations are in *general position* (Rousseeuw, 1984), in which case $\hat{\gamma} = d - 1$. However, because $\mathscr{D} > 0$ if one (or more) of the regressors other than the constant is discrete, the general position property then occurs with probability approaching zero as $n \to \infty$. Our breakdown point result is hence for generic $\hat{\gamma}$.

Theorem 1. *If $\hat{\gamma} + 1 < N < n$ then the breakdown point of $\hat{\theta}$ satisfies $\hat{b} \geq \{\min(n - N, N - \hat{\gamma} - 1) + 1\}/n$.*

Although theorem 1 only provides a lower bound to the breakdown point, it is straightforward to construct counter examples to the breakdown point being better than the one stated in theorem 1.

Theorem 1 has several implications. First, for $q = 0.5$, the breakdown point when the observations are in general position is $(\lfloor n/2 \rfloor - d + 2)/n$ if $d > 1$,[5] which is exactly the same as in Rousseeuw (1984, theorem 1). The best breakdown point is achieved when q is chosen to make $N = \lfloor (n + \hat{\gamma} + 1)/2 \rfloor$, which results in a breakdown point of $\lfloor (n - \hat{\gamma} + 1)/2 \rfloor / n$. If the observations are in general position then the breakdown point equals $\lfloor (n - d)/2 \rfloor + 1$, which is the same as that in the remark following theorem 1 of Rousseeuw (1984) and hence also as that of Siegel (1982).

[5] It is $\lfloor (n + 1)/2 \rfloor / n$ if $d = 1$.

Asymptotically, the optimal choice of q in terms of breakdown properties is

$$q = \frac{1 + \mathscr{Y}}{2}, \tag{6}$$

resulting in a breakdown point converging to $(1 - \mathscr{Y})/2$ as $n \to \infty$, which is the best achievable for any regression equivariant estimator (Rousseeuw, 1984). The rationale for the choice of q in (6) is that $\hat{\mathscr{Y}}_n \xrightarrow{p} \mathscr{Y}$ and hence that $\hat{\gamma} \approx n\mathscr{Y}$, resulting in an optimal N of $\approx n(1 + \mathscr{Y})/2$.

4. Asymptotics

We now turn to a discussion of the properties of $\hat{\boldsymbol{\theta}}$ absent contamination. Throughout we assume that $\{(\boldsymbol{x_i}, \boldsymbol{y_i})\}$ is an i.i.d. sequence of random variables and that $0 < q < 1$.

We start by establishing identification. Let $m_0 = \mathscr{Q}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal \theta_0|) = \mathscr{Q}(|\boldsymbol{u_i}|)$.

Assumption A. *The conditional density $f(\cdot|\cdot)$ of $\boldsymbol{u_i}$ given $\boldsymbol{x_i} = x$ is for any x even, continuous, positive on the entire real line, weakly decreasing at all $u > 0$, and strictly decreasing at m_0.*

Assumption A is strong, but for $q = 0.5$ weaker than Kim and Pollard (1990, example 6.3) because we allow $\boldsymbol{u_i}$ and $\boldsymbol{x_i}$ to be dependent and do not assume the existence of derivatives for consistency.

Recall from section 3 that $\mathscr{Y} = \sup_{\|t\|=1} \mathbb{P}(|\boldsymbol{x_i}^\intercal t| = 0)$.

Assumption B. $\mathscr{Y} < 1$.

Assumption B requires that the regressors are perfectly collinear with probability less than one. It is implied by the requirement that $0 < \mathbb{E}(\boldsymbol{x_i}\boldsymbol{x_i}^\intercal) < \infty$ (Kim and Pollard, 1990, example 6.3), but does not assume the existence of moments for $\boldsymbol{x_i}$. Given that for $\mathscr{Y} > 0$ regressors are in general position with probability approaching zero (see section 3), assumption B is weak.

Theorem 2. *Under assumptions A and B, θ_0 defined in (3) is unique.*

We need one additional condition for consistency.

Assumption C. $\mathscr{Y} < q$.

Assumption C is the population equivalent (for $q = 0.5$) of the requirement in Rousseeuw (1984) that no *vertical hyperplane* (passing through the origin) contains more than $\lfloor n/2 \rfloor$ observations.

Assumption C can be restrictive. Indeed, with both a constant and a binary regressor it is violated when $q \leq 0.5$. But if $\mathscr{Y} > q$ then with probability approaching one $\hat{\mathscr{Y}}_n > q$, also, and the condition imposed on N in theorem 1 is violated. Consequently, none of the LMS–type estimators, Rousseeuw (1984); Zinde-Walsh (2002) and ours, will then have a breakdown point any better than the OLS estimator. So if assumption C is violated, it just means that q is chosen too small. In particular, if q is chosen according to (6) then assumption C is equivalent to assumption B.

Theorem 3. *Under assumptions A to C,* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \theta_0$.

We now proceed with a discussion of the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Let

$$m(t) = \mathscr{Q}(|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t|), \qquad m_\infty(t) = \mathscr{Q}(|\boldsymbol{x_i}^\intercal t|). \tag{7}$$

The notation m_∞ is inspired by the fact that for any $t \neq 0$, $\lim_{\lambda\to\infty}\{m(\lambda t)/\lambda\} = m_\infty(t)$ provided that $m_\infty(t)$ is unique.

Let $f(\cdot)$ and $F(\cdot)$ be the unconditional counterparts of $f(\cdot|\cdot)$ and $F(\cdot|\cdot)$, and let $\mathscr{X} = \{x : \exists \|t\| = 1, \epsilon > 0 : \big||x^\intercal t| - m_\infty(t)\big| < \epsilon\}$.

Assumption D. *(i)*

$$\lim_{\eta\downarrow 0} \frac{\inf_{\|t\|=1} \mathbb{P}\{|\boldsymbol{x_i}^\intercal t| \leq m_\infty(t) + \eta\} - q}{\eta} > 0, \quad \lim_{\eta\downarrow 0} \frac{\inf_{\|t\|=1} \mathbb{P}\{|\boldsymbol{x_i}^\intercal t| \geq m_\infty(t) - \eta\} - 1 + q}{\eta} > 0,$$

where each inequality is taken to hold if the limit is infinite. Moreover, (ii) for some $\epsilon > 0$, $0 < \inf_x f(\epsilon|x) \leq \sup_x f(0|x) < \infty$ *and for some* $2 < r < \infty$, *(iii)* $\lim_{s\to\infty}\{\inf_{x\in\mathscr{X}} f(s|x)/f^r(s)\} \geq 1$, *and (iv)* $\lim_{s\to\infty}[f\{s + F(-s)/f(s)\}/f^r(s)] > 0$.

Conditions (ii) and (iii) in assumption D are automatically satisfied if $\boldsymbol{u_i}$ and $\boldsymbol{x_i}$ are independent and can be seen as mild conditions restricting their dependence. We have verified condition (iv) for a number of distributions satisfying assumption A, including (symmetrized versions of) the Normal, Gumbel, Laplace, and Cauchy distributions.

Finally, condition (i) is satisfied when all regressors other than the constant are continuous. For discrete distributions, (i) is satisfied for most, but not all, choices of q. Condition (i) assumes away the possibility that the q–quantile of $|\boldsymbol{x_i}^\intercal t|$ is ambiguous for any vector t of length one. Condition (i) is unique to our paper.

Because q can be chosen to satisfy (i), condition (i) is more a nuisance than a serious obstacle for our estimator. Nevertheless, we highlight two alternatives that can be used to replace assumption D. The first solution is to assume that the tails of the conditional error density are sufficiently thick, i.e. declining more slowly than those of the density of an exponential distribution, which is not desirable. The second solution is to use a different estimator, a possibility which is discussed in section 6. We do not provide a formal justification for either solution.[6]

Finally, we need a condition on the derivative of f.

Assumption E. $\sup_{u,x} f'(u|x) < \infty$.

Assumption E is strong, but the assumption of the existence of the first derivative is also used in Kim and Pollard (1990); Hossjer (1994); Zinde-Walsh (2002), among others. Please note that assumption E is only used to establish asymptotic normality.

Let $A(t,m) = \mathbb{P}(|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| \leq m)$, $\mathscr{D}(t) = \partial_m A\{t, m(t)\}$,[7]

$$H(t,s) = \mathrm{Cov}[I\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| \leq m(t)\}, I\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal s| \leq m(s)\}], \tag{8}$$

and

$$\mathscr{V} = 4\frac{\iint ts^\intercal \frac{m(t)m(s)}{\mathscr{D}(t)\mathscr{D}(s)} H(t,s)\exp[-\{m^2(t)+m^2(s)\}]dtds}{\left[\int \exp\{-m^2(t)\}dt\right]^2}. \tag{9}$$

Theorem 4. *Let assumptions A to E hold. Then* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \theta_0) \xrightarrow{d} N(0, \mathscr{V})$.

So our estimator has good breakdown properties and is both $\sqrt{n}$–consistent and asymptotically normal. For the sake of completeness, we now provide a consistent estimator $\hat{\mathscr{V}}$ of $\mathscr{V}$. Let

$$\begin{aligned}\hat{\boldsymbol{H}}^*(\theta,\tilde{\theta}) = n^{-1}\sum_{i=1}^n I\{|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta| \leq \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta|)\} I\{|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\tilde{\theta}| \leq \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\tilde{\theta})\} \\ - \left(n^{-1}\sum_{i=1}^n I\{|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta| \leq \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta|)\}\right)\left(n^{-1}\sum_{i=1}^n I\{|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta| \leq \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\tilde{\theta}|)\}\right),\end{aligned} \tag{10}$$

and for some scalar $h^*(\theta)$,

$$\hat{\mathscr{D}}^*(\theta) = \frac{1}{2nh^*(\theta)}\sum_{i=1}^n I\{\big||\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta| - \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^\intercal\theta|)\big| \leq h^*(\theta)\}. \tag{11}$$

[6] Both alternatives ensure that $t\exp\{-m^2(t)\}/\mathscr{D}(t)$ in (9) is integrable, which is needed.

[7] ∂_m denotes the partial derivative with respect to m.

Then our estimator of the asymptotic variance is

$$\hat{\mathscr{V}} = 4\frac{\int\int(\theta-\hat{\boldsymbol{\theta}})(\tilde{\theta}-\hat{\boldsymbol{\theta}})^{\intercal}\frac{\hat{\mathscr{Q}}(|\boldsymbol{y_i}-\boldsymbol{x_i}^{\intercal}\theta|)\,\hat{\mathscr{Q}}(|\boldsymbol{y_i}-\boldsymbol{x_i}^{\intercal}\tilde{\theta}|)}{\hat{\mathscr{D}}^*(\theta)\hat{\mathscr{D}}^*(\tilde{\theta})}\hat{\boldsymbol{H}}^*(\theta,\tilde{\theta})e^{-\{\hat{\mathscr{Q}}(|\boldsymbol{y_i}-\boldsymbol{x_i}^{\intercal}\theta|^2)+\hat{\mathscr{Q}}(|\boldsymbol{y_i}-\boldsymbol{x_i}^{\intercal}\tilde{\theta}|^2)\}}d\theta d\tilde{\theta}}{\left[\int\exp\{-\hat{\mathscr{Q}}(|\boldsymbol{y_i}-\boldsymbol{x_i}^{\intercal}\theta|^2)\}d\theta\right]^2}. \tag{12}$$

The use of a uniform *kernel* in (11) is not essential but simplifies the proofs.

We need a single additional assumption, relating to the choice of *bandwidth* h^*.

Assumption F. *The bandwidth function h^* satisfies $h^*(\theta) = (1+\|\theta\|^{p^*})h_0$ for some $2 < p^* < \infty$ and $h_0 \prec 1 \prec n^{(1-1/p^*)/\sigma}h_0$ for some $\sigma > 4$.*

Because h_0 and p^* are chosen by the practitioner, assumption F is not restrictive. Assumption F permits the bandwidth to converge at the 'optimal' $n^{-1/5}$ rate if $p^* > 5$. We are now in a position to state the final theorem of this section.

Theorem 5. *Let assumptions A to F hold. Then $\hat{\mathscr{V}} \xrightarrow{p} \mathscr{V}$.*

5. Influence Function

From the proof of theorem 4[8] it is apparent that the dominant asymptotic term is

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\frac{\int t\frac{m(t)}{\mathscr{D}(t)}\left[I\{|\boldsymbol{u_i}-\boldsymbol{x_i}^{\intercal}t|\le m(t)\}-q\right]\exp\{-m^2(t)\}dt}{\int\exp\{-m^2(t)\}dt},$$

resulting in the influence function (Hampel, 1974)[9]

$$\mathscr{I}(y,x) = 2\frac{\int t\frac{m(t)}{\mathscr{D}(t)}\left[I\{|y-x^{\intercal}(\theta_0+t)|\le m(t)\}-q\right]\exp\{-m^2(t)\}dt}{\int\exp\{-m^2(t)\}dt}. \tag{13}$$

Since y, x enter (13) only through an indicator function, $\mathscr{I}$ is uniformly bounded and the GES[10] of our estimator is hence finite.

The LSS is more complicated to determine. We now show that even in the constant–only case, our estimator does not have a finite LSS when the tails of the error distribution are thin.

Theorem 6. *Suppose that $\boldsymbol{x_i}$ consists of only a constant and that F is the distribution function of a mean zero normal random variable with variance ς^2. Then if $q = 0.5$, $\varsigma^2 > 2 \iff \sup_y|\partial_y\mathscr{I}(y)| < \infty$.*

[8] See (48).

[9] In Hampel (1974) the influence function is defined as a functional derivative, which generally equals an element in the sum in the first order asymptotic term (Reeds, 1976; Boos and Serfling, 1980; Fernholz, 1983). We do not establish such equivalence here.

[10] The GES is the supremum over x, y of the norm of $\mathscr{I}$.

As the proof of theorem 6 illustrates, the LSS is infinite for thin–tailed error distributions because $\exp(-m^2)$ does then not decrease fast enough as m increases. This problem can be remedied by replacing exp in (4) by another smooth function which equals zero whenever its argument is sufficiently large negative. We do not investigate such a modification in this paper because of reasons outlined at the end of section 6.

6. SCALING

Like Laplace estimators (Chernozhukov and Hong, 2003), the proposed estimator is not invariant to scaling, or indeed monotonic transformations, of the objective function. In our case scaling the objective function is equivalent to scaling the data, so consider the estimator $\hat{\boldsymbol{\theta}}_{\alpha}$ below in which the scaling is made explicit by means of a scalar $0 < \alpha < \infty$.

$$\hat{\boldsymbol{\theta}}_{\alpha} = \frac{\int \theta \exp\{-\alpha\, \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^{\mathsf{T}}\theta|^2)\}d\theta}{\int \exp\{-\alpha\, \hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}^{\mathsf{T}}\theta|^2)\}d\theta}. \tag{14}$$

Having α be finite and nonzero is important for our results. It is apparent that $\lim_{\alpha\to\infty} \hat{\boldsymbol{\theta}}_{\alpha}$ (for $q = 0.5$) yields Rousseeuw's LMS estimator, which is not $\sqrt{n}$–consistent and lacks a bounded influence function unless the supremum is only taken over the slope regressors (Davies, 1993).

To obtain the limit of $\hat{\boldsymbol{\theta}}_{\alpha}$ as $\alpha \to 0$ is somewhat more complicated. We limit ourselves to the case with scalar–valued nonnegative $\boldsymbol{x_i}$ and $q = 0.5$, which is nonetheless instructive.[11] Let μ denote the index corresponding to the median observation if the data are ordered by $\boldsymbol{x_i}$–value and in the case of ties by $\boldsymbol{y_i}$–value.[12] So if the data are arranged such that $\boldsymbol{x_i} < \boldsymbol{x_j} \Rightarrow i < j$ and $\boldsymbol{x_i} \leq \boldsymbol{x_j},\ \boldsymbol{y_i} < \boldsymbol{y_j} \Rightarrow i < j$ and n is odd then $\mu = (n+1)/2$.

Theorem 7. *Suppose that n is odd, $d = 1$, $q = 0.5$, and that there are no ties in the $\boldsymbol{y_i}$–values. If all $\boldsymbol{x_i}$'s are nonnegative and $\boldsymbol{x_\mu} > 0$, then*

$$\lim_{\alpha\to 0} \hat{\boldsymbol{\theta}}_{\alpha} = \frac{\boldsymbol{y_\mu}}{\boldsymbol{x_\mu}}. \tag{15}$$

Theorem 7 has two interesting implications. First, if $\boldsymbol{x_i}$ is a constant then $\boldsymbol{y_\mu}/\boldsymbol{x_\mu}$ equals the sample median, which has excellent properties. In most other cases, however, $\boldsymbol{y_\mu}/\boldsymbol{x_\mu}$ is an inconsistent estimator of θ_0. Indeed, if $\boldsymbol{x_i}$ is continuously distributed then $\boldsymbol{y_\mu}/\boldsymbol{x_\mu}$ is the ratio of the y and x values of the observation corresponding to the sample median of the $\boldsymbol{x_i}$'s.

[11] Nonnegativity is innocuous since $\boldsymbol{x_i}, \boldsymbol{y_i}$ can be replaced with $-\boldsymbol{x_i}, -\boldsymbol{y_i}$ if $\boldsymbol{x_i}$ is negative.

[12] We ignore the possibility of ties in the $\boldsymbol{y_i}$'s given that they are assumed continuous throughout the paper.

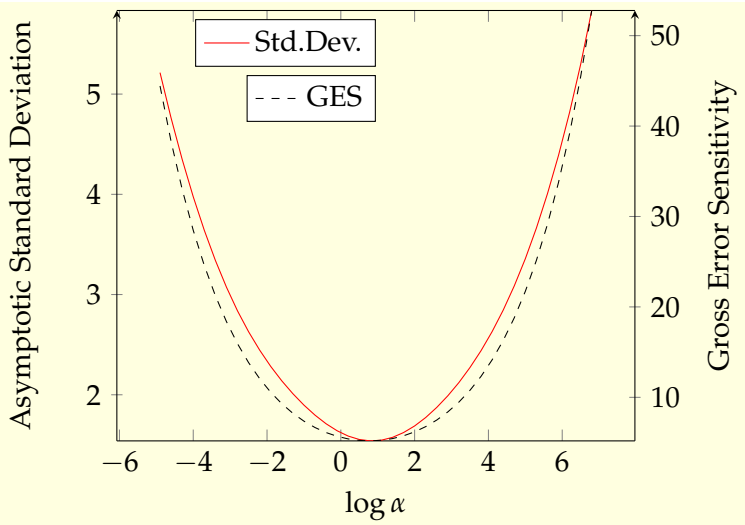

FIGURE 1. Asymptotic Variance and Gross Error Sensitivity

So the value of α that minimizes the asymptotic variance is generally different from zero and infinity and the same is true for the value that minimizes the gross error sensitivity. Other estimators in this literature, including Krasker (1980); Mallows (1983), also require the choice of an input parameter but in those papers the input parameter represents a choice between efficiency (as measured by the trace of the asymptotic variance matrix) and robustness (as measured by the gross error sensitivity).[13]

Figure 1 demonstrates that in our case there need not be a tradeoff between efficiency and robustness. The design used to generate the graph is one in which $\boldsymbol{u_i}, \boldsymbol{x_i}$ are independent standard normal random variables and there is no constant term in the model. We chose this design because it results in an explicit expression for $m(t)$.

The question, then, is how one should choose α. One possibility is to use a first stage estimator (be it ours with a fixed α or some other robust method), estimate the asymptotic variance $\mathscr{V}_\alpha$ (or indeed the gross error sensitivity),

$$\mathscr{V}_\alpha = 4\alpha^2 \frac{\int\int ts^\intercal \frac{m(t)m(s)}{\mathscr{D}(t)\mathscr{D}(s)} H(t,s) \exp[-\alpha\{m^2(t)+m^2(s)\}]dtds}{\left[\int \exp\{-\alpha m^2(t)\}dt\right]^2}.$$

and choose the value of α that minimizes one's estimate of (the trace of) $\mathscr{V}_\alpha$. We do not provide results for such a data–dependent choice of α.

[13] See Krasker (1980).

Finally, we mentioned in section 4 that after replacing $\hat{\mathscr{Q}}(|y_i - x_i^\top\theta|^2)$ in (4) with a monotonic transformation thereof assumption D can be weakened. The main reason is that if one makes the exponents in (4) tend to $-\infty$ faster as $\|\theta\| \to \infty$ then anything multiplying the exponents can increase faster as $\|\theta\| \to \infty$, also. We have chosen not to take this route in this paper because $\hat{\boldsymbol{\theta}}$ defined in (4) is closer to the original LMS estimator, it is convenient, it appears to work well, and assumption D seems weaker than some other assumptions made in this paper.

7. Computation

There are several ways of computing our estimator; all involve numerical integration. Especially for low–dimensional (d small) problems, Gaussian quadrature works well. For high–dimensional problems, a Monte Carlo–based approach usually works better.

Because the computation of a sample median requires $O(n)$ operations (Knuth, 1997, chapter 6), the computation of each of the integrands in (4) requires $O(n)$ operations, also. Hence if $\hat{\boldsymbol{\theta}}$ is computed using the (classical) Monte Carlo method, with or without importance sampling (Robert and Casella, 2004, Definition 3.9), or indeed using quadrature, then the total number of operations needed is linear in n.

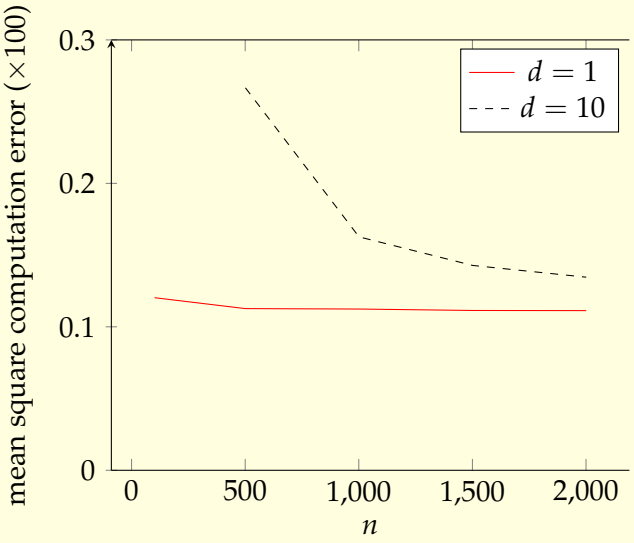

Figure 2. Computational Accuracy as a Function of n

To illustrate, consider figure 2, for which we used the Monte Carlo method with importance sampling using a normal distribution with variance chosen to match the tails of $\exp\{-\hat{\mathscr{Q}}(|y_i -$

$x_i^\top\theta|^2)\}$ as an instrumental distribution. For each (n,d)–combination, we constructed 1,000 samples $s = 1,\ldots,1000$, computed $\hat{\boldsymbol{\theta}}_{s(\infty)} = \hat{\boldsymbol{\theta}}$ using 1,000,000 draws. We then computed $\hat{\boldsymbol{\theta}}_{sr}$ 1000 times ($r = 1,\ldots,1000$) using 1000 draws in each case. Finally, we use the average mean square deviation for each (n,d)–combination, $\sum_{s=1}^{1000}\sum_{r=1}^{1000}(\hat{\boldsymbol{\theta}}_{sr} - \hat{\boldsymbol{\theta}}_{s(\infty)})^2/1000^2$, as a measure of the computational accuracy of using 1000 draws.

If the number of operations needed to achieve the same level of accuracy were to increase with n, both curves in figure 2 would be increasing. The reason that they are initially decreasing is due to our choice of an instrumental distribution, which is a better match for the integrand for large n than it is for small n.

Although the results depicted in figure 2 are encouraging, some words of caution are in order. First, it is conceivable that performance is different for designs different from the one chosen here. Second, although computation is linear in n, it could be slow for any n if a large number of random draws is needed to achieve a desired level of accuracy, which arises when the instrumental distribution used is a bad match for the integrand. Indeed, the best choice of it depends on the shape of $\exp\{-\hat{\mathscr{Q}}(|y_i - x_i^\top\theta|^2)\}$, in particular, on the unknown parameter vector θ_0. Likewise, the number of draws needed to achieve the same level of accuracy need not go up linearly in d.

For these reasons, it can be preferable to use other numerical integration methods such as Gibbs sampling (Geman and Geman, 1984). A simple scheme, which requires $O(n^2)$ operations for a draw, is described in appendix F. A faster algorithm is available from the authors.

References Cited

Boos, D., and R. Serfling (1980): "A note on differentials and the CLT and LIL for statistical functions, with application to M-estimates," *The Annals of Statistics*, 8(3), 618–624.

Chang, W., J. McKean, J. Naranjo, and S. Sheather (1999): "High-Breakdown Rank Regression.," *Journal of the American Statistical Association*, 94(445), 205–219.

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

Cizek, P. (2008): "General trimmed estimation: robust approach to nonlinear and limited dependent variable models," *Econometric Theory*, 24(06), 1500–1529.

Croux, C., P. Rousseeuw, and O. Hossjer (1994): "Generalized S-Estimators.," *Journal of the American Statistical Association*, 89(428).

DAVID, H. (1986): "Inequalities for ordered sums," *Annals of the Institute of Statistical Mathematics*, 38, 551–555.

DAVIES, P. (1993): "Aspects of robust linear regression," *The Annals of Statistics*, 21(4), 1843–1899.

DONOHO, D., AND P. HUBER (1982): "The notion of breakdown point," *A Festschrift for Erich L. Lehmann in Honor of His Sixty-fifth Birthday*, pp. 157–184.

FERNHOLZ, L. (1983): *Von Mises calculus for statistical functionals*, vol. 19 of *Lecture Notes in Statistics*. Springer.

GEMAN, S., AND D. GEMAN (1984): "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

GLICK, N. (1974): "Consistency conditions for probability estimators and integrals of density estimators," *Utilitas Mathematica*, 6, 61–74.

HADI, A., AND A. LUCENO (1997): "Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms," *Computational statistics & data analysis*, 25(3), 251–272.

HAMPEL, F. (1968): "Contributions to the theory of robust estimation," Ph.D. thesis, University of California at Berkeley.

——— (1971): "A general qualitative definition of robustness," *The Annals of Mathematical Statistics*, 42(6), 1887–1896.

——— (1974): "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, 69(346), 383–393.

HAMPEL, F., E. RONCHETTI, P. ROUSSEEUW, AND W. STAHEL (1986): *Robust Statistics*. Wiley New York.

HOSSJER, O. (1994): "Rank–based estimates in the linear model with high breakdown point," *Journal of the American Statistical Association*, 89(425), 149–158.

HUBER, P. (1973): "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, 1(5), 799–821.

JUN, S. J., J. PINKSE, AND Y. WAN (2009): "Cube root n and faster convergence, Laplace estimators, and uniform inference," Discussion paper, The Pennsylvania State University.

KIM, J., AND D. POLLARD (1990): "Cube root asymptotics," *Annals of Statistics*, 18(1), 191–219.

KNUTH, D. E. (1997): *The art of computer programming*, vol. 1. Addison–Wesley, 3 edn.

KOENKER, R., AND G. BASSETT (1978): "Regression quantiles," *Econometrica*, 46(1), 33–50.

KRASKER, W. (1980): "Estimation in linear regression models with disparate data points," *Econometrica*, 48(6), 1333–1346.

MALLOWS, C. L. (1983): "Minimax Aspects of Bounded-Influence Regression: Comment," *Journal of the American Statistical Association*, 78(381), 77.

PORTNOY, S., AND R. KOENKER (1997): "The Gaussian hare and the Laplacian tortoise: computability of squred–error versus absolute–error estimators, with discussion," *Statistical Science*, 12(4), 279–300.

REEDS, J. (1976): "On the definition of von Mises functionals," Ph.D. thesis, Harvard University.

ROBERT, C., AND G. CASELLA (2004): *Monte Carlo statistical methods*. Springer.

ROUSSEEUW, P. (1984): "Least median of squares regression," *Journal of the American Statistical Association*, 79(388), 871–880.

ROUSSEEUW, P., AND A. LEROY (1987): *Robust regression and outlier detection*. Wiley.

ROUSSEEUW, P., AND V. YOHAI (1984): "Robust regression by means of S-estimators," *Robust and nonlinear time series analysis*, 26, 256–272.

SAKATA, S., AND H. WHITE (1995): "An alternative definition of finite–sample breakdown point with applications to regression model estimators.," *Journal of the American Statistical Association*, 90(431).

——— (1998): "High breakdown point conditional dispersion estimation with application to S & P 500 daily returns volatility," *Econometrica*, 66(3), 529–567.

SIEGEL, A. (1982): "Robust regression using repeated medians," *Biometrika*, 69(1), 242.

YOHAI, V. (1987): "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, 15(2), 642–656.

YOHAI, V., AND R. ZAMAR (1988): "High breakdown-point estimates of regression by means of the minimization of an efficient scale," *Journal of the American Statistical Association*, 83(402), 406–413.

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

APPENDIX A. BASICS

Lemma A1. $\forall t : m(-t) = m(t)$.

Proof. Note that

$$q = \mathbb{P}\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| \leq m(t)\} = \mathbb{E}[F\{m(t) + \boldsymbol{x_i}^\intercal t|\boldsymbol{x_i}\} - F\{-m(t) + \boldsymbol{x_i}^\intercal t|\boldsymbol{x_i}\}]$$
$$= \mathbb{E}[1 - F\{-m(t) - \boldsymbol{x_i}^\intercal t|\boldsymbol{x_i}\} - 1 + F\{m(t) - \boldsymbol{x_i}^\intercal t|\boldsymbol{x_i}\}]$$
$$= \mathbb{E}[F\{m(t) + \boldsymbol{x_i}^\intercal(-t)|\boldsymbol{x_i}\} - F\{-m(t) + \boldsymbol{x_i}^\intercal(-t)|\boldsymbol{x_i}\}], \quad (16)$$

where the penultimate equality follows from the symmetry of F. Hence $m(t) = m(-t)$. □

Appendix B. Consistency

The results in appendix B presume assumptions A and B to hold. Let for arbitrary scalar λ and $t \in \mathbb{R}^d$ and any $2 \leq p < \infty$, $\boldsymbol{v_i}(\lambda, t) = |\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t|/(1 + \lambda^p)$, $\boldsymbol{a_i} = [\boldsymbol{u_i}|\boldsymbol{x_i}^\intercal]^\intercal$, $\mathscr{Q}_a = \mathscr{Q}(\|\boldsymbol{a_i}\|)$, $\hat{\mathscr{Q}}_{\boldsymbol{a}} = \hat{\mathscr{Q}}(\|\boldsymbol{a_i}\|)$, $\hat{\boldsymbol{m}}_v(\lambda, t) = \hat{\mathscr{Q}}\{\boldsymbol{v_i}\}(\lambda, t)$, $m_v(\lambda, t) = \mathscr{Q}\{\boldsymbol{v_i}(\lambda, t)\}$. We moreover use m, m_∞ as defined in (7) and $\mathscr{L}^\infty(\mathbb{R}^d)$ as the collection of bounded functions on $\mathbb{R}^d$ equipped with a sup–norm.

Lemma B1. *For all $\epsilon_1 > 0$ there exists a $C_1 < \infty$ such that (i) $\sup_{\lambda > C_1} \sup_{\|t\|=1} m_v(\lambda, t) \leq \epsilon_1$ and (ii)*

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \sup_{\lambda > C_1} \sup_{\|t\|=1} \hat{\boldsymbol{m}}_v(\lambda, t) > \epsilon_1 \Big\} = 0.$$

Proof. We show (ii), where we establish along the way that (i) holds, also. Take $C_1 = \max(2\mathscr{Q}_a/\epsilon_1, 1)$. Then $\sup_{\lambda > C_1} \sup_{\|t\|=1} \hat{\boldsymbol{m}}_v(\lambda, t) \leq \hat{\mathscr{Q}}_{\boldsymbol{a}} C_1/(1 + C_1^p)$ and

$$\mathbb{P}\{\hat{\mathscr{Q}}_{\boldsymbol{a}} C_1 > (1 + C_1^p)\epsilon_1\} \leq \underbrace{\mathbb{P}\{|\hat{\mathscr{Q}}_{\boldsymbol{a}} - \mathscr{Q}_a| C_1 > (1 + C_1^p)\epsilon_1/2\}}_{\prec 1} + \underbrace{I\{\mathscr{Q}_a C_1 > (1 + C_1^p)\epsilon_1/2\}}_{=0}. \quad \square$$

Lemma B2. *For any $C_1 < \infty$ there exists a $C_2 < \infty$ such that*

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \sup_{0 \leq \lambda \leq C_1} \sup_{\|t\|=1} \hat{\boldsymbol{m}}_v(\lambda, t) > C_2 \Big\} = 0.$$

Proof. Take $C_2 = 4\mathscr{Q}_a > 0$. Noting that $\max(1, \lambda)/(1 + \lambda^p) \leq 1$,

$$\mathbb{P}\Big\{ \sup_{0 \leq \lambda \leq C_1} \sup_{\|t\|=1} \hat{\boldsymbol{m}}_v(\lambda, t) > C_2 \Big\} \leq \mathbb{P}(\hat{\mathscr{Q}}_{\boldsymbol{a}} > C_2) \leq \mathbb{P}(|\hat{\mathscr{Q}}_{\boldsymbol{a}} - \mathscr{Q}_a| > C_2/2) \prec 1. \quad \square$$

Let $\hat{\boldsymbol{A}}_v(\lambda, t, m) = n^{-1} \sum_{i=1}^n I\{\boldsymbol{v_i}(\lambda, t) \leq m\}$ and $A_v(\lambda, t, m) = \mathbb{P}\{\boldsymbol{v_i}(\lambda, t) \leq m\}$.

Lemma B3. $\sqrt{n}(\hat{\boldsymbol{A}}_v - A_v) \xrightarrow{w} \mathbb{G}_v$ *in $\mathscr{L}^\infty(\mathbb{R}^{d+2})$ for a zero mean Gaussian process $\mathbb{G}_v$ with covariance kernel $H_v(\lambda, t, m, \tilde{\lambda}, \tilde{t}, \tilde{m}) = \mathrm{Cov}[I\{|\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t| \leq (1 + \lambda^p)m\}, I\{|\boldsymbol{u_i} - \tilde{\lambda} \boldsymbol{x_i}^\intercal \tilde{t}| \leq (1 + \lambda^p)\tilde{m}\}]$.*

Proof. Let $\mathscr{C}$ be the collection of sets of $(u, x^\intercal)^\intercal$ indexed by $(a, b^\intercal, m)^\intercal \in \mathbb{R}^{d+2}$ such that $|au + x^\intercal b| \leq m$. Since the collection of half spaces is a Vapnik–Chervonenkis (VC) class and $\mathscr{C}$ is the collection of intersections of half spaces, $\mathscr{C}$ is a VC class. Therefore, $\mathscr{F} = \{I\{C\} : C \in \mathscr{C}\}$ is a VC subgraph class of functions that are indexed by $(a, b^\intercal, m)^\intercal \in \mathbb{R}^{d+2}$. Since $\mathbb{R}^{d+2}$ is separable, $\mathscr{F}$ is a pointwise measurable class. Therefore, $\mathscr{F}$ is a Donsker class in $\mathscr{L}^\infty(\mathbb{R}^{d+2})$. Reparametrizing by $a = 1/(1+\lambda^p)$ and $b = \lambda t/(1+\lambda^p)$ does not affect the Donsker property, and therefore the weak convergence of $\sqrt{n}(\hat{\boldsymbol{A}}_v - A_v)$ in $\mathscr{L}^\infty(\mathbb{R}^{d+2})$ follows. Apply a central limit theorem to arbitrary finite marginals and the Gaussian limit process and covariance kernel follow. □

Let $c_{pt} = 1 + \|t\|^p$ and $\mathbb{G}_{\boldsymbol{p}}$ a Gaussian process with covariance kernel $H_p^*(t,s) = H(t,s)/c_{pt}c_{ps}$, where H is as defined in (8).

Lemma B4. *Let* $\mathscr{S}_{\boldsymbol{n1}}(t,m) = \sqrt{n}\{\hat{\boldsymbol{A}}_{\boldsymbol{n}}(t,m) - A(t,m)\}$, $\mathscr{S}_{\boldsymbol{n2}}(t) = \sqrt{n}[\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, m(t)\} - A\{t, m(t)\}$, *and* $\mathscr{S}_{\boldsymbol{n3p}}(t) = \sqrt{n}[\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, m(t)\} - A\{t, m(t)\}]/c_{pt}$. *For* $\mathbb{G}_{\boldsymbol{p}}$ *as defined above and some other Gaussian processes* $\mathscr{G}_{\boldsymbol{1}}, \mathscr{G}_{\boldsymbol{2}}$, *(i)* $\mathscr{S}_{\boldsymbol{n1}} \xrightarrow{w} \mathscr{G}_{\boldsymbol{1}}$ *in* $\mathscr{L}^\infty(\mathbb{R}^{d+1})$, *(ii)* $\mathscr{S}_{\boldsymbol{n2}} \xrightarrow{w} \mathscr{G}_{\boldsymbol{2}}$ *in* $\mathscr{L}^\infty(\mathbb{R}^d)$, *(iii)* $\mathscr{S}_{\boldsymbol{n3p}} \xrightarrow{w} \mathbb{G}_{\boldsymbol{p}}$ *in* $\mathscr{L}^\infty(\mathbb{R}^d)$.

Proof. First (i). Since the collection of half spaces in a Euclidean space is a VC class, the indicator functions $\mathbb{F}^* = \{I(|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| \leq m) : (t,m) \in \mathbb{R}^{d+1}\}$ form a VC subgraph class, because $\mathbb{F}^*$ is generated by using a finite intersection of half spaces. Since $\mathbb{R}^{d+1}$ is separable, $\mathbb{F}^*$ is a pointwise measurable class and is hence Donsker.

Since the derivations for (ii) and (iii) are similar to each other, we only consider (iii). Since the Gaussianity of finite marginals follows from a central limit theorem, we focus on the stochastic equicontinuity of $\mathscr{S}_{\boldsymbol{n3p}}$. Note that

$$\begin{aligned}\big|\mathscr{S}_{\boldsymbol{n3p}}(t) - \mathscr{S}_{\boldsymbol{n3p}}(\tilde{t})\big| \leq \sup_m \big|\mathscr{S}_{\boldsymbol{n1}}(t,m) - \mathscr{S}_{\boldsymbol{n1}}(\tilde{t},m)\big| + \sup_{t,m}\big|\mathscr{S}_{\boldsymbol{n1}}(t,m)\big|\Big|\frac{1}{c_{pt}} - \frac{1}{c_{p\tilde{t}}}\Big| \\ + \sup_{t^*}\big|\mathscr{S}_{\boldsymbol{n1}}\{t^*, m(t)\} - \mathscr{S}_{\boldsymbol{n1}}\{t^*, m(\tilde{t})\}\big|,\end{aligned}$$

where the RHS converges in probability to 0 as $\|t - \tilde{t}\| \to 0$, because of (i) and since $1/c_{pt}$ and m are continuous in t. □

Lemma B5. $\sup_{\lambda,t} |A_v\{\lambda, t, \hat{\boldsymbol{m}}_v(\lambda,t)\} - A_v\{\lambda, t, m_v(\lambda,t)\}| \preceq 1/\sqrt{n}$.

Proof. By the triangle inequality and the definition of m_v,

$$\sup_{\lambda,t} |A_v\{\lambda, t, \hat{\boldsymbol{m}}_v(\lambda,t)\} - A_v\{\lambda, t, m_v(\lambda,t)\}| \leq$$
$$\sup_{\lambda,t} |A_v\{\lambda, t, \hat{\boldsymbol{m}}_v(\lambda,t)\} - \hat{\boldsymbol{A}}_v\{\lambda, t, \hat{\boldsymbol{m}}_v(\lambda,t)\}| + \sup_{\lambda,t} |\hat{\boldsymbol{A}}_v\{\lambda, t, \hat{\boldsymbol{m}}_v(\lambda,t)\} - q|. \quad (17)$$

The first right hand side term (RHS1) in (17) is $\preceq 1/\sqrt{n}$ by lemma B3 and RHS2 is $\prec 1/\sqrt{n}$ by the definition of $\hat{\boldsymbol{m}}_v$. □

Lemma B6. *For any* $C_1 < \infty$, $\sup_{0\leq\lambda\leq C_1} \sup_{\|t\|=1} |\hat{\boldsymbol{m}}_v(\lambda,t) - m_v(\lambda,t)| \preceq 1/\sqrt{n}$.

Proof. By lemma B5 and the mean value theorem, for some $\hat{\boldsymbol{m}}_v^*(\lambda,t)$ between $\hat{\boldsymbol{m}}_v(\lambda,t)$ and $m_v(\lambda,t)$,

$$\sup_{0\leq\lambda\leq C_1} \sup_{\|t\|=1} |\partial_m A_v\{\lambda, t, \hat{\boldsymbol{m}}_v^*(\lambda,t)\}\{\hat{\boldsymbol{m}}_v(\lambda,t) - m_v(\lambda,t)\}| \preceq 1/\sqrt{n}.$$

It hence suffices to show that for some $C_3 > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big[\inf_{0\leq\lambda\leq C_1} \inf_{\|t\|=1} \partial_m A_v\{\lambda, t, \hat{\boldsymbol{m}}_v^*(\lambda,t)\} > C_3\Big] = 1.$$

By lemma B2 it suffices to show that

$$\inf_{0\leq\lambda\leq C_1} \inf_{\|t\|=1} \inf_{0\leq m\leq C_2} |\partial_m A_v(\lambda, t, m)| > 0.$$

Because $A_v(\lambda,t,m) = \mathbb{E}[F\{\lambda \boldsymbol{x_i}^\intercal t + (1+\lambda^p)m|\boldsymbol{x_i}\} - F\{\lambda \boldsymbol{x_i}^\intercal t - (1+\lambda^p)m|\boldsymbol{x_i}\}]$, it follows that for sufficiently large but finite C_4,

$$\partial_m A_v(\lambda,t,m) = \mathbb{E}[f\{\lambda \boldsymbol{x_i}^\intercal t + (1+\lambda^p)m|\boldsymbol{x_i}\} + f\{\lambda \boldsymbol{x_i}^\intercal t - (1+\lambda^p)m|\boldsymbol{x_i}\}]$$
$$\geq 2\,\mathbb{E}[f\{C_1C_4 + (1+C_1^p)C_2|\boldsymbol{x_i}\} I(\|\boldsymbol{x_i}\| \leq C_4)] > 0, \quad (18)$$

by assumption A. □

Lemma B7. $\sup_{\lambda\geq 0,\|t\|=1} |\hat{\boldsymbol{m}}_v(\lambda,t) - m_v(\lambda,t)| \prec 1$.

Proof. We show that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big\{\sup_{\lambda\geq 0,\|t\|=1} |\hat{\boldsymbol{m}}_v(\lambda,t) - m_v(\lambda,t)| > \epsilon\Big\} = 0. \quad (19)$$

In lemma B1, take $\epsilon_1 = \epsilon/4$ to show that for the choice of C_1 given there,

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \sup_{\lambda > C_1, \|t\|=1} |\hat{\boldsymbol{m}}_v(\lambda, t) - m_v(\lambda, t)| > \epsilon/2 \Big\} = 0.$$

The case $0 \le \lambda \le C_1$ is dealt with in lemma B6. □

Lemma B8. $\inf_{t\in\mathbb{R}^d}\{m(t)/(1+\|t\|)\} > 0$.

Proof. We show equivalently that

$$\exists \epsilon, c > 0 : \sup_{\lambda \ge 0, \|t\|=1} \mathbb{P}\{|\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t| \le c(1+\lambda)\} \le q - \epsilon. \tag{20}$$

By assumption C it follows that for any sufficiently small $\epsilon > 0$,

$$\sup_{\|t\|=1} \mathbb{P}(|\boldsymbol{x_i}^\intercal t| \le \epsilon) \le q - 2\epsilon. \tag{21}$$

Now choose $K_1 = -F^{-1}(\epsilon/2), K_2 = \max(2K_1/\epsilon, 1), c = \min\big(\epsilon/4, (q-\epsilon)/[2(1+K_2)\,\mathbb{E}\{f(0|\boldsymbol{x_i})\}]\big)$. Then

$$\begin{aligned}
&\sup_{\lambda > K_2, \|t\|=1} \mathbb{P}\{|\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t| \le c(1+\lambda)\} \le \\
&\qquad \mathbb{P}(|\boldsymbol{u_i}| > K_1) + \sup_{\lambda > K_2, \|t\|=1} \mathbb{P}\{|\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t| \le c(1+\lambda),\ |\boldsymbol{u_i}| \le K_1\} \\
&\le \epsilon + \sup_{\lambda > K_2, \|t\|=1} \mathbb{P}\{\lambda |\boldsymbol{x_i}^\intercal t| \le c(1+\lambda) + K_1\} \le \epsilon + \sup_{\|t\|=1} \mathbb{P}(|\boldsymbol{x_i}^\intercal t| \le 2c + K_1/K_2) \\
&\qquad \le \epsilon + \sup_{\|t\|=1} \mathbb{P}(|\boldsymbol{x_i}^\intercal t| \le \epsilon) \le q - \epsilon,
\end{aligned} \tag{22}$$

by (21). Further,

$$\begin{aligned}
&\sup_{\lambda \le K_2, \|t\|=1} \mathbb{P}\{|\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t| \le c(1+\lambda)\} \\
&\qquad \le \sup_{\lambda \le K_2, \|t\|=1} \mathbb{E}[F\{\lambda \boldsymbol{x_i}^\intercal t + c(1+\lambda)|\boldsymbol{x_i}\} - F\{\lambda \boldsymbol{x_i}^\intercal t - c(1+\lambda)|\boldsymbol{x_i}\}] \\
&\qquad \le 2c(1+K_2)\,\mathbb{E}\{f(0|\boldsymbol{x_i})\} \le q - \epsilon,
\end{aligned} \tag{23}$$

by assumption A and the choice of c. Combining (22) and (23) yields (20). □

Lemma B9. $\inf_{t\in\mathbb{R}^d}\{\hat{\boldsymbol{m}}_n(t)/(1+\|t\|)\} \succeq 1$.

Proof. We show the equivalent result that for any sufficiently small $\epsilon, c > 0$,

$$\mathbb{P}\Big[\sup_{\lambda\geq 0,\|t\|=1} \hat{\boldsymbol{A}}_{\boldsymbol{n}}\{\lambda t, c(1+\lambda)\} > q - 2\epsilon\Big] \prec 1.$$

Now,

$$\mathbb{P}\Big[\sup_{\lambda\geq 0,\|t\|=1} \hat{\boldsymbol{A}}_{\boldsymbol{n}}\{\lambda t, c(1+\lambda)\} > q - 2\epsilon\Big] \leq$$
$$I[A\{\lambda t, c(1+\lambda)\} \geq q\epsilon] + \mathbb{P}\Big[\sup_{\lambda\geq 0,\|t\|=1} \big|\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{\lambda t, c(1+\lambda)\} - A\{\lambda t, c(1+\lambda)\}\big| > \epsilon\Big]. \quad (24)$$

RHS2 in (24) is $\prec 1$ by lemma B4 and RHS1 is exactly (20). □

Lemma B10. $\sup_{t\in\mathbb{R}^d}\{m(t)/(1+\|t\|)\} < \infty$.

Proof. The result follows immediately from the fact that $\mathscr{Q}(|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t|) \leq (1+\|t\|)\mathscr{Q}_a$, with $\mathscr{Q}_a$ defined at the beginning of appendix B. □

Lemma B11. $\sup_{t\in\mathbb{R}^d}\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)/(1+\|t\|)\} \preceq 1$.

Proof. The result follows immediately from the fact that $\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) \leq (1+\|t\|)\hat{\mathscr{Q}}_{\boldsymbol{a}}$, with $\hat{\mathscr{Q}}_{\boldsymbol{a}}$ defined at the beginning of appendix B. □

Appendix C. Asymptotic Normality

The assumptions of theorem 4 are taken to hold for the lemmas below. The suprema and infima in this section are taken over $\lambda \in [0,\infty)$ and $\|t\| = 1$, unless otherwise noted. Let $f^{-1} : [0, f(0)] \to [0,\infty)$ be a function (not necessarily unique) such that $f\{f^{-1}(t)\} = t$ for all t.

Lemma C1. *For some $C > 0$ and all sufficiently large λ,*

$$\sup_t |m(\lambda t) - \lambda m_\infty(t)| \leq f^{-1}\Big(\frac{1}{C\lambda}\Big) + C\lambda F\Big\{-f^{-1}\Big(\frac{1}{C\lambda}\Big)\Big\}.$$

Proof. For C to be chosen, set $\epsilon = 2F\{-f^{-1}(1/C\lambda)\}$. By David (1986, theorem 1) and assumption D for some finite C independent of λ,

$$\sup_t\{m(\lambda t) - \lambda m_\infty(t)\} \leq \mathscr{Q}^*(|\boldsymbol{u_i}|; 1-\epsilon) + \lambda \sup_t\{\mathscr{Q}^*(|\boldsymbol{x_i}^\intercal t|; q+\epsilon) - m_\infty(t)\}$$
$$\leq \mathscr{Q}^*(|\boldsymbol{u_i}|; 1-\epsilon) + C\lambda\epsilon/2 = f^{-1}(1/C\lambda) + C\lambda F\{-f^{-1}(1/C\lambda)\}.$$

The case $\sup_t\{\lambda m_\infty(t) - m(\lambda t)\}$ is similar. □

Lemma C2. $\inf_{\lambda,t} \partial_m A_v\{\lambda, t, m_v(\lambda,t)\} > 0$.

Proof. For all λ belonging to a compact set, the result was established in (18). We now show that the result also holds for large λ. Noting that by the symmetry of the conditional distribution of $\boldsymbol{u_i}$ given $\boldsymbol{x_i}$,

$$A_v(\lambda,t,m) = \mathbb{P}\{|\boldsymbol{u_i} - \lambda \boldsymbol{x_i}^\intercal t| \le (1+\lambda^p)m\} = \mathbb{E}[F\{\lambda \boldsymbol{x_i}^\intercal t + (1+\lambda^p)m|\boldsymbol{x_i}\} - F\{\lambda \boldsymbol{x_i}^\intercal t - (1+\lambda^p)m|\boldsymbol{x_i}\}]$$
$$= \mathbb{E}[F\{\lambda|\boldsymbol{x_i}^\intercal t| + (1+\lambda^p)m|\boldsymbol{x_i}\} - F\{\lambda|\boldsymbol{x_i}^\intercal t| - (1+\lambda^p)m|\boldsymbol{x_i}\}].$$

Hence

$$\partial_m A_v\{\lambda,t,m_v(\lambda,t)\} = (1+\lambda^p)\,\mathbb{E}[f\{\lambda|\boldsymbol{x_i}^\intercal t| + m(\lambda t)|\boldsymbol{x_i}\} + f\{\lambda|\boldsymbol{x_i}^\intercal t| - m(\lambda t)|\boldsymbol{x_i}\}]$$
$$\ge (1+\lambda^p)\,\mathbb{E}[f\{\lambda|\boldsymbol{x_i}^\intercal t| - m(\lambda t)|\boldsymbol{x_i}\}]$$
$$= (1+\lambda^p)\,\mathbb{E}\big(f[\lambda\{|\boldsymbol{x_i}^\intercal t| - m_\infty(t)\} + \{\lambda m_\infty(t) - m(\lambda t)\}|\boldsymbol{x_i}]\big). \quad (25)$$

Note first that $\mathbb{P}\{-\epsilon/\lambda < |\boldsymbol{x_i}^\intercal t| - m_\infty(t) \le 0\}$ and $\mathbb{P}\{0 \le |\boldsymbol{x_i}^\intercal t| - m_\infty(t) < \epsilon/\lambda\}$ both exceed C^*/λ by assumption D for any fixed $\epsilon > 0$ and some C^* independent of t, λ. Hence the RHS in (25) is for sufficiently large λ bounded below by

$$\lambda^p\,\mathbb{E}\Big[f\Big\{\sup_t |\lambda m_\infty(t) - m(\lambda t)|\ |\boldsymbol{x_i}\Big\} I\big(\big||\boldsymbol{x_i}^\intercal t| - m_\infty(t)\big| \le \epsilon/\lambda\big)\Big]$$
$$\ge C^*\lambda^{p-1}\Big[f\Big\{\sup_t |\lambda m_\infty(t) - m(\lambda t)|\Big\}\Big]^r$$
$$\ge \frac{C^*}{C^{p-1}}(C\lambda)^{p-1} f\Big[f^{-1}\Big(\frac{1}{C\lambda}\Big) + C\lambda F\Big\{-f^{-1}\Big(\frac{1}{C\lambda}\Big)\Big\}\Big]^r, \quad (26)$$

where the first inequality in (26) follows from assumption D and the second from lemma C1 and where C is as chosen in lemma C1. Since p can be chosen arbitrarily large, it hence suffices that for some p^*,

$$\lim_{s\to\infty} \frac{f\{s + F(-s)/f(s)\}}{\{f(s)\}^{p^*}} > 0,$$

which was assumed in assumption D. □

For some $\sigma > 4$, let

$$\psi_n = n^{1/\sigma p}. \quad (27)$$

Lemma C3.

$$\sup_{0\le\lambda\le\psi_n}\sup_{\|t\|=1}|\hat{\boldsymbol{m}}_{\boldsymbol{v}}(\lambda,t)-m_v(\lambda,t)|\preceq 1/\sqrt{n}.$$

Proof. We use the shorthand $\hat{\boldsymbol{A}}_{\boldsymbol{v}}\{\hat{\boldsymbol{m}}_{\boldsymbol{v}}\}$ for $\hat{\boldsymbol{A}}_{\boldsymbol{v}}\{\lambda,t,\hat{\boldsymbol{m}}_{\boldsymbol{v}}(\lambda,t)\}$ and likewise for similar symbols. By the mean value theorem and assumption E for some function $\hat{\boldsymbol{m}}^*_{\boldsymbol{v}}$ between $\hat{\boldsymbol{m}}_{\boldsymbol{v}}$ and m_v,

$$A_v(\hat{\boldsymbol{m}}_{\boldsymbol{v}})-A_v(m_v)=\partial_m A_v(m_v)(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)+\partial_m^2 A_v(\hat{\boldsymbol{m}}^*_{\boldsymbol{v}})(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)^2/2. \tag{28}$$

Further, by the triangle inequality,

$$\begin{aligned}\sup_{\lambda\ge0}\sup_{\|t\|=1}|A_v(\hat{\boldsymbol{m}}_{\boldsymbol{v}})-A_v(m_v)|\le\\ \sup_{\lambda\ge0}\sup_{\|t\|=1}|A_v(\hat{\boldsymbol{m}}_{\boldsymbol{v}})-\hat{\boldsymbol{A}}_{\boldsymbol{v}}(\hat{\boldsymbol{m}}_{\boldsymbol{v}})-A_v(m_v)+\hat{\boldsymbol{A}}_{\boldsymbol{v}}(m_v)|+\sup_{\lambda\ge0}\sup_{\|t\|=1}|A_v(m_v)-\hat{\boldsymbol{A}}_{\boldsymbol{v}}(m_v)|\\ +\sup_{\lambda\ge0}\sup_{\|t\|=1}|\hat{\boldsymbol{A}}_{\boldsymbol{v}}(\hat{\boldsymbol{m}}_{\boldsymbol{v}})-q|+\sup_{\lambda\ge0}\sup_{\|t\|=1}|q-A_v(m_v)|.\end{aligned} \tag{29}$$

RHS3 and RHS4 in (29) are $\prec 1/\sqrt{n}$ by construction and RHS1 is $\prec 1/\sqrt{n}$ by lemma B3. RHS2 is $\preceq 1/\sqrt{n}$, also, by lemma B3. Combining (28) and (29) yields

$$\sup_{\lambda\ge0}\sup_{\|t\|=1}\left|\partial_m A_v(m_v)(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)+\partial_m^2 A_v(\hat{\boldsymbol{m}}^*_{\boldsymbol{v}})(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)^2/2\right|\preceq 1/\sqrt{n}. \tag{30}$$

Now let $\kappa_n\prec 1$ be such that

$$\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}|\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v|\preceq\kappa_n. \tag{31}$$

Such κ_n exist by lemma B7; we will choose it later. Then for any such κ_n it follows from (30) that

$$\begin{aligned}\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}\left|\partial_m A_v(m_v)(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)\right|\preceq\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}\left|\partial_m^2 A_v(\hat{\boldsymbol{m}}^*_{\boldsymbol{v}})(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)^2/2\right|+1/\sqrt{n}\\ \preceq\sup_{u,x}f'(u|x)\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}\left|(1+\lambda^p)^2(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)^2\right|+1/\sqrt{n}.\end{aligned} \tag{32}$$

RHS1 in (32) is $\prec\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}\left|\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v\right|$ by (31) and assumption E whereas the LHS in (32) is $\succeq\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}\left|\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v\right|$, by lemma C2. Hence

$$\sup_{0\le\lambda\le\kappa_n^{-2/\sigma p}}\sup_{\|t\|=1}|\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v|\preceq 1/\sqrt{n},$$

so we can choose $\kappa_n=1/\sqrt{n}$, which corresponds to $\psi_n=n^{1/\sigma p}$, as defined in (27). □

Lemma C4.

$$\sup_{0\le\lambda\le\psi_n}\sup_{\|t\|=1}\left|\hat{\boldsymbol{A}}_{\boldsymbol{v}}\{\lambda,t,m_v(\lambda,t)\}-A_v\{\lambda,t,m_v(\lambda,t)\}-\partial_m A_v\{\lambda,t,m_v(\lambda,t)\}\{\hat{\boldsymbol{m}}_{\boldsymbol{v}}(\lambda,t)-m_v(\lambda,t)\}\right| \prec 1/\sqrt{n}.$$

Proof. Using the same short hand notation as in lemma C3 we have by the triangle inequality that

$$\begin{aligned}|\hat{\boldsymbol{A}}_{\boldsymbol{v}}(m_v)-A_v(m_v)+\partial_m A_v(m_v)(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)| &\le |\hat{\boldsymbol{A}}_{\boldsymbol{v}}(m_v)-A_v(m_v)-\hat{\boldsymbol{A}}_{\boldsymbol{v}}(\hat{\boldsymbol{m}}_{\boldsymbol{v}})+A_v(\hat{\boldsymbol{m}}_{\boldsymbol{v}})|\\ &+|\hat{\boldsymbol{A}}_{\boldsymbol{v}}(\hat{\boldsymbol{m}}_{\boldsymbol{v}})-q|+|q-A_v(m_v)|+|A_v(m_v)-A_v(\hat{\boldsymbol{m}}_{\boldsymbol{v}})+\partial_m A_v(m_v)(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)|. \quad (33)\end{aligned}$$

RHS2 and RHS3 in (33) are $\prec 1/\sqrt{n}$ by construction and RHS1 is $\prec 1/\sqrt{n}$ by lemma B3, all uniformly in λ, t. By the mean value theorem, for some $\hat{\boldsymbol{m}}_{\boldsymbol{v}}^*$ between $m_v, \hat{\boldsymbol{m}}_{\boldsymbol{v}}$, RHS4 in (33) is

$$\sup_{0\le\lambda\le\psi_n}\sup_{\|t\|=1}|\partial_m^2 A_v(\hat{\boldsymbol{m}}_{\boldsymbol{v}}^*)(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)^2/2| \le (1+\psi_n^p)^2\sup_{u,x}f'(u|x)\sup_{0\le\lambda\le\psi_n}\sup_{\|t\|=1}(\hat{\boldsymbol{m}}_{\boldsymbol{v}}-m_v)^2 \preceq n^{2/\sigma-1}\prec 1/\sqrt{n},$$

by lemma C3 and because $\sigma > 4$; see (27). □

Lemma C5. *Recalling that* $c_{pt} = 1+\|t\|^p$,

$$\sup_{\|t\|\le\psi_n}\left|\frac{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)-m(t)}{c_{pt}}+\frac{\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t,m(t)\}-A\{t,m(t)\}}{c_{pt}\partial_m A\{t,m(t)\}}\right| \prec \frac{1}{\sqrt{n}}. \quad (34)$$

Proof. Note that for $\lambda = \|t\|$ and $t_0 = t/\|t\|$,

$$\begin{aligned}\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) &= c_{pt}\hat{\boldsymbol{m}}_{\boldsymbol{v}}(\lambda,t_0), & m(t) &= c_{pt}m_v(\lambda,t_0),\\ A\{t,m(t)\} &= A_v\{\lambda,t_0,m_v(\lambda,t_0)\}=q, & \hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t,m(t)\} &= \hat{\boldsymbol{A}}_{\boldsymbol{v}}\{\lambda,t_0,m_v(\lambda,t_0)\},\\ \partial_m A\{t,m(t)\} &= \partial_m A_v\{\lambda,t_0,m_v(\lambda,t_0)\}/c_{pt}.\end{aligned}$$

The stated result then follows from lemmas C2 and C4. □

APPENDIX D. VARIANCE MATRIX ESTIMATION

In this appendix the assumptions of theorem 5 are used. Let $h(t) = h^*(\theta_0 + t)$, set $p = p^* - 1$, with p^*, σ as defined in assumption F, and let

$$\begin{aligned}
\hat{\mathscr{D}}(t) &= \frac{1}{2nh(t)} \sum_{i=1}^{n} I\big\{\big||\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| - \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)\big| \le h(t)\big\}, \\
\hat{\boldsymbol{H}}(t,s) &= n^{-1} \sum_{i=1}^{n} I\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| \le \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)\} I\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal s| \le \hat{\boldsymbol{m}}_{\boldsymbol{n}}(s)\} \\
&\quad - \left(n^{-1} \sum_{i=1}^{n} I\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal t| \le \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)\}\right)\left(n^{-1} \sum_{i=1}^{n} I\{|\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal s| \le \hat{\boldsymbol{m}}_{\boldsymbol{n}}(s)\}\right).
\end{aligned} \tag{35}$$

Lemma D1. $\sup_{t,s\in\mathbb{R}^d} |\hat{\boldsymbol{H}}(t,s) - H(t,s)| \prec 1$.

Proof. We will show the uniform convergence of RHS1 in the definition of $\hat{\boldsymbol{H}}$ in (35), because the second term is similar. Reparametrize the first term of $\hat{\boldsymbol{H}}(t,s)$ in terms of $\lambda_0, \lambda_1, t_0, t_1$ to obtain

$$n^{-1} \sum_{i=1}^{n} I\{(1+\lambda_1^p)^{-1} |\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal \lambda_1 t_0| \le \hat{\boldsymbol{m}}_{\boldsymbol{v}}(\lambda_1, t_0)\} I\{(1+\lambda_2^p)^{-1} |\boldsymbol{u_i} - \boldsymbol{x_i}^\intercal \lambda_2 s_0| \le \hat{\boldsymbol{m}}_{\boldsymbol{v}}(\lambda_2, s_0)\},$$

where $\lambda_1, \lambda_2 \ge 0$ and $\|t_0\| = \|s_0\| = 1$. Since $\hat{\boldsymbol{m}}_{\boldsymbol{v}}$ converges uniformly to m_v by lemma B7, it suffices to show that for $\boldsymbol{a_i}$ defined at the beginning of appendix B,

$$n^{-1} \sum_{i=1}^{n} I(|\boldsymbol{a_i}^\intercal t^*| \le m^*) I(|\boldsymbol{a_i}^\intercal \tilde{t}^*| \le \tilde{m}^*) \xrightarrow{p} \mathbb{E}\{I(|\boldsymbol{a_i}^\intercal t^*| \le m^*) I(|\boldsymbol{a_i}^\intercal \tilde{t}^*| \le \tilde{m}^*)\} \tag{36}$$

uniformly in $(t^*, m^*, \tilde{t}^*, \tilde{m}^*) \in \mathbb{R}^{d+1} \times \mathbb{R}_+ \times \mathbb{R}^{d+1} \times \mathbb{R}_+$. Since the collection of half spaces in a Euclidean space is a VC class, the collection of indicator functions $\mathscr{F} = \{I(|\boldsymbol{a_i}^\intercal t^*| \le m^*) I(|\boldsymbol{a_i}^\intercal \tilde{t}^*| \le \tilde{m}^*) : (t^*, m^*, \tilde{t}^*, \tilde{m}^*) \in \mathbb{R}^{d+1} \times \mathbb{R}_+ \times \mathbb{R}^{d+1} \times \mathbb{R}_+\}$ generated by finite intersections of half spaces form a VC subgraph class. Since $\mathbb{R}^{d+1}$ and $\mathbb{R}_+$ are separable, $\mathscr{F}$ is a pointwise measurable class, and it follows that $\mathscr{F}$ is Glivenko–Cantelli. □

Lemma D2. *For some some $\epsilon > 0$ and recalling that $c_{pt} = 1 + \|t\|^p$,*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\inf_{t\in\mathbb{R}^d} c_{pt} \hat{\mathscr{D}}(t) < \epsilon\Big\} = 0.$$

Proof. Note that $\hat{\mathscr{D}}(t) = 2\big[\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) + h(t)\} - \hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - h(t)\}\big]/h(t)$, such that by lemma B4,

$$\sup_t c_{pt} \Big| 2\hat{\mathscr{D}}(t) - \frac{A\{t, \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) + h(t)\} - A\{t, \hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - h(t)\}}{h(t)} \Big| \preceq \frac{1}{\sqrt{n} h_0} \prec 1. \tag{37}$$

Using $\hat{\mathscr{Q}}_{\boldsymbol{a}}, \mathscr{Q}_a$ from appendix B it follows from assumption F that

$$\sup_{\|t\|>\psi_n} \frac{|\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m(t)|}{h(t)} \leq \frac{\hat{\mathscr{Q}}_{\boldsymbol{a}} + \mathscr{Q}_a}{h_0} \sup_{\|t\|>\psi_n} \frac{1+\|t\|}{1+\|t\|^p} \preceq \frac{n^{(1-p)/p\sigma}}{h_0} \prec 1. \tag{38}$$

Likewise, by lemma C3 and assumption F,

$$\sup_{\|t\|\leq\psi_n} \frac{|\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m(t)|}{h(t)} = \sup_{\|t\|\leq\psi_n} \frac{|\hat{\boldsymbol{m}}_{\boldsymbol{v}} - m_v(t)|}{h_0} \preceq \frac{1}{\sqrt{n}h_0} \prec 1. \tag{39}$$

From (37) to (39) it follows that it suffices to show that

$$\inf_t \frac{A\{t, m(t) + h(t)/2\} - A\{t, m(t) - h(t)/2\}}{h_0} > 0. \tag{40}$$

The LHS in (40) is by the mean value theorem for some $0 < \omega(t) < 1$ and $\epsilon^*(t) = \omega(t)h(t)/2$ equal to

$$\begin{aligned} c_{pt}\partial_m A\{t, m(t) + \epsilon^*(t)\} + c_{pt}\partial_m A\{t, m(t) - \epsilon^*(t)\} = \\ c_{pt}\,\mathbb{E}\big[f\{\boldsymbol{x_i}^\intercal t + m(t) + \epsilon^*(t)|\boldsymbol{x_i}\} + f\{\boldsymbol{x_i}^\intercal t - m(t) - \epsilon^*(t)|\boldsymbol{x_i}\} \\ + f\{\boldsymbol{x_i}^\intercal t + m(t) - \epsilon^*(t)|\boldsymbol{x_i}\} + f\{\boldsymbol{x_i}^\intercal t - m(t) + \epsilon^*(t)|\boldsymbol{x_i}\}\big], \end{aligned}$$

which by assumption A equals

$$\begin{aligned} c_{pt}\,\mathbb{E}\big[f\{|\boldsymbol{x_i}^\intercal t| + m(t) + \epsilon^*(t)|\boldsymbol{x_i}\} + f\{|\boldsymbol{x_i}^\intercal t| - m(t) - \epsilon^*(t)|\boldsymbol{x_i}\} \\ + f\{|\boldsymbol{x_i}^\intercal t| + m(t) - \epsilon^*(t)|\boldsymbol{x_i}\} + f\{|\boldsymbol{x_i}^\intercal t| - m(t) + \epsilon^*(t)|\boldsymbol{x_i}\}\big] \geq \\ c_{pt}\,\mathbb{E}\big[I\{|\boldsymbol{x_i}^\intercal t| - m(t) \geq 0\}f\{|\boldsymbol{x_i}^\intercal t| - m(t) - \epsilon^*(t)|\boldsymbol{x_i}\} + I\{|\boldsymbol{x_i}^\intercal t| - m(t) < 0\}f\{|\boldsymbol{x_i}^\intercal t| - m(t) + \epsilon^*(t)|\boldsymbol{x_i}\}\big] \\ \geq c_{pt}\,\mathbb{E}\big[I\{|\boldsymbol{x_i}^\intercal t| - m(t) \geq 0\}f\{|\boldsymbol{x_i}^\intercal t| - m(t)|\boldsymbol{x_i}\} + I\{|\boldsymbol{x_i}^\intercal t| - m(t) < 0\}f\{|\boldsymbol{x_i}^\intercal t| - m(t)|\boldsymbol{x_i}\}\big] \\ = c_{pt}\,\mathbb{E}\big[f\{|\boldsymbol{x_i}^\intercal t| - m(t)|\boldsymbol{x_i}\}\big]. \end{aligned}$$

Now,

$$\inf_t\Big(c_{pt}\,\mathbb{E}\big[f\{|\boldsymbol{x_i}^\intercal t| - m(t)|\boldsymbol{x_i}\}\big]\Big) = \inf_{\lambda\geq 0}\inf_{\|t\|=1}\Big((1+\lambda^p)\,\mathbb{E}\big[f\{\lambda|\boldsymbol{x_i}^\intercal t| - m(\lambda t)|\boldsymbol{x_i}\}\big]\Big). \tag{41}$$

The RHS in (41) is the infimum of the middle expression in (25), which is shown to be bounded away from zero, uniformly in λ, t, in lemma C2. □

Lemma D3. $\forall t \in \mathbb{R}^d : |\hat{\mathscr{D}}(t) - \mathscr{D}(t)| \prec 1.$

Proof. Choose t. From (37) and assumption F it follows that

$$\left|2\hat{\mathscr{D}}(t) - \frac{A\{t, \hat{\boldsymbol{m}}_n(t) + h(t)\} - A\{t, \hat{\boldsymbol{m}}_n(t) - h(t)\}}{h(t)}\right| \preceq \frac{1}{\sqrt{n}h_0} \prec 1.$$

Take n large enough to ensure that $\|t\| \leq \psi_n$. We have by the mean value theorem that

$$\frac{A\{t, \hat{\boldsymbol{m}}_n(t) + h(t)\} - A\{t, \hat{\boldsymbol{m}}_n(t) - h(t)\}}{h(t)} - 2\mathscr{D}(t) = \frac{\partial_m^2 A(t,\cdot)}{2h(t)}\{\hat{\boldsymbol{m}}_n(t) - m(t) + h(t)\}^2 - \frac{\partial_m^2 A(t,\cdot)}{2h(t)}\{\hat{\boldsymbol{m}}_n(t) - m(t) - h(t)\}^2. \quad (42)$$

By lemma C3 and assumptions E and F both RHS terms in (42) are $\simeq (1/n + h_0^2)/h_0 \prec 1$. □

Lemma D4.

$$\forall t, s \in \mathbb{R}^d : \frac{\hat{\boldsymbol{H}}(t,s)}{\hat{\mathscr{D}}(t)\hat{\mathscr{D}}(s)} - \frac{H(t,s)}{\mathscr{D}(t)\mathscr{D}(s)} \prec 1.$$

Proof. Follows immediately from lemmas D1 to D3 and the fact that $\hat{\boldsymbol{H}}$ and H are bounded. □

Appendix E. Theorems

Proof of theorem 1. The proof of this theorem is inspired by that of Rousseeuw (1984, theorem 1); see also Rousseeuw and Leroy (1987, section 3.4). Let $\mathscr{Z} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}$ be the original sample, $\mathscr{Z}^* = \{(\boldsymbol{x_i^*}, \boldsymbol{y_i^*})\}$ be the contaminated sample and $\mathscr{Z}^\dagger$ the sample consisting of observations shared between the two.

Suppose that the number of observations contaminated is at most $\boldsymbol{b^*} = \min(n - N, N - \hat{\boldsymbol{\gamma}} - 1) > 0$. We show that $\hat{\boldsymbol{\theta}}$ is bounded, for which it suffices to show that (i) $\int \|\theta\| \exp\{-\hat{\mathscr{Q}}(|\boldsymbol{y_i^*} - \theta^\mathsf{T}\boldsymbol{x_i^*}|^2)\}d\theta < \infty$, (ii) $\int \exp\{-\hat{\mathscr{Q}}(|\boldsymbol{y_i^*} - \theta^\mathsf{T}\boldsymbol{x_i^*}|^2)\}d\theta > 0$. Let $\bar{\boldsymbol{y}} = \max_{\mathscr{Z}} |\boldsymbol{y_i}|$ and $\bar{\boldsymbol{x}} = \max_{\mathscr{Z}} \|\boldsymbol{x_i}\|$.

First (ii). Since $\boldsymbol{b^*} > 0$ there exists for each θ at least one observation $\left(\boldsymbol{x_i^\dagger}(\theta), \boldsymbol{y_i^\dagger}(\theta)\right) \in \mathscr{Z}^\dagger$ for which $\hat{\mathscr{Q}}(|\boldsymbol{y_i^*} - \theta^\mathsf{T}\boldsymbol{x_i^*}|) \leq |\boldsymbol{y_i^\dagger}(\theta) - \theta^\mathsf{T}\boldsymbol{x_i^\dagger}(\theta)| \leq \bar{\boldsymbol{y}} + \bar{\boldsymbol{x}}\|\theta\|$. Hence the left hand side (LHS) in (ii) is bounded below by $\exp(-2\bar{\boldsymbol{y}}^2) \int \exp(-2\bar{\boldsymbol{x}}^2\|\theta\|^2)d\theta > 0$.

Now (i). Let $\mathscr{B}(B, \rho)$ be the ρ–expansion of a $(d-1)$–dimensional subspace B of $\mathbb{R}^d$ and let $\rho(\mathscr{Z})$ be the smallest ρ for which $\mathscr{B}(B, \rho)$ contains at least $\hat{\boldsymbol{\gamma}} + 1$ of the $\boldsymbol{x_i}$'s. By the definition of $\hat{\boldsymbol{\gamma}}$, $\rho(\mathscr{Z}) > 0$.

Since at most b^* observations are contaminated, $\mathscr{Z}^\dagger$ contains at least $\hat{\boldsymbol{\gamma}} + 1$ observations indexed $i_1, \ldots, i_{\hat{\boldsymbol{\gamma}}+1}$ for which $\hat{\mathscr{Q}}(|\boldsymbol{y_i^*} - \theta^\mathsf{T}\boldsymbol{x_i^*}|) \geq |\boldsymbol{y}_{i_j} - \boldsymbol{x}_{i_j}^\mathsf{T}\theta|$. Take $B(\theta) = \{x \in \mathbb{R}^d : x^\mathsf{T}\theta = 0\}$. Then for at least one $j = j(\theta)$, $\boldsymbol{x}_{i_j} \notin \mathscr{B}\{B(\theta), \rho(\mathscr{Z})/2\}$ by the definition of $\rho(\mathscr{Z})$. For this value of j, let $\boldsymbol{x_i^\dagger}(\theta) = \boldsymbol{x}_{i_j}$, $\boldsymbol{y_i^\dagger}(\theta) = \boldsymbol{y}_{i_j}$, and let $\bar{\boldsymbol{y}}$ be as defined above. Then since $\hat{\mathscr{Q}}(|\boldsymbol{y_i^*} - \theta^\mathsf{T}\boldsymbol{x_i^*}|) \geq |\boldsymbol{y_i^\dagger} -$

$\theta^{\mathsf{T}}\boldsymbol{x}_i^{\dagger}(\theta)| \geq \rho(\mathscr{Z})\|\theta\|/2 - \bar{\boldsymbol{y}}$ and $\hat{\mathscr{Q}}(|\boldsymbol{y}_i^* - \theta^{\mathsf{T}}\boldsymbol{x}_i^*|^2) \geq \rho^2(\mathscr{Z})\|\theta\|^2/4 - \rho(\mathscr{Z})\bar{\boldsymbol{y}}\|\theta\|$ for all θ, the LHS in (i) is bounded above by $\exp(\bar{\boldsymbol{y}}^2)\int \|\theta\| \exp\big[-\rho^2(\mathscr{Z})\{\|\theta\| - 2\bar{\boldsymbol{y}}/\rho(\mathscr{Z})\}^2/4\big]d\theta < \infty$. □

Proof of theorem 4. We work on the numerator and denominator in (5) separately.

First the denominator. Let $\omega(m) = \exp(-m^2)$. Expanding $\omega\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)\}$ around $m(t)$ yields for some $\hat{\boldsymbol{m}}_{\boldsymbol{n}}^*(t)$ between $m(t)$ and $\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)$,

$$\int \exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\}dt = \int \exp\{-m^2(t)\}dt - 2\int \hat{\boldsymbol{m}}_{\boldsymbol{n}}^*(t)\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m(t)\}\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{*\boldsymbol{2}}(t)\}dt. \tag{43}$$

For RHS2 in (43) we have by lemmas B7 to B11 for some $\epsilon > 0$ that

$$\int \hat{\boldsymbol{m}}_{\boldsymbol{n}}^*(t)\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m(t)\}\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{*\boldsymbol{2}}(t)\}dt \prec \int c_{pt}(1 + \|t\|)\exp(-\epsilon^2\|t\|^2)dt \preceq 1.$$

Now the numerator. Let ψ_n be as defined in (27), $\mathscr{T}_n = \{t \in \mathbb{R}^d : \|t\| \leq \psi_n\}$, and $\mathscr{T}_n^c = \{t \in \mathbb{R}^d : \|t\| > \psi_n\}$. Then by lemma A1,

$$\begin{aligned}\int t\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\}dt = \\ \int_{\mathscr{T}_n} t[\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\} - \exp\{-m^2(t)\}]dt + \int_{\mathscr{T}_n^c} t\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\}dt - \int_{\mathscr{T}_n^c} t\exp\{-m^2(t)\}dt.\end{aligned} \tag{44}$$

For RHS2 in (44) note that by lemma B9 for some $\epsilon > 0$ and since ψ_n is polynomial in n,

$$\int_{\mathscr{T}_n^c} \|t\|\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\}dt \preceq \int_{\mathscr{T}_n^c} \|t\|\exp(-\epsilon^2 t^2)dt \prec 1/\sqrt{n}. \tag{45}$$

RHS3 in (44) can similarly be shown to be $\prec 1/\sqrt{n}$ using lemma B8. For RHS1 in (44) we have

$$\begin{aligned}\int_{\mathscr{T}_n} t[\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\} - \exp\{-m^2(t)\}]dt = -2\int_{\mathscr{T}_n} tm(t)\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m(t)\}\exp\{-m^2(t)\}dt \\ - \int_{\mathscr{T}_n} t\{1 - 2\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{*\boldsymbol{2}}(t)\}\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m(t)\}^2\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{*\boldsymbol{2}}(t)\}dt.\end{aligned} \tag{46}$$

RHS2 in (46) is $\prec 1/\sqrt{n}$ by lemmas B8 to B11, C2 and C3. Further, RHS1 in (46) equals

$$\begin{aligned}-2\int_{\mathscr{T}_n} tc_{pt}m(t)\Big\{\frac{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t) - m_t}{c_{pt}} + \frac{\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, m(t)\} - A\{t, m(t)\}}{c_{pt}\partial_m A\{t, m(t)\}}\Big\}\exp\{-m^2(t)\}dt \\ -2\int_{\mathscr{T}_n^c} tc_{pt}m(t)\frac{\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, m(t)\} - A\{t, m(t)\}}{c_{pt}\partial_m A\{t, m(t)\}}\exp\{-m^2(t)\}dt. \\ +2\int tm(t)\frac{\hat{\boldsymbol{A}}_{\boldsymbol{n}}\{t, m(t)\} - A\{t, m(t)\}}{\mathscr{D}(t)}\exp\{-m^2(t)\}dt\end{aligned} \tag{47}$$

The first term in (47) is $\prec 1/\sqrt{n}$ by lemma C5. The second term in (47) is also $\prec 1/\sqrt{n}$, which can be established along the lines of (45), noting that $c_{pt}\partial_m A$ is bounded away from zero by lemma C2 and that $tc_{pt}m(t)$ is polynomial in t.

Finally, $\sqrt{n}$ times the last term in (47) equals

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\int t\frac{m(t)}{\mathscr{D}(t)}\left[I\{|\boldsymbol{u_i}-\boldsymbol{x_i}^{\intercal}t|\leq m(t)\}-q\right]\exp\{-m^2(t)\}dt. \tag{48}$$

Apply the Lindeberg–Levy central limit theorem. □

Proof of theorem 5. We show consistency of the numerator; the denominator is easier. Recall the definitions in (35) and let

$$\hat{\boldsymbol{R}}(t,s)=\frac{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)\hat{\boldsymbol{m}}_{\boldsymbol{n}}(s)}{\hat{\mathscr{D}}(t)\hat{\mathscr{D}}(s)}\hat{\boldsymbol{H}}(t,s)\exp\{-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(s)\},$$
$$R(t,s)=\frac{m(t)m(s)}{\mathscr{D}(t)\mathscr{D}(s)}H(t,s)\exp\{-m^2(t)-m^2(s)\}.$$

Substituting $t=\theta-\theta_0$ and $s=\tilde{\theta}-\theta_0$ in the numerator in (12) yields

$$\iint\{t+(\theta_0-\hat{\boldsymbol{\theta}})\}\{s+(\theta_0-\hat{\boldsymbol{\theta}})\}^{\intercal}\hat{\boldsymbol{R}}(t,s)dtds. \tag{49}$$

To establish that (49) converges in probability to the numerator in (9), it suffices by the consistency of $\hat{\boldsymbol{\theta}}$ to show that *(i)* $\iint ts^{\intercal}\{\hat{\boldsymbol{R}}(t,s)-R(t,s)\}dtds\prec 1$, *(ii)* $\iint t\hat{\boldsymbol{R}}(t,s)dtds\preceq 1$, *(iii)* $\iint s^{\intercal}\hat{\boldsymbol{R}}(t,s)dtds\preceq 1$, *(iv)* $\iint\hat{\boldsymbol{R}}(t,s)dtds\preceq 1$. Since establishing (ii) to (iv) is similar to but easier than (i), we only establish (i) here.

We use the weak version of Glick (1974, theorem A).[14] Pointwise convergence in probability of $\hat{\boldsymbol{R}}$ to R follows from lemmas B7 and D4 and Slutsky.

We now only need to find a convergent upper bound to $\|t\|\|s\||\hat{\boldsymbol{R}}(t,s)|$ whose limit is integrable. Because $\sup_m\{|m|\exp(-|m|)\}=1/e$, $\hat{\boldsymbol{H}}$ is bounded, it suffices by lemma D2 to show that for any fixed order polynomial P, $\int|P(t)|\exp\{\hat{\boldsymbol{m}}_{\boldsymbol{n}}(t)-\hat{\boldsymbol{m}}_{\boldsymbol{n}}^{\boldsymbol{2}}(t)\}dt$ is convergent, which follows from lemma B9. □

Proof of theorem 6. Let $u=y-\theta_0$. It follows from (13) that we need to determine when

$$\partial_u\int t\frac{m(t)}{\mathscr{D}(t)}I\{|u-t|\leq m(t)\}\exp\{-m^2(t)\}dt \tag{50}$$

[14] See the comment on page 67 in Glick (1974).

is uniformly bounded in u.[15] Let $t^- = t^-(u), t^+ = t^+(u)$ be such that $u - t^- - m(t^-) = 0$, $u - t^+ + m(t^+) = 0$ if such t^-, t^+ exist; set t^- and/or t^+ equal to $-\infty, \infty$, respectively if no solution exists. So $|u - t| \leq m(t) \iff t^-(u) \leq t \leq t^+(u)$.

From the definition of m it follows that if solutions for t^+, t^- exist they solve

$$F(2t^+ - u) - F(u) = 1/2, \qquad F(u) - F(2t^- - u) = 1/2.$$

Hence $t^+ = \infty$ if $u \geq 0$ and $t^- = -\infty$ if $u \leq 0$. We concentrate on the case $u < 0$; the case $u > 0$ is similar and the left derivative at $u = 0$ (if it is finite) equals the limit as $u \uparrow 0$. The expression in (50) then equals

$$\partial_u \int_{-\infty}^{t^+(u)} t \frac{m(t)}{\mathscr{D}(t)} \exp\{-m^2(t)\} dt = t^+ \frac{m(t^+)}{\mathscr{D}(t^+)} \exp\{-m^2(t^+)\} \partial_u t^+. \tag{51}$$

By the implicit function theorem $\partial_u t^+ = \mathscr{D}(t^+)/2f(2t^+ - u)$, such that twice the RHS in (51) equals

$$t^+ \frac{m(t^+)}{f(2t^+ - u)} \exp\{-m^2(t^+)\} = \frac{t^+(t^+ - u)}{f(2t^+ - u)} \exp\{-(t^+ - u)^2\} =$$
$$\varsigma\sqrt{2\pi}\, t^+(t^+ - u) \exp\{-(1 - 2/\varsigma^2)t^{+2} + 2(1 - 1/\varsigma^2)ut^+ - (1 - 1/2\varsigma^2)u^2\}.$$

Take the limit as $u \uparrow 0$ noting that for fixed negative u, t^+ is finite and that (tedious but simple derivations yield) $\lim_{u\uparrow 0}\{ut^+(u)\} = 0$. □

Proof of theorem 7. Note that $\hat{\mathscr{D}}(|\boldsymbol{y_i} - \boldsymbol{x_i}\theta|)$ is a piecewise linear function of θ. Indeed, on each such segment $[\underline{\theta}, \bar{\theta}]$, $\hat{\mathscr{D}}(|\boldsymbol{y_i} - \boldsymbol{x_i}\theta|) = |\boldsymbol{y_j} - \boldsymbol{x_j}\theta|$ for some $j = 1, \ldots, n$. If $\underline{\theta}, \bar{\theta}$ are both finite then for $s = 0, 1$,

$$\lim_{\alpha\to 0} \int_{\underline{\theta}}^{\bar{\theta}} \theta^s \exp\{-\alpha\, \hat{\mathscr{D}}\, |\boldsymbol{y_i} - \boldsymbol{x_i}\theta|^2\} d\theta = \lim_{\alpha\to 0} \int_{\underline{\theta}}^{\bar{\theta}} \theta^s \exp\{-\alpha(\boldsymbol{y_j} - \boldsymbol{x_j}\theta)^2\} d\theta = (\bar{\theta}^{s+1} - \underline{\theta}^{s+1})/(s+1),$$

which is finite. The limit is hence determined by the terms for which $\bar{\theta}$ or $\underline{\theta}$ is infinite. Note that $\hat{\mathscr{D}}(|\boldsymbol{y_i} - \boldsymbol{x_i}\theta|) = |\boldsymbol{y_\mu} - \boldsymbol{x_\mu}\theta|$ for any sufficiently large $|\theta|$. Now, for the denominator,

$$\lim_{\alpha\to 0}\Big(\sqrt{\alpha/\pi} \int_{-\infty}^{\bar{\theta}} \exp\{-\alpha(\boldsymbol{y_\mu} - \boldsymbol{x_\mu}\theta)^2\} d\theta\Big) = 1/2\boldsymbol{x_\mu}. \tag{52}$$

[15] There is a slight abuse of notation here since the left and right derivatives at $u = 0$ can differ. If that is the case, let for the remainder of this proof ∂_u denote the greater in absolute value of the left and right derivatives.

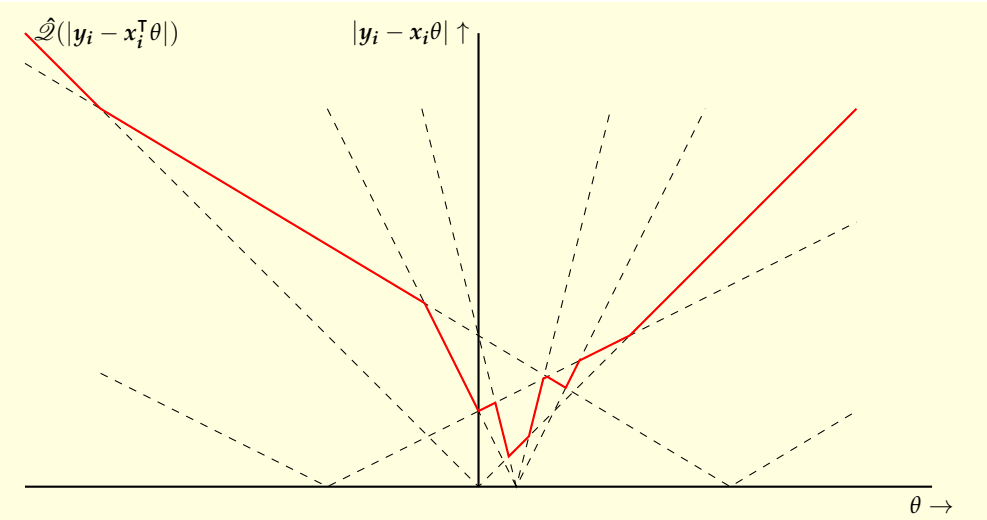

FIGURE 3. Gibbs sampling — single draws

By symmetry, the same limit applies to $\int_{\bar\theta}^{\infty}$, so the denominator converges to $1/x_\mu$. For the numerator and arbitrary $\underline\theta, \bar\theta$ by substitution of $t = \sqrt{2\alpha}(x_\mu\theta - y_\mu)$,

$$\begin{aligned}\sqrt{\frac{\alpha}{\pi}}\int_{\underline\theta}^{\bar\theta}\theta\exp\{-\alpha(y_\mu - x_\mu\theta)^2\}d\theta &= \frac{1}{x_\mu^2}\int_{\sqrt{2\alpha}(x_\mu\underline\theta - y_\mu)}^{\sqrt{2\alpha}(x_\mu\bar\theta - y_\mu)}\Big(\frac{t}{\sqrt{2\alpha}} + y_\mu\Big)\phi(t)dt \\ &= \frac{1}{\sqrt{2\alpha}x_\mu^2}\big[\phi\{\sqrt{2\alpha}(x_\mu\underline\theta - y_\mu)\} - \phi\{\sqrt{2\alpha}(x_\mu\bar\theta - y_\mu)\}\big] \\ &\quad + \frac{y_\mu}{x_\mu^2}\big[\Phi\{\sqrt{2\alpha}(x_\mu\bar\theta - y_\mu)\} - \Phi\{\sqrt{2\alpha}(x_\mu\underline\theta - y_\mu)\}\big].\end{aligned}$$

Hence adding the terms for $\underline\theta = -\infty$ with arbitrary $\bar\theta$ and $\bar\theta = \infty$ with arbitrary $\underline\theta$ and taking $\alpha \to 0$, we obtain

$$\frac{y_\mu}{x_\mu^2} + \lim_{\alpha\to 0}\frac{\phi\{\sqrt{2\alpha}(x_\mu\underline\theta - y_\mu)\} - \phi\{\sqrt{2\alpha}(x_\mu\bar\theta - y_\mu)\}}{\sqrt{2\alpha}} = \frac{y_\mu}{x_\mu^2}, \tag{53}$$

since ϕ has derivative zero at zero. Hence $\lim_{\alpha\to 0}\hat{\boldsymbol\theta}_\alpha = (y_\mu/x_\mu^2)/(1/x_\mu) = y_\mu/x_\mu$. □

APPENDIX F. COMPUTATION

The method we describe here is Gibbs sampling (Geman and Geman, 1984); for other possibilities see section 7. Because $\hat{\boldsymbol\theta}$ can be thought of as the mean of a distribution with density function $\psi(\theta) \propto \exp\{-\hat{\mathscr{Q}}(|y_i - x_i^\intercal\theta|^2)\}$, all we need is a method to draw random numbers from that distribution. The idea is to draw random numbers from the conditional distribution of each element θ^* of θ given the remaining elements.

Since the linear combination of remaining regressors and corresponding coefficients can be absorbed into $\boldsymbol{y_i}$, the remainder of our discussion presumes $d = 1$ and n odd.[16] To simplify the discussion further, we will presume that all regressors are positive–valued and that there are no two regressors taking the same value. Zeroes and ties require minor adjustments to the procedure and negative values can be accommodated by replacing $(\boldsymbol{x_i}, \boldsymbol{y_i})$ with $(-\boldsymbol{x_i}, -\boldsymbol{y_i})$.

Figure 3 represents the way $\hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}\theta|)$ varies with the value of θ. Since $\hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}\theta|)$ is piecewise linear in θ, ψ is a different normal density in each of a number of intervals $(\theta_{(j)}, \theta_{(j+1)}]$, $j = 0, \ldots, J$. It can be shown that $J \leq 2n + 1$. Thus, since for some observation i_j, $\hat{\mathscr{Q}}(|\boldsymbol{y_i} - \boldsymbol{x_i}\theta|) = |\boldsymbol{y}_{i_j} - \boldsymbol{x}_{i_j}\theta|$ for all $\theta \in (\theta_{(j)}, \theta_{(j+1)}]$ we get for $\boldsymbol{\Phi_j}(\theta) = \Phi\{\sqrt{2}(\boldsymbol{x}_{i_j}\theta_{(j+1)} - \boldsymbol{y}_{i_j})\}$ that

$$\Psi(\theta) \propto \sum_{j=0}^{J} \Big\{ I(\theta_{(j+1)} \leq \theta) \Big(\frac{\boldsymbol{\Phi_j}(\theta_{(j+1)}) - \boldsymbol{\Phi_j}(\theta_{(j)})}{\boldsymbol{x}_{i_j}} \Big) + I(\theta_{(j)} \leq \theta < \theta_{(j+1)}) \Big(\frac{\boldsymbol{\Phi_j}(\theta) - \boldsymbol{\Phi_j}(\theta_{(j)})}{\boldsymbol{x}_{i_j}} \Big) \Big\}.$$

So one only needs to find the boundary points $\theta_{(j)}$ and corresponding observation i_j in order to compute Ψ. Once the boundaries $\theta_{(j)}$ are known, $\Psi^{-1}(\zeta)$ can be computed for any value $\zeta \in (0, 1)$ by identifying the value of j for which $\Psi(\theta_{(j)}) \leq \zeta < \Psi(\theta_{(j+1)})$ and then computing the inverse of a univariate normal distribution function.

What remains to be done, therefore, is to find the corner points $\theta_{(j)}$. The simplest way of achieving this is to compute all intersection points by brute force, which takes $O(n^2)$ operations. Noting that both i_0 and i_J equal the index of the median of $\boldsymbol{x_i}$, one can start on the left and look at all intersection points of the downward sloping line corresponding to i_0 with the remaining observations. The observation whose line intersects first will be i_1, and so forth. Such a brute force approach suffices for most applications.

For the simulations in section 6 we used a more complicated algorithm, which is significantly faster for large n. A C program is available from the authors.

[16] To accommodate even n, just add an observation with both $\boldsymbol{x_i}, \boldsymbol{y_i}$ set to zero.