Nuffield College
UNIVERSITY OF OXFORD

ECONOMICS DISCUSSION PAPERS

Paper No: 2026-W01

Robustness Properties of Least Trimmed Squares with Categorial Covariates

By Otso Hao and Bent Nielsen

April 2026

Robustness Properties of Least Trimmed Squares with Categorical Covariates

Otso Hao*‡ & Bent Nielsen†‡

18 March 2026

Abstract

We study outlier robustness properties of Least Trimmed Squares (LTS) estimators in linear models with categorical covariates. In a sample of size n, LTS minimises the sum of h smallest squared residuals, where the number $h \leq n$ is chosen by the user. We find conditions under which LTS is bounded (in probability) when a positive share of observations are outliers. Our boundedness guarantees are uniform in h and apply to a chosen sub-coefficient of interest. We show that LTS is robust in a wider range of settings with categorical regressors than suggested by existing boundedness and breakdown point results. We also propose a new data-driven approach to choosing an initial h, which is useful for methods that estimate the number of outliers.

1 Introduction

Applied researchers are often concerned about 'outliers': observations that deviate from the majority and induce distortions in methods such as least squares. We study outlier robustness properties of the Least Trimmed Squares (LTS) estimator (Rousseeuw, 1984) in regression models where some covariates are categorical. In a sample of size n, LTS minimises the sum of h smallest squared residuals, where $h \leq n$ is chosen by the user. LTS possesses desirable properties, including scale equivariance and resistance to bad leverage points, which distinguish it from other robust methods such as M-estimators.

In the past, LTS estimators have been mainly studied under an assumption that all covariates are continuous, or satisfying a property known as *general position*. Yet in practice, covariates are often categorical. An example is the data on identical twins

*Nuffield College & Department of Economics, University of Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: otso.hao@nuffield.ox.ac.uk.

†Nuffield College & Department of Economics, University of Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: bent.nielsen@nuffield.ox.ac.uk.

‡Support from Aarhus Center for Econometrics (ACE) funded by the Danish National Foundation [grant DNRF186], Economic and Social Research Council [grant 2261272], and The Ella and Georg Ehrnrooth Foundation is gratefully acknowledged. Authors thank Vanessa Berenguer-Rico, Jurgen Doornik, James Duffy, and Frank Windmeijer for helpful comments.

from Bonjour et al. (2003), which contain individual wages (w_{if}), years of schooling (s_{if}), and additional controls (c_{if}) for twins i in families f. The equation to be estimated is $w_{1f} - w_{2f} = \beta(s_{1f} - s_{2f}) + \alpha'(c_{1f} - c_{2f}) + (\varepsilon_{1f} - \varepsilon_{2f})$, where all regressors are categorical and β, the parameter of interest, is interpreted as the wage return to education. Amin (2011) revisited the analysis and raised concerns of outliers. Our theory sheds light on the properties of LTS estimators in this setting, offering a systematic approach to outlier detection and robust estimation.

We find conditions for LTS estimators to be bounded (in probability) when a positive share of observations are 'outliers'. Boundedness is a probabilistic counterpart to finite sample breakdown point (Donoho and Huber, 1983), where the latter is defined as the smallest share of contamination in a *given sample* that can create unbounded distortions in an estimator. Our definition of 'outliers' is motivated by a statistical model where LTS is maximum likelihood (Berenguer-Rico et al., 2023).

The conditions involve the concentration of regressors on 'hyperplane strips', which are thin neighbourhoods around hyperplanes. Berenguer-Rico and Nielsen (2025b) analysed boundedness of LTS estimators in a general setting with strip conditions on the full sample regressors. Related strip conditions have been used to study M-estimators (Johansen and Nielsen, 2019; Chen and Wu, 1988) and S-estimators (Lopuhaä et al., 2023; Davies, 1990). Strip conditions are most binding for categorical regressors, which is our focus.

Our approach to boundedness contains two new features. First, we find conditions to guarantee boundedness for a subcomponent of the full regression slope vector. This is relevant when only some of the parameters are of interest as in Bonjour et al. (2003). Second, our boundedness results are uniform over a range of values for h, which is useful if LTS is computed iteratively over different h, as is common in practice.

We show that LTS is outlier robust in a wider range of settings with categorical covariates than suggested by existing results. As an example, consider a model with an intercept and a binary regressor z_i, with $z_i = 1$ for half of the observations. Berenguer-Rico and Nielsen (2025b) and existing breakdown point results (e.g. Mili and Coakley, 1996) suggest LTS is robust up to share $1/4$ of outliers. In contrast, we show how the robustness of LTS depends on the nature of outlier contamination. If extremely 'malicious' forms of contamination, where outliers concentrate on a single value of z_i, are ruled out, up $1/3$ of outliers can be allowed. If outliers further have a large magnitude, as in the LTS model of Berenguer-Rico et al. (2023), up to share $1/2$ of outliers can be allowed. Our findings resonate with Huber and Ronchetti (2009, Section 11.4), who note that breakdown point measures of robustness are overly pessimistic when some regressors are categorical.

We also establish a relationship between strip conditions and 'hyperplane conditions', where the latter play a key role in breakdown point theory (Davies and Gather, 2005). This result has practical relevance, as hyperplane conditions are easier to interpret and check from data.

We then propose a new approach to choosing an initial h for the LTS estimator. An initial h is needed to estimate the number of outliers using, for example, an index plot (Rousseeuw and Leroy, 1987) or forward search (Hadi and Simonoff, 1993; Atkinson and Riani, 2000). An initial choice is also used to initialise MM-estimators (Yohai, 1987)

and reweighted LTS estimators (Alfons et al., 2013; Cížek, 2013). The current standard practice is to use a breakdown point optimal h as an initial value. In fact, the initial choice $h \approx n/2$ is often used, even though this has lacked robustness guarantees in models with categorical regressors.

We write down a data-driven algorithm for choosing an initial h. This algorithm allows a user to exploit information about the type of outlier contamination suspected in their application. The breakdown point optimal choice of initial h is covered as a special case when no information about outliers is available. The algorithms are given robustness guarantees using our boundedness results.

The paper is structured as follows. Section 2 defines LTS estimators and explains the role of an initial h. Section 3 gives the boundedness results. In section 4, we use these results to write down an algorithm for choosing an initial h. Sections 5 and 6 contain simulations and an empirical illustration. Section 7 concludes. The appendix consists of proofs and details on algorithms for exploring the regressor space.

2 LTS estimator and the initial h

LTS estimators depend on a tuning parameter h that determines how many observations are trimmed from the sample. An initial choice of h is often used to estimate the number of outliers in a data generating process.

2.1 LTS Estimator

Consider a linear equation $y_i = x_i'\beta + \sigma\varepsilon_i$ for $i = 1, \ldots, n$ with a scalar y_i and a p-vector x_i which may include an intercept. Following Rousseeuw and van Driessen (2000), the LTS estimator is defined as follows. The user first chooses a number $h \leq n$. For an h-subset ζ of $\{1, \ldots, n\}$, define least squares estimators

$$\hat{\beta}_\zeta = \arg\min_\beta \sum_{i\in\zeta} (y_i - x_i'\beta)^2, \quad \hat{\sigma}^2_\zeta = \frac{1}{h}\sum_{i\in\zeta}(y_i - x_i'\hat{\beta}_\zeta)^2.$$

LTS estimator is the triplet $\hat{\zeta} = \arg\min_{\zeta:|\zeta|=h} \hat{\sigma}^2_\zeta, \hat{\beta} = \hat{\beta}_{\hat{\zeta}}, \hat{\sigma}^2 = \hat{\sigma}^2_{\hat{\zeta}}$. The estimator may not be unique, so we let $\mathcal{M}_n$ denote the set of solutions $\mathcal{M}_h = \arg\min_{\zeta:|\zeta|=h} \hat{\sigma}^2_\zeta$.

2.2 Role of an initial h

LTS estimators with an initial h are used to provide slope and scale estimates for various robust methods. We discuss methods that estimate the number of outliers using an initial h. For clarity, let $\underline{h}$ denote an initial choice and write $h_\circ$ for the number of 'good' observations in a data generating process.

The index plot method (Rousseeuw and Leroy, 1987) begins by computing an LTS estimator with an initial choice $h = \underline{h}$, which gives a slope $\hat{\beta}_{(0)}$ and scale $\hat{\sigma}_{(0)}$. A potential estimator for $h_\circ$ is then $\hat{h} = \sum_{i=1}^n \mathbb{I}\{|y_i - x_i'\hat{\beta}_{(0)}|/\hat{\sigma}_{(0)} \leq c\}$, which counts the number of scaled residuals below a cut-off $c > 0$ chosen by the user. The cut-off is typically calibrated to a case where all errors are standard normal. Reweighted LTS estimators

(Cížek, 2013; Alfons et al., 2013) are extensions of the index plot method, and likewise depend on an initial h.

For the index plot, $\underline{h}$ is typically chosen to be slightly above $n/2$. The standard choice following Rousseeuw and Leroy (1987, p.134) is $\underline{h} = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$, where p is the number of covariates. This maximises the finite sample breakdown point of LTS if all regressors are continuous, but has so far lacked robustness guarantees when some regressors are categorical. We will show that the choice $\underline{h} \approx n/2$ is justified if the regressors of interest are continuous.

The forward search algorithm (Hadi and Simonoff, 1993; Atkinson et al., 2010) likewise starts with an initial choice $h = \underline{h}$. The number h is then iteratively increased until a stopping rule is satisfied. The stopped value is an estimate of $h_\circ$. Atkinson and Riani (2000) suggest stopping rules based on recursively computed residuals and t-statistics.

For forward search, it is common to choose $\underline{h}$ to be much smaller than $n/2$, as the algorithm requires a 'burn-in' for evaluating the stopping rule. Choosing $\underline{h}$ smaller than $n/2$ has so far lacked a formal theory, even though it has been found to work well for various types of data (Atkinson and Riani, 2000). We will show that the choice $\underline{h} < n/2$ has robustness guarantees as long as the number of outliers is not too large.

Recent studies have also suggested estimating $h_\circ$ by minimising an objective function over a range of h starting from an initial value $\underline{h}$. The proposed objective functions include a normality test statistic (Berenguer-Rico et al., 2023) and a stability index based on bootstrapped LTS estimators (Heng and Lange, 2025).

3 LTS boundedness

We develop conditions for boundedness of LTS estimators with three goals in mind: first, to allow for a wider range of categorical regressors; second, to choose an initial h; and third, to obtain boundedness for sub-coefficients.

3.1 A sequence of data generating processes

Consider a sequence of data generating processes indexed by n. For each n, we have random variables y_i and p-vectors x_i for $i = 1, ..., n$. Let x_{in} be a normalised version of x_i, allowing indicator variables $x_{in} = \mathbb{I}\{i \geq \tau n\}$ for $\tau \in (0, 1)$ and trending regressors $x_{in} = i/n$. For each n, the equation of interest is $y_i = x_{in}'\beta + \sigma\varepsilon_i$.

For each n, let $h_\circ$ be the number of good observations and $\zeta_\circ$ an $h_\circ$-subset of $1, \ldots, n$ giving the indices of the good observations. Thus, $h_\circ$, $\zeta_\circ$ are deterministic sequences in n, but this is suppressed in the notation.

For boundedness results, we need only little structure on the 'good' observations. In analogy with least squares estimation, we will require that good errors have bounded sample second moments.

Assumption 1. *Suppose the good errors satisfy* $h_\circ^{-1} \sum_{i \in \zeta_\circ} \varepsilon_i^2 = O_p(1)$.

For the first few results in this section, we place no structure on the outlier errors. Later, we will use properties of the outlier errors to improve the boundedness results.

Write '$|\cdot|$' for the Euclidean norm on vectors and cardinality when applied to sets. Define $\lim_{(a,n)\to(0,\infty)} s_{a,n} = s$ to mean $\forall \epsilon > 0$, $\exists a_0, n_0$, $\forall a \leq a_0, n \geq n_0$: $|s_{a,n} - s| < \epsilon$.

3.2 A first boundedness result

Boundedness of the LTS estimator in regression has recently been explored by Berenguer-Rico and Nielsen (2025b). Their result requires that regressors are not too concentrated. We start by presenting a slightly modified, but equivalent result.

The full sample OLS estimator is unique when $\sum_{i=1}^{n} \delta' x_{in} x_{in}' \delta > 0$ for any $|\delta| = 1$. This is equivalent to requiring that $n^{-1} \sum_{i=1}^{n} \mathbb{I}\{|x_{in}'\delta| = 0\} < 1$. We generalize the latter to get boundedness results for LTS. To this end, define a function counting the share of regressors in 'strips' of width $a \geq 0$ around a hyperplane through

$$F_n(a) = \max_{\delta:|\delta|=1} n^{-1} \sum_{i\in\zeta} \mathbb{I}\{|x_{in}'\delta| \leq a\}. \tag{1}$$

The function F_n is a discrete distribution function. In particular, the maximum in its definition is attained. Johansen and Nielsen (2019) use F_n to show boundedness of M-estimators with non-convex objective functions, with related results in Chen and Wu (1988). They also relate it to a boundedness condition for S-estimators; see also Davies (1990) and Lopuhaä et al. (2023). We state the first boundedness result.

Assumption 2. *Suppose $h_\circ = \lfloor \lambda_\circ n \rfloor$ where $1/2 < \lambda_\circ \leq 1$. There exists $0 < \xi < 2\lambda_\circ - 1$ such that*

$$\lim_{(a,n)\to(0,\infty)} P\left(F_n(a) > \xi\right) = 0. \tag{2}$$

Theorem 1. *Suppose Assumptions 1, 2 and that $h = h_\circ$. Then the set of LTS estimators retaining h observations satisfies $\max_{\zeta\in\mathcal{M}_h} |\hat{\beta}_\zeta| = O_p(1)$.*

Examples of regressors satisfying Assumption 2 are discussed in Johansen and Nielsen (2019) and Berenguer-Rico and Nielsen (2025b). We mention a couple of examples.

Example 3.1 (Cointegration)**.** *(Berenguer-Rico and Nielsen, 2025a, Section B.3) If the regressors x_{in} include some continuous I(0) components and some continuous I(1) components satisfying certain regularity conditions, then $F_n(a)$ vanishes for small a and large n. Hence Assumption 2 holds for any $\lambda_\circ > 2/3$.*

Example 3.2 (One binary regressor)**.** *(Johansen and Nielsen, 2019, Example 3.1) Suppose $x_{in}' = (1, z_{in})$ with z_{in} binary, but not necessarily i.i.d. Let $m = |\{i \leq n : z_{in} = 1\}|$. Then $F_n(a) = n^{-1} \max(m, n-m)$ for small $a > 0$. Assumption 2 now constrains the range of possible values of $\lambda_\circ$. In the best case $m = n/2$ and $F_n(a) = 1/2$, we must choose $\xi > 1/2$, and thus the condition $2\lambda_\circ - 1 > \xi$ requires $\lambda_\circ > 3/4$.*

Going forward, we focus on models with some discrete regressors, where conditions for boundedness are most restrictive. We mainly consider cross sectional regressions. We will present variations of Theorem 1. *First*, we allow for an unknown h and give uniform boundedness guarantees. *Second*, we show how conditions can be relaxed if the user is only interested in subcoefficients. *Third*, we show that the strip conditions can be simplified to hyperplane conditions in special cases assuming i.i.d. regressors or discrete regressors with a finite support. *Fourth*, we show how the conditions needed for oracle inference also improve boundedness results.

3.3 Uniformity and subcoefficients

We improve the boundedness result in two aspects. First, we allow h to be unknown and show uniformity over $\underline{h} \leq h \leq h_\circ$. Second, if only subcoefficients are of interest, the strip condition can be weakened accordingly.

Let $R'\beta$ be the parameter of interest, where R is a $p \times s$ selection matrix. Define, for a given h-subset $\zeta \subseteq \{1, \ldots, n\}$, the strip function

$$F^R_{n\zeta}(a) = \max_{\delta:|\delta|=1,\, |R'\delta|>0} h^{-1} \sum_{i\in\zeta} \mathbb{I}\{|x'_{in}\delta| \leq a|R'\delta|\}. \tag{3}$$

When there is no selection, that is $R = I$ and $\zeta = \{1, ..., n\}$, this function reduces to $F_n(a)$ in (1). Due to the inequality $F^R_{n\zeta}(a) \leq F^I_{n\zeta}(a)$, we will see that boundedness for a subcoefficient $R'\beta$ can often be obtained under weaker conditions than for the full parameter β. This insight is useful when some regressors are only included as controls, such as in our empirical illustration.

Assumption 3. *Let R be a $p \times s$ matrix with $R'R = I_s$ and $s \leq p$. Suppose the data are generated with $h_\circ = \lfloor \lambda_\circ n \rfloor$ 'good' observations where $1/2 < \lambda_\circ \leq 1$. Let $\underline{h} = \lfloor \underline{\lambda} n \rfloor$ be an initial choice with $1 - \lambda_\circ < \underline{\lambda} \leq \lambda_\circ$. There exists $0 < \xi < \underline{\lambda} + \lambda_\circ - 1$ such that*

$$\lim_{(a,n)\to(0,\infty)} P\left(\lambda_\circ F^R_{n\zeta_\circ}(a) > \xi\right) = 0. \tag{4}$$

Theorem 2. *Suppose Assumption 1, 3. Then* $\max_{\underline{h}\leq h\leq h_\circ} \max_{\zeta\in\mathcal{M}_h} |\hat{\beta}_\zeta| = O_p(1)$.

Theorem 2 implies Theorem 1 since Assumption 3 is weaker than Assumption 2. To see this, let $\underline{\lambda} = \lambda_\circ$ and $R = I$. Then the inequality $h_\circ F^I_{n\zeta_\circ}(a) \leq nF_n(a)$ implies $\lambda_\circ F^I_{n\zeta_\circ}(a) \leq 1/n + F_n(a)$. Thus, if $F_n(a) \leq \xi$ for some $\xi < 2\lambda_\circ - 1$ then $\lambda_\circ F^I_{n\zeta_\circ}(a) \leq \xi^\dagger$ for some $\xi < \xi^\dagger < 2\lambda_\circ - 1$ and large n.

Assumptions 2 and 3 are identical when $h_\circ F^R_{n\zeta_\circ}(a) = nF^R_n(a)$. This occurs when a strip contains many good observations and no outliers. Guarding against such 'malicious' contamination can be overly pessimistic in applications (Huber and Ronchetti, 2009, section 17.4). Instead, it can be reasonable to assume outliers are more evenly spread out, as illustrated in the next example. In Section 4, we propose an algorithm for choosing an initial h that allows a user to exploit such assumptions.

Example 3.3 (One binary regressor as in Example 3.2)**.** *Suppose $x'_{in} = (1, z_{in})$ with z_{in} binary while $\underline{\lambda} = \lambda_\circ$ and $R = I_2$. Suppose $|\{i \leq n : z_{in} = 1\}| = n/2$ so that $F_n(a) = 1/2$.*

In the worst case, all outlying regressors are zero so that $m_\circ = |\{i \in \zeta_\circ : z_{in} = 1\}| = n/2$ and $F^I_{n\zeta_\circ}(a) = n/(2h_\circ) \approx 1/(2\lambda_\circ)$. We must choose $\xi > \lambda_\circ/(2\lambda_\circ) = 1/2$ and require $2\lambda_\circ - 1 > 1/2$, that is $\lambda_\circ > 3/4$ as in Example 3.2.

In the best case, outlying regressors are balanced so that $m_\circ = h_\circ/2$ and $F^I_{n\zeta_\circ}(a) = 1/2$. We then need $\xi > \lambda_\circ/2$ and require $2\lambda_\circ - 1 > \lambda_\circ/2$, that is $\lambda_\circ > 2/3$. Thus, we can tolerate more outliers if we have knowledge of the outlying regressors.

We will explore cross section regressions with continuous and discrete regressors. To facilitate, we will find convenient sufficient conditions for Assumption 3.

3.4 Imposing structure on the good regressors

We will now require convergence of the empirical measure for good regressors. This is satisfied for i.i.d. good regressors. The strip condition in Assumption 3 involved a limit $a \to 0$. This now reduces to an evaluation at $a = 0$ so that the strip condition turns into a simpler hyperplane condition, which is easier to handle in practice.

Assumption 4. *Suppose there is a probability measure P_{good} such that*

$$h_\circ^{-1} \sum\nolimits_{i \in \zeta_\circ} \mathbb{I}\{x_{in} \in A\} \to P_{good}(A) \tag{5}$$

almost surely for every measurable set A. Let $p_{good,\delta} = P_{good}\{x : x'\delta = 0\}$.

Proposition 3.1. *Suppose Assumption 4. Then on a set with probability one it holds*

$$\lim_{(a,n)\to(0,\infty)} F^R_{n\zeta_\circ}(a) = \lim_{n\to\infty} F^R_{n\zeta_\circ}(0) = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} p_{good,\delta},$$

Assumption 4 holds for i.i.d. or ergodic regressors. We give some examples below. The quantity $p_{good,\delta}$ is the probability of a hyperplane. Due to Proposition 3.1, Assumption 3 can be evaluated using the inequality $\lambda_\circ F^R_{n\zeta_\circ}(0) \le F^R_n(0) + 1/n$, where $F^R_n(0)$ can be computed using algorithms in Appendix B.

Assumption 5. *Suppose* $\sup_{\delta:|\delta|=1,|R'\delta|>0} \lambda_\circ p_{good,\delta} < \underline{\lambda} + \lambda_\circ - 1$.

Proposition 3.2. *Assumptions 4, 5 imply Assumption 3. In particular, if Assumptions 1, 4, 5 hold then* $\max_{\underline{h}\le h\le h_\circ} \max_{\zeta\in\mathcal{M}_h} |\hat\beta_\zeta| = O_p(1)$.

Proposition 3.2 connects our boundedness results with existing breakdown point theory where largest hyperplanes play a central role (Davies, 1993). For example, the finite sample breakdown point of LTS is maximised by choosing $h = \lfloor (n+nF^I_n(0)+1)/2 \rfloor$, which is also the largest possible breakdown point of a regression equivariant estimator.

Example 3.4 (Continuous regressors)**.** *Consider continuous, i.i.d. regressors x_{in} for $1 \le i \le n$. Then $p_{good,\delta} = 0$, $\forall \delta \ne 0$. By Proposition 3.2, $F_n(a)$ vanishes and Assumption 2 holds if $\underline{\lambda} + \lambda_\circ > 1$. An initial choice $\underline{\lambda} < 1/2$ is then allowed if $\lambda_\circ > 1/2$. This justifies existing practices in the Forward Search (Atkinson and Riani, 2000).*

Example 3.5 (Mixed regressor)**.** *If a regressor mixes continuous and discrete features, the strip function does not vanish. Consider i.i.d. regressors $x_{in} = z_i w_i$ with independent binary z_i and continuous w_i. Then $\sup_\delta p_{good,\delta} = P(z_i = 0)$, which is generally non-zero.*

Example 3.6 (Stationary and ergodic regressors)**.** *Let $y_i = x_i'\beta + \varepsilon_i$ where $(x_i)_{i=1}^\infty$ is stationary and ergodic. Then $\mathbb{I}\{x_i \in A\}$ is stationary, ergodic, and integrable (Breiman, 1992, Proposition 6.31). By the ergodic theorem (Breiman, 1992, Theorem 6.28),* (5) *holds with P_{good} equal to the distribution of x_1.*

Our final example illustrates that if the regressors of interest are continuous then up to 50% contamination can be allowed even if there are discrete controls. This justifies the use of the conventional initial value $\underline{h} \approx h/2$ in models with categorical regressors. A related insight applies if the regressor of interest is discrete but not very concentrated. These findings will be used in our empirical application.

Example 3.7 (Subcoefficients). *Consider i.i.d. regressors* $x_{in}' = (1, z_{1i}, z_{2i}, w_i)$ *and* $\beta' = (\beta_0, \beta_1, \beta_2, \beta_w)$. *Let* w_i *be continuous and* z_{i1}, z_{i2} *be independent, binary with* $P(z_{i1} = 1) = 0.5$ *and* $P(z_{i2} = 1) = 0.9$. *Let* $\underline{\lambda} = \lambda_\circ$ *for simplicity. We apply Proposition 3.2.*

If β_w *is of interest then* $R_w' = (0, 0, 0, 1)$. *By continuity,* $p_{good,\delta} = 0$ $\forall \delta$*:* $R_w'\delta \neq 0$. *Thus, Assumption 3 requires* $\lambda_\circ \times 0 < 2\lambda_\circ - 1$, *that is* $\lambda_\circ > 0.5$ *as in Example 3.1.*

Note that if $\delta^\dagger = (1, 0, -1, 0)/\sqrt{2}$ *then* $p_{good,\delta^\dagger} = P(z_{i2} = 1) = 0.9$ *and if* $\delta^\ddagger = (0, 1, 0, 0)$ *then* $p_{good,\delta^\ddagger} = P(z_{i1} = 0) = 0.5$.

If β_1 *is of interest then* $R_1' = (0, 1, 0, 0)$. *Then* $\delta^\dagger$ *is not relevant as* $R_1'\delta^\dagger = 0$ *and we get* $p_{good,\delta} \leq p_{good,\delta^\ddagger} = 0.5$ $\forall \delta$*:* $R_1'\delta \neq 0$. *We require* $\lambda_\circ \times 0.5 < 2\lambda_\circ - 1$, *that is* $\lambda_\circ > 2/3$ *as in Example 3.3.*

If β_2 *is of interest, then* $R_2' = (0, 0, 1, 0)$. *If* β_0 *is of interest, then* $R_0' = (1, 0, 0, 0)$. *If* β_0, β_1 *are of interest, then* $R_{01}' = (I_2, 0_{2\times 2})$. *In all cases* $R'\delta^\dagger \neq 0$. *Thus,* $p_{good,\delta} \leq p_{good,\delta^\dagger} = 0.9$ $\forall \delta$*:* $R'\delta \neq 0$ *and we require* $\lambda_\circ \times 0.9 < 2\lambda_\circ - 1$, *that is* $\lambda_\circ > 10/11 \approx 0.91$.

3.5 Categorical regressors with a finite support

For the second result, we consider regressors that have a fixed and finite support. No further restrictions are then needed on the distribution of regressors.

Proposition 3.3. *Suppose* x_{in} *has a finite support not depending on* n. *There exists* $a^* > 0$ *such that* $F^R_{n\zeta}(a) = F^R_{n\zeta}(0)$ *for all* $\zeta \subseteq \{1, ..., n\}$ *and* $0 \leq a < a^*$.

Example 3.8 (Two binary regressors). *Consider binary regressors* $x_{in}' = (1, z_{i1}, z_{i2})$. *By Proposition 3.3, for all small* $a > 0$ *it holds* $F^I_{n\zeta_\circ}(a) = h_\circ^{-1} \max_{S\in\mathcal{P}_2} \sum_{i\in\zeta_\circ} \mathbb{I}\{(z_{1i}, z_{2i}) \in S\}$, *where* $\mathcal{P}_2$ *are the 2-element subsets of* $\{0, 1\}^2$. *Thus, Assumption 3 depends on the probabilities of the sets* $\{z_{i1} = 1\}$, $\{z_{i1} = 0\}$, $\{z_{i2} = 1\}$, $\{z_{i2} = 0\}$, $\{z_{i1} = z_{i2}\}$ *and* $\{z_{i1} + z_{i2} = 1\}$.

3.6 Exploiting conditions for oracle inference

Usually, we are interested in statistical inference as well as boundedness. Inference requires more structure on the data generating processes, which we will now exploit.

Berenguer-Rico et al. (2023) show LTS is maximum likelihood in a model where good and outlier errors are separated. Further, LTS has the oracle property that asymptotic inference is the same as for OLS when the errors are weakly separated (Berenguer-Rico and Nielsen, 2025b). These assumptions contrast with the (Huber, 1964) model where errors are i.i.d. draws from a mixture distribution and the limiting distribution of LTS depends on the unknown contamination.

We now let the good and outlying regressors be i.i.d. within each group. This special case of Assumption 4 is the simplest setting that can generate 'bad leverage points', where errors depend on regressors.

Assumption 6. *Let* $h_\circ = \lfloor \lambda_\circ n \rfloor$ *with* $1/2 < \lambda_\circ \leq 1$. *Let* $x_{in} = x_i$ *be independent for* $1 \leq i \leq n$ *such that* $x_i \sim P_{good}$ *for* $i \in \zeta_\circ$ *and* $x_j \sim P_{out}$ $j \notin \zeta_\circ$.

For oracle inference, the outlier errors must diverge at a certain rate (Berenguer-Rico and Nielsen, 2025b). Here, it suffices that they diverge at any rate. We also rule out exact fit hyperplanes, a condition that is not binding in most applications.

Assumption 7. *Suppose*
(*i*) $1/\min_{j\in\zeta_\circ^c}|\varepsilon_j| = o_p(1)$.
(*ii*) $\lim_{n\to\infty} P(\min_{\zeta:|\zeta|=\underline{h}} \hat{\sigma}_\zeta^2 > 0) = 1$.

We relax Assumption 5 to a weaker condition on $p_{good,\delta} = P_{good}\{x : x'\delta = 0\}$. Define $p_{out,\delta} = P_{out}\{x : x'\delta = 0\}$, $P_{full} = \lambda_\circ P_{good} + (1-\lambda_\circ)P_{out}$, and $p_{full,\delta} = P_{full}\{x : x'\delta = 0\}$.

Assumption 8. *Suppose* $\sup_{\delta:|\delta|=1,|R'\delta|>0}\{\lambda_\circ p_{good,\delta} - (1-\lambda_\circ)p_{out,\delta}\} < \underline{\lambda} + \lambda_\circ - 1$.

Theorem 3. *Suppose Assumptions 1, 6, 7, 8. Then* $\max_{\underline{h}\le h\le h_\circ}\max_{\zeta\in\mathcal{M}_h}|\hat{\beta}_\zeta| = O_p(1)$.

Assumption 8 reduces to Assumption 5 when $p_{out,\delta} = 0\ \forall\delta$. Thus, to take advantage of the present condition, we require some additional information that outliers are not 'malicious', just as in Section 3.3. The following example illustrates this.

Example 3.9 (One binary regressor as in Examples 3.2, 3.3)**.** *Suppose* $x_{in}' = (1, z_{in})$, *with* z_{in} *binary and satisfying Assumptions 6, 7, 8 with* $P_{full}\{z_{in} = 1\} = 1/2$. *Let* $p_{out} = P_{out}\{z_i = 1\}$ *so that* $p_{good} = P_{good}\{z_i = 1\}$ *satisfies* $\lambda_\circ p_{good} = 1/2 - (1-\lambda_\circ)p_{out}$.

We have non-zero probability for the two hyperplanes defined by $\delta^\dagger = (1,-1)/\sqrt{2}$ *and* $\delta^\ddagger = (0,1)$. *We get* $p_{s,\delta^\dagger} = 1 - p_{s,\delta^\ddagger} = p_s$ *for* $s \in \{good, out\}$. *Thus, Assumption 8 is* $3/2 < \underline{\lambda} + \lambda_\circ + 2(1-\lambda_\circ)\min\{p_{out}, 1-p_{out}\}$.

In the worst case, if $p_{out} = 1$ *we require* $\underline{\lambda} + \lambda_\circ > 3/2$ *as in Examples 3.2, 3.3.*

In the best case, if $p_{out} = 1/2$ *we require* $\underline{\lambda} > 1/2$, *improving on Examples 3.2,3.3 by further exploiting knowledge of outlying regressors.*

4 Choosing an initial h

We propose algorithms for choosing an initial h for LTS. The algorithms allow a user to restrict the type of outliers and guard against a higher share of contamination through feasible versions of Assumptions 5, 8. The initial values from these algorithms have robustness guarantees based on our boundedness results.

4.1 Imposing restrictions on the regressors

We consider the setting of Assumption 6, which implies Assumption 4. We seek an initial $\underline{h} = \lfloor\underline{\lambda}n\rfloor$ guaranteeing uniform boundedness.

As before, write $p_{s,\delta} = P_s\{x : x'\delta = 0\}$ for $s \in \{good, out, full\}$, where $p_{full,\delta} = \lambda_\circ p_{good,\delta} + (1-\lambda_\circ)p_{out,\delta}$. We will provide a feasible version of Assumption 5, that is of

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0}\big[p_{full,\delta} - \big(1-\lambda_\circ\big)p_{out,\delta}\big] < \underline{\lambda} + \lambda_\circ - 1. \tag{6}$$

Although both $\lambda_\circ$ and $p_{out,\delta}$ are unknown, so that (6) cannot be directly verified, in some applications it is reasonable to restrict the range of values these quantities can take. Restrictions on $p_{out,\delta}$ can be used to rule out extremely concentrated outlying regressors, which are often unlikely in applications. Restricting $\lambda_\circ$ to be close to 'one' can be reasonable if outliers are suspected to be isolated observations instead of, say, a misspecified group structure in the data. These ideas motivate the following algorithm.

Algorithm 1.
(1) *Choose parameter of interest $R'\beta$, lower bounds $b_{out,\delta} \geq 0$ on $p_{out,\delta}$ $\forall \delta : |R'\delta| > 0$, a lower bound on the share of good observations $\underline{\lambda}_\circ \geq 1/2$, and an $\epsilon > 0$.*
(2) *Compute estimates $\hat{p}_{full,\delta} = n^{-1} \sum_{i=1}^n \mathbb{I}\{x_{in}'\delta = 0\}$.*
(3) *Compute*

$$\underline{\lambda}^{(1)} = \max_{\delta : |\delta|=1,\, |R'\delta|>0} \{\hat{p}_{full,\delta} + (1 - \underline{\lambda}_\circ)(1 - b_{out,\delta})\}. \tag{7}$$

(4) *The initial choice $\underline{h} = \lfloor \underline{\lambda} n \rfloor$ is $\underline{\lambda} = \max\{\underline{\lambda}^{(1)} + \epsilon, 1 - \underline{\lambda}_\circ\}$.*

Computing $\underline{\lambda}^{(1)}$ can be difficult. For example, for $R = I$ and binary regressors, the complexity is NP-hard (Amaldi and Kann, 1995, Corollary 2). An algorithm that approximates $\underline{\lambda}^{(1)}$ by drawing random subsets of $p - 1$ observations is described in Appendix B and implemented in the empirical illustration.

Although there are uncountably many orthogonal vectors δ, we foresee that only a small number of non-zero restrictions $b_{out,\delta}$ would typically be needed. As $\underline{\lambda}^{(1)}$ is decreasing in $b_{out,\delta}$, restrictions only influence the initial choice when $\hat{p}_{full,\delta}$ is large. We thus suggest first using an algorithm to detect large values of $\hat{p}_{full,\delta}$, and then considering non-zero restrictions $b_{out,\delta}$ only on these hyperplanes. This idea is demonstrated in the empirical application of Section 6.

Algorithm 1 coincides asymptotically with the breakdown point optimal choice $\underline{\lambda} \approx \max_{\delta:|\delta|=1}(1 + \hat{p}_{full,\delta} + n^{-1})/2$ from Mili and Coakley (1996) when $R = I$, $b_{out,\delta} = 0$ $\forall \delta$ and $\underline{\lambda}_\circ = \underline{\lambda}_\circ^{(1)}$.

We have the following robustness guarantee for Algorithm 1. As in our previous boundedness results, the conditions link the lower bound $\underline{h}$ in the search to the lower bound on the true outlier proportion $\underline{\lambda}_\circ$.

Theorem 4. *Consider initial choice $\underline{h}$ and $\epsilon > 0$ from Algorithm 1. Suppose Assumptions 1, 6 hold, $b_{out,\delta} \leq p_{out,\delta}$ $\forall \delta : |R'\delta| > 0$, and $\underline{\lambda}_\circ^{(1)} \leq \underline{\lambda}_\circ \leq \lambda_\circ - \epsilon$ where*

$$\underline{\lambda}_\circ^{(1)} = \sup_{\delta : |\delta|=1,\, |R'\delta|>0} \frac{1 + p_{full,\delta} - b_{out,\delta}}{2 - b_{out,\delta}} \tag{8}$$

Then $\max_{\underline{h} \leq h \leq h_\circ} \max_{\zeta \in \mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$.

The hyperplane restriction $b_{out,\delta} \leq p_{out,\delta}$ cannot, in general, be tested. However, it often has an application-specific interpretation that can be defended. Our empirical illustration provides an example.

The restriction $\underline{\lambda}_\circ \geq \underline{\lambda}_\circ^{(1)}$ can be evaluated by computing (8) with $p_{full,\delta}$ replaced by $\hat{p}_{full,\delta}$. The restriction $\lambda_\circ \leq \underline{\lambda}_\circ - \epsilon$ is again hard to test, but it is easy to interpret and communicate.

Theorem 4 highlights a trade-off between hyperplane restrictions $b_{out,\delta}$ and the rate of contamination allowed by the initial choice. The assumption $b_{out,\delta} \leq p_{out,\delta}$ always holds under the trivial restriction $b_{out,\delta} = 0$ for all δ, as in the breakdown point optimal initial choice. However, by choosing valid restrictions $b_{out,\delta} > 0$, the initial choice can permit a higher level of contamination, as $\underline{\lambda}_\circ^{(1)}$ decreases with $b_{out,\delta}$. This is illustrated with the following example.

Example 4.1 (One binary regressor as in Examples 3.2, 3.3, 3.9)**.** *Suppose* $x'_{in} = (1, z_i)$ *where* z_i *is binary with* $p_{full} = P_{full}(z_i = 1) = 1/2$ *and take* $R = I$*. Here, the choice of restrictions* $b_{out,\delta}$ *amounts to choosing an upper and lower bound on* $p_{out} = P_{out}(z_i = 1)$*.*

If $\delta^\dagger = (1, -1)/\sqrt{2}$ *and* $\delta^\ddagger = (0, 1)$ *then* $p_{full,\delta^\dagger} = p_{full}$ *and* $p_{full,\delta^\ddagger} = 1 - p_{full}$*.*

If $0 \leq p_{out} \leq 1$ *is unconstrained then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 0$ *so that* $\underline{\lambda}_\circ^{(1)} = 3/4$ *by* (8)*. This is the breakdown point optimal choice.*

If $1/3 \leq p_{out} \leq 2/3$ *is appropriate then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 1/3$ *so that* $\underline{\lambda}_\circ^{(1)} = 0.7$*.*

If $p_{out} = 1/2$ *is appropriate then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 1/2$ *so that* $\underline{\lambda}_\circ^{(1)} = 2/3$*.*

For algorithms such as the Forward Search (Atkinson and Riani, 2000), the initial choice should satisfy $\underline{h} < h_\circ$ so that a stopping rule can be evaluated relative to a 'burn-in'. If $\underline{\lambda}_\circ = \underline{\lambda}_\circ^{(1)}$ and $\hat{p}_{full,\delta} = p_{full,\delta}$ for all δ, then $\underline{\lambda} = \underline{\lambda}_\circ + \epsilon$. Since $\underline{\lambda}^{(1)}$ is decreasing in $\underline{\lambda}_\circ$, we thus need to choose $\underline{\lambda}_\circ$ above the minimum level $\underline{\lambda}_\circ^{(1)}$ in order to guarantee a burn-in.

4.2 Exploiting oracle conditions

We adapt Algorithm 1 and the boundedness result in Theorem 4 to the setting where outliers satisfy the conditions required for oracle inference (Section 3.6). By exploiting these conditions, we obtain initial choices $\underline{h}$ that can tolerate a higher share of contamination. We provide a feasible version of Assumption 8, that is of

$$\sup_{\delta:|\delta|=1,|R'\delta|>0} \{\lambda_\circ p_{good,\delta} - (1 - \lambda_\circ) p_{out,\delta}\} < \underline{\lambda} + \lambda_\circ - 1. \tag{9}$$

The adapted algorithm and the associated boundedness result have a structure similar to Algorithm 1 and Theorem 4, and the points raised in Section 4.1 continue to apply.

Algorithm 2. *Follow steps* 1,2 *and choice of* $\epsilon > 0$ *in Algorithm 1.*
(3) *Compute*

$$\underline{\lambda}^{(2)} = \max_{\delta:|\delta|=1,\,|R'\delta|>0} [\hat{p}_{full,\delta} + (1 - \underline{\lambda}_\circ)(1 - 2b_{out,\delta}) \mathbb{I}\{b_{out,\delta} < 1/2\}]. \tag{10}$$

(4) *The initial choice* $\underline{h} = \lfloor \underline{\lambda} n \rfloor$ *is* $\underline{\lambda} = \max\{\underline{\lambda}^{(2)} + \epsilon, 1 - \underline{\lambda}_\circ\}$.

Theorem 5. *Consider initial values* $\underline{h}$ *and* $\epsilon > 0$ *from Algorithm 2. Suppose Assumptions 1, 6, 7 hold,* $b_{out,\delta} \leq p_{out,\delta}$ $\forall \delta : |R'\delta| > 0$*, and* $\underline{\lambda}_\circ^{(2)} \leq \underline{\lambda}_\circ \leq \lambda_\circ - \epsilon$ *where*

$$\underline{\lambda}_\circ^{(2)} = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} \Big[\frac{1 + p_{full,\delta} - 2b_{out,\delta}}{2(1 - b_{out,\delta})} \mathbb{I}\{b_{out,\delta} < 1/2\} + p_{full,\delta} \mathbb{I}\{b_{out,\delta} \geq 1/2\} \Big] \tag{11}$$

Then $\max_{\underline{h} \leq h \leq h_\circ} \max_{\zeta \in \mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$*.*

If $b_{out,\delta} = 0$ $\forall \delta$ then $\underline{\lambda}_\circ^{(2)} = \underline{\lambda}_\circ^{(1)}$, and we again get the breakdown point optimal choice. If some information about the outlying regressors is available, we can improve over the robustness of Algorithm 1, as the following example shows.

Example 4.2 (One binary regressor as in Examples 3.2, 3.3, 3.9, 4.1)**.** *The setup and the values of* $p_{full,\delta}$ *and* $b_{out,\delta}$ *are as in Example 4.1. We have* $p_{full,\delta^\dagger} = p_{full,\delta^\ddagger} = 1/2$*.*

If $0 \leq p_{out} \leq 1$ *then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 0$ *so that* $\underline{\lambda}_\circ^{(1)} = \underline{\lambda}_\circ^{(2)} = 3/4$ *by* (8)*,* (11)*.*

If $1/3 \leq p_{out} \leq 2/3$ *then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 1/3$ *so that* $\underline{\lambda}_\circ^{(1)} = 0.7 > 5/8 = \underline{\lambda}_\circ^{(2)}$*.*

If $p_{out} = 1/2$ *then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 1/2$ *so that* $\underline{\lambda}_\circ^{(1)} = 2/3 > 1/2 = \underline{\lambda}_\circ^{(2)}$*.*

5 Simulations

We use a Monte Carlo simulation with one binary and one continuous regressor to evaluate our boundedness results against the properties of LTS in finite samples.

5.1 Data generating processes

We simulate data from six data-generating processes (DGPs). The equation of interest is $y_i = \beta_0 + \beta_1 z_i + \beta_2 w_i + \varepsilon_i$, where z_i is binary and w_i has a continuous distribution. We compute LTS and least squares estimators for $\beta = (\beta_0, \beta_1, \beta_2)'$ in 10000 generated samples. Sample sizes are $n \in \{25, 100, 400, 1600, 6400\}$, and we compute LTS estimators with $h = \lfloor \lambda n \rfloor$ for $\lambda \in \{0.6, 0.7, 0.75, 0.8, 0.85\}$. Data are generated with $\beta = (0, 0, 0)'$.

Computations are implemented in R (4.3.2) using the *robustbase* package (0.99-2). We made small adjustments to the FAST-LTS algorithm in *robustbase* to improve numerical stability. These adjustments relate to a known limitation of FAST-LTS in the presence of categorical covariates (Koller and Stahel, 2017).

Inspired by the theoretical analysis, each sample consists of 'good' observations $\zeta_\circ = \{1, ..., h_\circ\}$ and 'outliers' $\zeta_\circ^c = \{h_\circ + 1, ..., n\}$. The share of 'good' observations is fixed by having $h_\circ = \lfloor 0.8n \rfloor$. We describe the data generating processes in three parts.

1. Binary regressor z_i. Distribution of z_i differs across good observations and outliers. We let $z_i = 1$ for the first $\lfloor h_\circ p_{good} \rfloor$ good observations and first $\lfloor (n - h_\circ) p_{out} \rfloor$ outliers. The share of observations with $z_i = 1$ in the full sample is thus approximately $p_{full} = 0.8 p_{good} + 0.2 p_{out}$. Table 1 shows p_{good}, p_{out}, and p_{full} across the six DGPs.

In DGPs 1-3, we set $p_{full} = 0.7$ so that the full sample distribution of z_i is unbalanced. In DGP1, we set $p_{out} = p_{good} = 0.7$ so that good and outlying regressors have the same distribution. In DGP2, we set $p_{out} = 1$ so that outlying regressors are 'benign' and located where the majority of 'good' observations are. In DGP3, we set $p_{out} = 0$, so that outliers are 'malicious' and located where good observations are scarce.

In DGPs 4-6, we set $p_{full} = 0.52$, so that the full sample distribution of z_i is relatively balanced. In DGP4, we set $p_{out} = 0$ so that the outlying regressors are again 'malicious'. In DGPs 5 and 6, we set $p_{out} = p_{good} = 0.52$.

2. Continuous regressor w_i and errors ε_i. In DGPs 1-5, we draw w_i for $i = 1, ..., n$ independently from a uniform distribution on $[-10, 10]$. The good errors are generated as $\varepsilon_i \sim N(0, 1)$ while the outlying errors are generated as $\varepsilon_j \sim U[0, 0.5] + \sqrt{2 \log n}$. Outlying errors are thus growing, so that Assumption $7(i)$ holds. The rate of growth matches a consistency condition from Berenguer-Rico and Nielsen (2025b).

In DGP6, we introduce bad leverage by drawing $w_j \sim U[-10, 10] + n$ if $z_j = 0$ and $w_j \sim U[-10, 10]$ if $z_j = 1$ for $j \in \zeta_\circ^c$. This violates Assumption 6, where outlying regressors have a distribution that does not depend on n. To make the unboundedness of LTS in DGP6 apparent, we also adjust the distribution of good errors and draw $\varepsilon_i \sim N(0, 3)$ if $z_i = 0$ and $\varepsilon_i \sim N(0, 1/5)$ if $z_j = 1$ for $i \in \zeta_\circ$.

3. Lower bounds on λ. Table 1 shows $\underline{\lambda}^{(1)}$ and $\underline{\lambda}^{(2)}$, calculated from (7) and (11) with $\hat{p}_{full,\delta} = p_{full,\delta}$, $b_{out,\delta} = p_{out,\delta}$, and $\underline{\lambda}_\circ = 0.8$. DGPs 1-5 satisfy Assumptions 1, 6 so LTS estimators with $\lfloor n\underline{\lambda}^{(2)} \rfloor \leq h \leq \lfloor 0.8n \rfloor$ should be bounded by Theorem 5. In DGP6, Assumption 6 is violated so Theorem 5 does not apply, but by Proposition 3.2

Table 1: Distribution of z_i and initial retention shares in the DGPs

	Distribution of z_i			Lower bounds		
DGP	p_{good}	p_{out}	p_{full}	$\underline{\lambda}^{(1)}$	$\underline{\lambda}^{(2)}$	$\underline{\lambda}^{\text{bp}}$
1	0.70	0.70	0.70	0.76	0.70	0.85
2	0.62	1.00	0.70	0.70	0.70	0.85
3	0.88	0.00	0.70	0.90	0.90	0.85
4	0.65	0.00	0.52	0.72	0.72	0.76
5	0.52	0.52	0.52	0.62	0.52	0.76
6	0.52	0.52	0.52	0.62	0.52	0.76

LTS should be bounded for $\lfloor n\underline{\lambda}^{(1)} \rfloor \leq h \leq \lfloor 0.8n \rfloor$. For comparison, Table 1 also shows the breakdown point optimal value $\underline{\lambda}_{bp} = 1/2 + p_{full}/2$.

5.2 Simulation results

Table 2 shows mean squared errors for LTS and least squares estimators computed in the six DGPs described above.

In DGP1, mean squared errors converge to zero for $\lambda \in \{0.7, 0.75, 0.8\}$. This is in line with Theorem 5 and the lower bound $\underline{\lambda}^{(2)} = 0.7$ in Table 1. Errors also converge for $\lambda = 0.6$, indicating that our conditions are not necessary. The breakdown point optimal choice $\lambda_{bp} = 0.85$ suffers from a growing bias.

In DGP2, outliers are 'benign'. LTS estimators with $\lambda \in [0.6, 0.8]$ are again bounded. The bias of the breakdown point optimal choice $\lambda_{bp} = 0.85$ is reduced.

In DGP3, outliers are 'malign'. LTS estimators with all choices of λ are unbounded. This corresponds to lower bounds $\underline{\lambda}^{(1)} = \underline{\lambda}^{(2)} = 0.9$ being above the rate of contamination $\lambda_\circ = 0.8$, and there being no value of λ satisfying Assumptions 3 or 7.

In DGP4, outliers are again malicious, but the binary regressor is more balanced among 'good' observations. Estimators with $\lambda \in \{0.75, 0.8\}$ are bounded, in line with the lower bound $\underline{\lambda}_\circ^{(2)} = 0.72$. In contrast with DGP1 and DGP2, there is not much slack in our boundedness conditions, and estimates with $\lambda = 0.7$ seem unbounded.

In DGP5, estimates with $\lambda \in [0.6, 0.8]$ are bounded, in line with $\underline{\lambda}^{(2)} = 0.52$. In DGP6, Assumption 6 is violated, and the estimator with $\lambda = 0.6$ becomes unbounded. Meanwhile, the estimators with $\lambda \in [0.7, 0.8]$ are still bounded, in line with the lower bound $\underline{\lambda}^{(1)} = 0.62$ and Proposition 3.2, which does not require Assumption 6.

6 Empirical Illustration

We revisit the Bonjour et al. (2003) study of returns to education using a UK sample of identical twins. Within-twin differences in earnings are related to differences in education and other labour market characteristics through

$$w_{1f} - w_{2f} = \mu + \beta(school_{1f} - school_{2f}) + \alpha'(c_{1f} - c_{2f}) + (\varepsilon_{1f} - \varepsilon_{2f}), \qquad (12)$$

where 'if' indexes twin $i = 1, 2$ in family $f = 1, \ldots, 183$, w_{if} is log-earnings, and *school* is years of education. The vector c_{if} includes years of work experience (exp) and dummy

Table 2: Mean squared errors in the simulation

		λ						λ				
n	DGP	0.6	0.7	0.75	0.8	0.85	DGP	0.6	0.7	0.75	0.8	0.85
25	1	1.1	0.8	0.6	0.5	1.7	4	16.7	9.3	5.6	4.7	6.8
100		0.3	0.2	0.1	0.1	1.3		34.8	11.0	1.0	0.7	7.2
400		0.1	0.1	0.0	0.0	1.5		44.4	14.9	0.0	0.0	2.2
1600		0.0	0.0	0.0	0.0	2.0		51.8	15.2	0.0	0.0	1.0
6400		0.0	0.0	0.0	0.0	2.3		59.1	25.9	0.1	0.0	1.9
25	2	0.9	0.6	0.5	0.4	0.7	5	0.9	0.7	0.6	0.6	1.1
100		0.3	0.2	0.1	0.1	0.3		0.2	0.1	0.1	0.1	0.6
400		0.1	0.0	0.0	0.0	0.3		0.1	0.0	0.0	0.0	0.5
1600		0.0	0.0	0.0	0.0	0.3		0.0	0.0	0.0	0.0	0.6
6400		0.0	0.0	0.0	0.0	0.3		0.0	0.0	0.0	0.0	0.7
25	3	46.0	46.0	45.9	45.9	45.8	6	5.1	4.3	4.0	3.7	3.3
100		56.0	56.0	56.0	55.9	55.9		4.9	1.0	0.8	0.6	0.5
400		65.3	65.3	65.3	65.3	65.3		16.5	0.4	0.2	0.1	0.1
1600		74.2	74.2	74.2	74.2	74.2		28.7	0.1	0.1	0.0	0.0
6400		82.9	82.9	82.9	82.9	82.9		29.8	0.0	0.0	0.0	0.0

Notes. Columns correspond to LTS estimators with different choices of $h = \lfloor \lambda n \rfloor$*. Mean squared errors for OLS[n=6400]: DGP1=1.66, DGP2=3.38, DGP3=36.83, DGP4=10.26, DGP5=1.08, DGP6=2.37.*

variables for living in London or South-East England (*LonSE*), part-time employment (*part*), marital status (*marry*), and self-employment (*self*). The parameter of interest is β, which is interpreted as the wage return to an additional year of education.

A sensitivity analysis by Amin (2011) found that instrumental variable estimates of β are sensitive to the exclusion of a small number of observations in the data, whereas OLS estimates are more stable. We reassess robustness of OLS estimates using LTS.

6.1 Selecting an initial h for twins data

We use Algorithms 1 and 2 to choose an initial h for LTS.

Table 3: Five largest hyperplanes in twins data

	Orthogonal vector							Share of
	1	school	LonSE	marry	exp	part	self	obs. (%)
δ_1	0	0	0	0	0	0	1	94.5
δ_2	0	0	1	0	0	0	0	82.5
δ_3	0	0	1	0	0	0	-1	79.2
δ_4	0	0	0	0	0	1	0	71.6
δ_5	0	0	0	0	0	1	1	70.5

Notes: The rows refer to the i-th largest hyperplane among the regressors. Orthogonal vectors δ_i and shares of observations are indicated.

Table 4: Selection of initial h for twins data

	Hyperplane restrictions	Coefficients	$\underline{h}_\circ^{(1)}$	$\underline{h}_\circ^{(2)}$	$\underline{h}_\circ$	$\underline{h}^{(1)}$	$\underline{h}^{(2)}$
(1)	None	All coefficients	178	178	178	178	178
(2)	$p_{out,\delta_1} \geq 0.5$	All coefficients	177	173	178	175	173
(3)	None	Excluding α_{self}	167	167	168	166	166
(4)	$p_{out,\delta_2} \geq 0.3$, $p_{out,\delta_3} \geq 0.15$	Excluding α_{self}	165	161	168	161	157

Notes: Column 2 shows restrictions b_{out,δ_i} relating to the orthogonal vectors in Table 3. Column 3 shows coefficients of interest for which boundedness is required. Columns 4-8 show minimal choices of h under various assumptions.

Step 0: Hyperplane detection in the covariate space. We use a randomised algorithm (Appendix B) to find the five hyperplanes with most observations. Their orthogonal vectors and associated observation shares are shown in Table 3. The largest hyperplane contains 95% of the sample. Observation shares decline sharply thereafter, and the fifth-largest hyperplane has 71% of the observations.

Step 1: Choice of R and $b_{out,\delta}$. In step 1 of Algorithm 1, we choose hyperplane restrictions on outliers and the coefficient of interest. Columns 2 and 3 of Table 4 show four alternative specifications. Row 1 has no hyperplane restrictions and requires boundedness for all coefficients. Row 2 requires at least half of outliers to be on the hyperplane orthogonal to the vector δ_1 from Table 3. In row 3, boundedness is no longer required for the coefficient on self-employment. Row 4 adds restrictions on the hyperplanes orthogonal to δ_2 and δ_3.

We interpret the hyperplane restriction in row 2. The bound $p_{out,\delta_1} \geq 0.5$ in row 2 rules out all 'outliers' being twins where one is self-employed and the other is not. This restriction could be violated if, say, the data contained many high-earning entrepreneurs. We have gauged this restriction heuristically by plotting the outcome variable for twins with $self_f = 1$ against those with $self_f = 0$ (not shown), finding no evidence of 'outliers' being more prevalent among twins with $self_f = 1$.

Step 2: Choice of $\underline{\lambda}_\circ$. Algorithms 1 and 2 require a lower bound $\underline{h}_\circ = n\underline{\lambda}_\circ$ on the number of good observations. Columns 4, 5 of Table 4 show $\underline{h}_\circ^{(1)} = \lceil n\underline{\lambda}_\circ^{(1)} \rceil$ and $\underline{h}_\circ^{(2)} = \lceil n\underline{\lambda}_\circ^{(2)} \rceil$, computed from (8),(11) with $p_{full,\delta} = \hat{p}_{full,\delta}$. Theorems 4,5 suggest we can choose $\underline{h}_\circ$ as low as $\underline{h}_\circ^{(1)}$ or $\underline{h}_\circ^{(2)}$ depending on whether we impose the conditions for oracle inference. Choosing it slightly higher allows a search for h to be initialised with a burn-in. In rows 1,2 we choose $\underline{h}_\circ = 178$ as shown in column 6. Since $n = 183$, this allows up to 5 outliers as in Amin (2011). In rows 3,4, we choose $\underline{h}_\circ = 168$, allowing up to 15 outliers.

Step 3: Initial choice $\underline{h}$. The last two columns of Table 4 show initial choices $\underline{h}^{(1)} = \lfloor n\underline{\lambda}^{(1)} \rfloor$ and $\underline{h}^{(2)} = \lfloor n\underline{\lambda}^{(2)} \rfloor$, calculated from (7) and (10) with $\epsilon = 0$. In rows 2,4 the initial values $\underline{h}^{(i)}$ are below the lower bound $\underline{h}_\circ$, giving a burn-in that could be used for searching. This reflects the choice $\underline{h}_\circ \geq \underline{h}_\circ^{(i)}$ from step 2.

6.2 Results from LTS estimation of twins model

We assess the number of outliers using an index plot (Rousseeuw and Leroy, 1987). Figure 1 shows residuals from an LTS estimator with $h = 168$, scaled by the standard deviation of the selected 'good' residuals. This initial estimator is bounded under rows $3, 4$ from Table 4. Observations 8, 28, and 89 stand out as potential outliers, with four additional observations (39, 106, 111, 119) somewhat separated from the majority of the data. The index plot thus suggests at most seven outliers.

Table 5 shows LTS estimates of equation (12), trimming 0 to 15 observations. The bracketed terms display asymptotically standard normal t-statistics. The square brackets require large outlying errors and a correctly specified h (Berenguer-Rico and Nielsen, 2025b). The round brackets require that there are no outliers and all errors are independent normal (Cížek, 2004). For $n - h = 0$, LTS=OLS and standard theory applies.

Table 5 suggests that OLS estimates (for $n - h = 0$) of returns to education are not driven by 'outliers'. This conclusion holds under all four sets of restrictions in Table 4, as the LTS estimates remain qualitatively similar to the OLS estimate of 0.037 across all trimming rates. Both of the reported t-statistics reject the null of no returns at the 5% level when $n - h = 15$. Given that trimming 15 observations probably exceeds the true number of outliers, this finding should be interpreted with caution.

By contrast, the estimates for self-employment are highly sensitive to the choice of h. As trimming increases from one to four observations, point estimates jump from 0.05 to over 0.3. Since we cannot give theoretical robustness guarantees for this coefficient across all the reported trimming rates, we cannot rule out this jump being spurious.

The coefficient on work experience increases from the OLS estimate of 0.001 to about 0.01 when more than four observations are trimmed. At the same time, the t-statistics turn from insignificant to being significant at the 5% level. Under restrictions (3), (4) in Table 4, LTS is bounded across all displayed trimming rates. The results thus suggest that the small and insignificant least squares estimate on work experience may be driven by a few outliers, which LTS detects.

7 Conclusion

We have studied conditions for the boundedness of LTS estimators in the presence of outlier contamination, focusing on models with categorical covariates where these conditions are most binding.

We showed that LTS can tolerate more contamination than suggested by existing results on boundedness and breakdown point. Our approach shows how robustness properties of LTS depend on whether outliers are 'maliciously' concentrated or more evenly spread out. Our theory uses 'strip conditions', and we have established cases where these are equivalent to 'hyperplane conditions'. This equivalence connects results on boundedness to existing breakdown point theory and allows conditions for boundedness to be checked in practice.

Our boundedness results are uniform between an initial choice of h and the true share of 'good' observations. This is relevant when iterated LTS estimators, such as reweighted LTS (Cížek, 2013), are computed. Further, we found conditions for LTS to be bounded

Figure 1: Index plot of scaled residuals with LTS(h=168)

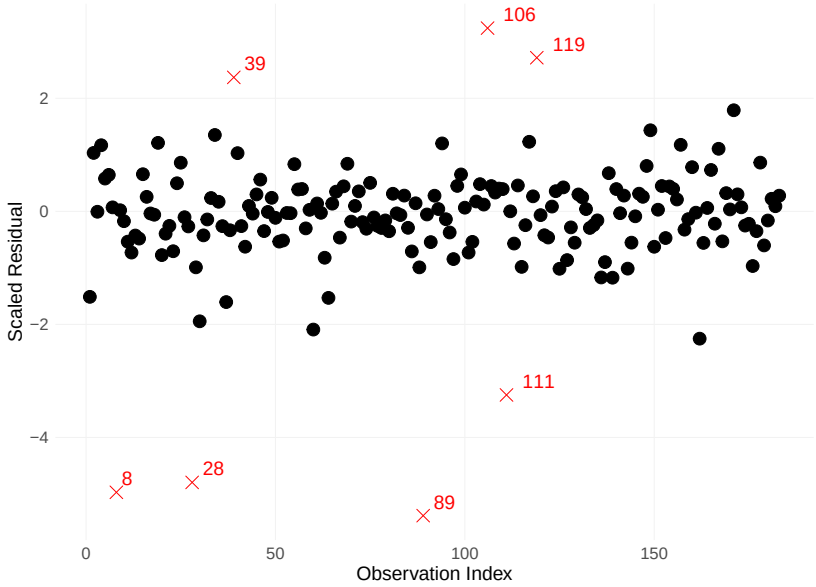

Table 5: Results from LTS estimation of twins model

$n-h$	15	10	8	5	4	1	0
school	0.037	0.031	0.030	0.033	0.031	0.031	0.037
	[2.71]	[2.12]	[1.89]	[1.89]	[1.70]	[1.35]	[1.48]
	(2.12)	(1.78)	(1.63)	(1.71)	(1.57)	(1.32)	(1.48)
LonSE	0.123	0.081	0.084	0.084	0.079	0.089	0.085
	[1.79]	[1.09]	[1.07]	[0.96]	[0.86]	[0.78]	[0.68]
	(1.40)	(0.92)	(0.92)	(0.87)	(0.79)	(0.76)	(0.68)
marry	0.054	0.024	0.023	0.019	0.051	-0.054	-0.068
	[0.99]	[0.41]	[0.39]	[0.28]	[0.72]	[-0.64]	[-0.72]
	(0.78)	(0.34)	(0.34)	(0.26)	(0.66)	(-0.62)	(-0.72)
exp	0.016	0.013	0.014	0.012	0.010	0.003	-0.001
	[4.65]	[3.64]	[3.51]	[2.84]	[2.17]	[0.58]	[-0.09]
	(3.64)	(3.06)	(3.03)	(2.57)	(2.00)	(0.56)	(-0.09)
part	-0.087	-0.105	-0.074	-0.074	-0.081	-0.093	-0.094
	[-1.55]	[-1.76]	[-1.18]	[-1.06]	[-1.09]	[-1.03]	[-0.94]
	(-1.22)	(-1.48)	(-1.02)	(-0.96)	(-1.00)	(-1.00)	(-0.94)
self	0.345	0.353	0.383	0.371	0.376	0.053	0.075
	[2.64]	[2.44]	[2.50]	[2.17]	[2.08]	[0.25]	[0.32]
	(2.06)	(2.05)	(2.16)	(1.96)	(1.92)	(0.24)	(0.32)

Notes: Columns show LTS estimates with different trimming numbers $n-h$. When $n-h=0$ then LTS=OLS. Squared brackets give t-statistics valid under the LTS model of Berenguer-Rico and Nielsen (2025b), and round brackets give t-statistics valid under uncontaminated normal errors (Víšek, 2006).

for a subcomponent of the full regression slope vector. Requiring boundedness only

for a parameter of interest allows LTS to be used with more concentrated categorical regressors. For example, if the regressor of interest is continuous, we can justify the conventional initial choice $h \approx n/2$ even in models with categorical regressors.

We used our boundedness results to propose an algorithm for selecting an initial h for LTS estimators. This algorithm could be used to initialise methods such as the index plot, forward search, and two-step LTS estimators. Our algorithms improve on a breakdown point optimal initial choice, which is covered as a special case, by allowing a user to exploit information about the nature of outlier contamination.

The ideas presented here could be used to study boundedness and tuning parameter selection for other robust regression methods. Methods closely related to LTS include least trimmed absolute deviations (Tableman, 1994) and trimmed likelihood estimators (Hadi and Luceño, 1997).

Appendix

A Proofs for Main Results

A.1 Proof of Theorem 1

Berenguer-Rico and Nielsen (2025b) derive a boundedness result using the strip function

$$F_{n,h}(a) = \max_{\zeta:|\zeta|=h} F^I_{n\zeta}(a) = \max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1} h^{-1} \sum_{i\in\zeta} \mathbb{I}\{|x'_{in}\delta| \leq a\}, \tag{13}$$

where $F^R_{n\zeta}$ is defined in (3). We show equivalence between $F_{n,h}$ and F_n from (1).

Lemma A.1. *Let $0 \leq \xi < h/n \leq 1$, $a \geq 0$. Then $\{F_n(a) > \xi\} = \{(h/n)F_{n,h}(a) > \xi\}$.*

Proof. Let $b_i = \mathbb{I}\{|x'_{in}\delta| \leq a\}$. Then $n^{-1}\sum_{i=1}^n b_i \geq (h/n)\max_{\zeta:|\zeta|=h} h^{-1}\sum_{i\in\zeta} b_i$ for any $0 < h \leq n$. Thus, comparing $F_n(a)$, $F_{n,h}(a)$ from (1), (13), we find $F_n(a) \geq (h/n)F_{n,h}(a)$. Therefore, for any $0 < h \leq n$ and $\xi \geq 0$ we have $\{(h/n)F_{n,h}(a) > \xi\} \subseteq \{F_n(a) > \xi\}$.

Suppose then $F_n(a) > \xi$ and $h/n > \xi \geq 0$. Since the maximum is attained, $\exists\delta$: $|\delta| = 1$ such that $nF_n(a) = |\{i : |x'_i\delta| \leq a\}|$. Consider two cases.

Case 1: $nF_n(a) \geq h$. If ζ^* is an h-subset of $\{i : |x'_{in}\delta| \leq a\}$ then

$$(h/n)F_{n,h}(a) \geq n^{-1}\sum_{i\in\zeta^*} \mathbb{I}\{|x'_{in}\delta| \leq a\} = h/n > \xi.$$

Case 2: $nF_n(a) < h$. Let ζ^* be the union of $\{i : |x'_{in}\delta| \leq a\}$ and any subset of $h - nF_n(a)$ elements from $\{i : |x'_{in}\delta| > a\}$. Then ζ^* is an h-subset of $\{1,\ldots,n\}$ and

$$(h/n)F_{n,h}(a) \geq n^{-1}\sum_{i\in\zeta^*} \mathbb{I}\{|x'_{in}\delta| \leq a\} = F_n(a) > \xi.$$

In both cases, $\{F_n(a) > \xi\} \subseteq \{(h/n)F_{n,h}(a) > \xi\}$. □

Proof of Theorem 1. The boundedness result in Berenguer-Rico and Nielsen (2025b) requires $P\{F_{n,h_\circ}(a) > \xi^\dagger\} \to 0$ as $(a,n) \to (0,\infty)$ for some $0 < \xi^\dagger < 2 - \lambda_\circ^{-1}$. This is equivalent to $P\{(h_\circ/n)F_{n,h_\circ}(a) > \xi\} \to 0$ for $\xi = \xi^\dagger h_\circ/n$. By Lemma A.1 this is equivalent to $P\{F_n(a) > \xi\} \to 0$. Since $h_\circ/n \to \lambda_\circ$, it is sufficient for the final condition to hold for some $0 < \xi < (2 - \lambda_\circ^{-1})\lambda_\circ = 2\lambda_\circ - 1$. □

A.2 Proof of Theorem 2

Lemma A.2. *Consider a set $\zeta \subseteq \{1,\dots,n\}$ with $|\zeta| = h$ and a $p \times s$ matrix R. Stack y_i and x_{in}' for $i \in \zeta$ to get an h-vector y_ζ and an $h \times p$ matrix X_ζ. Let $\hat\beta_\zeta$ be the set of least squares solutions for the equation $y_\zeta = X_\zeta\beta + \varepsilon_\zeta$. It holds*

$$\max_{\delta:|\delta|=1,\,|R'\delta|>0} h^{-1}\sum_{i\in\zeta}\mathbb{I}\{x_{in}'\delta = 0\} < 1 \implies R'\hat\beta_\zeta \text{ is unique.} \tag{14}$$

Proof. Least squares solutions satisfy the normal equations $(X_\zeta'X_\zeta)\beta = X_\zeta'y_\zeta$. Therefore, if $\hat\beta_\zeta^*, \hat\beta_\zeta^\dagger$ are solutions then $(X_\zeta'X_\zeta)(\hat\beta_\zeta^\dagger - \hat\beta_\zeta^*) = 0$. This means $\hat\beta_\zeta^\dagger = \hat\beta_\zeta^* + v$ for some v with $X_\zeta v = 0$. The assumption in (14) gives

$$\{\delta : X_\zeta\delta = 0\} = \{\delta : h^{-1}\sum_{i\in\zeta}\mathbb{I}\{x_{in}'\delta = 0\} = 1\} \subseteq \{\delta : R'\delta = 0\}.$$

Since $X_\zeta v = 0$ then also $R'v = 0$, giving $R'\hat\beta_\zeta^\dagger = R'(\hat\beta_\zeta^* + v) = R'\hat\beta_\zeta^*$. As this holds for any solutions $\hat\beta_\zeta^*$ and $\hat\beta_\zeta^\dagger$, then $R'\hat\beta_\zeta$ is unique. □

Define $\hat\delta_\zeta = (\hat\beta_\zeta - \beta)/|\hat\beta_\zeta - \beta|$ if $|\hat\beta_\zeta - \beta| > 0$ and otherwise take $\hat\delta_\zeta$ to be a vector with each element equal to $1/\sqrt{p}$. We allow $\hat\beta_\zeta$ and $\hat\delta_\zeta$ to be set valued.

Lemma A.3. *Let R be a $p \times s$ matrix such that $R'R = I$, $1/2 < \lambda_\circ \le 1$, and $0 < \underline\lambda \le \lambda_\circ$. Define $h_\circ = \lfloor\lambda_\circ n\rfloor$ and $\underline h = \lfloor\underline\lambda n\rfloor$. Suppose $h_\circ^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 = O_p(1)$. Then for any $a,\epsilon,\tau > 0$ there exists $B > 0$ and sets (Ω_n) with $P(\Omega_n) \ge 1 - \epsilon$ such that on Ω_n for*

$$Z_{\tau ah}^R = \left\{\zeta : |\zeta| = h, R'\hat\beta_\zeta \text{ is unique, and } \max_{\delta\in\hat\delta_\zeta} h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} \ge \tau\right\}$$

and all $\underline h \le h \le h_\circ$ it holds

$$\min_{\zeta\in Z_{\tau ah}^R:|R'\hat\beta_\zeta - R'\beta|/\sigma > B}\hat\sigma_\zeta^2 > \max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h}\hat\sigma_{\zeta_h}^2. \tag{15}$$

Proof. 1. Construction of Ω_n and B. Let $a,\epsilon,\tau > 0$ be given. Denote $\underline h = \lfloor\underline\lambda n\rfloor$. Since $h_\circ^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 = O_p(1)$ and $h_\circ/\underline h \to \lambda_\circ/\underline\lambda < \infty$, we have $\underline h^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 = O_p(1)$. Therefore, we can find a constant $A_0 > 0$ and sets Ω_n with $P(\Omega_n) \ge 1 - \epsilon$ such that on Ω_n it holds $\underline h^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 < A_0$. Then on Ω_n it holds for all $\underline h \le h \le h_\circ$ that

$$\max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h}\hat\sigma_{\zeta_h}^2/\sigma^2 \le \max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h} h^{-1}\sum_{i\in\zeta_h}\varepsilon_i^2 \le \underline h^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 < A_0. \tag{16}$$

We also have on Ω_n for any $\eta > 0$ and all $\underline{h} \leq h \leq h_\circ$ that

$$A_0 > h^{-1} \sum_{i \in \zeta_\circ} \varepsilon_i^2 \geq h^{-1} \sum_{i \in \zeta_\circ} \varepsilon_i^2 \mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} > (A_0/\eta) h^{-1} \sum_{i \in \zeta_\circ} \mathbb{I}\{\varepsilon_i^2 > A_0/\eta\},$$

implying $h^{-1} \sum_{i \in \zeta_\circ} \mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} < \eta$. Fix an arbitrary $0 < \eta < \tau$ and take

$$B \geq \frac{1}{a} \left[\left(\frac{A_0}{\eta} \right)^{1/2} + \left(\frac{A_0}{\tau - \eta} \right)^{1/2} \right]. \tag{17}$$

2. Showing inequality (15) *holds on* Ω_n. Consider any $\underline{h} \leq h \leq h_\circ$ and $\zeta \in Z^R_{\tau a h}$ with $|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B$. While $R'\hat{\beta}_\zeta$ is unique for any $\zeta \in Z^R_{\tau a h}$, the solution $\hat{\beta}_\zeta$ may not be. However, since $\zeta \in Z^R_{\tau a h}$ there exists $\hat{\delta}^*_\zeta \in \hat{\delta}_\zeta$ such that

$$h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|\} \geq \tau. \tag{18}$$

Since $|R'\hat{\beta}_\zeta - R'\beta|/\sigma > 0$ by assumption, it also holds $|\hat{\beta}^*_\zeta - \beta| > 0$. Therefore, $\hat{\delta}^*_\zeta = (\hat{\beta}^*_\zeta - \beta)/|\hat{\beta}^*_\zeta - \beta|$ for some $\hat{\beta}^*_\zeta \in \hat{\beta}_\zeta$. Denoting $\hat{\ell}_\zeta = |R'\hat{\beta}^*_\zeta - R'\beta|/\sigma$, then

$$\begin{aligned}(y_i - x_{in}'\hat{\beta}^*_\zeta)/\sigma &= \varepsilon_i - x_{in}'(\hat{\beta}^*_\zeta - \beta)/\sigma \\ &= \varepsilon_i - \frac{|R'\hat{\beta}^*_\zeta - R'\beta|}{\sigma} \Big\{ \frac{x_{in}'(\hat{\beta}^*_\zeta - \beta)}{|\hat{\beta}^*_\zeta - \beta|} \Big\} \frac{|\hat{\beta}^*_\zeta - \beta|}{|R'\hat{\beta}^*_\zeta - R'\beta|} = \varepsilon_i - \hat{\ell}_\zeta x_{in}'\hat{\delta}^*_\zeta/|R'\hat{\delta}^*_\zeta|.\end{aligned}$$

Use this to write

$$\begin{aligned}\hat{\sigma}^2_\zeta/\sigma^2 &= h^{-1} \sum_{i \in \zeta} (\varepsilon_i - \hat{\ell}_\zeta x_{ni}'\hat{\delta}^*_\zeta/|R'\hat{\delta}^*_\zeta|)^2 \\ &\geq h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{ni}'\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|\} \mathbb{I}\{\varepsilon_i^2 \leq A_0/\eta\} (\varepsilon_i - \hat{\ell}_\zeta x_{ni}'\hat{\delta}^*_\zeta/|R'\hat{\delta}^*_\zeta|)^2. \end{aligned} \tag{19}$$

When $|x_{in}'\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|$, $\hat{\ell}_\zeta > B$, and $\varepsilon_i^2 \leq A_0/\eta$ we have using (17) that

$$\hat{\ell}_\zeta |x_{ni}'\hat{\delta}^*_\zeta|/|R'\hat{\delta}^*_\zeta| > aB \geq \left(\frac{A_0}{\eta} \right)^{1/2} + \left(\frac{A_0}{\tau - \eta} \right)^{1/2} \geq |\varepsilon_i| + \left(\frac{A_0}{\tau - \eta} \right)^{1/2}.$$

Combining with the reverse triangle inequality gives

$$(\varepsilon_i - \hat{\ell}_\zeta x_{ni}'\hat{\delta}^*_\zeta/|R'\hat{\delta}^*_\zeta|)^2 \geq (\hat{\ell}_\zeta |x_{ni}'\hat{\delta}^*_\zeta|/|R'\hat{\delta}^*_\zeta| - |\varepsilon_i|)^2 > A_0/(\tau - \eta).$$

Plugging this in (19) gives the lower bound

$$\hat{\sigma}^2_\zeta/\sigma^2 > \left(\frac{A_0}{\tau - \eta} \right) h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{ni}'\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|\} \mathbb{I}\{\varepsilon_i^2 \leq A_0/\eta\}.$$

Next, use that for two sets S and T it holds $\mathbb{I}_{S\cap T} \geq \mathbb{I}_S - \mathbb{I}_{T^c}$ to see

$$\hat{\sigma}^2_\zeta/\sigma^2 > \left(\frac{A_0}{\tau-\eta}\right) h^{-1} \sum_{i\in\zeta\cap\zeta_\circ} \left(\mathbb{I}\{|x'_{ni}\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|\} - \mathbb{I}\{\varepsilon_i^2 > A_0/\eta\}\right).$$

Using (18) and that on Ω_n it holds $h^{-1}\sum_{i\in\zeta_\circ}\mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} < \eta$, we get

$$\hat{\sigma}^2_\zeta/\sigma^2 > \left(\frac{A_0}{\tau-\eta}\right)(\tau-\eta) = A_0.$$

Thus, combined with (16), on Ω_n it holds $\hat{\sigma}^2_\zeta/\sigma^2 > \max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h}\hat{\sigma}^2_{\zeta_h}/\sigma^2$, so (15) holds. □

Proof of Theorem 2. The proof has three parts. First, we describe a constant $B > 0$ and a sequence of sets Ω_n in the probability space. We then show that on Ω_n and n large it holds $\max_{\zeta\in\mathcal{M}_h}|R'\hat{\beta}_\zeta - R'\beta|/\sigma \leq B$ for all $\underline{h} \leq h \leq h_\circ$. Finally, we show that for any $\epsilon > 0$, Ω_n and B can be constructed so that $P(\Omega_n) \geq 1-\epsilon$ for all n large.

1. Construction of Ω_n and $B > 0$. Let Ω_n be a set such that

$$\lambda_\circ\left(\max_{\delta:|\delta|=1,\,|R'\delta|>0} h_\circ^{-1}\sum_{i\in\zeta_\circ}\mathbb{I}\{|x'_{in}\delta| \leq a|R'\delta|\}\right) \leq \xi. \tag{20}$$

for some $0 < \xi < \underline{\lambda} + \lambda_\circ - 1$ and $a > 0$. Define $2\tau = \lambda_\circ^{-1}(\underline{\lambda} + \lambda_\circ - 1 - \xi) > 0$ and

$$Z^R_{\tau a h} = \Big\{\zeta : (i)\,|\zeta| = h, (ii)\,R'\hat{\beta}_\zeta \text{ is unique, } (iii)\,\max_{\delta\in\hat{\delta}_\zeta} h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\delta| > a|R'\delta|\} \geq \tau\Big\}.$$

On Ω_n there is $B > 0$ such that for all $\underline{h} \leq h \leq h_\circ$

$$\min_{\zeta\in Z^R_{\tau a h}:|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B}\hat{\sigma}^2_\zeta > \max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h}\hat{\sigma}^2_{\zeta_h}, \tag{21}$$

2.1. Deterministic Analysis on Ω_n — A set inclusion. We show that on Ω_n it holds for all $\underline{h} \leq h \leq h_\circ$ and n large

$$\Xi = \{\zeta : |\zeta| = h,\, |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B\} \subseteq \{\zeta \in Z^R_{\tau a h} : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B\}. \tag{22}$$

To do this, we check that for all n large, $\underline{h} \leq h \leq h_\circ$, and $\zeta \in \Xi$ the three properties in the definition of $Z^R_{\tau a h}$ hold. Property (i) is immediate. We check properties (ii) and (iii).

We check property (ii). Use the identity

$$R_{\delta,h,\zeta} = h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\delta| > a|R'\delta|\} = h^{-1}|\zeta\cap\zeta_\circ| - h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\delta| \leq a|R'\delta|\}. \tag{23}$$

Bound the first term in (23) from below. Using $|\zeta| = h \geq \underline{h}$ and $|\zeta_\circ^c| = n - h_\circ$ we get

$$h^{-1}|\zeta\cap\zeta_\circ| \geq h^{-1}(|\zeta| - |\zeta_\circ^c|) \geq h^{-1}(\underline{h} + h_\circ - n).$$

For the second term in (23), inequality (20) gives for all $\delta \in \Delta = \{\delta : |\delta| = 1, |R'\delta| > 0\}, h, \zeta$

$$h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\delta| \leq a|R'\delta|\} \leq (h_\circ/h)(\xi/\lambda_\circ).$$

Combining, bound (23) for any $\delta \in \Delta$, $\underline{h} \leq h \leq h_\circ$, and $|\zeta| = h$ with

$$R_{\delta,h,\zeta} \geq h^{-1}(\underline{h} + h_\circ - n) - (h_\circ/h)(\xi/\lambda_\circ) = (n/h)\{\underline{\lambda} + \lambda_\circ - 1 - \xi + o(1)\},$$

where the final equality uses $h_\circ/n = \lambda_\circ + o(1)$ and $\underline{h}/n = \underline{\lambda} + o(1)$ from Assumption 3. Since $h \leq h_\circ$ and $h_\circ/n = \lambda_\circ + o(1)$, it holds $n/h \geq n/h_\circ = \lambda_\circ^{-1} + o(1)$. Thus, using our choice of τ, conclude for all large n, $\delta \in \Delta$, $\underline{h} \leq h \leq h_\circ$, and $|\zeta| = h$

$$\begin{aligned} R_{\delta,h,\zeta} \geq \lambda_\circ^{-1}\{\underline{\lambda} + \lambda_\circ - 1 - \xi + o(1)\} + o(1) =& \lambda_\circ^{-1}(\underline{\lambda} + \lambda_\circ - 1 - \xi) + o(1) \\ \geq& 2\tau + o(1) \geq \tau. \end{aligned} \tag{24}$$

Property (ii) then follows from Lemma A.2, since by (24) for all $\delta \in \Delta$

$$h^{-1} \sum_{i \in \zeta} \mathbb{I}\{x_{in}'\delta = 0\} \leq 1 - h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} = 1 - R_{\delta,h,\zeta} \leq 1 - \tau < 1. \tag{25}$$

We check property (iii). Since ζ satisfies $|\hat{\beta}_\zeta - \beta| \geq |R'\hat{\beta}_\zeta - R'\beta| > B\sigma > 0$ by assumption, we have $\hat{\delta}_\zeta = (\hat{\beta}_\zeta - \beta)/|\hat{\beta}_\zeta - \beta|$ by definition. It follows that all $\delta \in \hat{\delta}_\zeta$ have norm one and satisfy $|R'\delta| = |R'\hat{\beta}_\zeta - R'\beta|/|\hat{\beta}_\zeta - \beta| > 0$. Thus, using (24), conclude that (iii) holds since

$$\max_{\delta \in \hat{\delta}_\zeta} R_{\delta,h,\zeta} \geq \min_{\delta : |\delta| = 1, |R'\delta| > 0} R_{\delta,h,\zeta} \geq \tau. \tag{26}$$

2.2 Deterministic Analysis on Ω_n — Conclusion. Using first (22) and then (21), we have for large n and any $\underline{h} \leq h \leq h_\circ$

$$\min_{\zeta : |\zeta| = h,\, |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B} \hat{\sigma}_\zeta^2 \geq \min_{\zeta \in Z_{\tau a h}^R : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B} \hat{\sigma}_\zeta^2 > \max_{\zeta_h \subseteq \zeta_\circ : |\zeta_h| = h} \hat{\sigma}_{\zeta_h}^2. \tag{27}$$

Thus, any h-subset ζ with $|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B$ is not in $\mathcal{M}_h$, i.e. $\max_{\zeta \in \mathcal{M}_h}\{|R'\hat{\beta}_\zeta - R'\beta|/\sigma\} \leq B$.

3. Probability analysis. Construct $\Omega_n = \Omega_{1n} \cap \Omega_{2n}$ as follows. Using the strip condition in Assumption 3, we can find $0 < \xi < \underline{\lambda} + \lambda_\circ - 1$, $a > 0$, and Ω_{1n} such that (20) holds on Ω_{1n} and $P(\Omega_{1n}) \geq 1 - \epsilon$ for large n. Since $\sum_{i \in \zeta_\circ} \varepsilon_i^2 = O_p(1)$ by Assumption 1, then by Lemma A.3 there exist $B > 0$ and Ω_{2n} such that (21) holds for all $\underline{h} \leq h \leq h_\circ$ and $P(\Omega_{2n}) \geq 1 - \epsilon$ for large n. The conclusion in part 2 is then obtained on Ω_n for n large such that (24) holds. □

A.3 Proof of Proposition 3.2

The following definitions are from Roman (2005, Chapter 16). For a linear subspace S and a vector x, the set $x + S$ is called a flat (affine set). The affine hull of a non-empty

set A is the smallest flat containing A. For a set A with affine hull $x + S$, the affine dimension $\dim(A)$ of A is the dimension of S. For a set A, let $\text{span}(A)$ be the set of all finite linear combinations of vectors from A. The orthogonal complement of a linear subspace S is denoted $S^{\perp} = \{x : x'\delta = 0, \forall \delta \in S\}$.

Lemma A.4 (Rao, 1962, Lemma 7.3)**.** *Let P_{good} be a probability measure on $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$. There exist measures $\{\mu_k\}_{k=1}^{\infty}$ and distinct affine sets $\{A_k\}_{k=1}^{\infty}$ of affine dimensions $\{m_k\}_{k=1}^{\infty}$ such that (i) $P_{good} = \sum_{k=1}^{\infty} \mu_k$, (ii) $\mu_k(\mathbb{R}^p \setminus A_k) = 0$, and (iii) $\mu_k(B) = 0$ for any measurable set B with affine dimension less than m_k.*

Remark A.1. *For two linear subspaces S and S'. If S has dimension strictly greater than zero, it can be checked (e.g. using Theorem 1.14. of Roman (2005))*

$$S \not\subseteq S' \implies \dim(S \cap S') < \dim(S). \tag{28}$$

An analogous property holds for flats. Let $X = x + S$ and $X' = x' + S'$ be flats. If X has affine dimension strictly greater than zero then

$$X \not\subseteq X' \implies \dim(X \cap X') < \dim(X). \tag{29}$$

Property (29) *is immediate if $X \cap X' = \varnothing$. If $X \cap X' \neq \varnothing$, then $X \cap X' = x^* + S \cap S'$ for any $x^* \in X \cap X'$ (Roman (2005), Theorem 16.5). Since $X \not\subseteq X'$, it holds $S \not\subseteq S'$. By* (28)*, conclude that $\dim(X \cap X') = \dim(S \cap S') < \dim(S) = \dim(X)$.*

Lemma A.5. *Let P_{good} be a probability measure and $\{\mu_k, A_k, m_k\}_{k=1}^{\infty}$ given by Lemma A.4. For every $k \in \mathbb{N}$ and with the convention $\sup \varnothing = 0$*

$$\sup_{\delta \in \text{span}(A_k) : |\delta| = 1} \mu_k\{x : |x'\delta| \leq a\} \xrightarrow{a \to 0} 0.$$

Proof. Let $k \in \mathbb{N}$. Write $\mu = \mu_k$, $A = A_k$, $m = m_k$, and $\Delta = \text{span}(A) \cap \{\delta : |\delta| = 1\}$.

Consider first $m = 0$, in which case A is a singleton $\{z\} \subset \mathbb{R}^p$. If further $z = 0$ then $\Delta = \emptyset$, giving $\sup_{\delta \in \Delta} \mu\{x : |x'\delta| \leq a\} = 0$ for all $a > 0$. If $z \neq 0$ then $\Delta = \{z/|z|, -z/|z|\}$ and $\sup_{\delta \in \Delta} \mu\{x : |x'\delta| \leq a\} = \mu\{x : |x'z|/|z| \leq a\}$. Further, by Lemma A.4$(ii)$ it holds $\mu\{x : |x'z|/|z| \leq a\} = \mathbb{I}\{|z'z|/|z| \leq a\}\mu\{z\}$ and thus for all $0 < a < |z'z|/|z|$ we conclude $\sup_{\delta \in \Delta} \mu_k\{x : |x'\delta| \leq a\} = 0$.

Consider then $m \geq 1$. For a contradiction, suppose $k \in \mathbb{N}$ is such that the claim is not true. From the contradiction assumption, $\exists \epsilon > 0$, and sequences a_n with $a_n \downarrow 0$ and $\delta_n \in \Delta$ such that $\mu\{x : |x'\delta_n| \leq a_n\} \geq \epsilon$ for all n.

As $\{\delta : |\delta| = 1\}$ is compact while $\text{span}(A)$ is closed, then Δ is compact. Thus, δ_n has a converging subsequence $\delta_{n'} \to \delta^* \in \Delta$.

Since $\delta^* \in \text{span}(A)$, there exists an element $a \in A$ with $a'\delta^* \neq 0$. Therefore, $A \not\subseteq \{x : x'\delta^* = 0\}$. From Remark A.1, it follows that the set $\dim(A \cap \{x : x'\delta^* = 0\}) < \dim(A) = m$. By Lemma A.4$(ii, iii)$, we get $\mu\{x : x'\delta^* = 0\} = \mu(\{x : x'\delta^* = 0\} \cap A) = 0$.

Let $A^* = \limsup_{n \to \infty}\{x : |x'\delta_{n'}| \leq a_{n'}\} = \bigcap_{N \in \mathbb{N}} \bigcup_{n \geq N}\{x : |x'\delta_{n'}| \leq a_{n'}\}$. It holds $A^* \subseteq \{x : x'\delta^* = 0\}$. To see this, consider any x such that $|x'\delta^*| > 0$. With the reverse triangle and Cauchy-Schwarz inequalities we get $|x'\delta_{n'}| \geq |x'\delta^*| - |x||\delta_{n'} - \delta^*|$. Since $\delta_{n'} \to \delta^*$ and $a_{n'} \to 0$, we have $|x'\delta_{n'}| > |x'\delta^*|/2 > a_{n'}$ for large n. This shows $x \notin A^*$.

Use $\liminf_{n\to\infty}\mu(S_n) \le \mu(\limsup_{n\to\infty} S_n)$ (Billingsley, 1995, Theorem 4.1.) to get the contradiction

$$\begin{aligned} 0 < \epsilon &\le \liminf_{n\to\infty} \mu\left\{x : |x'\delta_{n'}| \le a_{n'}\right\} \\ &\le \mu\left(\limsup_{n\to\infty}\left\{x : |x'\delta_{n'}| \le a_{n'}\right\}\right) \le \mu\{x : x'\delta^* = 0\} = 0. \end{aligned}$$ □

For a matrix B with columns $B_1, \ldots, B_J$, write $\text{span}(B)$ for $\text{span}\{B_1, \ldots, B_J\}$. For a set S, write $S^\perp = \{\delta : x'\delta = 0,\ \forall x \in S\}$ while for a matrix B write $B^\perp = \{\delta : B'\delta = 0\}$. For a linear subspace $S \subseteq \mathbb{R}^p$ and vector $\delta \in \mathbb{R}^p$, denote $\text{proj}_S(\delta) = \arg\min_{x\in S}|x - \delta|$. If B is a matrix with full column rank then $\text{proj}_{\text{span}(B)}(\delta) = B(B'B)^{-1}B'\delta$.

Remark A.2. *If M is a matrix with $M'M = I$ then $|proj_{\text{span}(M)}(\delta)| = |M'\delta|$ for any vector δ. This follows since $\text{proj}_M(\delta) = M(M'M)^{-1}(M'\delta) = M(M'\delta)$, and therefore*

$$|\text{proj}_{\text{span}(M)}(\delta)|^2 = (\delta'M)(M'M)(M'\delta) = (\delta'M)(M'\delta) = |M'\delta|^2.$$

Lemma A.6. *Consider a $p \times s$ matrix R with $R'R = I$ and a finite collection of linear subspaces $\{S_k\}_{k\in N} \subseteq \mathbb{R}^p$ with $S = \text{span}(\cup_{k\in N} S_k) \neq \{0\}$. There exists a matrix B of full column rank such that $\text{span}(B) = S$ and every column of B is an element of $\cup_{k\in N} S_k$ with norm one. For $c = \{\text{mineig}(B'B)/p\}^{1/2}$ and any $\delta \in \mathbb{R}^p$ it holds*

$$|proj_S(\delta)| \le \max_{k\in N} |proj_{S_k}(\delta)|/c. \tag{30}$$

Further, if for some vector δ of norm one

$$0 \le \max_{k\in N} |\text{proj}_{S_k}(\delta)| < |R'\delta|c \tag{31}$$

then $S^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \varnothing$.

Proof. 1. Existence of matrix B. For every S_k with $S_k \neq \{0\}$, there exists a matrix B_k whose columns are elements of S_k with norm one and $\text{span}(B_k) = S_k$. Let B be a matrix whose columns consist of the largest possible number of linearly independent columns from $\{B_k\}_{k\in N}$. By construction, each column B_j of B has norm one and is an element of $\cup_{k\in N} S_k$. Since B has the maximal number of linearly independent columns from $\{B_k\}_{k\in N}$, then $\text{span}(B) = S$.

2. Proof of (30). We have (Axler (2015), 6.55) that for two linear subspaces $U, V \subseteq \mathbb{R}^p$

$$\text{if } U \subseteq V \text{ then } |\text{proj}_U(\delta)| \le |\text{proj}_V(\delta)| \text{ for all } \delta. \tag{32}$$

With $\text{maxeig}\{(B'B)^{-1}\} = 1/\text{mineig}(B'B)$ we get, for any $\delta \in \mathbb{R}^p$,

$$|\text{proj}_S(\delta)|^2 = |\delta'B(B'B)^{-1}B'\delta| \le |B'\delta|^2/\text{mineig}(B'B).$$

Let B_j denote the j-th column of B. We can bound $|B'\delta|^2 \le \max_j(B_j'\delta)^2 p$. Thus, we get

$$|\text{proj}_S(\delta)|^2 \le \max_j(B_j'\delta)^2 p/\text{mineig}(B'B) = \max_j(B_j'\delta)^2/c^2.$$

As each column B_j has norm one, it holds $(B_j'\delta)^2 = |\text{proj}_{\text{span}(B_j)}(\delta)|^2$ for all j from Remark A.2. Since for all j it holds $\text{span}(B_j) \subseteq S_k$ for some k by construction, use property (32) to conclude

$$|\text{proj}_S(\delta)|^2 \leq \max_j |\text{proj}_{\text{span}(B_j)}(\delta)|^2/c^2 \leq \max_{k\in N} |\text{proj}_{S_k}(\delta)|^2/c^2.$$

Taking square roots gives (30).

3. Proof of $S^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \varnothing$. Let δ satisfy (31). Combine (30, 31) to get $|\text{proj}_S(\delta)| < |R'\delta|$. Since $R'R = I$, then $|R'\delta| = |\text{proj}_{\text{span}(R)}(\delta)|$ by Remark A.2. Thus, $|\text{proj}_S(\delta)| < |\text{proj}_{\text{span}(R)}(\delta)|$. Contraposition of (32) shows $\text{span}(R) \not\subseteq S$.

Since $\text{span}(R) \not\subseteq S$, a vector π exists such that $R\pi \notin S$. Writing $S = \{x \in \mathbb{R}^p : x'\delta = 0,\ \forall \delta \in S^\perp\}$, we get $R\pi \notin S \iff \pi'R'\delta^* \neq 0$ for some $\delta^* \in S^\perp$. We see that $\delta^*/|\delta^*|$ is an element of the set $S^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\}$, which is thus non-empty. □

Lemma A.7. *Let P_{good} be a probability measure and R a $p \times s$ matrix such that $R'R = I$. For all $\epsilon > 0$ there exists $a > 0$ such that*

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : |x'\delta| \leq |R'\delta|a\} \leq \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\} + \epsilon. \tag{33}$$

Proof. 1. Construction of measure μ. Let measures $\{\mu_k\}_{k=1}^\infty$ and distinct affine sets $\{A_k\}_{k=1}^\infty$ of affine dimensions $\{m_k\}_{k=1}^\infty$ be given by Lemma A.4. They satisfy (i) $P_{good} = \sum_{k=1}^\infty \mu_k$, (ii) $\mu_k(\mathbb{R}^p \setminus A_k) = 0$, and (iii) $\mu_k(B) = 0$ for any measurable set B with affine dimension less than m_k. By (i), there exists $\mu = \sum_{k=1}^K \mu_k$ for some finite K such that $\mu(B) \leq P_{good}(B) \leq \mu(B) + \epsilon$ for any measurable set B. It suffices to show (33) for μ.

2. Choice of constant c. Define $S_k = \text{span}(A_k)$ for $1 \leq k \leq K$. By Lemma A.6, for every non-empty subset $N \subseteq \{1, \ldots, K\}$ with $S_N = \text{span}(\cup_{k\in N} S_k) \neq \{0\}$ there exists a matrix B_N and a constant $0 < c_N = \{\text{mineig}(B_N'B_N)/p\}^{1/2}$ such that

$$\begin{aligned} 0 \leq \max_{k\in N} |\text{proj}_{S_k}(\delta)| < |R'\delta|c_N \text{ for some } \delta \text{ of norm one} \\ \implies S_N^\perp \cap \{\delta : |\delta| = 1,\ |R'\delta| > 0\} \neq \varnothing. \end{aligned} \tag{34}$$

Take $0 < c = \min_{N\subseteq\{1,\ldots,K\}:S_N\neq\{0\}} c_N$.

3. Choice of constant a. By Lemma A.5, there exists $a > 0$ such that

$$\max_{k\leq K} \sup_{\delta\in S_k:|\delta|=1} \mu_k\{x : |x'\delta| \leq a/c\} \leq \epsilon/K. \tag{35}$$

4. Main argument. Let $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$. Write $\delta_k = \text{proj}_{S_k}(\delta)$, $\delta_k^\perp = \text{proj}_{S_k^\perp}(\delta)$, and $N_\delta = \{1 \leq k \leq K : |\delta_k| \leq |R'\delta|c\}$, and $N_\delta^c = \{1, \ldots, K\} \setminus N_\delta$. We have the identity

$$\mu\{x : |x'\delta| \leq a|R'\delta|\} = \sum_{k\in N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\} + \sum_{k\notin N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\}. \tag{36}$$

Consider the latter sum in (36). Since $\delta = \delta_k + \delta_k^\perp$, if $x \in S_k$ then $x'\delta = x'\delta_k + x'\delta_k^\perp = x'\delta_k$. Using $\mu_k(\mathbb{R}^p \setminus A_k) = 0$ from property (ii) and $A_k \subseteq S_k$, it follows $\mu_k\{x : |x'\delta| \leq$

$a|R'\delta|\} = \mu_k\{x : |x'\delta_k| \leq a|R'\delta|\}$. As $k \notin N_\delta$, it holds $|\delta_k| > |R'\delta|c > 0$ by construction. Since $\delta_k \in S_k$ and $\delta_k/|\delta_k|$ has norm one, use (35) to bound

$$\sum_{k \notin N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\} \leq \sum_{k \notin N_\delta} \mu_k\{x : |x'\delta_k|/|\delta_k| \leq a/c\} \leq K\epsilon/K = \epsilon. \tag{37}$$

Consider the first sum in (36). Since $\mu_k(\mathbb{R}^p \setminus S_k) \leq \mu_k(\mathbb{R}^p \setminus A_k) = 0$, it holds $\mu_k(B) \leq \mu_k(S_k)$ for any measurable set B. Thus, bound $\mu_k\{x : |x'\delta| \leq a|R'\delta|\} \leq \mu_k(S_k)$. Let $S_\delta = \text{span}(\cup_{k \in N_\delta} S_k)$. Since $S_k \subseteq S_\delta$ for all $k \in N_\delta$, it holds $\mu_k(S_k) \leq \mu_k(S_\delta)$. Using the definition $\mu = \sum_{k \leq K} \mu_k$ we conclude

$$\sum_{k \in N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\} \leq \sum_{k \in N_\delta} \mu_k(S_\delta) \leq \sum_{k \leq K} \mu_k(S_\delta) = \mu(S_\delta). \tag{38}$$

Insert bounds (37) and (38) into (36) to get

$$M = \mu\{x : |x'\delta| \leq a|R'\delta|\} \leq \mu(S_\delta) + \epsilon. \tag{39}$$

If N_δ is empty then the first sum in (36) is empty, so (36) and (37) combine to

$$M \leq \epsilon \leq \sup_{\delta : |\delta| = 1,\, |R'\delta| > 0} \mu\{x : x'\delta = 0\} + \epsilon.$$

If N_δ is non-empty and $S_\delta = \{0\}$ then (39) gives

$$M \leq \mu\{0\} + \epsilon \leq \sup_{\delta : |\delta| = 1, |R'\delta| > 0} \mu\{x : x'\delta = 0\} + \epsilon. \tag{40}$$

Suppose N_δ is non-empty and $S_\delta \neq \{0\}$. Then δ satisfies $|\delta| = 1$ and $\max_{k \in N_\delta} |\delta_k| \leq |R'\delta|c$, and so that by (34) there exists $\delta^* \in S_\delta^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\}$. In particular, $S_\delta \subseteq \{x : x'\delta^* = 0\}$ and from (39)

$$M \leq \mu(S_\delta) + \epsilon \leq \mu\{x : x'\delta^* = 0\} + \epsilon \leq \sup_{\delta : |\delta| = 1, |R'\delta| > 0} \mu\{x : x'\delta = 0\} + \epsilon.$$

□

Lemma A.8. *Let $\{x_{in}\}_{i \leq n, n \in \mathbb{N}}$ be defined on a probability space (Ω, Σ, P). Let $\zeta_n \subseteq \{1, \ldots, n\}$ be sets with $|\zeta_n| = h_n$. Suppose there exists a probability measure P^* on $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ such that for any measurable set B it holds $h_n^{-1} \sum_{i \in \zeta_n} \mathbb{I}\{x_{in} \in B\} \to P^*(B)$, P-almost surely. Then P-almost surely*

$$\sup_{\delta : |\delta| = 1,\, a \geq 0} \left| h_n^{-1} \sum_{i \in \zeta_n} \mathbb{I}\{|x_{in}'\delta| \leq a\} - P^*\{x : |x'\delta| \leq a\} \right| \xrightarrow{n \to \infty} 0$$

Proof. Let $P_n^* = h_n^{-1} \sum_{i \in \zeta_n} \mathbb{I}\{x_{in} \in B\}$. Theorem 1 and Example 11 of Elker et al. (1979) show that if $P_n^*(B) \to P^*(B)$ P-almost surely for all Borel sets B then

$$\sup_{\delta : |\delta| = 1,\, a \in \mathbb{R}} |P_n^*\{x : x'\delta \leq a\} - P^*\{x : x'\delta \leq a\}| \xrightarrow{n \to \infty} 0. \tag{41}$$

We rewrite $\mathbb{I}\{x : |x'\delta| \leq a\}$. Since $\{|x'\delta| \leq a\} = \{x'\delta \leq a\} \backslash \{x'\delta < -a\}$ and the latter set complements $\{x'(-\delta) \leq a\}$ then $\mathbb{I}\{x : |x'\delta| \leq a\} = \mathbb{I}\{x : x'\delta \leq a\} + \mathbb{I}\{x : x'(-\delta) \leq a\} - 1$. Thus, $P_n^*\{x : |x'\delta| \leq a\} - P^*\{x : |x'\delta| \leq a\} = P_n^*\{x : x'\delta \leq a\} - P_n^*\{x : x'\delta \leq a\} + P_n^*\{x : x'(-\delta) \leq a\} - P^*\{x : x'(-\delta) \leq a\}$. The claim follows by the triangle inequality and (41). □

Proof of Proposition 3.2. It holds $\lim_{(a,n)\to(0,\infty)} F^R_{n\zeta_\circ}(a) = \lim_{n\to\infty} F^R_{n\zeta_\circ}(0)$ if the first limit exists. We show that

$$\lim_{(a,n)\to(0,\infty)} F^R_{n\zeta_\circ}(a) = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\}. \tag{42}$$

Let $\epsilon > 0$ be given. By condition (5) and Lemma A.8, we can find a set of probability one where, pointwise, for large n

$$\sup_{\delta:|\delta|=1,\,a\geq 0} \left| h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{|x_i'\delta| \leq a\} - P_{good}\{x : |x'\delta| \leq a\} \right| \leq \epsilon. \tag{43}$$

In particular, for each element of this set, for $a \geq 0$ and n large it holds

$$\begin{aligned} F^R_{n\zeta_\circ}(a) &= \max_{\delta:|\delta|=1,\,|R'\delta|>0} h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{|x_i'\delta| \leq a|R'\delta|\} \\ &\leq \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : |x'\delta| \leq a|R'\delta|\} + \epsilon. \end{aligned}$$

and

$$F^R_{n\zeta_\circ}(a) \geq \max_{\delta:|\delta|=1,\,|R'\delta|>0} h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{x_i'\delta = 0\} \geq \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\} - \epsilon.$$

By Lemma A.7, we can find $a^* > 0$ such that for $0 \leq a \leq a^*$

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : |x'\delta| \leq a|R'\delta|\} \leq \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\} + \epsilon.$$

Putting the inequalities together gives that for all $0 \leq a \leq a^*$ and n large

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\} - \epsilon \leq F^R_{n\zeta_\circ}(a) \leq \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\} + 2\epsilon.$$

Thus, (42) holds.

For the second part of the statement, if $\sup_{\delta:|\delta|=1,\,|R'\delta|>0} \lambda_\circ P_{good}\{x : x'\delta = 0\} < \lambda_\circ + \underline{\lambda} - 1$ then from $\lim_{(a,n)\to(0,\infty)} \lambda_\circ F^R_{n\zeta_\circ}(a) = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} \lambda_\circ P_{good}\{x : x'\delta = 0\}$ holding almost surely we can find $\xi < \lambda_\circ + \underline{\lambda} - 1$ so that Assumption 3 is satisfied. □

A.4 Proof of Proposition 3.3

Proof. It holds $F^R_{n\zeta}(0) \leq F^R_{n\zeta}(a)$ for all $a > 0$ and $\zeta \subseteq \{1, \ldots, n\}$ by definition. We show that $\exists a^* > 0$ such that $F^R_{n\zeta}(a) \leq F^R_{n\zeta}(0)$ for all $\zeta \subseteq \{1, \ldots, n\}$ and $0 \leq a < a^*$.

Let the finite set $\mathcal{X} = \{x_1, \ldots, x_K\} \subseteq \mathbb{R}^p$ denote the support of x_{in}. Define $\mathcal{X}_{a\delta R} = \{x \in \mathcal{X} : |x'\delta| \leq a|R'\delta|\}$ so that for any $a \geq 0$

$$F^R_{n\zeta}(a) = \max_{\delta:|\delta|=1,\,|R'\delta|>0} h^{-1} \sum_{i\in\zeta} \mathbb{I}\{|x_{in}'\delta| \leq a|R'\delta|\} = \max_{\delta:|\delta|=1,\,|R'\delta|>0} h^{-1} \sum_{i\in\zeta} \mathbb{I}\{x_{in} \in \mathcal{X}_{a\delta R}\}.$$

To show $F^R_{n\zeta}(a) \leq F^R_{n\zeta}(0)$ for, we show that for every $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ there exists $\delta^* \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ such that $\mathcal{X}_{a\delta R} \subseteq \mathcal{X}_{0\delta^* R} = \{x \in \mathcal{X} : x'\delta^* = 0\}$.

If $\mathcal{X} = \{0\}$ then $\mathcal{X}_{a\delta R} = \mathcal{X}_{0\delta R} = \{0\}$ for all δ, a. Consider then $\mathcal{X} \neq \{0\}$ and let $\underline{x} \in \mathcal{X}$ be such that $|\underline{x}| = \min_{x \in \mathcal{X}: |x|>0} |x|$. For every non-empty subset $N \subseteq \{1, \dots, K\}$ with $\{0\} \neq \{x_k\}_{k \in N} \subseteq \mathcal{X}$, by Lemma A.6 there exists a matrix B_N such that every column of B_N has norm one and $\text{span}(B_N) = \text{span}(\cup_{k \in N} \{x_k\})$. Furthermore, by Lemma A.6, the constant $0 < c_N = \{\text{mineig}(B_N' B_N)/p\}^{1/2}$ satisfies

$$\begin{aligned} 0 \leq \max_{k \in N} |\text{proj}_{\text{span}(x_k)}(\delta)| < |R'\delta| c_N \text{ for some } \delta \text{ of norm one} \\ \implies \text{span}(\cup_{k \in N} \{x_k\})^\perp \cap \{\delta : |\delta| = 1,\ |R'\delta| > 0\} \neq \varnothing. \end{aligned} \tag{44}$$

Take $c = \min_{N \subseteq \{1,\dots,K\}: \{x_k\}_{k \in N} \neq \{0\}} c_N$ and $0 < a^* = c|\underline{x}|$.

Let $\zeta \subseteq \{1, \dots, n\}$ and $0 < a < a^*$. For $\delta \in \{\delta : |\delta| = 1,\ |R'\delta| > 0\}$ with $\mathcal{X}_{a\delta R} = \varnothing$ or $\mathcal{X}_{a\delta R} = \{0\}$ it holds $\mathcal{X}_{a\delta R} = \mathcal{X}_{0\delta^* R}$ for any $\delta^* \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$.

Consider then δ such that $\mathcal{X}_{a\delta R}$ is not empty and not equal to $\{0\}$. We have $\text{proj}_{\{0\}}(\delta) = 0$ and therefore

$$\max_{x \in \mathcal{X}_{a\delta R}} |\text{proj}_{\text{span}(x)}(\delta)| = \max_{x \in \mathcal{X}_{a\delta R}: x \neq 0} |\text{proj}_{\text{span}(x)}(\delta)|.$$

For any non-zero x it holds $|\text{proj}_{\text{span}(x)}(\delta)| = |x(x'\delta)/(x'x)| = |x'\delta|/|x|$. Since for $x \in \mathcal{X}_{a\delta R}$ it holds $|x'\delta| \leq a|R'\delta|$ by definition, we get

$$\max_{x \in \mathcal{X}_{a\delta R}: x \neq 0} |\text{proj}_{\text{span}(x)}(\delta)| = \max_{x \in \mathcal{X}_{a\delta R}: x \neq 0} \frac{|x'\delta|}{|x|} \leq \max_{x \in \mathcal{X}_{a\delta R}: x \neq 0} \frac{a|R'\delta|}{|x|}.$$

Bounding $a < a^* = c|\underline{x}|$ and $|x| \geq |\underline{x}|$ in the final term we conclude

$$\max_{x \in \mathcal{X}_{a\delta R}} |\text{proj}_{\text{span}(x)}(\delta)| < |R'\delta| c.$$

Thus, by (44), there exists $\delta^* \in \text{span}(\cup_{x \in \mathcal{X}_{a\delta R}} \{x\})^\perp \cap \{\delta : |\delta| = 1,\ |R'\delta| > 0\}$, giving $\mathcal{X}_{a\delta R} \subseteq \{x : x'\delta^* = 0\}$. Since $\mathcal{X}_{a\delta R} \subseteq \mathcal{X}$ then $\mathcal{X}_{a\delta R} \subseteq \{x \in \mathcal{X} : x'\delta^* = 0\} = \mathcal{X}_{0\delta^* R}$, as desired. □

A.5 Proof of Theorem 3

Lemma A.9. *Let R be a $p \times s$ matrix and $h \leq n$. Suppose $\min_{\zeta: |\zeta| = h} \hat{\sigma}_\zeta^2 > 0$ and*

$$\max_{\delta: |\delta|=1,\ |R'\delta|>0} \sum_{i=1}^{n} \mathbb{I}\{x_{in}'\delta = 0\} < n \tag{45}$$

Then $R'\hat{\beta}_\zeta$ is unique for all minimisers $\zeta \in \mathcal{M}_h$.

Proof. By Lemma A.2, if ζ with $|\zeta| = h$ is such that

$$\max_{\delta: |\delta|=1,\ |R'\delta|>0} h^{-1} \sum_{i \in \zeta} \mathbb{I}\{x_{in}'\delta = 0\} < 1 \tag{46}$$

then $R'\hat{\beta}_\zeta$ is unique. Therefore, (46) holds with equality for any ζ with $R'\hat{\beta}_\zeta$ non-unique. We show that if (46) holds with equality then $\zeta \notin \mathcal{M}_h$, which establishes the claim.

Let ζ be such that $h^{-1}\sum_{i\in\zeta}\mathbb{I}\{x_{in}'\delta=0\}=1$ for some $\delta\in\{\delta:|\delta|=1,|R'\delta|>0\}$. Let $\hat{\beta}_\zeta$ denote a solution to the least squares problem for observations in ζ. By (45), there exists $j\in\{1,\dots,n\}\setminus\zeta$ such that $x_{jn}'\delta\neq 0$. Furthermore, from the assumption $\min_{\zeta:|\zeta|=h}\hat{\sigma}_\zeta^2>0$, there exists $i^*\in\zeta$ such that $(y_{i^*}-x_{i^*n}'\hat{\beta}_\zeta)^2>0$. Let ζ^* be the h-set of indices consisting of j and all elements of ζ except for i^*. Let $\tilde{\beta}_{\zeta^*}=\hat{\beta}_\zeta+\delta t$ where $t=(y_j-x_{jn}'\hat{\beta}_\zeta)/(x_{jn}'\delta)$ so that $y_j-x_{jn}'\tilde{\beta}_{\zeta^*}=(y_j-x_{jn}'\hat{\beta}_\zeta)-(y_j-x_{jn}'\hat{\beta}_\zeta)=0$. We thus see that $\zeta\notin\mathcal{M}_h$ since

$$\begin{aligned}\hat{\sigma}_{\zeta^*}^2\leq h^{-1}\sum_{i\in\zeta^*}(y_i-x_{in}'\tilde{\beta}_{\zeta^*})^2&=h^{-1}(y_j-x_{jn}'\tilde{\beta}_{\zeta^*})^2+h^{-1}\sum_{i\in\zeta\setminus\{i^*\}}(y_i-x_{in}'\tilde{\beta}_{\zeta^*})^2\\&=h^{-1}\sum_{i\in\zeta\setminus\{i^*\}}(y_i-x_{in}'\hat{\beta}_\zeta)^2<\hat{\sigma}_\zeta^2.\end{aligned}$$

□

We now prove a general result regarding LTS boundedness when the magnitude of the 'outlying' errors grows. The result uses the following assumption.

Assumption 9. *Let R be a $p\times s$ matrix with $R'R=I$. Suppose the following hold.*
(i) The data are generated with $h_\circ=\lfloor\lambda_\circ n\rfloor$ 'good' observations with $1/2<\lambda_\circ\leq 1$. Let the lower bound $\underline{h}=\lfloor\underline{\lambda}n\rfloor$ be such that $1-\lambda_\circ<\underline{\lambda}\leq\lambda_\circ$.
(ii) $h_\circ^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2=O_p(1)$.
(iii) $\lim_{n\to\infty}P(\min_{\zeta:|\zeta|=\underline{h}}\hat{\sigma}_\zeta^2>0)=1$.
(iv) There exists a deterministic sequence $b_n\to\infty$ such that $b_n^2/\min_{j\in\zeta_\circ^c}\varepsilon_j^2=o_p(1)$.
(v) For $\zeta\subset\{1,\dots,n\}$, $\delta\in\{\delta:|\delta|=1\}$, $a,\tau\geq 0$, and $t\leq|\zeta\cap\zeta_\circ|$ define

$$H_{\delta\zeta}(a)=\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\delta|\leq a|R'\delta|\},\quad \bar{G}_{\delta\zeta}(a)=\sum_{i\in\zeta\cap\zeta_\circ^c}\mathbb{I}\{|x_{in}'\delta|>a|R'\delta|\},$$

$$H_{\delta\zeta}^{-1}(t)=\inf\{a\geq 0:H_{\delta\zeta}(a)\geq t\},\quad \kappa_{\delta\zeta\tau}=H_{\delta\zeta}^{-1}(|\zeta\cap\zeta_\circ|-\tau n),$$

$$K_{n\tau}(a)=\max_{\underline{h}\leq h\leq h_\circ}\max_{\zeta:|\zeta|=h}\max_{\delta:|\delta|=1,\,|R'\delta|>0}\left\{H_{\delta\zeta}(a)+\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n)-h\right\}.$$

There exists $\nu,\tau>0$ such that $\lim_{(a,n)\to(0,\infty)}P\left(K_{n\tau}(a)>-\nu n\right)=0$.

Remark A.3. *$H_{\delta\zeta}$ is a non-decreasing, right-continuous step function with limits from the right. $H_{\delta\zeta}^{-1}$ is a non-decreasing, left-continuous step function with limits from the right. It holds $H_{\delta\zeta}(H_{\delta\zeta}^{-1}(t))\geq t$ by right-continuity of $H_{\delta\zeta}$.*

Lemma A.10. *Suppose Assumption 9. Consider the LTS estimator selecting h observations and let $\mathcal{M}_h$ be the set of minimising ζ. Then $\max_{\zeta\in\mathcal{M}_h}|R'\hat{\beta}_\zeta|=O_p(1)$ uniformly for $\underline{h}\leq h\leq h_\circ$.*

Proof of Lemma A.10. Structure of the proof is similar to that of Theorem 2. We first construct a sequence of sets Ω_n and a constant B_0. We then show that on Ω_n and for large n it holds $\max_{\zeta\in\mathcal{M}_h}|\hat{\beta}_\zeta-\beta|/\sigma\leq B_0$ for all $\underline{h}\leq h\leq h_\circ$. Finally, we show that for any $\epsilon>0$ the sequence Ω_n can be constructed so that $P(\Omega_n)\geq 1-\epsilon$ for all large n.

1. Construction of Ω_n. There are $a,\nu,\tau>0$ such that on Ω_n it holds for all $\underline{h}\leq h\leq h_\circ$

$$\max_{\zeta:|\zeta|=h}\max_{\delta:|\delta|=1,\,|R'\delta|>0}\left\{H_{\delta\zeta}(a)+\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n)\right\}\leq h-\nu n.\tag{47}$$

Since $\kappa_{\delta\zeta\tau}$ is decreasing in τ and $\bar{G}_{\delta\zeta}(a)$ is decreasing in a, we can adjust τ downwards while maintaining the inequality (47). Take $0 < 2\tau < \min\{\nu, \underline{\lambda} + \lambda_\circ - 1\}$.

We require that on Ω_n, there are $A_0 > 0$ and $0 < \eta < \tau/2$ such that

$$\underline{h}^{-1} \sum_{i\in\zeta_\circ} \varepsilon_i^2 < A_0, \quad \underline{h}^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\left\{\varepsilon_i^2 > A_0/\eta\right\} \leq \eta, \tag{48}$$

where the first inequality implies the second by Chebyshev's inequality.

For $\underline{h} \leq h \leq h_\circ$, let $\zeta_h \subseteq \zeta_\circ$ be sets with $|\zeta_h| = h$. Let $B_0 > 0$ be such that on Ω_n and for

$$Z^R_{\tau a h} = \Big\{\zeta : (i)\, |\zeta| = h, (ii)\, R'\hat{\beta}_\zeta \text{ is unique}, (iii) \max_{\delta\in\hat{\delta}_\zeta} h^{-1} \sum_{i\in\zeta\cap\zeta_\circ} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} \geq \tau\Big\}$$

it holds

$$\min_{\zeta\in Z^R_{\tau a h}: |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0} \hat{\sigma}^2_\zeta > \hat{\sigma}^2_{\zeta_h}. \tag{49}$$

We can adjust B_0 upwards while maintaining (49), so we take

$$B_0 \geq \big(\frac{A_0}{\eta}\big)^{1/2} + \big(\frac{A_0}{\tau - \eta}\big)^{1/2}. \tag{50}$$

Finally, we require that on Ω_n it holds

$$\min_{j\in\zeta_\circ^c} |\varepsilon_j| > \sqrt{A_0/\tau} + b_n B_0, \ \min_{\zeta:|\zeta|=\underline{h}} \hat{\sigma}^2_\zeta > 0. \tag{51}$$

2.1. Showing $\mathcal{M}_h \subseteq \{\zeta : R'\hat{\beta}_\zeta$ is unique$\}$ on Ω_n. For $\underline{h} \leq h \leq h_\circ$, let $\zeta_h \subseteq \zeta_\circ$ have $|\zeta_\circ| = h$ as above. Since $\zeta_h \cap \zeta_\circ = \zeta_h$ it holds $H_{\delta\zeta_h}(a) = \sum_{i\in\zeta_h} \mathbb{I}\{|x_{in}'\delta| \leq a|R'\delta|\}$. Using $\min \sum_i \mathbb{I}_{A_i} = h - \max \sum_i \mathbb{I}_{A_i^c}$ we get

$$\min_{\delta:|\delta|=1,\, |R'\delta|>0} \sum_{i\in\zeta_h} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} = h - \max_{\delta:|\delta|=1,\, |R'\delta|>0} H_{\delta\zeta_h}(a). \tag{52}$$

Since $|\zeta_h \cap \zeta_\circ^c| = 0$ then $\bar{G}_{\delta\zeta_h}(\kappa_{\delta\zeta\tau} b_n) = 0$. Thus, (47) gives $\max_{\delta:|\delta|=1,\, |R'\delta|>0} H_{\delta\zeta_h}(a) \leq h - \nu n$. Insert this in (52), and use $n \geq h$ and our choice of τ to get uniformly in $\underline{h} \leq h \leq h_\circ$

$$\min_{\delta:|\delta|=1,\, |R'\delta|>0} \sum_{i\in\zeta_h} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} \geq \nu n \geq \nu h > \tau h > 0. \tag{53}$$

Use $\max \sum_{i=1}^n \mathbb{I}_{A_i} = n - \min \sum_{i=1}^n \mathbb{I}_{A_i^c}$, (53), $\tau > 0$, and $h \geq \underline{h} > 0$ to get

$$\max_{\delta:|\delta|=1, |R'\delta|>0} \sum_{i=1}^n \mathbb{I}\{x_{in}'\delta = 0\} \leq n - \min_{\delta:|\delta|=1, |R'\delta|>0} \sum_{i\in\zeta_h} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} \leq n - \tau h < n.$$

Therefore, (45) is satisfied. From (51) we have $\min_{\zeta:|\zeta|=h} \hat{\sigma}^2_\zeta \geq (\underline{h}/h) \min_{\zeta:|\zeta|=\underline{h}} \hat{\sigma}^2_\zeta > 0$ for all $\underline{h} \leq h \leq h_\circ$. Then Lemma A.9 shows $\mathcal{M}_h \subseteq \{\zeta : R'\hat{\beta}_\zeta$ is unique$\}$ for all $\underline{h} \leq h \leq h_\circ$.

2.2. Identity on Ω_n. Define $\hat{\ell}_\zeta = |R'\hat{\beta}_\zeta - R'\beta|/\sigma$. We will consider ζ in the set $S = \{\zeta : |\zeta| = h,\ R'\hat{\beta}_\zeta \text{ unique},\ \hat{\ell}_\zeta > B_0\}$. Fix $\hat{\beta}^*_\zeta \in \hat{\beta}_\zeta$ for each ζ and write $\hat{\delta}_\zeta = (\hat{\beta}^*_\zeta - \beta)/|\hat{\beta}^*_\zeta - \beta|$. Define

$$H_\zeta = H_{\hat{\delta}_\zeta,\zeta}\,,\ \bar{H}_\zeta = |\zeta\cap\zeta_\circ| - H_\zeta\,,\ \bar{G}_\zeta = \bar{G}_{\hat{\delta}_\zeta,\zeta}\,, G_\zeta = |\zeta\cap\zeta_\circ^c| - \bar{G}_\zeta\,, H_\zeta^{-1} = H^{-1}_{\hat{\delta}_\zeta,\zeta},\ \kappa_{\zeta\tau} = \kappa_{\hat{\delta}_\zeta,\zeta,\tau}.$$

For all $\underline{h} \le h \le h_\circ$ and ζ with $|\zeta| = h$ we have the identity

$$H_\zeta(a) + \bar{H}_\zeta(a) + G_\zeta(\kappa_{\zeta\tau} b_n) + \bar{G}_\zeta(\kappa_{\zeta\tau} b_n) = h. \tag{54}$$

2.3. Bounding argument on Ω_n. Define $S = \{\zeta : |\zeta| = h,\ R'\hat{\beta}_\zeta \text{ unique},\ \hat{\ell}_\zeta > B_0\}$ as above and write $S = S_1^h \cup S_2^h \cup S_3^h$ where

$$\begin{aligned}
S_1^h &= S \cap \{\zeta : \bar{H}_\zeta(a) \ge n\tau\},\\
S_2^h &= S \cap \{\zeta : \bar{H}_\zeta(a) < n\tau\} \cap \{\zeta : \hat{\ell}_\zeta \kappa_{\zeta\tau} \le B_0\},\\
S_3^h &= S \cap \{\zeta : \bar{H}_\zeta(a) < n\tau\} \cap \{\zeta : \hat{\ell}_\zeta \kappa_{\zeta\tau} > B_0\}.
\end{aligned}$$

We show $\min_{\zeta\in S_j^h} \hat{\sigma}^2_\zeta > \hat{\sigma}^2_{\zeta_h}$ for $\underline{h} \le h \le h_\circ$ and $j = 1, 2, 3$.

The set S_1^h. Let $\underline{h} \le h \le h_\circ$ and $\zeta \in S_1^h$. By definitions of $\bar{H}_\zeta$ and S_1^h it holds

$$\sum_{i\in\zeta\cap\zeta_\circ} \mathbb{I}\{|x'_{in}\hat{\delta}_\zeta| > a|R'\hat{\delta}|\} = \bar{H}_\zeta(a) \ge n\tau \ge h\tau.$$

Thus, $S_1^h \subseteq Z_{\tau ah} \cap \{\zeta : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0\}$. From (49) it follows

$$\min_{\zeta\in S_1^h} \hat{\sigma}^2_\zeta \ge \min_{\zeta\in Z^R_{\tau ah} : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0} \hat{\sigma}^2_\zeta > \hat{\sigma}^2_{\zeta_h}.$$

The set S_2^h. Let $\underline{h} \le h \le h_\circ$ and $\zeta \in S_2^h$. As in the proof of Lemma *A*.3, we have $(y_i - x'_{in}\hat{\beta}_\zeta)/\sigma = \varepsilon_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|$. Use this to write

$$\frac{\hat{\sigma}^2_\zeta}{\sigma^2} = \frac{1}{h}\sum_{j\in\zeta} \frac{(y_i - x'_{in}\hat{\beta}_\zeta)^2}{\sigma^2} \ge \frac{1}{h}\sum_{j\in\zeta\cap\zeta_\circ^c} \mathbb{I}\{|x'_{jn}\hat{\delta}_\zeta| \le \kappa_{\zeta\tau} b_n |R'\hat{\delta}_\zeta|\}\big(\varepsilon_j - \frac{\hat{\ell}_\zeta x'_{jn}\hat{\delta}_\zeta}{|R'\hat{\delta}_\zeta|}\big)^2. \tag{55}$$

For $\zeta \in S_2^h$ it holds $\hat{\ell}_\zeta \kappa_{\zeta\tau} \le B_0$ by definition. Combine with $|x'_{jn}\hat{\delta}_\zeta| \le \kappa_{\zeta\tau} b_n |R'\hat{\delta}_\zeta|$ to get $|\hat{\ell}_\zeta x'_{jn}\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| \le b_n B_0$. Combine with (51) to get

$$\min_{j\in\zeta_\circ^c} |\varepsilon_j| > \sqrt{A_0/\tau} + b_n B_0 \ge \sqrt{A_0/\tau} + |\hat{\ell}_\zeta x'_{jn}\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta|. \tag{56}$$

The reverse triangle inequality and (56) give that for $j \in \zeta_\circ^c$ it holds $|\varepsilon_j - \hat{\ell}_\zeta x'_{jn}\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|| \ge |\varepsilon_j| - |\hat{\ell}_\zeta x'_{jn}\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| > \sqrt{A_0/\tau}$ and therefore $(\varepsilon_j - \hat{\ell}_\zeta x'_{jn}\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|)^2 > A_0/\tau$. Insert this in (55) and use the definition of G_ζ to bound

$$\hat{\sigma}^2_\zeta/\sigma^2 > h^{-1}\sum_{j\in\zeta\cap\zeta_\circ^c} \mathbb{I}\{|x'_{jn}\hat{\delta}_\zeta| \le \kappa_{\zeta\tau} b_n |R'\hat{\delta}_\zeta|\}(A_0/\tau) = h^{-1} G_\zeta(\kappa_{\zeta\tau} b_n)(A_0/\tau). \tag{57}$$

It holds $\bar{H}_\zeta(a) < n\tau$ by definition of S_2^h. By (47), we have $H_\zeta(a)+\bar{G}_\zeta(\kappa_{\zeta\tau}b_n) \le h-\nu n$. Combine these with identity (54) to get

$$G_\zeta(\kappa_{\zeta\tau}b_n) = h - \bar{H}_\zeta(a) - H_\zeta(a) - \bar{G}_\zeta(\kappa_{\zeta\tau}b_n) > n(\nu-\tau) > h\tau, \tag{58}$$

where the last inequality uses $n \ge h$ and the choice $2\tau < \nu$. Inserting (58) into (57) shows $\hat{\sigma}_\zeta^2/\sigma^2 > A_0$. Since $A_0 > \underline{h}^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2$ from (48) and $\underline{h}^{-1}\sum_{i\in\zeta_\circ}\varepsilon_i^2 \ge h^{-1}\sum_{i\in\zeta_h}\varepsilon_i^2 \ge \hat{\sigma}_{\zeta_h}^2$, we conclude $\min_{\zeta\in S_2^h}\hat{\sigma}_\zeta^2 > \hat{\sigma}_{\zeta_h}^2$ for all $\underline{h} \le h \le h_\circ$.

The set S_3^h. Let $\underline{h} \le h \le h_\circ$ and $\zeta \in S_3^h$. Using $(y_i - x_{in}'\hat{\beta}_\zeta)/\sigma = \varepsilon_i - \hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|$ bound

$$\hat{\sigma}_\zeta^2/\sigma^2 \ge h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| \ge \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\}\mathbb{I}\{\varepsilon_i^2 \le A_0/\eta\}(\varepsilon_i - \hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|)^2. \tag{59}$$

It holds $\hat{\ell}_\zeta\kappa_{\zeta\tau} > B_0$ by definition of S_3^h. When $|x_{in}'\hat{\delta}_\zeta| \ge \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|$ then $|\hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| > B_0$. Using (50) and $|\varepsilon_i| \le (A_0/\eta)^{1/2}$ we get further

$$|\hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| > \Big(\frac{A_0}{\eta}\Big)^{1/2} + \Big(\frac{A_0}{\tau-\eta}\Big)^{1/2} \ge |\varepsilon_i| + \Big(\frac{A_0}{\tau-\eta}\Big)^{1/2}. \tag{60}$$

Thus, reverse triangle inequality and (60) give

$$\Big|\varepsilon_i - \hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|\Big| \ge |\hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| - |\varepsilon_i| > \Big(\frac{A_0}{\tau-\eta}\Big)^{1/2},$$

and $(\varepsilon_i - \hat{\ell}_\zeta x_{in}'\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|)^2 > A_0/(\tau-\eta)$. Insert this in (59) to bound

$$\hat{\sigma}_\zeta^2/\sigma^2 > h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| \ge \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\}\mathbb{I}\{\varepsilon_i^2 \le A_0/\eta\}\Big(\frac{A_0}{\tau-\eta}\Big).$$

With the inequality $\mathbb{I}_S\mathbb{I}_T \ge \mathbb{I}_S - \mathbb{I}_{T^c}$ holding for any two sets S, T we get

$$\hat{\sigma}_\zeta^2/\sigma^2 > \frac{A_0}{\tau-\eta}\Big(h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| \ge \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} - h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{\varepsilon_i^2 > A_0/\eta\}\Big). \tag{61}$$

We bound the first sum in (61). Using $\min\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}_{A_i} = |\zeta\cap\zeta_\circ| - \max h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}_{A_i^c}$

$$\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| \ge \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} = |\zeta\cap\zeta_\circ| - \sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| < \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\}. \tag{62}$$

It holds $\mathbb{I}\{|x_{in}'\delta| < \kappa|R'\delta|\} = \lim_{a\uparrow\kappa}\mathbb{I}\{|x_{in}'\delta| \le a|R'\delta|\}$ and thus

$$M_\zeta = \sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| < \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} = \lim_{a\uparrow\kappa_{\zeta\tau}}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| \le a|R'\hat{\delta}_\zeta|\} = \lim_{a\uparrow\kappa_{\zeta\tau}} H_\zeta(a).$$

By definition of $\kappa_{\zeta\tau} = H_\zeta^{-1}(|\zeta\cap\zeta_\circ| - \tau n)$ we have $H_\zeta(a) \le \max\{|\zeta\cap\zeta_\circ| - \tau n, 0\}$ for all $a < \kappa_{\zeta\tau}$, and therefore $\lim_{a\uparrow\kappa_{\zeta\tau}} H_\zeta(a) \le \max\{|\zeta\cap\zeta_\circ| - \tau n, 0\}$. Use $|\zeta| \ge \underline{h}$ and $|\zeta_\circ^c| = n - h_\circ$ to bound $|\zeta\cap\zeta_\circ| \ge |\zeta| - |\zeta_\circ^c| \ge \underline{h} + h_\circ - n$. Since $h_\circ/n = \lambda_\circ + o(1)$

and $\underline{h}/n = \underline{\lambda} + o(1)$ by assumption, our choice of τ implies that for large n it holds $|\zeta \cap \zeta_\circ| \geq n\{\underline{\lambda} + \lambda_\circ - 1 + o(1)\} > \tau n$. Thus $\max\{|\zeta \cap \zeta_\circ| - \tau n, 0\} = |\zeta \cap \zeta_\circ| - \tau$ for large n and we conclude $M_\zeta = \lim_{a \uparrow \kappa_{\zeta\tau}} H_\zeta(a) \leq \max\{|\zeta \cap \zeta_\circ| - \tau n, 0\} = |\zeta \cap \zeta_\circ| - \tau n$. Plug this in (62) and use $n \geq h$ to get

$$\sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\hat{\delta}_\zeta| \geq \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} \geq \tau n \geq \tau h. \tag{63}$$

For the second sum in (61), use (48) to get for all $\underline{h} \leq h \leq h_\circ$

$$\sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} \leq \sum_{i \in \zeta_\circ} \mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} \leq \eta \underline{h} \leq \eta h. \tag{64}$$

Insert (63) and (64) in (61) to conclude $\hat{\sigma}_\zeta^2/\sigma^2 \geq A_0 > \hat{\sigma}_{\zeta_h}^2/\sigma^2$. Since the bounds are uniform in $\zeta \in S_3^h$ and $\underline{h} \leq h \leq h_\circ$, conclude $\min_{\zeta \in S_3^h} \hat{\sigma}_\zeta^2 > \hat{\sigma}_{\zeta_h}^2$ for all $\underline{h} \leq h \leq h_\circ$.

2.4. Conclusion on Ω_n. Part 2.1 shows $\mathcal{M}_h \subseteq \{\zeta : R'\hat{\beta}_\zeta \text{ unique}\}$ for all $\underline{h} \leq h \leq h_\circ$. Part 2.3. shows for all $\underline{h} \leq h \leq h_\circ$ and $\zeta \in \{\zeta : |\zeta| = h, R'\hat{\beta}_\zeta \text{ unique}, |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0\}$ it holds $\hat{\sigma}_\zeta^2 > \hat{\sigma}_{\zeta_h}^2$. Thus, any h-subset ζ with $|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0$ is not in $\mathcal{M}_h$.

3. Probability Analysis. Let $\epsilon > 0$. By Assumption $9(v)$, $\exists a, \nu, \tau > 0$ and Ω_{1n} such that (47) holds on Ω_{1n} and $P(\Omega_{1n}) \geq 1 - \epsilon$ for large n. By Assumption $9(i, ii)$, $\exists A_0 > 0$ and Ω_{2n} such that the first inequality in (48) holds on Ω_{2n} and $P(\Omega_{2n}) \geq 1 - \epsilon$ for all n. By Assumption $9(i, ii)$ and Lemma A.3, $B_0 > 0$ and Ω_{3n} exist such that for any $\zeta_h \subseteq \zeta_\circ$ with $|\zeta_h| = h$, (49) holds on Ω_{3n} and $P(\Omega_{3n}) \geq 1 - \epsilon$ for all n. Finally, by Assumption $9(iii, iv)$, Ω_{4n} exists such that (51) holds on Ω_{4n} and $P(\Omega_{4n}) \geq 1 - \epsilon$ for large n. Take $\Omega_n = \Omega_{1n} \cap \Omega_{2n} \cap \Omega_{3n} \cap \Omega_{4n}$. □

Remark A.4. *A finite union $V = \cup_{j \in J} V_k$ of linear subspaces $V_j \subseteq \mathbb{R}^p$ is a linear subspace only if $V = V_j$ for some $j \in J$. If $V = \{0\}$ the claim is immediate. If $V \neq \{0\}$ is a linear subspace, then V is a non-trivial vector space over the infinite field $\mathbb{R}$. The claim then follows from the following fact (Roman (2005), Theorem 1.2): A non-trivial vector space over an infinite field is not the union of a finite number of proper subspaces.*

Lemma A.11. *Let R be a $p \times s$ matrix and $S \subseteq \mathbb{R}^p$ a linear subspace with $S_\delta^\perp = S^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \varnothing$. For any finite subset $\{x_j\}_{j \in J} \subset \mathbb{R}^p \setminus S$, a vector $\delta^* \in S_\delta^\perp$ exists such that $x_j'\delta^* \neq 0$ for all $j \in J$.*

Proof. We show $\exists \delta^\dagger \in S^\perp$ with $|\delta^\dagger| = 1$ such that $x_j'\delta^\dagger \neq 0 \ \forall \in J$.It is equivalent to show $S^\perp \not\subseteq V$, where $V = \bigcup_{j \in J} x_j^\perp$. If V is not a linear subspace, $S^\perp \not\subseteq V$ is immediate. Suppose V is a linear subspace. By Remark A.4, it follows $V = x_j^\perp$ for some j. Since $x_j \notin S$ by assumption, conclude $S^\perp \not\subseteq x_j^\perp = V$.

By assumption, $\exists \bar{\delta} \in S_\delta^\perp$. Consider $\delta_t = (1-t)\delta^\dagger + t\bar{\delta}$ for $t \in [0, 1]$. It holds $\delta_t \in S^\perp$ $\forall t$ since $S^\perp$ is a linear subspace. As $x_j'\delta^\dagger \neq 0$ for $j \in J$ by construction, there is at most one $t_j \in [0, 1]$ such that $x_j'\delta_{t_j} = 0$. Likewise, since $R'\bar{\delta} \neq 0$, there is at most one $t_R \in [0, 1]$ such that $R'\delta_{t_R} = 0$. Thus, as J is finite and $[0, 1]$ is a continuum, $\exists t^* \in [0, 1]$ such that $t^* \neq t_j$ for all $j \in J$ and $t^* \neq t_R$ and such that $\delta_{t^*} \in S_\delta^\perp$. □

Lemma A.12. *Suppose $1/\min_{j \in \zeta_\circ^c} \varepsilon_j^2 = o_p(1)$. Then there exists a deterministic sequence $b_n^2 \to \infty$ such that $b_n^2/\min_{j \in \zeta_\circ^c} \varepsilon_j^2 = o_p(1)$.*

Proof. Write $z_n^2 = 1/\min_{j\in\zeta_\circ^c}\varepsilon_j^2$. Since $z_n^2 \to_p 0$ by assumption, for every $k\in\mathbb{N}$ there exists $N_k\in\mathbb{N}$ such that $P(z_n^2 \le 1/k) \ge 1-1/k$ for $n\ge N_k$. Define the sequence k_n by $k_n = k$ for $n = N_k,\dots,N_{k+1}-1$. It holds $k_n\to\infty$ and therefore $b_n^2 = \log k_n\to\infty$. For any $\epsilon>0$, there exists k^* such that $\log(k)/k\le\epsilon$ for all $k\ge k^*$. Then for all n such that $k_n\ge k^*$

$$P(b_n^2 z_n^2\le\epsilon) = P(z_n^2\le\epsilon/\log k_n)\ge P(z_n^2\le 1/k_n)\ge 1-1/k_n\to 1. \qquad \square$$

Proof of Theorem 3. We check that Assumptions 1 and 7 imply Assumption 9. Theorem 3 then follows from Lemma A.10.

Assumptions $9(i,ii,iii)$ are given in the statements of Assumptions 1 and 7.

Assumption $9(iv)$ holds by Lemma A.12 and Assumption7(i).

We show that Assumption 6, 7(i), 8. imply Assumption $9(v)$. The proof of this is closely related to that of Lemma A.7. The argument is structured as follows. First, we fix constants $\nu,\tau>0$. Second, we define a decomposition of the measure P_{good}. Third, we construct sets Ω_n with a prescribed set of properties. Fourth, we show that on the sets Ω_n it holds

$$K_{n\tau}(a) = \max_{\underline{h}\le h\le h_\circ}\max_{\zeta:|\zeta|=h}\max_{\delta:|\delta|=1,\,|R'\delta|>0}\left\{H_{\delta\zeta}(a)+\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n)-h\right\}\le -n\nu \tag{65}$$

for all $a>0$ small and n large. Finally, we argue that for any $\epsilon>0$ the sets (Ω_n) can be constructed so that $P(\Omega_n)\ge 1-\epsilon$ for large n. This shows that Assumption $9(v)$ holds.

0. Choice of constants $\nu,\tau>0$. By Assumption 8, we can choose $\nu>0$ such that

$$\sup_{\delta:|\delta|=1,|R'\delta|>0}\{\lambda_\circ P_{good}\{x: x'\delta=0\}-(1-\lambda_\circ)P_{out}\{x:x'\delta=0\}\}<\underline{\lambda}+\lambda_\circ-1-3\nu. \tag{66}$$

Let $\tau,\epsilon'>0$ be small so that $\tau+5\epsilon'<\nu$.

1. Construction of measure μ. Let measures $\{\mu_k\}_{k=1}^\infty$ and distinct affine sets $\{A_k\}_{k=1}^\infty$ of affine dimensions $\{m_k\}_{k=1}^\infty$ be given by Lemma A.4. They satisfy (i) $P_{good}=\sum_{k=1}^\infty\mu_k$, (ii) $\mu_k(\mathbb{R}^p\setminus A_k)=0$, and (iii) $\mu_k(B)=0$ for any measurable set B with affine dimension less than m_k. From property (i), there exists $\mu=\sum_{k=1}^K\mu_k$ for some finite K such that $\mu(B)\le P_{good}(B)\le\mu(B)+\epsilon'$ for any measurable set B.

2. Choice of constant c. Define $S_k=\mathrm{span}(A_k)$ for $1\le k\le K$. By Lemma A.6, for every non-empty subset $N\subseteq\{1,\dots,K\}$ with $S_N=\mathrm{span}(\cup_{k\in N}S_k)\ne\{0\}$ there exists a matrix B_N and a constant $0<c_N=\{\mathrm{mineig}(B_N'B_N)/p\}^{1/2}$ such that for any δ

$$|\mathrm{proj}_S(\delta)|\le\max_{k\in N}|\mathrm{proj}_{S_k}(\delta)|/c_N. \tag{67}$$

and

$$\begin{aligned}0\le\max_{k\in N}|\mathrm{proj}_{S_k}(\delta)|<|R'\delta|c_N\text{ for some }|\delta|=1\\ \implies S_N^\perp\cap\{\delta:|\delta|=1,\,|R'\delta|>0\}\ne\varnothing.\end{aligned} \tag{68}$$

Take $0<c=\min_{N\subseteq\{1,\dots,K\}:S_N\ne\{0\}}c_N$.

3. Choice of constant a^*. By Lemma A.5, there exists $a^* > 0$ such that

$$\max_{k\le K} \sup_{\delta\in S_k:|\delta|=1} \mu_k\{x : |x'\delta| \le a^*/c\} \le \epsilon'/K. \tag{69}$$

4. Construction of Ω_n. We require that on Ω_n it holds for all $a \ge 0$ and δ with $|\delta| = 1$

$$h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{|x_{in}'\delta| \le a\} \le P_{good}\{x : |x'\delta| \le a\} + \epsilon' \tag{70}$$

$$(n - h_\circ)^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{x_{jn}'\delta = 0\} \ge P_{out}\{x : |x'\delta| \le a\} - \epsilon'. \tag{71}$$

Let $\{|x_{jn}|\}_{j\in\zeta_\circ^c}$ have ordered values $\gamma_1, \ldots, \gamma_{n-h_\circ}$. There exists $b_n \to \infty$ such that on Ω_n

$$\gamma_{\lceil (n-h_\circ)(1-\epsilon')\rceil} \le b_n a^*. \tag{72}$$

5.1. Bound for $H_{\delta\zeta}(a)$ on Ω_n. Consider any $0 < a < a^*$. We find a uniform bound on the term $H_{\delta\zeta}(a)$ in (65). Define $\kappa_{\delta\zeta\tau}(a) = H_{\delta\zeta}^{-1}(H_{\delta\zeta}(a) - \tau n)$. It holds $H(\kappa_{\delta\zeta\tau}(a)) \ge H_{\delta\zeta}(a) - \tau n$ by Remark A.3, and therefore $H_{\delta\zeta}(a) - H_{\delta\zeta}(\kappa_{\delta\zeta\tau}(a)) \le \tau n$. For all $\underline{h} \le h \le h_\circ$, h-subsets ζ, and $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ we get

$$H_{\delta\zeta}(a) = H_{\delta\zeta}(\kappa_{\delta\zeta\tau}(a)) + H_{\delta\zeta}(a) - H_{\delta\zeta}(\kappa_{\delta\zeta\tau}(a)) \le H_{\delta\zeta}(\kappa_{\delta\zeta\tau}(a)) + \tau n.$$

Using first the definition of $H_{\zeta\delta}$ and then (70) gives

$$H_{\delta\zeta}(\kappa_{\delta\zeta\tau}(a)) \le \sum_{i\in\zeta_\circ} \mathbb{I}\{|x_{in}'\delta| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\} \le h_\circ P_{good}\{x : |x'\delta| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\} + h_\circ\epsilon'.$$

Since $P_{good}(B) \le \sum_{k\le K} \mu_k(B) + \epsilon'$ holds for any measurable B (part 1), then for all h, ζ, δ

$$H_{\delta\zeta}(a) \le h_\circ \sum_{k\le K} \mu_k(\{x : |x'\delta| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\}) + (2h_\circ\epsilon' + \tau n). \tag{73}$$

Let $\delta_k = \mathrm{proj}_{S_k}(\delta)$, $\delta_k^\perp = \mathrm{proj}_{S_l^\perp}(\delta)$, and $N_{\delta\zeta} = \{k \le K : |\delta_k| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|(c/a^*)\}$. Consider first $k \notin N_{\delta\zeta}$. Since $\delta = \delta_k + \delta_k^\perp$, if $x \in S_k$ then $x'\delta = x'\delta_k$. As $\mu_k(\mathbb{R}^p \setminus S_k) \le \mu_k(\mathbb{R}^p \setminus A_k) = 0$ by property (ii) in part 1, it follows

$$\mu_k\{x : |x'\delta| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\} = \mu_k\{x : |x'\delta_k| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\}.$$

For $k \notin N_{\delta\zeta}$ it holds $|\delta_k| > 0$ and we write $\delta_k^\circ = \delta_k/|\delta_k|$. Divide both sides of $|x'\delta_k| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|$ by $|\delta_k|$ and use $\kappa_{\delta\zeta\tau}(a)|R'\delta|/|\delta_k| < a^*/c$ holding by definition of $N_{\delta\zeta}$ to get

$$\mu_k\{x : |x'\delta_k| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\} \le \mu_k\{x : |x'\delta_k^\circ| \le a^*/c\}.$$

Bound the right-hand-side using (69) to conclude that for all $k \notin N_{\delta\zeta}$

$$\mu_k\{x : |x'\delta| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\} \le \epsilon'/K.$$

Insert this into (73) and use $|N^c_{\delta\zeta}| \le K$ to get

$$H_{\delta\zeta}(a) \le h_\circ \sum_{k\in N_{\delta\zeta}} \mu_k(\{x : |x'\delta| \le \kappa_{\delta\zeta\tau}(a)|R'\delta|\}) + (3h_\circ\epsilon' + \tau n).$$

Writing $S_{\delta\zeta} = span(\cup_{k\in N_{\delta\zeta}} S_k)$ and using properties (i, ii) of μ show that for any measurable B it holds $\sum_{k\in N_{\delta\zeta}} \mu_k(B) \le \sum_{k\in N_{\delta\zeta}} \mu_k(S_{\delta\zeta}) \le P_{good}(S_{\delta\zeta})$. Thus, for all $\underline{h} \le h \le h_\circ$, h-subsets ζ, and $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ it holds

$$H_{\delta\zeta}(a) \le h_\circ P_{good}(\mathcal{S}_{\delta\zeta}) + (3h_\circ\epsilon' + n\tau). \tag{74}$$

5.2. Bound for $\bar{G}_{\delta\zeta}$ on Ω_n. Use $\kappa_{\delta\zeta\tau} = H^{-1}_{\delta\zeta}(|\zeta \cap \zeta_\circ| - n\tau) \ge H^{-1}_{\delta\zeta}(H_{\delta\zeta}(a) - n\tau) = \kappa_{\delta\zeta\tau}(a)$ to get

$$\begin{aligned}\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n) &= \sum_{j\in\zeta\cap\zeta^c_\circ} \mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}b_n|R'\delta|\} \le \sum_{j\in\zeta^c_\circ} \mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|\} \\ &\le \sum_{j\in\zeta^c_\circ} \mathbb{I}\{x_j \in S_{\delta\zeta}\}\mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|\} + \sum_{i\in\zeta^c_\circ} \mathbb{I}\{x_j \notin S_{\delta\zeta}\}. \end{aligned} \tag{75}$$

Consider the first sum in (75). If $x \in S_{\delta\zeta}$ then $|x'\delta| = |x'\mathrm{proj}_{S_{\delta\zeta}}(\delta)|$ and by Cauchy-Schwarz inequality $|x'\delta| \le |x||\mathrm{proj}_{S_{\delta\zeta}}(\delta)|$. By (67) and definition of $N_{\delta\zeta}$ it holds $|\mathrm{proj}_{S_{\delta\zeta}}(\delta)| \le \max_{k\in N_{\delta\zeta}} |\delta_k|/c \le \kappa_{\delta\zeta\tau}(a)|R'\delta|/a^*$. Thus, for $x \in S_{\delta\zeta}$

$$|x'\delta| \le |x|\,|\mathrm{proj}_{S_{\delta\zeta}}(\delta)| \le |x|\kappa_{\delta\zeta\tau}(a)|R'\delta|/a^*. \tag{76}$$

From (76), for $x \in S_{\delta\zeta}$ it holds $|x'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|$ only if $|x| > b_n a^*$. Thus,

$$\mathcal{I}_{\delta\zeta\tau} = \sum_{i\in\zeta^c_\circ} \mathbb{I}\{x_j \in S_{\delta\zeta}\}\mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|\} \le \sum_{i\in\zeta^c_\circ} \mathbb{I}\{|x_j| > b_n a^*\}. \tag{77}$$

Use $\gamma_{\lceil(n-h_\circ)(1-\epsilon')\rceil} \le b_n a^*$ from (72) to get

$$\mathcal{I}_{\delta\zeta\tau} \le \sum_{i\in\zeta^c_\circ} \mathbb{I}\{|x_j| > b_n a^*\} \le \mathbb{I}\{|x_j| > \gamma_{\lceil(n-h_\circ)(1-\epsilon')\rceil}\} \le (n - h_\circ)\epsilon'. \tag{78}$$

Insert (78) back into (75) to conclude that for all h, ζ, δ

$$\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n) \le (n-h_\circ)\epsilon' + \sum_{i\in\zeta^c_\circ} \mathbb{I}\{x_j \notin S_{\delta\zeta}\} = (n-h_\circ) - \sum_{i\in\zeta^c_\circ} \mathbb{I}\{x_j \in S_{\delta\zeta}\} + (n-h_\circ)\epsilon'. \tag{79}$$

5.3. Deterministic analysis on Ω_n - Conclusion. For $\underline{h} \le h \le h_\circ$ define

$$K_{n\tau h}(a) = \max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1,|R'\delta|>0} \left\{H_{\delta\zeta}(a) + \bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n) - h\right\} = \max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1,|R'\delta|>0} \mathcal{E}_{\delta\zeta}.$$

Plug-in (74) and (79) to see that with $\epsilon^*_n = 3h_\circ\epsilon' + \tau n + (n - h_\circ)\epsilon'$ it holds

$$\mathcal{E}_{\delta\zeta} = h_\circ P_{good}(S_{\delta\zeta}) + (n - h_\circ) - \sum_{j\in\zeta^c_\circ} \mathbb{I}\{x_j \in S_{\delta\zeta}\} - h + \epsilon^*_n, \tag{80}$$

for all $\zeta \in \{\zeta : |\zeta| = h\}$ and $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$.

By construction of $N_{\delta\zeta}$, it holds $\max_{k\in N_{\delta\zeta}} |\delta_k| \leq \kappa_{\delta\zeta\tau}(a)|R'\delta|(c/a^*)$. Since $\kappa_{\delta\zeta\tau}(a) \leq a$ by definition and $a < a^*$ we get further $\max_{k\in N_{\delta\zeta}} |\delta_k| < |R'\delta|c$. Therefore, by our choice of c and (68), it follows that $S^{\perp}_{\delta\zeta} \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \emptyset$. Further, by Lemma A.11, there exists $\delta^*_\zeta \in S^{\perp}_{\delta\zeta} \cap \{\delta : |\delta| = 1, |R'\delta| > 0\}$ such that $x_j'\delta^*_\zeta \neq 0$ for all $j \in \zeta^c_\circ$ with $x_j \notin S_{\delta\zeta}$. The vector δ^*_ζ then satisfies

$$P_{good}(S_{\delta\zeta}) \leq P_{good}\{x : x'\delta^*_\zeta = 0\}, \sum_{j\in\zeta^c_\circ} \mathbb{I}\{x_j \in S_{\delta\zeta}\} = \sum_{j\in\zeta^c_\circ} \mathbb{I}\{x_j'\delta^*_\zeta = 0\}. \tag{81}$$

Plug (81) in (80) and use (71) to conclude

$$\begin{aligned}\mathcal{E}_{\delta\zeta} &\leq h_\circ P_{good}\{x : x'\delta^*_\zeta = 0\} + (n - h_\circ) - \sum_{j\in\zeta^c_\circ} \mathbb{I}\{x_j'\delta^*_\zeta = 0\} - h + \epsilon^*_n \\ &\leq h_\circ P_{good}\{x : x'\delta^*_\zeta = 0\} + (n - h_\circ) - (n - h_\circ)P_{out}\{x : x'\delta^*_\zeta = 0\} - h + 2\epsilon^*_n. \end{aligned} \tag{82}$$

Since $h_\circ/n = \lambda_\circ + o(1)$ by assumption, using (66) we can bound $h_\circ P_{good}\{x : x'\delta^*_\zeta = 0\} - (n - h_\circ)P_{out}\{x : x'\delta^*_\zeta = 0\} \leq n\{\underline{\lambda} + \lambda_\circ - 1 - 3v + o(1)\}$. Insert this in (82) and use $n - h_\circ - h \leq -n\{\underline{\lambda} + \lambda_\circ - 1 + o(1)\}$ to conclude

$$K_{n\tau h}(a) = \max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1,\,|R'\delta|>0} \mathcal{E}_{\delta\zeta} \leq n\{2\epsilon^*_n/n - 3\nu + o(1)\}.$$

It holds $2\epsilon^*_n/n \leq 8\epsilon' + 2\tau + o(1) \leq \nu$ for large n by our choice of τ, ϵ'. Conclude that for large n

$$K_{n\tau h}(a) \leq n\{-2\nu + o(1)\} \leq -n\nu.$$

Since the bounds are uniform in $\underline{h} \leq h \leq h_\circ$, we get (65).

3. Probability Analysis Construct $\Omega_n = \Omega_{1n} \cap \Omega_{2n} \cap \Omega_{3n}$. Using Assumption 6 and Lemma A.8, we can find sets (Ω_{2n}) with $P(\Omega_{2n}) \geq 1 - \epsilon$ for n so that (70) and (71) hold on Ω_{2n}. Assumption 6 further gives that $\{|x_j|\}_{j\in\zeta^c_\circ}$ is drawn independently from some common distribution. Therefore, the order statistic $\gamma_{\lceil(n-h_\circ)(1-\epsilon')\rceil}$ is bounded in probability. Combining with $b_n \to \infty$, we see that for every $\epsilon > 0$ we can find (Ω_{3n}) so that $P(\Omega_{3n}) \geq 1 - \epsilon$ for large n so that (72) holds on Ω_{3n}. □

A.6 Proof of Theorem 4

Lemma A.13. *Suppose Assumption 6. Then* $\sup_{\delta:|\delta|=1} |\hat{p}_{full,\delta} - p_{full,\delta}| \to 0$ *a.s.*

Proof. Define $\hat{P}_{full}(B) = n^{-1}\sum_{i=1}^n \mathbb{I}\{x_{in} \in B\}$ for any measurable B. Using $h_\circ/n = \lambda_\circ + o(1)$, for any measurable set B it holds

$$\begin{aligned}P_{full}(B) &= \frac{h_\circ}{n} h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{x_{in} \in B\} + \frac{n - h_\circ}{n}(n - h_\circ)^{-1} \sum_{i\in\zeta^c_\circ} \mathbb{I}\{x_{in} \in B\} \\ &= \lambda_\circ h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{x_{in} \in B\} + (1 - \lambda_\circ)(n - h_\circ)^{-1} \sum_{i\in\zeta^c_\circ} \mathbb{I}\{x_{in} \in B\} + o(1).\end{aligned}$$

By Assumption 6, for any B it holds with probability one that

$$\sum_{i\in\zeta_\circ}\mathbb{I}\{x_{in}\in B\}=\sum_{i=1}^{h_\circ}\mathbb{I}\{x_i^{\text{good}}\in B\}\to P_{good}(B) \tag{83}$$

$$\sum_{i\in\zeta_\circ^c}\mathbb{I}\{x_{in}\in B\}=\sum_{i=1}^{n-h_\circ}\mathbb{I}\{x_i^{\text{out}}\in B\}\to P_{out}(B).$$

Thus, $\hat{P}_{good}(B)\to\lambda_\circ P_{good}(B)+(1-\lambda_\circ)P_{out}=P_{full}(B)$ with probability one. By Lemma A.8, we conclude $\sup_{\delta:|\delta|=1}|\hat{p}_{full,\delta}-p_{full,\delta}|\to 0$ a.s. □

Proof of Theorem 4. 1. Population counterpart. Let $\underline{\boldsymbol{\lambda}}=\max\{\underline{\boldsymbol{\lambda}}^{(1)}+\epsilon/2,1-\underline{\lambda}_\circ\}$, where

$$\underline{\boldsymbol{\lambda}}^{(1)}=\sup_{\delta:|\delta|=1,|R'\delta|>0}\{p_{full,\delta}+(1-\lambda_\circ)(1-b_{out,\delta})\}. \tag{84}$$

We show $\underline{\boldsymbol{\lambda}}$ satisfies Assumption 3. To do this, we check $1-\lambda_\circ<\underline{\boldsymbol{\lambda}}\leq\lambda_\circ$ and (4).

Begin by checking (4). Assumption 6 implies Assumption 4. Thus, by Proposition 3.2, (4) holds if Assumption 5 is satisfied, that is

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0}\lambda_\circ p_{good,\delta}<\underline{\boldsymbol{\lambda}}+\lambda_\circ-1. \tag{85}$$

Using identity $\lambda_\circ p_{good,\delta}=p_{full,\delta}-(1-\lambda_\circ)p_{out,\delta}$ and subtracting $(\lambda_\circ-1)$ from both sides, (85) holds if

$$\sup_{\delta:|\delta|=1,|R'\delta|>0}\{p_{full,\delta}+(1-\lambda_\circ)(1-p_{out,\delta})\}<\underline{\boldsymbol{\lambda}}. \tag{86}$$

The left-hand-side is decreasing in $\lambda_\circ$ and $p_{out,\delta}$. Since $\underline{\lambda}_\circ<\lambda_\circ$ and $b_{out,\delta}\leq p_{out,\delta}$ $\forall\delta:|R'\delta|>0$ by assumption, (84) shows

$$\{p_{full,\delta}+(1-\lambda_\circ)(1-p_{out,\delta})\}\leq\{p_{full,\delta}+(1-\underline{\lambda}_\circ)(1-b_{out,\delta})\}\leq\underline{\boldsymbol{\lambda}}^{(1)}$$

$\forall\delta:|R'\delta|>0$. Since $\underline{\boldsymbol{\lambda}}=\max\{\underline{\boldsymbol{\lambda}}^{(1)}+\epsilon/2,1-\underline{\lambda}_\circ\}$, (86) holds.

It remains to check (i) $1-\lambda_\circ<\underline{\boldsymbol{\lambda}}$ and (ii) $\underline{\boldsymbol{\lambda}}\leq\lambda_\circ$. Since $1-\underline{\lambda}_\circ\leq\underline{\boldsymbol{\lambda}}$ by definition and $\lambda_\circ>\underline{\lambda}_\circ$ by assumption, (i) holds. By definition of $\underline{\boldsymbol{\lambda}}$, (ii) holds if $\underline{\boldsymbol{\lambda}}^{(1)}+\epsilon/2\leq\lambda_\circ$ and $1-\underline{\lambda}_\circ\leq\lambda_\circ$. The latter holds as $1-\underline{\lambda}_\circ\leq 1/2<\lambda_\circ$ by step 1 and Assumption 6. By the assumption $\underline{\lambda}_\circ\leq\lambda_\circ-\epsilon$, the former holds if $\underline{\boldsymbol{\lambda}}^{(1)}\leq\underline{\lambda}_\circ$, that is if, by (84),

$$p_{full,\delta}+(1-\underline{\lambda}_\circ)(1-b_{out,\delta})\leq\underline{\lambda}_\circ \tag{87}$$

$\forall\delta:|R'\delta|>0$. Rearranging, (87) is equivalent to

$$\frac{1+p_{full,\delta}-b_{out,\delta}}{2-b_{out,\delta}}\leq\underline{\lambda}_\circ. \tag{88}$$

The supremum of the left-hand-side in (88) over $\delta\in\{\delta:|\delta|=1,|R'\delta|>0\}$ is denoted $\underline{\lambda}_\circ^{(1)}$ in (8), and satisfies $\underline{\lambda}_\circ^{(1)}\leq\underline{\lambda}_\circ$ by assumption.

2. Conclusion. Consider $\underline{h}=\lfloor\underline{\lambda}n\rfloor$ and $\underline{\lambda}=\max\{\underline{\lambda}^{(1)}+\epsilon,1-\underline{\lambda}_\circ\}$, where $\underline{\lambda}^{(1)}$ is given by (7). Define $\underline{\boldsymbol{h}}=\lfloor\underline{\boldsymbol{\lambda}}n\rfloor$, where $\underline{\boldsymbol{\lambda}}=\max\{\underline{\boldsymbol{\lambda}}^{(1)}+\epsilon/2,1-\underline{\lambda}_\circ\}$ as above. By Assumption 6 and Lemma A.13, $\sup_\delta|\hat{p}_{full,\delta}-\hat{p}_{full,\delta}|\to 0$ a.s. Thus also $\underline{\lambda}^{(1)}\to\underline{\boldsymbol{\lambda}}^{(1)}$ a.s., implying $P(\underline{h}\geq\underline{\boldsymbol{h}})\to 1$. By part 1, $\underline{\boldsymbol{\lambda}}^{(1)}$ satisfies Assumption 3. Assumption 1 and Theorem 2 then show $\max_{\underline{\boldsymbol{h}}\leq h\leq h_\circ}\max_{\zeta\in\mathcal{M}_h}|R'\hat{\beta}_\zeta|=O_p(1)$. Since $P(\underline{h}\geq\underline{\boldsymbol{h}})\to 1$, conclude $\max_{\underline{h}\leq h\leq h_\circ}\max_{\zeta\in\mathcal{M}_h}|R'\hat{\beta}_\zeta|=O_p(1)$. □

A.7 Proof of Theorem 5

Proof. 1. Population counterpart. Let $\underline{\boldsymbol{\lambda}} = \max\{\underline{\boldsymbol{\lambda}}^{(2)} + \epsilon/2, 1 - \underline{\lambda}_\circ\}$, where

$$\underline{\boldsymbol{\lambda}}^{(2)} = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} \left[p_{full,\delta} + \mathbb{I}\{b_{out,\delta} < 1/2\}(1 - \underline{\lambda}_\circ)(1 - 2b_{out,\delta})\right]. \tag{89}$$

We show $\underline{\boldsymbol{\lambda}}$ satisfies Assumption 8. We need to check $1 - \lambda_\circ < \underline{\boldsymbol{\lambda}} \leq \lambda_\circ$ and

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} \{\lambda_\circ p_{good,\delta} - (1 - \lambda_\circ)p_{out,\delta}\} < \underline{\boldsymbol{\lambda}} + \lambda_\circ - 1. \tag{90}$$

We check (90). Insert $\lambda_\circ p_{good,\delta} = p_{full,\delta} - (1 - \lambda_\circ)p_{out,\delta}$ and subtract $\lambda_\circ - 1$ from both sides to see (90) holds if

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} \{p_{full,\delta} + (1 - \lambda_\circ)(1 - 2p_{out,\delta})\} < \underline{\boldsymbol{\lambda}}. \tag{91}$$

The left-hand-side is decreasing in $p_{out,\delta}$. Since $b_{out,\delta} \leq p_{out,\delta}$ $\forall \delta : |R'\delta| > 0$ by assumption, (91) holds if

$$\sup_{\delta:|\delta|=1,|R'\delta|>0} W_\delta = \sup_{\delta:|\delta|=1,|R'\delta|>0} \{p_{full,\delta} + (1 - \lambda_\circ)(1 - 2b_{out,\delta})\} < \underline{\boldsymbol{\lambda}}. \tag{92}$$

For $b_{out,\delta} < 1/2$, W_δ is decreasing in $\lambda_\circ$. Since $\underline{\lambda}_\circ \leq \lambda_\circ$ by assumption then $W_\delta \leq p_{full,\delta} + (1 - \underline{\lambda}_\circ)(1 - 2b_{out,\delta})$. For $b_{out,\delta} \geq 1/2$, W_δ is increasing in $\lambda_\circ$. Since $\lambda_\circ \leq 1$ then $W_\delta \leq p_{full,\delta}$. Put together, it follows $W_\delta \leq \underline{\boldsymbol{\lambda}}^{(2)}$. Since $\underline{\boldsymbol{\lambda}} \geq \underline{\boldsymbol{\lambda}}^{(2)} + \epsilon/2$, (92) holds.

It remains to check (i) $1 - \lambda_\circ < \underline{\boldsymbol{\lambda}}$ and (ii) $\underline{\boldsymbol{\lambda}} \leq \lambda_\circ$. Since $1 - \underline{\lambda}_\circ \leq \underline{\boldsymbol{\lambda}}$ by definition and $\lambda_\circ > \underline{\lambda}_\circ$ by assumption, (i) holds. By definition of $\underline{\boldsymbol{\lambda}}$, (ii) holds if $\underline{\boldsymbol{\lambda}}^{(2)} + \epsilon/2 \leq \lambda_\circ$ and $1 - \underline{\lambda}_\circ \leq \lambda_\circ$. The latter holds since $1 - \underline{\lambda}_\circ \leq 1/2 < \lambda_\circ$ by assumption. By the assumption $\underline{\lambda}_\circ \leq \lambda_\circ - \epsilon$, the former holds if $\underline{\boldsymbol{\lambda}}^{(2)} \leq \underline{\lambda}_\circ$, that is, by (89), if

$$p_{full,\delta} + \mathbb{I}\{b_{out,\delta} < 1/2\}(1 - \underline{\lambda}_\circ)(1 - 2b_{out,\delta}) \leq \underline{\lambda}_\circ \tag{93}$$

$\forall \delta : |R'\delta| > 0$. For δ with $b_{out,\delta} < 1/2$, rearrange to see that (93) holds if

$$\frac{1 + p_{full,\delta} - 2b_{out,\delta}}{2(1 - b_{out,\delta})} \leq \underline{\lambda}_\circ.$$

For $b_{out,\delta} \geq 1/2$, (93) holds if $p_{full,\delta} \leq \lambda_\circ$. Combined, (93) holds if $\forall \delta : |R'\delta| > 0$

$$\mathbb{I}\{b_{out,\delta} < 1/2\}\frac{1 + p_{full,\delta} - 2b_{out,\delta}}{2(1 - b_{out,\delta})} + \mathbb{I}\{b_{out,\delta} \geq 1/2\}p_{full,\delta} \leq \underline{\lambda}_\circ. \tag{94}$$

Supremum of the left-hand-side in (94) over $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ is denoted $\underline{\lambda}_\circ^{(2)}$ in (11). This satisfies $\underline{\lambda}_\circ^{(2)} \leq \underline{\lambda}_\circ$ by assumption.

2. Conclusion. Consider $\underline{h} = \lfloor \underline{\lambda} n \rfloor$ and $\underline{\lambda} = \max\{\underline{\lambda}^{(2)} + \epsilon, 1 - \underline{\lambda}_\circ\}$, where $\underline{\lambda}^{(2)}$ is given by (10). Define $\underline{\boldsymbol{h}} = \lfloor \underline{\boldsymbol{\lambda}} n \rfloor$, where $\underline{\boldsymbol{\lambda}} = \max\{\underline{\boldsymbol{\lambda}}^{(2)} + \epsilon/2, 1 - \underline{\lambda}_\circ\}$ as above. By Assumption 6 and Lemma A.13, $\sup_\delta |\hat{p}_{full,\delta} - \hat{p}_{full,\delta}| \to 0$ a.s. Thus also $\underline{\lambda}^{(2)} \to \underline{\boldsymbol{\lambda}}^{(2)}$ a.s., implying $P(\underline{h} \geq \underline{\boldsymbol{h}}) \to 1$. By part 1, $\underline{\boldsymbol{\lambda}}^{(2)}$ satisfies Assumption 8. Assumptions 1, 6, 7 and Theorem 3 then show $\max_{\underline{\boldsymbol{h}} \leq h \leq h_\circ} \max_{\zeta \in \mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$. Since $P(\underline{h} \geq \underline{\boldsymbol{h}}) \to 1$, conclude $\max_{\underline{h} \leq h \leq h_\circ} \max_{\zeta \in \mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$. □

B Hyperplane search algorithms

We mention some algorithms for locating hyperplanes with many observations. Such algorithms are needed to check conditions for boundedness of LTS estimators (see Section 3.4) and to choose an initial h together with Algorithms 1 and 2. Implementations in R can be found in the replication materials for the empirical illustration.

B.1 Exhaustive search

Let $\mathcal{X} = \{x \in \mathbb{R}^p : x_i = x \text{ for some } 1 \leq i \leq n\}$ be the support of the regressors. We can exhaustively search over all hyperplanes containing regressors by enumerating subsets of size $p-1$ from $\mathcal{X}$. Mili and Coakley (1996) mentions this in passing, but we are not aware of the details or of an available implementation of their suggestion.

Exhaustive search is outlined in Algorithm 3. Algorithm 3 returns the number of observations in the largest hyperplane, but it could be adjusted to save the k largest hyperplanes for $k > 1$ and their orthogonal vectors.

The number of $(p-1)$-subsets is of order $|\mathcal{X}|^{p-1}$, so Algorithm 3 quickly becomes infeasible as the dimension grows. Two further issues arise. First, some $(p-1)$-subsets do not span a hyperplane, so we include dimension checking in the algorithm. Second, different $(p-1)$-subsets may span the same hyperplane. To avoid repeated checks, we record all subsets that lie in a hyperplane once it has been visited, at the cost of increased memory usage.

Algorithm 3 (Exhaustive hyperplane search)**.**
(0) Initialise $m^* \leftarrow 0$ *and* $\mathcal{I}_{skip} \leftarrow \emptyset$.
(1) Let $\iota : \{1, \dots, \binom{|\mathcal{X}|}{p-1}\} \to [\mathcal{X}]^{p-1}$ *be a bijection for enumerating* $(p-1)$*-subsets of* $\mathcal{X}$ *and* $m(x)$ *the frequency of value* $x \in \mathcal{X}$. *Initialise* $\mathcal{I}_{skip} \leftarrow \emptyset$ *to track subsets contained in a previously visited hyperplane.*
(2) For $i = 1, \dots, \binom{|\mathcal{X}|}{p-1}$*:*
  If $i \in \mathcal{I}_{skip}$ *or* $\dim \iota(i) < p-1$, *continue.*
  Let $X_H \leftarrow span\{\iota(i)\} \cap \mathcal{X}$ *and* $m_H \leftarrow \sum_{x \in X_H} m(x)$.
  Get the $(p-1)$*-subsets* $\mathcal{I}_{add} \leftarrow \{\iota^{-1}(W) : W \in [X_H]^{p-1}\}$ *contained in the current hyperplane. Update* $\mathcal{I}_{skip} \leftarrow \mathcal{I}_{skip} \cup \mathcal{I}_{add}$.
  If $m_H > m^*$ *then update* $m^* \leftarrow m_H$.
(3) Return m^*.

B.2 Randomised search

Algorithm 4 searches for the number of observations on the largest hyperplane by drawing random $(p-1)$-subsets from the support $\mathcal{X}$. To locate hyperplanes with many observations, it weights $x \in \mathcal{X}$ by its frequency $m(x) = \sum_{i=1}^n \mathbb{I}\{x_i = x\}$ in the random sampling. Algorithm 4 could be adapted to return the k largest hyperplanes for $k > 0$ and orthogonal vectors.

Algorithm 4 (Randomised hyperplane search)**.**
(0) Initialise $m^* \leftarrow 0$.

(1) Let $m(x)$ be the frequency and $w(x) = m(x)/\sum_{x\in\mathcal{X}} m(x)$ the share of value $x \in \mathcal{X}$.
(2) For $1, \ldots, n_{iter}$:
Initialise current subset $S \leftarrow \emptyset$ and rank $r \leftarrow 0$.
While $r < p - 1$:
Let $S^{\perp} \leftarrow \mathcal{X} \setminus span(S)$ be values linearly independent of S.
Draw a random element of $S^{\perp}$ with weights w and add this to S.
Update $r \leftarrow r + 1$.
Let $X_H = span(S) \cap \mathcal{X}$ and $m_H \leftarrow \sum_{x\in X_H} m(x)$
If $m_H > m^*$ then update $m^* \leftarrow m_H$.
(3) Return m^*.

Algorithm 4 ensures that each randomly drawn subset spans a unique hyperplane by sequentially drawing linearly independent elements. The 'non-singular subsampling' algorithm of Koller and Stahel (2017) could also be adapted for this purpose, but our implementation opts for a simpler solution.

B.3 Heuristics

We note the following heuristics for finding hyperplanes with many observations.

(1) It is often useful to tabulate the frequency of each value from the support of the regressors. With p binary regressors, any hyperplane contains at most 2^{p-1} points. The sum of 2^{p-1} most frequent values then upper bounds the largest hyperplane.

(2) Marginal distributions often help locate large hyperplanes. They reveal 'continuous' variables, which can typically be ignored in the search. Under mutual independence of the regressors, marginal distributions find the largest hyperplane exactly, as formalised in the following.

Remark B.1. *Consider $x \overset{iid}{\sim} P$, where $x' = (1, z')$ and $z' = (z_1, \ldots, z_k)$ has finite support. If the components of z are mutually independent then* $\sup_{\delta:|\delta|=1} P\{x : x'\delta = 0\} = \max_{1\le j\le k} \max_{a\in\mathbb{R}} P(z_j = a)$.

Proof of Remark B.1. Let $\mathcal{X} = \{x : P(x) > 0\}$ and $\mathcal{X}_j = \{a : P(z_j = a) > 0\}$. Let $a_j = \arg\max_{a\in\mathcal{X}_j} P(z_j = a)$ and $\pi_j(x) = (1, z_1, \ldots, z_{j-1}, a_j, z_{j+1}, \ldots, z_k)$ for $x \in \mathcal{X}$.

We argue that if for $S \subseteq \mathcal{X}$ it holds $|\{\pi_j(x) : x \in S\}| = |\pi_j(S)| < |S|$ $\forall j \in \{1, \ldots, k\}$ then $\text{span}(S)$ has dimension $k + 1$. Note that if $|\pi_j(S)| < |S|$ then $\exists v, w \in S$ differing only in the $(j + 1)$-th coordinate. Thus, e_{j+1}, the $(j + 1)$-th standard basis vector, is in $\text{span}(S)$ $\forall j = 1, \ldots, k$. It follows $\{e_2, \ldots, e_{k+1}\} \subset \text{span}(S)$ and then also $e_1 \in \text{span}(S)$. Conclude that $\text{span}(S)$ has dimension $k + 1$.

Define $S_\delta = \mathcal{X} \cap \{x : x'\delta = 0\}$ for all δ with $|\delta| = 1$. By mutual independence, $P(x) \le P\{\pi_j(x)\}$ $\forall x \in \mathcal{X}$. Since S_δ has at most dimension k, by the above observation $\exists j$ such that $|\pi_j(S_\delta)| = |S_\delta|$. Conclude

$$P(S_\delta) = \sum_{x\in S_\delta} P(x) \le \sum_{x\in S_\delta} P\{\pi_j(x)\} \le P(z_j = a_j) \le \max_{1\le j\le k} P(z_j = a_j),$$

where the second inequality is by the definition of a marginal distribution and $|\pi_j(S_\delta)| = |S_\delta|$ implying there are no repeated terms in the sum $\sum_{x\in S_\delta} P\{\pi_j(x)\}$. □

References

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics*, 7(1):226–248.

Amaldi, E. and Kann, V. (1995). The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Computer Science*, 147(1.2):181–210.

Amin, V. (2011). Returns to education: Evidence from UK twins: Comment. *American Economic Review*, 101(4):1629–1635.

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York, 2nd edition.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society*, 39:117–134.

Axler, S. (2015). *Linear Algebra Done Right*. Springer, New York, 3rd edition.

Berenguer-Rico, V., Johansen, S., and Nielsen, B. (2023). A model where the least trimmed squares estimator is maximum likelihood. *Journal of the Royal Statistical Society, Series B*, 85(3):886–912.

Berenguer-Rico, V. and Nielsen, B. (2025a). Least trimmed squares asymptotics: Cointegration and outliers. Discussion paper, Nuffield College.

Berenguer-Rico, V. and Nielsen, B. (2025b). Least trimmed squares: Nuisance parameter free asymptotics. *Econometric Theory*, pages 1–39.

Billingsley, P. (1995). *Probability and Measure*. John Wiley and Sons, 3rd edition.

Bonjour, D., Cherkas, L. F., Haskel, J. E., Hawkes, D. D., and Spector, T. D. (2003). Returns to education: Evidence from UK twins. *American Economic Review*, 93(5):1799–1812.

Breiman, L. (1992). *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Chen, X. and Wu, Y. (1988). Strong consistency of m-estimates in linear models. *Journal of Multivariate Analysis*, 27(1):116–130.

Cížek, P. (2004). Asymptotics of least trimmed squares regression. CentER Discussion Paper 2004-72, Tilburg University.

Cížek, P. (2013). Reweighted least trimmed squares: An alternative to one-step estimators. *Test*, 22(3):514–533.

Davies, P. L. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics*, 18(4):1651–1675.

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

Davies, P. L. and Gather, U. (2005). Breakdown and groups. *Annals of Statistics*, 33(3):977–1035.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In Bickel, P., Doksum, K., and Hodges, J., editors, *A Festschrift for Erich L. Lehmann*, pages 157–184. Wadsworth, Belmont.

Elker, J., Pollard, D., and Stute, W. (1979). Glivenko–Cantelli theorems for classes of convex sets. *Advances in Applied Probability*, 11(4):820–833.

Hadi, A. S. and Luceño, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*,

25(3):251–272.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424):1264–1272.

Heng, Q. and Lange, K. (2025). Bootstrap estimation of the proportion of outliers in robust regression. *Statistics and Computing*, 35(1):3.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73 – 101.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. John Wiley & Sons.

Johansen, S. and Nielsen, B. (2019). Boundedness of M-estimators for linear regression in time series. *Econometric Theory*, 35(3):653–683.

Koller, M. and Stahel, W. A. (2017). Nonsingular subsampling for regression S estimators with categorical predictors. *Computational Statistics*, 32:631–646.

Lopuhaä, H. P., Gares, V., and Ruiz-Gazen, A. (2023). S-estimation in linear models with structured covariance matrices. *Annals of Statistics*, 51(6):2415–2439.

Mili, L. and Coakley, C. W. (1996). Robust estimation in structured linear regression. *Annals of Statistics*, 24(6):2593–2607.

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics*, 33(2):659–680.

Roman, S. (2005). *Advanced Linear Algebra*, volume 135 of *Graduate Texts in Mathematics*. Springer, New York, 2nd edition.

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

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons.

Rousseeuw, P. J. and van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In Gaul, W., Opitz, O., and Schader, M., editors, *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346. Springer Verlag.

Tableman, M. (1994). The asymptotics of the least trimmed absolute deviations (LTAD) estimator. *Statistics & Probability Letters*, 19(5):387–398.

Víšek, J. Á. (2006). The least trimmed squares. Part III: Asymptotic normality. *Kybernetika*, 42(2):203–224.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15(20):642–656.