Essays on Outlier Robustness

Otso Hao

Nuffield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2026

Acknowledgements

This work would not have been possible without the guidance of my advisor Bent Nielsen. I am greatly indebted for the time and inspiration he has given me. I thank my college advisor Frank Windmeijer for comments on my research and for supporting me with career planning. I also want to thank my examiners James Duffy and Ines Wilms for thorough comments on the thesis.

Chapter 1 is joint work with B. Nielsen and improved with suggestions from V. Berenguer-Rico, J. Doornik, J. Duffy, and F. Windmeijer. Chapter 2 benefited of comments from J. Doornik and M. Meitz. Chapter 3 is joint work with M. Firat and was partially written during an internship at the International Monetary Fund. Chapter 3 was improved with input from Y. Carriere-Swallow, P. Deb, J. Duffy, S. Kothari, J. Peiris, and A. Scott.

I want to thank P. Chaudhury, M. Keloharju, A. de Paula, and A. Cabrales for inspiring me to pursue research during my undergraduate studies and for opening doors to graduate study. I am also grateful to have met other early career researchers in outlier robustness, particularly X. Jiao and J. Klooster.

I am deeply grateful to my parents and brother for their unwavering support. Finally, I thank Essu for sharing this journey with me.

My work was enabled by funding from the Economic and Social Research Council [grant 2261272] and The Ella and Georg Ehrnrooth Foundation.

Word count: 472 words on p.7 × 135 pages = 63 720.

Abstract

This thesis studies outlier robust statistical methods under contamination, motivated by empirical challenges in economics.

Chapter 1 examines the robustness of the Least Trimmed Squares estimator in linear models with categorical covariates. We find uniform boundedness guarantees that apply to a chosen sub-coefficient of interest. We show that LTS is robust in a wider range of settings than suggested by existing boundedness and breakdown point results. We also propose a data-driven approach to choosing an initial LTS tuning parameter, which is useful for methods that estimate the number of outliers.

Chapter 2 develops asymptotic theory for the Impulse Indicator Saturation (IIS) method under contamination. We show the asymptotic equivalence of IIS to an infeasible least squares estimator that perfectly removes all outliers. We use this equivalence to derive the distribution of IIS estimators in cross-sectional and time series models with outliers. We further find a limit theory for the number of misclassified 'clean' observations.

Chapter 3 decomposes inflation to demand- and supply-driven components in a panel of 32 countries, drawing on the theory of IIS estimators developed in Chapter 2 to guard against outliers around the COVID-19 pandemic. We validate the decomposed inflation series by examining their relationship to external measures of demand and supply shocks. The decompositions are used in applications to post-2020 inflation dynamics and Phillips curves analysis.

Contents

List of Figures

List of Tables

Introduction

Outliers are commonly defined as observations that 'deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism' (Hawkins, 1980, p. 1). Outliers often coincide with influential data points, which are observations that have a large effect on the result of a statistical calculation. Classical methods, such as the sample mean for location or least squares for linear models, can be heavily influenced by a single observation. This sensitivity has prompted interest in outlier-robust methods.

While the development of robust methods has received much attention, their analysis under contamination has been more limited and typically focused on robustness measures such as the finite-sample breakdown point (Donoho and Huber, 1983) and influence functions (Hampel, 1974). Meanwhile, consistency and inference are often derived only in the absence of outliers or under infinitesimal contamination.

This thesis studies robust estimators in their natural setting: data containing outliers. The work is inspired by a model with outlier contamination developed in Berenguer-Rico et al. (2023). In a location–scale variant of this model, 'good' observations are independent draws from a normal distribution, while 'outliers' are required to lie outside the range of the realised 'good' sample. The model is distinct from the influential Huber contamination model (Huber, 1964) and provides a framework for deriving new results for robust estimators under contamination.

Chapter 1 is motivated by data from Bonjour et al. (2003) on earnings and labour market characteristics of identical UK twins. Bonjour et al. (2003) use these data to estimate returns to education with a linear model that includes categorical covariates. Their estimates are sensitive to the removal of a small number of observations (Amin, 2011), highlighting the need for an outlier-robust method.

The properties of robust estimators for linear models with categorical covariates are not well understood. Aside from a few studies (Hubert and Rousseeuw, 1996; Hubert and Rousseeuw, 1997; Maronna and Yohai, 2000), existing literature typically

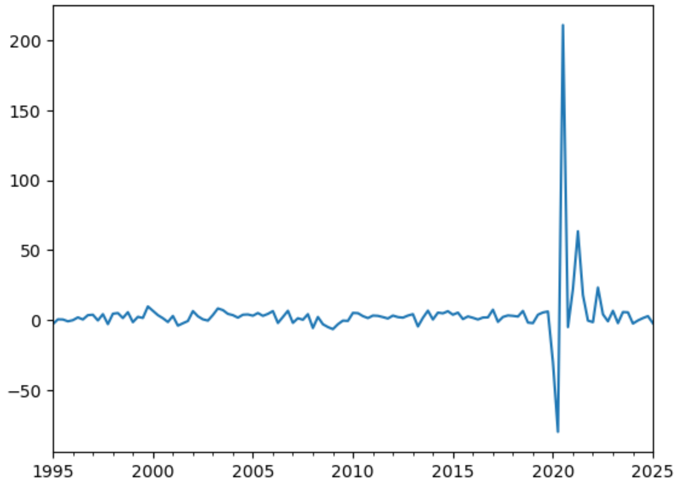


Figure 1: US real PCE on food and accommodation services

assumes all regressors are continuous. Even though just a single binary covariate can drive the breakdown point of a robust estimator to zero (Mili and Coakley, 1996), such properties are often ignored in practice. Chapter 1 refines this understanding, focusing on the Least Trimmed Squares (LTS) estimator (Rousseeuw, 1984).

We find conditions under which LTS is bounded (in probability) when a positive share of observations are outliers. Boundedness is a property that resembles the breakdown point notion of Hampel (1971). The conditions needed to prove boundedness are most binding when some covariates are categorical, which is our focus. We give uniform boundedness guarantees that apply to a chosen sub-coefficient of interest. We show that LTS is robust in a wider range of settings than suggested by existing boundedness and breakdown point results. We also propose a data-driven approach to choosing an initial LTS tuning parameter, which is useful for methods that estimate the number of outliers.

Chapter 2 is motivated by a practical problem in empirical macroeconomic research on how to deal with outliers around the COVID-19 pandemic. The challenge is illustrated by Figure 1, which plots quarterly log-differences of U.S. real personal consumption expenditure (PCE) on food and accommodation services. Fluctuations around the pandemic dwarf the series' typical variation elsewhere in the sample. Although several methods for addressing the pandemic-induced outliers have been proposed (Ng, 2021; Schorfheide and Song, 2021; Carriero et al., 2024), in practice it has been common to either ignore the problem or end the estimation sample in 2019. Chapter 2 works towards a more sustainable solution to this problem.

We study the Impulse Indicator Saturation (IIS) (Hendry et al., 2008) method for estimating a linear equation. We focus on a variant of IIS that resembles the 'split-half' algorithm proposed by Hendry (1999). The algorithm assumes a 'clean set' of observations is known and used to fit an initial estimator and detect outliers. IIS has been used for applications in economics and climate research (Koch et al., 2022; Stechemesser et al., 2024; Marczak and Proietti, 2016; Bennedsen et al., 2023; Durevall and Henrekson, 2011). However, its theoretical properties have only been established for uncontaminated data (Johansen and Nielsen, 2016).

We establish asymptotic properties of IIS under outlier contamination. Central to the analysis is an 'oracle property' showing that IIS is asymptotically equivalent to an infeasible least squares estimator that perfectly removes all outliers. We use the oracle property to derive the asymptotic distribution of IIS in cross-sectional and time series models with outliers. We also find a limit theory for the number of misclassified 'good' observations, providing guidance on choosing tuning parameters for IIS. An empirical illustration uses the PCE data introduced in Figure 1.

Chapter 3 decomposes inflation into demand- and supply-driven components in a panel of 32 countries. The decomposed series are updated frequently and can be used for policy monitoring. We employ a recent methodology (Shapiro, 2026) for constructing decompositions using disaggregated PCE data similar to those in Figure 1. Estimations are implemented using a version of the IIS algorithm studied in Chapter 2 to guard against outliers around the COVID-19 pandemic.

We validate the decomposed inflation series by examining their relationship to external measures of demand and supply shocks, extending the validation by Shapiro (2026) to a panel of countries. We then use the decompositions in two applications. First, we assess the relative contributions of demand and supply factors to the post-pandemic surge in inflation. Second, we examine whether the empirical Phillips curve relationship differs between demand- and supply-driven inflation.

The three chapters of this thesis are formatted as stand-alone articles. Appendices and references are provided after each chapter.

References for Introduction

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

Bennedsen, M., E. Hillebrand, and S. J. Koopman (2023). A multivariate dynamic statistical model of the global carbon budget 1959–2020. *Journal of the Royal Statistical Society: Series A* 186(1), 20–42.

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

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

Carriero, A., T. E. Clark, M. Marcellino, and E. Mertens (2024). Addressing COVID-19 outliers in BVARs with stochastic volatility. *The Review of Economics and Statistics* 106(5), 1403–1417.

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

Durevall, D. and M. Henrekson (2011). The futile quest for a grand explanation of long-run government expenditure. *Journal of Public Economics* 95(7), 708–722.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42(6), 1887–1896.

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

Hawkins, D. M. (1980). Identification of Outliers. Monographs on Applied Probability and Statistics. London: Chapman and Hall.

Hendry, D. F. (1999). An econometric analysis of US food expenditure 1931–1989. *Methodology & Tacit Knowledge: Two Experiments in Econometrics*. Ed. by J. Magnus and M. Morgan. Chichester, UK: John Wiley & Sons, 41–362.

Hendry, D. F., S. Johansen, and C. Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 23, 317–335.

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

Hubert, M. and P. J. Rousseeuw (1996). Robust regression with a categorical covariable. *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*. Ed. by H. Rieder. New York, NY: Springer, 215–224.

— (1997). Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference* 57(1), 153–163.

Johansen, S. and B. Nielsen (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics* 43(2), 321–348.

Koch, N., L. Naumann, F. Pretis, N. Ritter, and M. Schwarz (2022). Attributing agnostically detected large reductions in road CO2 emissions to policy mixes. *Nature Energy* 7(9), 844–853.

Marczak, M. and T. Proietti (2016). Outlier detection in structural time series models: The indicator saturation approach. *International Journal of Forecasting* 32(1), 180–202.

Maronna, R. A. and V. J. Yohai (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference* 89(1), 197–214.

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

Ng, S. (2021). Modeling macroeconomic variations after COVID-19. NBER Working Paper 29060. National Bureau of Economic Research.

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

Schorfheide, F. and D. Song (2021). Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic. NBER Working Paper 29535. National Bureau of Economic Research.

Shapiro, A. H. (2026). Decomposing supply and demand driven inflation. *Journal of Money, Credit and Banking* 58(2), 365–388.

Stechemesser, A., N. Koch, E. Mark, E. Dilger, P. Klösel, L. Menicacci, D. Nachtigall, F. Pretis, N. Ritter, M. Schwarz, H. Vossen, and A. Wenzel (2024). Climate policies that achieved major emission reductions: Global evidence from two decades. *Science* 385, 884–892.

Chapter 1

Robustness of Least Trimmed Squares with categorical covariates

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.*

1.1 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 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; Čížek, 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 1.2 defines LTS estimators and explains the role of an initial h. Section 1.3 gives the boundedness results. In section 1.4, we use these results to write down an algorithm for choosing an initial h. Sections 1.5 and 1.6 contain simulations and an empirical illustration. Section 1.7 concludes. The appendix consists of proofs and details on algorithms for exploring the regressor space.

1.2 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.

1.2.1 LTS Estimator

Consider a linear equation $y_i = x_i'\beta + \sigma\varepsilon_i$ for $i = 1, \dots, n$ 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 $h \le n$. For an h-subset ζ of $\{1, \dots, n\}$, 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 $\mathcal{M}_h = \arg\min_{\zeta:|\zeta|=h} \hat{\sigma}^2_\zeta$.

1.2.2 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 $\hat{h} = \sum_{i=1}^n \mathbb{I}\{|y_i - x_i'\hat{\beta}_{(0)}|/\hat{\sigma}_{(0)} \le c\}$, 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).

1.3 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.

1.3.1 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 1.1. *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 $\forall\epsilon > 0$, $\exists a_0, n_0$, $\forall a \leq a_0, n \geq n_0$: $|s_{a,n} - s| < \epsilon$.

1.3.2 A first boundedness result

Boundedness of the LTS estimator 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

$$F_n(a) = \max_{\delta:|\delta|=1} n^{-1} \sum_{i=1}^{n} \mathbb{I}\{|x_{in}'\delta| \leq a\}. \tag{1.1}$$

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 1.2. *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*

$$\lim_{(a,n)\to(0,\infty)} P\left(F_n(a) > \xi\right) = 0. \tag{1.2}$$

Theorem 1.1. *Suppose Assumptions 1.1, 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 1.2 are discussed in Johansen and Nielsen (2019) and Berenguer-Rico and Nielsen (2025b). We mention a couple of examples.

Example 1.3.1 (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 1.2 holds for any* $\lambda_\circ > 2/3$.

Example 1.3.2 (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 1.2 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 1.1. *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.

1.3.3 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 $\zeta \subseteq \{1, \dots, n\}$, the strip function

$$F^R_{n\zeta}(a) = \max_{\delta:|\delta|=1,\,|R'\delta|>0} h^{-1} \sum_{i\in\zeta} \mathbb{I}\{|x'_{in}\delta| \leq a|R'\delta|\}. \tag{1.3}$$

When there is no selection, that is $R = I$ and $\zeta = \{1, ..., n\}$, this function reduces to $F_n(a)$ in (1.1). 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 1.3. *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*

$$\lim_{(a,n)\to(0,\infty)} P\left(\lambda_\circ F^R_{n\zeta_\circ}(a) > \xi\right) = 0. \tag{1.4}$$

Theorem 1.2. *Suppose Assumption 1.1, 1.3. Then* $\max_{\underline{h}\leq h\leq h_\circ} \max_{\zeta\in\mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$.

Theorem 1.2 implies Theorem 1.1 since Assumption 1.3 is weaker than Assumption 1.2. 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 1.2 and 1.3 are identical when $h_\circ F^R_{n\zeta_\circ}(a) = nF^R_n(a)$ and $R = I$. 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 1.4, we propose an algorithm for choosing an initial h that allows a user to exploit such assumptions.

Example 1.3.3 (One binary regressor as in Example 1.3.2)**.** *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 1.3.2.*

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 1.3.

1.3.4 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 1.3 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 1.4. *Suppose there is a probability measure* P_{good} *such that*

$$h_\circ^{-1} \sum\nolimits_{i\in\zeta_\circ} \mathbb{I}\{x_{in} \in A\} \to P_{good}(A) \tag{1.5}$$

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

Proposition 1.3.1. *Suppose Assumption 1.4. 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 1.4 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 1.3.1, Assumption 1.3 can be evaluated using the inequality $\lambda_\circ F^R_{n\zeta_\circ}(0) \leq F^R_n(0)+1/n$, where $F^R_n(0)$ can be computed using algorithms in Appendix 1.B.

Assumption 1.5. *Suppose* $\sup_{\delta:|\delta|=1,|R'\delta|>0} \lambda_\circ p_{good,\delta} < \underline{\lambda} + \lambda_\circ - 1$.

Proposition 1.3.2. *Assumptions 1.4, 1.5 imply Assumption 1.3. In particular, if Assumptions 1.1, 1.4, 1.5 hold then* $\max_{\underline{h}\leq h\leq h_\circ} \max_{\zeta\in\mathcal{M}_h} |\hat{\beta}_\zeta| = O_p(1)$.

Proposition 1.3.2 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 1.3.4 (Continuous regressors)**.** *Consider continuous, i.i.d. regressors* x_{in} *for* $1 \leq i \leq n$*. Then* $p_{good,\delta} = 0$, $\forall \delta \neq 0$*. By Proposition 1.3.2,* $F_n(a)$ *vanishes and Assumption 1.2 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 1.3.5 (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 1.3.6 (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),* (1.5) *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 $\underline{n} \approx n/2$ 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 1.3.7 (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 1.3.2.*

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 1.3 requires* $\lambda_\circ \times 0 < 2\lambda_\circ - 1$, *that is* $\lambda_\circ > 0.5$ *as in Example 1.3.1.*

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 1.3.3.*

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* $\lambda_\circ > 10/11 \approx 0.91$.

1.3.5 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 1.3.3. *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 1.3.8 (Two binary regressors). *Consider binary regressors* $x'_{in} = (1, z_{i1}, z_{i2})$. *By Proposition 1.3.3,* $\forall$ *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 1.3 depends on the probabilities of the sets* $\{z_{i1} = 1\}$, $\{z_{i1} = 0\}$, $\{z_{i2} = 1\}$, $\{z_{i2} = 0\}$, $\{z_{i1} = z_{i2}\}$ *and* $\{z_{i1} + z_{i2} = 1\}$.

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 1.4 is the simplest setting that can generate 'bad leverage points', where errors depend on regressors.

Assumption 1.6. *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 1.7. *Suppose*
(i) $1/\min_{j \in \zeta_\circ^c} |\varepsilon_j| = o_p(1)$.
(ii) $\lim_{n\to\infty} P(\min_{\zeta:|\zeta|=\underline{h}} \hat{\sigma}^2_\zeta > 0) = 1$.

We relax Assumption 1.5 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 1.8. *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 1.3. *If Assumptions 1.1, 1.6, 1.7, 1.8 hold then* $\max_{\underline{h}\leq h \leq h_\circ} \max_{\zeta\in\mathcal{M}_h} |\hat{\beta}_\zeta| = O_p(1)$.

Assumption 1.8 reduces to Assumption 1.5 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 1.3.3. The following example illustrates this.

Example 1.3.9 (One binary regressor as in Examples 1.3.2, 1.3.3)**.** *Suppose* $x'_{in} = (1, z_{in})$*, with* z_{in} *binary and satisfying Assumptions 1.6, 1.7, 1.8 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 1.8 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 1.3.2, 1.3.3.*

In the best case, if $p_{out} = 1/2$ *we require* $\underline{\lambda} > 1/2$*, improving on Examples 1.3.2,1.3.3 by further exploiting knowledge of outlying regressors.*

1.4 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 1.5, 1.8. The initial values from these algorithms have robustness guarantees based on our boundedness results.

1.4.1 Imposing restrictions on the regressors

We consider the setting of Assumption 1.6, which implies Assumption 1.4. 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 1.5, that is of

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} \left[p_{full,\delta} - \left(1-\lambda_\circ\right)p_{out,\delta}\right] < \underline{\lambda} + \lambda_\circ - 1. \tag{1.6}$$

Although both $\lambda_\circ$ and $p_{out,\delta}$ are unknown, so that (1.6) 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.
(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*

$$\underline{\lambda}^{(1)} = \max_{\delta:|\delta|=1,\,|R'\delta|>0} \{\hat{p}_{full,\delta} + (1-\underline{\lambda}_\circ)(1-b_{out,\delta})\}. \tag{1.7}$$

(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 1.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 1.6.

Algorithm 1.1 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 1.4. *Consider initial choice* $\underline{h}$ *and* $\epsilon > 0$ *from Algorithm 1.1. Suppose Assumptions 1.1, 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*

$$\underline{\lambda}_\circ^{(1)} = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} \frac{1 + p_{full,\delta} - b_{out,\delta}}{2 - b_{out,\delta}} \tag{1.8}$$

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 (1.8) 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 1.4 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 1.4.1 (One binary regressor as in Examples 1.3.2, 1.3.3, 1.3.9)**.** *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 $0 \le p_{out} \le 1$ *is unconstrained then* $b_{out,\delta^\dagger} = b_{out,\delta^\ddagger} = 0$ *so that* $\underline{\lambda}_\circ^{(1)} = 3/4$ *by* (1.8)*. This is the breakdown point optimal choice.*

If $1/3 \le p_{out} \le 2/3$ *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.

1.4.2 Exploiting oracle conditions

We adapt Algorithm 1.1 and the boundedness result in Theorem 1.4 to the setting where outliers satisfy the conditions required for oracle inference (Section 1.3.6). 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 1.8, that is of

$$\sup_{\delta:|\delta|=1,|R'\delta|>0} \{\lambda_\circ p_{good,\delta} - (1-\lambda_\circ)p_{out,\delta}\} < \underline{\lambda} + \lambda_\circ - 1. \tag{1.9}$$

The adapted algorithm and the associated boundedness result have a structure similar to Algorithm 1 and Theorem 1.4, and the points raised in Section 1.4.1 continue to apply.

Algorithm 1.2. *Follow steps* 1,2 *and choice of* $\epsilon > 0$ *in Algorithm 1.1.*
(3) *Compute*

$$\underline{\lambda}^{(2)} = \max_{\delta:|\delta|=1,\,|R'\delta|>0} [\hat{p}_{full,\delta} + (1-\underline{\lambda}_\circ)(1-2b_{out,\delta})\mathbb{I}\{b_{out,\delta} < 1/2\}]. \tag{1.10}$$

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

Theorem 1.5. *Consider initial values* $\underline{h}$ *and* $\epsilon > 0$ *from Algorithm 2. Suppose Assumptions 1.1, 1.6, 1.7 hold,* $b_{out,\delta} \le p_{out,\delta}$ $\forall\delta : |R'\delta| > 0$*, and* $\underline{\lambda}_\circ^{(2)} \le \underline{\lambda}_\circ \le \lambda_\circ - \epsilon$

where

$$\underline{\lambda}_\circ^{(2)} = \sup_{\delta:|\delta|=1,\,|R'\delta|>0} \Big[\frac{1+p_{full,\delta}-2b_{out,\delta}}{2(1-b_{out,\delta})}\mathbb{I}\{b_{out,\delta}<1/2\}+p_{full,\delta}\mathbb{I}\{b_{out,\delta}\geq 1/2\}\Big] \quad (1.11)$$

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 1.4.2 (One binary regressor as in Examples 1.3.2, 1.3.3, 1.3.9, 1.4.1)**.** *The setup and the values of* $p_{full,\delta}$ *and* $b_{out,\delta}$ *are as in Example 1.4.1. 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* (1.8), (1.11).

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* $\underline{\lambda}_\circ^{(1)}=2/3>1/2=\underline{\lambda}_\circ^{(2)}$.

1.5 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.

1.5.1 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 1.1 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 1.7(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 1.6, 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 1.1 shows $\underline{\lambda}^{(1)}$ and $\underline{\lambda}^{(2)}$, calculated from (1.7) and (1.11) 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 1.1, 1.6 so LTS estimators with $\lfloor n \underline{\lambda}^{(2)} \rfloor \leq h \leq \lfloor 0.8n \rfloor$ should be bounded by Theorem 1.5. In DGP6, Assumption 1.6 is violated so Theorem 1.5 does not apply, but by Proposition 1.3.2 LTS should be bounded for $\lfloor n \underline{\lambda}^{(1)} \rfloor \leq h \leq \lfloor 0.8n \rfloor$. For comparison, Table 1.1 also shows the breakdown point optimal value $\underline{\lambda}_{bp} = 1/2 + p_{full}/2$.

1.5.2 Simulation results

Table 1.2 shows mean squared errors for LTS and least squares estimators computed in the six DGPs described above.

Table 1.1: 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

In DGP1, mean squared errors converge to zero for $\lambda \in \{0.7, 0.75, 0.8\}$. This is in line with Theorem 1.5 and the lower bound $\underline{\lambda}^{(2)} = 0.7$ in Table 1.1. 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 1.3 or 1.7.

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 1.6 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 1.3.2, which does not require Assumption 1.6.

1.6 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

Table 1.2: 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.*

education and other labour market characteristics through

$$w_{1f} - w_{2f} = \mu + \beta(school_{1f} - school_{2f}) + \alpha'(c_{1f} - c_{2f}) + (\varepsilon_{1f} - \varepsilon_{2f}), \qquad (1.12)$$

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 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.

Table 1.3: 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.

1.6.1 Selecting an initial h for twins data

We use Algorithms 1.1 and 1.2 to choose an initial h for LTS.

Step 0: Hyperplane detection in the covariate space. We use a randomised algorithm (Appendix 1.B) to find the five hyperplanes with most observations. Their orthogonal vectors and associated observation shares are shown in Table 1.3. 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 1.4 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 1.3. 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 1.4 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 (1.8),(1.11) with $p_{full,\delta} = \hat{p}_{full,\delta}$. Theorems 1.4,1.5 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

Table 1.4: 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 1.3. Column 3 shows coefficients of interest for which boundedness is required. Columns 4-8 show minimal choices of h under various assumptions.

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 1.4 show initial choices $\underline{h}^{(1)} = \lfloor n\underline{\lambda}^{(1)} \rfloor$ and $\underline{h}^{(2)} = \lfloor n\underline{\lambda}^{(2)} \rfloor$, calculated from (1.7) and (1.10) 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.

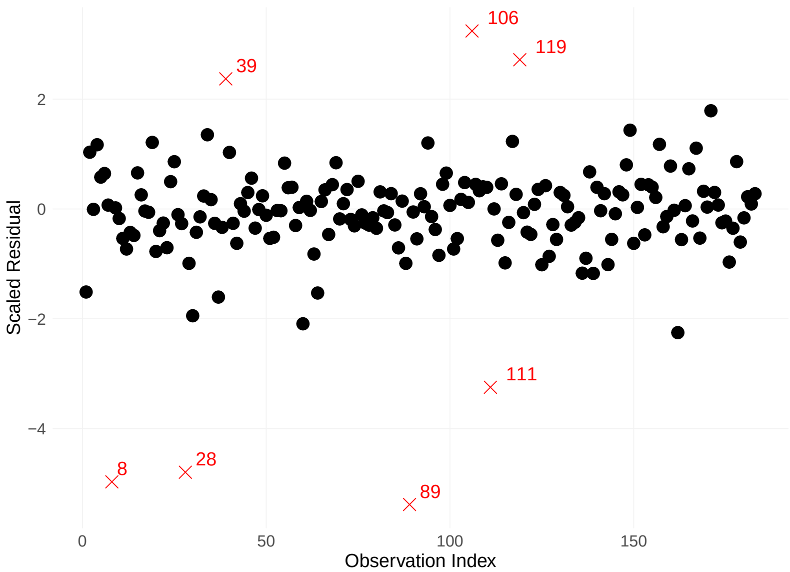
1.6.2 Results from LTS estimation of twins model

We assess the number of outliers using an index plot (Rousseeuw and Leroy, 1987). Figure 1.1 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 1.4. 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 1.5 shows LTS estimates of equation (1.12), 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 1.5 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 1.4, 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

Figure 1.1: Index plot of scaled residuals with LTS(h=168)



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 1.4, 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.

1.7 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.

Table 1.5: 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).

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 (Čížek, 2013), are computed. Further, we found conditions for LTS to be bounded 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 to Chapter 1

1.A Proofs for Main Results

1.A.1 Proof of Theorem 1.1

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

$$F_{n,h}(a) = \max_{\zeta:|\zeta|=h} F^I_{n\zeta}(a) = \max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1} h^{-1} \sum_{i\in\zeta} \mathbb{I}\{|x'_{in}\delta| \leq a\}, \tag{1.13}$$

where $F^R_{n\zeta}$ is defined in (1.3). We show equivalence between $F_{n,h}$ and F_n from (1.1).

Lemma 1.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 (1.1), (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, $\{F_n(a) > \xi\} \subseteq \{(h/n)F_{n,h}(a) > \xi\}$. $\square$

Proof of Theorem 1.1. 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 1.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 $0 < \xi < (2 - \lambda_\circ^{-1})\lambda_\circ = 2\lambda_\circ - 1$. $\square$

1.A.2 Proof of Theorem 1.2

Lemma 1.A.2. *Consider a set $\zeta \subseteq \{1, \ldots, n\}$ 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*

$$\max_{\delta:|\delta|=1,\, |R'\delta|>0} h^{-1} \sum_{i \in \zeta} \mathbb{I}\{x'_{in}\delta = 0\} < 1 \implies R'\hat{\beta}_\zeta \text{ is unique.} \tag{1.14}$$

Proof. Least squares solutions satisfy the normal equations $(X'_\zeta X_\zeta)\beta = X'_\zeta y_\zeta$. Therefore, if $\hat{\beta}^*_\zeta, \hat{\beta}^\dagger_\zeta$ are solutions then $(X'_\zeta X_\zeta)(\hat{\beta}^\dagger_\zeta - \hat{\beta}^*_\zeta) = 0$. This means $\hat{\beta}^\dagger_\zeta = \hat{\beta}^*_\zeta + v$ for some v with $X_\zeta v = 0$. The assumption in (1.14) 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 $R'\hat{\beta}^\dagger_\zeta = R'(\hat{\beta}^*_\zeta + v) = R'\hat{\beta}^*_\zeta$. As this holds for any solutions $\hat{\beta}^*_\zeta$ and $\hat{\beta}^\dagger_\zeta$, then $R'\hat{\beta}_\zeta$ is unique. $\square$

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 1.A.3. *Let R be a $p \times s$ matrix such that $R'R = I$, $1/2 < \lambda_\circ \leq 1$, and $0 < \underline{\lambda} \leq \lambda_\circ$. 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 $P(\Omega_n) \geq 1 - \epsilon$ such that on Ω_n for*

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

and all $\underline{h} \leq h \leq h_\circ$ it holds

$$\min_{\zeta \in Z^R_{\tau a h} : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B} \hat{\sigma}^2_\zeta > \max_{\zeta_h \subseteq \zeta_\circ : |\zeta_h| = h} \hat{\sigma}^2_{\zeta_h}. \tag{1.15}$$

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 $P(\Omega_n) \geq 1 - \epsilon$ 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 $\underline{h} \leq h \leq h_\circ$ that

$$\max_{\zeta_h \subseteq \zeta_\circ : |\zeta_h| = h} \hat{\sigma}^2_{\zeta_h} / \sigma^2 \leq \max_{\zeta_h \subseteq \zeta_\circ : |\zeta_h| = h} h^{-1} \sum_{i \in \zeta_h} \varepsilon_i^2 \leq \underline{h}^{-1} \sum_{i \in \zeta_\circ} \varepsilon_i^2 < A_0. \tag{1.16}$$

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

$$B \geq \frac{1}{a} \left[\left(\frac{A_0}{\eta} \right)^{1/2} + \left(\frac{A_0}{\tau - \eta} \right)^{1/2} \right]. \tag{1.17}$$

2. Showing inequality (1.15) *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

$$h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x'_{in}\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|\} \geq \tau. \tag{1.18}$$

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

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

Use this to write

$$
\begin{aligned}
\hat{\sigma}^2_\zeta/\sigma^2 &= h^{-1}\sum_{i\in\zeta}(\varepsilon_i - \hat{\ell}_\zeta x'_{ni}\hat{\delta}^*_\zeta/|R'\hat{\delta}^*_\zeta|)^2 \\
&\geq h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{ni}\hat{\delta}^*_\zeta| > a|R'\hat{\delta}^*_\zeta|\}\mathbb{I}\{\varepsilon_i^2 \leq A_0/\eta\}(\varepsilon_i - \hat{\ell}_\zeta x'_{ni}\hat{\delta}^*_\zeta/|R'\hat{\delta}^*_\zeta|)^2. \qquad (1.19)
\end{aligned}
$$

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 (1.17) 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

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

Inserting this in (1.19) 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 (1.18) 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

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

Thus, combined with (1.16), on Ω_n it holds $\hat{\sigma}_\zeta^2/\sigma^2 > \max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h}\hat{\sigma}_{\zeta_h}^2/\sigma^2$, so (1.15) holds. □

Proof of Theorem 1.2. 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 $\max_{\zeta\in\mathcal{M}_h}|R'\hat{\beta}_\zeta - R'\beta|/\sigma \le B$ for all $\underline{h} \le h \le h_\circ$. Finally, we show that for any $\epsilon > 0$, Ω_n and B can be constructed so that $P(\Omega_n) \ge 1-\epsilon$ for all n large.

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

$$\lambda_\circ\left(\max_{\delta:|\delta|=1,\,|R'\delta|>0} h_\circ^{-1}\sum_{i\in\zeta_\circ}\mathbb{I}\{|x_{in}'\delta| \le a|R'\delta|\}\right) \le \xi. \tag{1.20}$$

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

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

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

$$\min_{\zeta\in Z_{\tau a h}^R:|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B}\hat{\sigma}_\zeta^2 > \max_{\zeta_h\subseteq\zeta_\circ:|\zeta_h|=h}\hat{\sigma}_{\zeta_h}^2, \tag{1.21}$$

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

$$\Xi = \{\zeta : |\zeta| = h,\, |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B\} \subseteq \{\zeta \in Z_{\tau a h}^R : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B\}. \tag{1.22}$$

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

We check property (ii). Use the identity

$$R_{\delta,h,\zeta} = h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} = h^{-1}|\zeta \cap \zeta_\circ| - h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\delta| \leq a|R'\delta|\}. \tag{1.23}$$

Bound the first term in (1.23) 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 (1.23), inequality (1.20) 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 (1.23) 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 1.3. 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 $|\zeta| = h$

$$\begin{aligned} R_{\delta,h,\zeta} \geq \lambda_\circ^{-1}\{\underline{\lambda} + \lambda_\circ - 1 - \xi + o(1)\} + o(1) =& \lambda_\circ^{-1}(\underline{\lambda} + \lambda_\circ - 1 - \xi) + o(1) \\ \geq& 2\tau + o(1) \geq \tau. \end{aligned} \tag{1.24}$$

Property (ii) then follows from Lemma 1.A.2, since by (1.24) for all $\delta \in \Delta$

$$h^{-1}\sum_{i \in \zeta} \mathbb{I}\{x_{in}'\delta = 0\} \leq 1 - h^{-1} \sum_{i \in \zeta \cap \zeta_\circ} \mathbb{I}\{|x_{in}'\delta| > a|R'\delta|\} = 1 - R_{\delta,h,\zeta} \leq 1 - \tau < 1. \tag{1.25}$$

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 (1.24), conclude that (iii) holds since

$$\max_{\delta \in \hat{\delta}_\zeta} R_{\delta,h,\zeta} \geq \min_{\delta:|\delta|=1,|R'\delta|>0} R_{\delta,h,\zeta} \geq \tau. \tag{1.26}$$

2.2 Deterministic Analysis on Ω_n — Conclusion. Using first (1.22) and then (1.21), we have for large n and any $\underline{h} \leq h \leq h_\circ$

$$\min_{\zeta:|\zeta|=h,\, |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B} \hat{\sigma}^2_\zeta \geq \min_{\zeta \in Z^R_{\tau a h}:|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B} \hat{\sigma}^2_\zeta > \max_{\zeta_h \subseteq \zeta_\circ:|\zeta_h|=h} \hat{\sigma}^2_{\zeta_h}. \tag{1.27}$$

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 1.3, we can find $0 < \xi < \underline{\lambda} + \lambda_\circ - 1$, $a > 0$, and Ω_{1n} such that (1.20) 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 1.1, then by Lemma 1.A.3 there exist $B > 0$ and Ω_{2n} such that (1.21) 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 (1.24) holds. □

1.A.3 Proof of Proposition 1.3.2

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 1.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 1.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)*

$$S \not\subseteq S' \implies \dim(S \cap S') < \dim(S). \tag{1.28}$$

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

$$X \not\subseteq X' \implies \dim(X \cap X') < \dim(X). \tag{1.29}$$

Property (1.29) *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* (1.28)*, conclude that* $\dim(X \cap X') = \dim(S \cap S') < \dim(S) = \dim(X)$.

Lemma 1.A.5. *Let P_{good} be a probability measure and $\{\mu_k, A_k, m_k\}_{k=1}^{\infty}$ given by Lemma 1.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 1.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 1.A.1, it follows that the set $\dim(A \cap \{x : x'\delta^* = 0\}) < \dim(A) = m$. By Lemma 1.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 $\liminf_{n\to\infty} \mu(S_n) \leq \mu(\limsup_{n\to\infty} S_n)$ Billingsley, 1995, Theorem 4.1. to get the contradiction

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

For a matrix B with columns $B_1, \ldots, B_J$, write $\mathrm{span}(B)$ for $\mathrm{span}\{B_1, \ldots, B_J\}$. 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 $\mathrm{proj}_S(\delta) = \arg\min_{x\in S}|x - \delta|$. If B is a matrix with full column rank then $\mathrm{proj}_{\mathrm{span}(B)}(\delta) = B(B'B)^{-1}B'\delta$.

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

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

Lemma 1.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 $S = \mathrm{span}(\cup_{k\in N} S_k) \neq \{0\}$. There exists a matrix B of full column rank such that $\mathrm{span}(B) = S$ and every column of B is an element of $\cup_{k\in N} S_k$ with norm one. For $c = \{\mathrm{mineig}(B'B)/p\}^{1/2}$ and any $\delta \in \mathbb{R}^p$ it holds*

$$|proj_S(\delta)| \leq \max_{k\in N} |proj_{S_k}(\delta)|/c. \tag{1.30}$$

Further, if for some vector δ of norm one

$$0 \leq \max_{k\in N} |\mathrm{proj}_{S_k}(\delta)| < |R'\delta|c \tag{1.31}$$

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 $\mathrm{span}(B_k) = S_k$. 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 (1.30). We have (Axler, 2015, 6.55) that for two linear subspaces $U, V \subseteq \mathbb{R}^p$

$$\text{if } U \subseteq V \text{ then } |\text{proj}_U(\delta)| \leq |\text{proj}_V(\delta)| \text{ for all } \delta. \tag{1.32}$$

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

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

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

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

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 1.A.2. Since for all j it holds $\text{span}(B_j) \subseteq S_k$ for some k by construction, use property (1.32) 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 (1.30).

3. Proof of $S^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \varnothing$. Let δ satisfy (1.31). Combine (1.30, 1.31) to get $|\text{proj}_S(\delta)| < |R'\delta|$. Since $R'R = I$, then $|R'\delta| = |\text{proj}_{\text{span}(R)}(\delta)|$ by Remark 1.A.2. Thus, $|\text{proj}_S(\delta)| < |\text{proj}_{\text{span}(R)}(\delta)|$. Contraposition of (1.32) 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 1.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*

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : |x'\delta| \leq |R'\delta| a\} \leq \sup_{\delta:|\delta|=1,\,|R'\delta|>0} P_{good}\{x : x'\delta = 0\} + \epsilon. \tag{1.33}$$

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 1.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 (1.33) for μ.

2. Choice of constant c. Define $S_k = \text{span}(A_k)$ for $1 \leq k \leq K$. By Lemma 1.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

$$\begin{aligned} 0 \leq \max_{k \in N} |\text{proj}_{S_k}(\delta)| < |R'\delta| c_N \text{ for some } \delta \text{ of norm one} \\ \implies S_N^{\perp} \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \varnothing. \end{aligned} \tag{1.34}$$

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

3. Choice of constant a. By Lemma 1.A.5, there exists $a > 0$ such that

$$\max_{k \leq K} \sup_{\delta \in S_k : |\delta| = 1} \mu_k\{x : |x'\delta| \leq a/c\} \leq \epsilon/K. \tag{1.35}$$

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

$$\mu\{x : |x'\delta| \leq a|R'\delta|\} = \sum_{k \in N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\} + \sum_{k \notin N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\}. \tag{1.36}$$

Consider the latter sum in (1.36). 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 (1.35) to bound

$$\sum_{k \notin N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\} \leq \sum_{k \notin N_\delta} \mu_k\{x : |x'\delta_k|/|\delta_k| \leq a/c\} \leq K\epsilon/K = \epsilon. \tag{1.37}$$

Consider the first sum in (1.36). 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

$$\sum_{k\in N_\delta} \mu_k\{x : |x'\delta| \leq a|R'\delta|\} \leq \sum_{k\in N_\delta} \mu_k(S_\delta) \leq \sum_{k\leq K} \mu_k(S_\delta) = \mu(S_\delta). \tag{1.38}$$

Insert bounds (1.37) and (1.38) into (1.36) to get

$$M = \mu\{x : |x'\delta| \leq a|R'\delta|\} \leq \mu(S_\delta) + \epsilon. \tag{1.39}$$

If N_δ is empty then the first sum in (1.36) is empty, so (1.36) and (1.37) 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 (1.39) gives

$$M \leq \mu\{0\} + \epsilon \leq \sup_{\delta:|\delta|=1,|R'\delta|>0} \mu\{x : x'\delta = 0\} + \epsilon. \tag{1.40}$$

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 (1.34) 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 (1.39)

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

□

Lemma 1.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

$$\sup_{\delta:|\delta|=1,\, a\in\mathbb{R}} |P_n^*\{x : x'\delta \le a\} - P^*\{x : x'\delta \le a\}| \xrightarrow{n\to\infty} 0. \tag{1.41}$$

We rewrite $\mathbb{I}\{x : |x'\delta| \le a\}$. Since $\{|x'\delta| \le a\} = \{x'\delta \le a\} \backslash \{x'\delta < -a\}$ and the latter set complements $\{x'(-\delta) \le a\}$ then $\mathbb{I}\{x : |x'\delta| \le a\} = \mathbb{I}\{x : x'\delta \le a\} + \mathbb{I}\{x : x'(-\delta) \le a\} - 1$. Thus, $P_n^*\{x : |x'\delta| \le a\} - P^*\{x : |x'\delta| \le a\} = P_n^*\{x : x'\delta \le a\} - P_n^*\{x : x'\delta \le a\} + P_n^*\{x : x'(-\delta) \le a\} - P^*\{x : x'(-\delta) \le a\}$. The claim follows by the triangle inequality and (1.41). □

Proof of Proposition 1.3.2. 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

$$\lim_{(a,n)\to(0,\infty)} F^R_{n\zeta_\circ}(a) = \sup_{\delta:|\delta|=1,\, |R'\delta|>0} P_{good}\{x : x'\delta = 0\}. \tag{1.42}$$

Let $\epsilon > 0$ be given. By condition (1.5) and Lemma 1.A.8, we can find a set of probability one where, pointwise, for large n

$$\sup_{\delta:|\delta|=1,\, a\ge 0} \left| h_\circ^{-1} \sum_{i\in\zeta_\circ} \mathbb{I}\{|x_i'\delta| \le a\} - P_{good}\{x : |x'\delta| \le a\} \right| \le \epsilon. \tag{1.43}$$

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

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

and

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

By Lemma 1.A.7, we can find $a^* > 0$ such that for $0 \le a \le a^*$

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

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

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

Thus, (1.42) 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 1.3 is satisfied. □

1.A.4 Proof of Proposition 1.3.3

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

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

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

To show $F^R_{n\zeta}(a) \le F^R_{n\zeta}(0)$ 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 $\mathcal{X} \ne \{0\}$ 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 $\{0\} \ne \{x_k\}_{k\in N} \subseteq \mathcal{X}$, by Lemma 1.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 1.A.6, the constant $0 < c_N = \{\text{mineig}(B'_N B_N)/p\}^{1/2}$ satisfies

$$\begin{aligned} &0 \le \max_{k\in N} |\text{proj}_{\text{span}(x_k)}(\delta)| < |R'\delta| c_N \text{ for some } \delta \text{ of norm one} \\ &\Longrightarrow \text{span}(\cup_{k\in N}\{x_k\})^\perp \cap \{\delta : |\delta| = 1,\, |R'\delta| > 0\} \ne \varnothing. \end{aligned} \tag{1.44}$$

Take $c = \min_{N\subseteq\{1,\dots,K\}:\{x_k\}_{k\in N}\ne\{0\}} c_N$ 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 (1.44), 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. □

1.A.5 Proof of Theorem 1.3

Lemma 1.A.9. *Let R be a $p \times s$ matrix and $h \leq n$. Suppose $\min_{\zeta : |\zeta| = h} \hat{\sigma}^2_\zeta > 0$ and*

$$\max_{\delta : |\delta| = 1, \, |R'\delta| > 0} \sum_{i=1}^{n} \mathbb{I}\{x'_{in}\delta = 0\} < n \tag{1.45}$$

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

Proof. By Lemma 1.A.2, if ζ with $|\zeta| = h$ is such that

$$\max_{\delta : |\delta| = 1, \, |R'\delta| > 0} h^{-1} \sum_{i \in \zeta} \mathbb{I}\{x'_{in}\delta = 0\} < 1 \tag{1.46}$$

then $R'\hat{\beta}_\zeta$ is unique. Therefore, (1.46) holds with equality for any ζ with $R'\hat{\beta}_\zeta$ non-unique. We show that if (1.46) 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 (1.45), there exists $j\in\{1,\ldots,n\}\setminus\zeta$ 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

$$\hat\sigma_{\zeta^*}^2\le h^{-1}\sum_{i\in\zeta^*}(y_i-x_{in}'\tilde\beta_{\zeta^*})^2=h^{-1}(y_j-x_{jn}'\tilde\beta_{\zeta^*})^2+h^{-1}\sum_{i\in\zeta\setminus\{i^*\}}(y_i-x_{in}'\tilde\beta_{\zeta^*})^2$$
$$=h^{-1}\sum_{i\in\zeta\setminus\{i^*\}}(y_i-x_{in}'\hat\beta_\zeta)^2<\hat\sigma_\zeta^2. \qquad \square$$

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

Assumption 1.9. *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 $1/2<\lambda_\circ\le 1$. Let the lower bound $\underline{h}=\lfloor\underline{\lambda}n\rfloor$ be such that $1-\lambda_\circ<\underline{\lambda}\le\lambda_\circ$.
(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 $\zeta\subset\{1,\ldots,n\}$, $\delta\in\{\delta:|\delta|=1\}$, $a,\tau\ge 0$, and $t\le|\zeta\cap\zeta_\circ|$ define

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

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

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

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

Remark 1.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 $H_{\delta\zeta}(H_{\delta\zeta}^{-1}(t))\ge t$ by right-continuity of $H_{\delta\zeta}$.*

Lemma 1.A.10. *Suppose Assumption 1.9. 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 1.A.10. Structure of the proof is similar to that of Theorem 1.2. 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 $\underline{h} \leq h \leq h_\circ$

$$\max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1,\,|R'\delta|>0} \left\{H_{\delta\zeta}(a) + \bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau} b_n)\right\} \leq h - \nu n. \tag{1.47}$$

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 (1.47). 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

$$\underline{h}^{-1}\sum_{i\in\zeta_\circ} \varepsilon_i^2 < A_0, \quad \underline{h}^{-1}\sum_{i\in\zeta_\circ} \mathbb{I}\left\{\varepsilon_i^2 > A_0/\eta\right\} \leq \eta, \tag{1.48}$$

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

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

it holds

$$\min_{\zeta\in Z^R_{\tau a h}:|R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0} \hat{\sigma}^2_\zeta > \hat{\sigma}^2_{\zeta_h}. \tag{1.49}$$

We can adjust B_0 upwards while maintaining (1.49), so we take

$$B_0 \geq \left(\frac{A_0}{\eta}\right)^{1/2} + \left(\frac{A_0}{\tau - \eta}\right)^{1/2}. \tag{1.50}$$

Finally, we require that on Ω_n it holds

$$\min_{j\in\zeta_\circ^c}|\varepsilon_j| > \sqrt{A_0/\tau}+b_nB_0,\ \min_{\zeta:|\zeta|=\underline{h}}\hat{\sigma}_\zeta^2 > 0. \tag{1.51}$$

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

$$\min_{\delta:|\delta|=1,\,|R'\delta|>0}\sum_{i\in\zeta_h}\mathbb{I}\{|x_{in}'\delta|>a|R'\delta|\}=h-\max_{\delta:|\delta|=1,\,|R'\delta|>0}H_{\delta\zeta_h}(a). \tag{1.52}$$

Since $|\zeta_h\cap\zeta_\circ^c|=0$ then $\bar{G}_{\delta\zeta_h}(\kappa_{\delta\zeta\tau}b_n)=0$. Thus, (1.47) gives $\max_{\delta:|\delta|=1,\,|R'\delta|>0}H_{\delta\zeta_h}(a)\leq h-\nu n$. Insert this in (1.52), and use $n\geq h$ and our choice of τ to get uniformly in $\underline{h}\leq h\leq h_\circ$

$$\min_{\delta:|\delta|=1,\,|R'\delta|>0}\sum_{i\in\zeta_h}\mathbb{I}\{|x_{in}'\delta|>a|R'\delta|\}\geq\nu n\geq\nu h>\tau h>0. \tag{1.53}$$

Use $\max\sum_{i=1}^n\mathbb{I}_{A_i}=n-\min\sum_{i=1}^n\mathbb{I}_{A_i^c}$, (1.53), $\tau>0$, 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, (1.45) is satisfied. From (1.51) we have $\min_{\zeta:|\zeta|=h}\hat{\sigma}_\zeta^2\geq(\underline{h}/h)\min_{\zeta:|\zeta|=\underline{h}}\hat{\sigma}_\zeta^2>0$ for all $\underline{h}\leq h\leq h_\circ$. Then Lemma 1.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 $S=\{\zeta : |\zeta|=h,\, R'\hat{\beta}_\zeta$ 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

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

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

$$H_\zeta(a)+\bar{H}_\zeta(a)+G_\zeta(\kappa_{\zeta\tau}b_n)+\bar{G}_\zeta(\kappa_{\zeta\tau}b_n)=h. \tag{1.54}$$

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

$$S_1^h = S \cap \{\zeta : \bar{H}_\zeta(a) \geq n\tau\},$$

$$S_2^h = S \cap \{\zeta : \bar{H}_\zeta(a) < n\tau\} \cap \{\zeta : \hat{\ell}_\zeta \kappa_{\zeta\tau} \leq B_0\},$$

$$S_3^h = S \cap \{\zeta : \bar{H}_\zeta(a) < n\tau\} \cap \{\zeta : \hat{\ell}_\zeta \kappa_{\zeta\tau} > B_0\}.$$

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

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

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

Thus, $S_1^h \subseteq Z_{\tau a h} \cap \{\zeta : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0\}$. From (1.49) it follows

$$\min_{\zeta \in S_1^h} \hat{\sigma}_\zeta^2 \geq \min_{\zeta \in Z_{\tau a h}^R : |R'\hat{\beta}_\zeta - R'\beta|/\sigma > B_0} \hat{\sigma}_\zeta^2 > \hat{\sigma}_{\zeta_h}^2.$$

The set S_2^h. Let $\underline{h} \leq h \leq h_\circ$ and $\zeta \in S_2^h$. As in the proof of Lemma 1.*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

$$\frac{\hat{\sigma}_\zeta^2}{\sigma^2} = \frac{1}{h}\sum_{j \in \zeta} \frac{(y_i - x_{in}'\hat{\beta}_\zeta)^2}{\sigma^2} \geq \frac{1}{h}\sum_{j \in \zeta \cap \zeta_\circ^c} \mathbb{I}\{|x_{jn}'\hat{\delta}_\zeta| \leq \kappa_{\zeta\tau} b_n |R'\hat{\delta}_\zeta|\}\big(\varepsilon_j - \frac{\hat{\ell}_\zeta x_{jn}'\hat{\delta}_\zeta}{|R'\hat{\delta}_\zeta|}\big)^2. \quad (1.55)$$

For $\zeta \in S_2^h$ it holds $\hat{\ell}_\zeta \kappa_{\zeta\tau} \leq B_0$ by definition. Combine with $|x_{jn}'\hat{\delta}_\zeta| \leq \kappa_{\zeta\tau} b_n |R'\hat{\delta}_\zeta|$ to get $|\hat{\ell}_\zeta x_{jn}'\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| \leq b_n B_0$. Combine with (1.51) to get

$$\min_{j \in \zeta_\circ^c} |\varepsilon_j| > \sqrt{A_0/\tau} + b_n B_0 \geq \sqrt{A_0/\tau} + |\hat{\ell}_\zeta x_{jn}'\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta|. \quad (1.56)$$

For $j \in \zeta_\circ^c$, reverse triangle inequality and (1.56) show $|\varepsilon_j - \hat{\ell}_\zeta x_{jn}'\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|| \geq |\varepsilon_j| - |\hat{\ell}_\zeta x_{jn}'\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| > \sqrt{A_0/\tau}$ and therefore $(\varepsilon_j - \hat{\ell}_\zeta x_{jn}'\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|)^2 > A_0/\tau$. Insert this in

(1.55) and use the definition of G_ζ to bound

$$\hat{\sigma}_\zeta^2/\sigma^2 > h^{-1} \sum_{j\in\zeta\cap\zeta_\circ^c} \mathbb{I}\{|x'_{jn}\hat{\delta}_\zeta| \leq \kappa_{\zeta\tau} b_n |R'\hat{\delta}_\zeta|\}(A_0/\tau) = h^{-1} G_\zeta(\kappa_{\zeta\tau} b_n)(A_0/\tau). \quad (1.57)$$

It holds $\bar{H}_\zeta(a) < n\tau$ by definition of S_2^h. By (1.47), we have $H_\zeta(a) + \bar{G}_\zeta(\kappa_{\zeta\tau} b_n) \leq h - \nu n$. Combine these with identity (1.54) to get

$$G_\zeta(\kappa_{\zeta\tau} b_n) = h - \bar{H}_\zeta(a) - H_\zeta(a) - \bar{G}_\zeta(\kappa_{\zeta\tau} b_n) > n(\nu - \tau) > h\tau, \quad (1.58)$$

where the last inequality uses $n \geq h$ and the choice $2\tau < \nu$. Inserting (1.58) into (1.57) shows $\hat{\sigma}_\zeta^2/\sigma^2 > A_0$. Since $A_0 > \underline{h}^{-1} \sum_{i\in\zeta_\circ} \varepsilon_i^2$ from (1.48) and $\underline{h}^{-1} \sum_{i\in\zeta_\circ} \varepsilon_i^2 \geq h^{-1} \sum_{i\in\zeta_h} \varepsilon_i^2 \geq \hat{\sigma}_{\zeta_h}^2$, we conclude $\min_{\zeta\in S_2^h} \hat{\sigma}_\zeta^2 > \hat{\sigma}_{\zeta_h}^2$ for all $\underline{h} \leq h \leq h_\circ$.

The set S_3^h. Let $\underline{h} \leq h \leq h_\circ$ 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

$$\hat{\sigma}_\zeta^2/\sigma^2 \geq h^{-1} \sum_{i\in\zeta\cap\zeta_\circ} \mathbb{I}\{|x'_{in}\hat{\delta}_\zeta| \geq \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\}\mathbb{I}\{\varepsilon_i^2 \leq A_0/\eta\}(\varepsilon_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|)^2. \quad (1.59)$$

It holds $\hat{\ell}_\zeta \kappa_{\zeta\tau} > B_0$ by definition of S_3^h. When $|x'_{in}\hat{\delta}_\zeta| \geq \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|$ then $|\hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| > B_0$. Using (1.50) and $|\varepsilon_i| \leq (A_0/\eta)^{1/2}$ we get further

$$|\hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta|/|R'\hat{\delta}_\zeta| > \Big(\frac{A_0}{\eta}\Big)^{1/2} + \Big(\frac{A_0}{\tau - \eta}\Big)^{1/2} \geq |\varepsilon_i| + \Big(\frac{A_0}{\tau - \eta}\Big)^{1/2}. \quad (1.60)$$

Thus, reverse triangle inequality and (1.60) give

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

and $(\varepsilon_i - \hat{\ell}_\zeta x'_{in}\hat{\delta}_\zeta/|R'\hat{\delta}_\zeta|)^2 > A_0/(\tau - \eta)$. Insert this in (1.59) to bound

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

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

$$\hat{\sigma}^2_\zeta/\sigma^2 > \frac{A_0}{\tau-\eta}\Big(h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\hat{\delta}_\zeta| \geq \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} - h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{\varepsilon_i^2 > A_0/\eta\}\Big). \quad (1.61)$$

We bound the first sum in (1.61). Since $\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}_{A_i} = |\zeta\cap\zeta_\circ| - h^{-1}\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}_{A_i^c}$ then

$$\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\hat{\delta}_\zeta| \geq \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} = |\zeta\cap\zeta_\circ| - \sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\hat{\delta}_\zeta| < \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\}. \quad (1.62)$$

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

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

By definition of $\kappa_{\zeta\tau} = H_\zeta^{-1}(|\zeta\cap\zeta_\circ| - \tau n)$ we have $H_\zeta(a) \leq \max\{|\zeta\cap\zeta_\circ| - \tau n, 0\}$ for all $a < \kappa_{\zeta\tau}$, and therefore $\lim_{a\uparrow\kappa_{\zeta\tau}}H_\zeta(a) \leq \max\{|\zeta\cap\zeta_\circ| - \tau n, 0\}$. Use $|\zeta| \geq \underline{h}$ and $|\zeta_\circ^c| = n - h_\circ$ to bound $|\zeta\cap\zeta_\circ| \geq |\zeta| - |\zeta_\circ^c| \geq \underline{h} + h_\circ - n$. 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 (1.62) and use $n \geq h$ to get

$$\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{|x'_{in}\hat{\delta}_\zeta| \geq \kappa_{\zeta\tau}|R'\hat{\delta}_\zeta|\} \geq \tau n \geq \tau h. \quad (1.63)$$

For the second sum in (1.61), use (1.48) to get for all $\underline{h} \leq h \leq h_\circ$

$$\sum_{i\in\zeta\cap\zeta_\circ}\mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} \leq \sum_{i\in\zeta_\circ}\mathbb{I}\{\varepsilon_i^2 > A_0/\eta\} \leq \eta\underline{h} \leq \eta h. \quad (1.64)$$

Insert (1.63) and (1.64) in (1.61) to conclude $\hat{\sigma}^2_\zeta/\sigma^2 \geq A_0 > \hat{\sigma}^2_{\zeta_h}/\sigma^2$. Since the bounds are uniform in $\zeta \in S_3^h$ and $\underline{h} \leq h \leq h_\circ$, conclude $\min_{\zeta\in S_3^h}\hat{\sigma}^2_\zeta > \hat{\sigma}^2_{\zeta_h}$ 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 $\hat{\sigma}^2_\zeta > \hat{\sigma}^2_{\zeta_h}$. 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 1.9(v), $\exists a, \nu, \tau > 0$ and Ω_{1n} such that (1.47) holds on Ω_{1n} and $P(\Omega_{1n}) \geq 1 - \epsilon$ for large n. By Assumption 1.9(i, ii), $\exists A_0 > 0$ and Ω_{2n} such that the first inequality in (1.48) holds on Ω_{2n} and $P(\Omega_{2n}) \geq 1 - \epsilon$ for all n. By Assumption 1.9(i, ii) and Lemma 1.A.3, $B_0 > 0$ and Ω_{3n} exist such that for any $\zeta_h \subseteq \zeta_\circ$ with $|\zeta_h| = h$, (1.49) holds on Ω_{3n} and $P(\Omega_{3n}) \geq 1 - \epsilon$ for all n. Finally, by Assumption 1.9(iii, iv), Ω_{4n} exists such that (1.51) 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 1.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 1.A.11. *Let R be a $p \times s$ matrix and $S \subseteq \mathbb{R}^p$ a linear subspace with $S^\perp_\delta = S^\perp \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \varnothing$. For any finite subset $\{x_j\}_{j \in J} \subset \mathbb{R}^p \setminus S$, a vector $\delta^* \in S^\perp_\delta$ 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 1.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, $\exists \bar{\delta} \in S^\perp_\delta$. 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 $\delta_{t^*} \in S^\perp_\delta$. □

Lemma 1.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 $P(z_n^2 \leq 1/k) \geq 1 - 1/k$ for $n \geq N_k$. Define the sequence k_n by $k_n = k$ for $n = N_k, \ldots, N_{k+1} - 1$. 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 $\log(k)/k \leq \epsilon$ for all $k \geq k^*$. Then for all n such that $k_n \geq k^*$

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

Proof of Theorem 1.3. We check that Assumptions 1.1 and 1.7 imply Assumption 1.9. Theorem 1.3 then follows from Lemma 1.A.10.

Assumptions 1.9(i, ii, iii) are given in the statements of Assumptions 1.1 and 1.7.

Assumption 1.9(iv) holds by Lemma 1.A.12 and Assumption1.7(i).

We show that Assumption 1.6, 1.7(i), 1.8. imply Assumption 1.9(v). The proof of this is closely related to that of Lemma 1.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

$$K_{n\tau}(a) = \max_{\underline{h} \le h \le h_\circ} \max_{\zeta:|\zeta|=h} \max_{\delta:|\delta|=1,\, |R'\delta|>0} \left\{ H_{\delta\zeta}(a) + \bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau} b_n) - h \right\} \le -n\nu \tag{1.65}$$

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 1.9(v) holds.

0. Choice of constants $\nu, \tau > 0$. By Assumption 1.8, we can choose $\nu > 0$ such that

$$\sup_{\delta:|\delta|=1,|R'\delta|>0} \{\lambda_\circ P_{good}\{x : x'\delta = 0\} - (1-\lambda_\circ) P_{out}\{x : x'\delta = 0\}\} < \underline{\lambda} + \lambda_\circ - 1 - 3\nu. \tag{1.66}$$

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

$$|\text{proj}_S(\delta)| \le \max_{k \in N} |\text{proj}_{S_k}(\delta)|/c_N. \tag{1.67}$$

and

$$0 \le \max_{k \in N} |\text{proj}_{S_k}(\delta)| < |R'\delta| c_N \text{ for some } |\delta| = 1$$
$$\implies S_N^{\perp} \cap \{\delta : |\delta| = 1, \, |R'\delta| > 0\} \neq \varnothing. \tag{1.68}$$

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

3. Choice of constant a^*. By Lemma 1.A.5, there exists $a^* > 0$ such that

$$\max_{k \le K} \sup_{\delta \in S_k : |\delta| = 1} \mu_k\{x : |x'\delta| \le a^*/c\} \le \epsilon'/K. \tag{1.69}$$

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

$$h_\circ^{-1} \sum_{i \in \zeta_\circ} \mathbb{I}\{|x_{in}'\delta| \le a\} \le P_{good}\{x : |x'\delta| \le a\} + \epsilon' \tag{1.70}$$

$$(n - h_\circ)^{-1} \sum_{i \in \zeta_\circ} \mathbb{I}\{x_{jn}'\delta = 0\} \ge P_{out}\{x : |x'\delta| \le a\} - \epsilon'. \tag{1.71}$$

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

$$\gamma_{\lceil (n-h_\circ)(1-\epsilon') \rceil} \le b_n a^*. \tag{1.72}$$

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 (1.65). 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 1.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 (1.70) 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

$$H_{\delta\zeta}(a) \leq h_\circ \sum_{k \leq K} \mu_k(\{x : |x'\delta| \leq \kappa_{\delta\zeta\tau}(a)|R'\delta|\}) + (2h_\circ\epsilon' + \tau n). \tag{1.73}$$

Let $\delta_k = \text{proj}_{S_k}(\delta)$, $\delta_k^\perp = \text{proj}_{S_l^\perp}(\delta)$, and $N_{\delta\zeta} = \{k \leq K : |\delta_k| \leq \kappa_{\delta\zeta\tau}(a)|R'\delta|(c/a^*)\}$. 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 $\mu_k(\mathbb{R}^p \setminus S_k) \leq \mu_k(\mathbb{R}^p \setminus A_k) = 0$ by property (ii) in part 1, it follows

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

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 $|x'\delta_k| \leq \kappa_{\delta\zeta\tau}(a)|R'\delta|$ 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

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

Bound the right-hand-side using (1.69) to conclude that for all $k \notin N_{\delta\zeta}$

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

Insert this into (1.73) and use $|N_{\delta\zeta}^c| \leq K$ to get

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

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 $\sum_{k \in N_{\delta\zeta}} \mu_k(B) \leq \sum_{k \in N_{\delta\zeta}} \mu_k(S_{\delta\zeta}) \leq P_{good}(S_{\delta\zeta})$. Thus, for all $\underline{h} \leq h \leq h_\circ$, h-subsets ζ, and $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ it holds

$$H_{\delta\zeta}(a) \leq h_\circ P_{good}(\mathcal{S}_{\delta\zeta}) + (3h_\circ\epsilon' + n\tau). \tag{1.74}$$

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

$\kappa_{\delta\zeta\tau}(a)$ to get

$$\begin{aligned}\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n) &= \sum_{j\in\zeta\cap\zeta_\circ^c}\mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}b_n|R'\delta|\} \leq \sum_{j\in\zeta_\circ^c}\mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|\} \\ &\leq \sum_{j\in\zeta_\circ^c}\mathbb{I}\{x_j \in S_{\delta\zeta}\}\mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|\} + \sum_{i\in\zeta_\circ^c}\mathbb{I}\{x_j \notin S_{\delta\zeta}\}. \quad (1.75)\end{aligned}$$

Consider the first sum in (1.75). If $x \in S_{\delta\zeta}$ then $|x'\delta| = |x'\text{proj}_{S_{\delta\zeta}}(\delta)|$ and by Cauchy-Schwarz inequality $|x'\delta| \leq |x||\text{proj}_{S_{\delta\zeta}}(\delta)|$. By (1.67) and definition of $N_{\delta\zeta}$ it holds $|\text{proj}_{S_{\delta\zeta}}(\delta)| \leq \max_{k\in N_{\delta\zeta}}|\delta_k|/c \leq \kappa_{\delta\zeta\tau}(a)|R'\delta|/a^*$. Thus, for $x \in S_{\delta\zeta}$

$$|x'\delta| \leq |x|\,|\text{proj}_{S_{\delta\zeta}}(\delta)| \leq |x|\kappa_{\delta\zeta\tau}(a)|R'\delta|/a^*. \quad (1.76)$$

From (1.76), 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,

$$\mathcal{I}_{\delta\zeta\tau} = \sum_{i\in\zeta_\circ^c}\mathbb{I}\{x_j \in S_{\delta\zeta}\}\mathbb{I}\{|x_j'\delta| > \kappa_{\delta\zeta\tau}(a)b_n|R'\delta|\} \leq \sum_{i\in\zeta_\circ^c}\mathbb{I}\{|x_j| > b_n a^*\}. \quad (1.77)$$

Use $\gamma_{\lceil(n-h_\circ)(1-\epsilon')\rceil} \leq b_n a^*$ from (1.72) to get

$$\mathcal{I}_{\delta\zeta\tau} \leq \sum_{i\in\zeta_\circ^c}\mathbb{I}\{|x_j| > b_n a^*\} \leq \mathbb{I}\{|x_j| > \gamma_{\lceil(n-h_\circ)(1-\epsilon')\rceil}\} \leq (n-h_\circ)\epsilon'. \quad (1.78)$$

Insert (1.78) back into (1.75) to conclude that for all h, ζ, δ

$$\bar{G}_{\delta\zeta}(\kappa_{\delta\zeta\tau}b_n) \leq (n-h_\circ)\epsilon' + \sum_{i\in\zeta_\circ^c}\mathbb{I}\{x_j \notin S_{\delta\zeta}\} = (n-h_\circ) - \sum_{i\in\zeta_\circ^c}\mathbb{I}\{x_j \in S_{\delta\zeta}\} + (n-h_\circ)\epsilon'. \quad (1.79)$$

5.3. Deterministic analysis on Ω_n - Conclusion. For $\underline{h} \leq h \leq h_\circ$ 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 (1.74) and (1.79) to see that with $\epsilon_n^* = 3h_\circ\epsilon' + \tau n + (n-h_\circ)\epsilon'$ it holds

$$\mathcal{E}_{\delta\zeta} = h_\circ P_{good}(S_{\delta\zeta}) + (n-h_\circ) - \sum_{j\in\zeta_\circ^c}\mathbb{I}\{x_j \in S_{\delta\zeta}\} - h + \epsilon_n^*, \quad (1.80)$$

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 (1.68), it follows that $S_{\delta\zeta}^{\perp} \cap \{\delta : |\delta| = 1, |R'\delta| > 0\} \neq \emptyset$. Further, by Lemma 1.$A$.11, there exists $\delta_\zeta^* \in S_{\delta\zeta}^{\perp} \cap \{\delta : |\delta| = 1, |R'\delta| > 0\}$ such that $x_j'\delta_\zeta^* \neq 0$ for all $j \in \zeta_\circ^c$ with $x_j \notin S_{\delta\zeta}$. The vector δ_ζ^* then satisfies

$$P_{good}(S_{\delta\zeta}) \leq P_{good}\{x : x'\delta_\zeta^* = 0\}, \sum_{j \in \zeta_\circ^c} \mathbb{I}\{x_j \in S_{\delta\zeta}\} = \sum_{j \in \zeta_\circ^c} \mathbb{I}\{x_j'\delta_\zeta^* = 0\}. \tag{1.81}$$

Plug (1.81) in (1.80) and use (1.71) to conclude

$$\begin{aligned} \mathcal{E}_{\delta\zeta} &\leq h_\circ P_{good}\{x : x'\delta_\zeta^* = 0\} + (n - h_\circ) - \sum_{j \in \zeta_\circ^c} \mathbb{I}\{x_j'\delta_\zeta^* = 0\} - h + \epsilon_n^* \\ &\leq h_\circ P_{good}\{x : x'\delta_\zeta^* = 0\} + (n - h_\circ) - (n - h_\circ)P_{out}\{x : x'\delta_\zeta^* = 0\} - h + 2\epsilon_n^*. \end{aligned} \tag{1.82}$$

Since $h_\circ/n = \lambda_\circ + o(1)$ by assumption, using (1.66) 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 (1.82) 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 (1.65).

3. Probability Analysis Construct $\Omega_n = \Omega_{1n} \cap \Omega_{2n} \cap \Omega_{3n}$. Using Assumption 1.6 and Lemma 1.A.8, we can find sets (Ω_{2n}) with $P(\Omega_{2n}) \geq 1 - \epsilon$ for n so that (1.70) and (1.71) hold on Ω_{2n}. Assumption 1.6 further gives that $\{|x_j|\}_{j \in \zeta_\circ^c}$ 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 (1.72) holds on Ω_{3n}. □

1.A.6 Proof of Theorem 1.4

Lemma 1.A.13. *Suppose Assumption 1.6. 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

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

By Assumption 1.6, for any B it holds with probability one that

$$\sum_{i\in\zeta_\circ} \mathbb{I}\{x_{in} \in B\} = \sum_{i=1}^{h_\circ} \mathbb{I}\{x_i^{\text{good}} \in B\} \to P_{good}(B) \tag{1.83}$$

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

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

Proof of Theorem 1.4. 1. Population counterpart. Let $\underline{\boldsymbol{\lambda}} = \max\{\underline{\boldsymbol{\lambda}}^{(1)} + \epsilon/2, 1-\underline{\lambda}_\circ\}$, where

$$\underline{\boldsymbol{\lambda}}^{(1)} = \sup_{\delta:|\delta|=1, |R'\delta|>0} \{p_{full,\delta} + (1-\lambda_\circ)(1-b_{out,\delta})\}. \tag{1.84}$$

We show $\underline{\boldsymbol{\lambda}}$ satisfies Assumption 1.3. To do this, we check $1-\lambda_\circ < \underline{\boldsymbol{\lambda}} \le \lambda_\circ$ and (1.4).

Begin by checking (1.4). Assumption 1.6 implies Assumption 1.4. Thus, by Proposition 1.3.2, (1.4) holds if Assumption 1.5 is satisfied, that is

$$\sup_{\delta:|\delta|=1, |R'\delta|>0} \lambda_\circ p_{good,\delta} < \underline{\boldsymbol{\lambda}} + \lambda_\circ - 1. \tag{1.85}$$

Using identity $\lambda_\circ p_{good,\delta} = p_{full,\delta} - (1-\lambda_\circ)p_{out,\delta}$ and subtracting $(\lambda_\circ - 1)$ from both sides, (1.85) holds if

$$\sup_{\delta:|\delta|=1, |R'\delta|>0} \{p_{full,\delta} + (1-\lambda_\circ)(1-p_{out,\delta})\} < \underline{\boldsymbol{\lambda}}. \tag{1.86}$$

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, (1.84) 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 $\underline{\boldsymbol{\lambda}} = \max\{\underline{\boldsymbol{\lambda}}^{(1)} + \epsilon/2, 1 - \underline{\lambda}_\circ\}$, (1.86) 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 1.6. 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 (1.84),

$$p_{full,\delta} + (1 - \underline{\lambda}_\circ)(1 - b_{out,\delta}) \leq \underline{\lambda}_\circ \tag{1.87}$$

$\forall \delta : |R'\delta| > 0$. Rearranging, (1.87) is equivalent to

$$\frac{1 + p_{full,\delta} - b_{out,\delta}}{2 - b_{out,\delta}} \leq \underline{\lambda}_\circ. \tag{1.88}$$

The supremum of the left-hand-side in (1.88) over $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ is denoted $\underline{\lambda}_\circ^{(1)}$ in (1.8), 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 (1.7). 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 1.6 and Lemma 1.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 1.3. Assumption 1.1 and Theorem 1.2 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 $\max_{\underline{h} \leq h \leq h_\circ} \max_{\zeta \in \mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$. □

1.A.7 Proof of Theorem 1.5

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

$$\underline{\boldsymbol{\lambda}}^{(2)} = \sup_{\delta : |\delta| = 1, \, |R'\delta| > 0} \left[p_{full,\delta} + \mathbb{I}\{b_{out,\delta} < 1/2\}(1 - \underline{\lambda}_\circ)(1 - 2b_{out,\delta}) \right]. \tag{1.89}$$

We show $\underline{\boldsymbol{\lambda}}$ satisfies Assumption 1.8. We need to check $1-\lambda_\circ < \underline{\boldsymbol{\lambda}} \leq \lambda_\circ$ and

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} \{\lambda_\circ p_{good,\delta} - (1-\lambda_\circ)p_{out,\delta}\} < \underline{\boldsymbol{\lambda}} + \lambda_\circ - 1. \tag{1.90}$$

We check (1.90). 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 (1.90) holds if

$$\sup_{\delta:|\delta|=1,\,|R'\delta|>0} \{p_{full,\delta} + (1-\lambda_\circ)(1-2p_{out,\delta})\} < \underline{\boldsymbol{\lambda}}. \tag{1.91}$$

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, (1.91) holds if

$$\sup_{\delta:|\delta|=1,|R'\delta|>0} W_\delta = \sup_{\delta:|\delta|=1,|R'\delta|>0} \{p_{full,\delta} + (1-\lambda_\circ)(1-2b_{out,\delta})\} < \underline{\boldsymbol{\lambda}}. \tag{1.92}$$

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 $\underline{\boldsymbol{\lambda}} \geq \underline{\boldsymbol{\lambda}}^{(2)} + \epsilon/2$, (1.92) 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 (1.89), if

$$p_{full,\delta} + \mathbb{I}\{b_{out,\delta} < 1/2\}(1-\underline{\lambda}_\circ)(1-2b_{out,\delta}) \leq \underline{\lambda}_\circ \tag{1.93}$$

$\forall\delta : |R'\delta| > 0$. For δ with $b_{out,\delta} < 1/2$, rearrange to see that (1.93) holds if

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

For $b_{out,\delta} \geq 1/2$, (1.93) holds if $p_{full,\delta} \leq \lambda_\circ$. Combined, (1.93) holds if $\forall\delta : |R'\delta| > 0$

$$\mathbb{I}\{b_{out,\delta} < 1/2\}\frac{1+p_{full,\delta} - 2b_{out,\delta}}{2(1-b_{out,\delta})} + \mathbb{I}\{b_{out,\delta} \geq 1/2\}p_{full,\delta} \leq \underline{\lambda}_\circ. \tag{1.94}$$

Supremum of the left-hand-side in (1.94) over $\delta \in \{\delta : |\delta| = 1, |R'\delta| > 0\}$ is denoted $\underline{\lambda}_\circ^{(2)}$ in (1.11). 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 (1.10). 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 1.6 and Lemma 1.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 1.8. Assumptions 1.1, 1.6, 1.7 and Theorem 1.3 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 $\max_{\underline{h} \leq h \leq h_\circ} \max_{\zeta \in \mathcal{M}_h} |R'\hat{\beta}_\zeta| = O_p(1)$. □

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 1.3.4) 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.

1.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 1.3. Algorithm 1.3 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 1.3 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 1.3 (Exhaustive hyperplane search)**.**
(0) Initialise $m^* \leftarrow 0$ *and* $\mathcal{I}_{skip} \leftarrow \emptyset$.
(1) Let $\iota : \{1, \ldots, \binom{|\mathcal{X}|}{p-1}\} \to [\mathcal{X}]^{p-1}$ *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^**.*

1.B.2 Randomised search

Algorithm 1.4 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 1.4 could be adapted to return the k largest hyperplanes for $k > 0$ and orthogonal vectors.

Algorithm 1.4 (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 $1, \dots, n_{iter}$*:*

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 1.4 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.

1.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 1.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* $\sup_{\delta:|\delta|=1} P\{x : x'\delta = 0\} = \max_{1\leq j\leq k} \max_{a\in\mathbb{R}} P(z_j = a)$.

Proof of Remark 1.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, $P(x) \leq P\{\pi_j(x)\}\ \forall x \in \mathcal{X}$. Since S_δ has at most dimension k, by the above observation $\exists j$ such that $|\pi_j(S_\delta)| = |S_\delta|$. Conclude

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

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 for Chapter 1

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

Amaldi, E. and V. Kann (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 M. Riani (2000). Robust Diagnostic Regression Analysis. 2nd. New York: Springer-Verlag.

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

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

Berenguer-Rico, V. and B. Nielsen (2025a). Least Trimmed Squares Asymptotics: Cointegration and outliers. Discussion Paper. Nuffield College.

Berenguer-Rico, V., S. Johansen, and B. Nielsen (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 B. Nielsen (2025b). Least Trimmed Squares: Nuisance parameter free asymptotics. *Econometric Theory*, 1–39.

Billingsley, P. (1995). Probability and Measure. Third ed. New York, NY: John Wiley and Sons.

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

Breiman, L. (1992). Probability. Philadelphia, PA: Society for Industrial and Applied Mathematics. DOI: `10.1137/1.9781611971286`.

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

Čížek, P. (2004). Asymptotics of Least Trimmed Squares Regression. CentER Discussion Paper 2004-72. Tilburg University.

— (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.

— (1993). Aspects of robust linear regression. *Annals of Statistics* 21(4), 1843–1899.

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

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

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

Hadi, A. S. and A. Luceño (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis* 25(3), 251–272.

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

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

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

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

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

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

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

Mili, L. and C. W. Coakley (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. 2nd. Vol. 135. Graduate Texts in Mathematics. New York: Springer.

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

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

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

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.

Chapter 2

Asymptotic properties of Impulse Indicator Saturation under outlier contamination

Abstract: *Impulse Indicator Saturation (IIS) is an outlier robust method for estimating a linear equation. We study the asymptotic properties of a stylised IIS algorithm under contamination. This contrasts with previous work, where the properties of IIS have only been established for clean data. We show that IIS has an 'oracle property': it is asymptotically equivalent to an infeasible least squares estimator that perfectly removes all outliers. We use the oracle property to derive the asymptotic distribution of IIS in cross-sectional and time series models with outliers. We also find a limit theory for the number of misclassified 'good' observations, which may guide tuning parameter selection for IIS. Simulations and an empirical illustration with macroeconomic time series data are provided.*

2.1 Introduction

Impulse Indicator Saturation (IIS) (Hendry et al., 2008) is an outlier-robust method for estimating a linear equation $y_i = x_i'\beta + \varepsilon_i$. We analyse a stylised IIS algorithm inspired by Hendry (1999). The algorithm assumes data can be split into two parts: one free of outliers and the other potentially contaminated. The 'clean' part is used

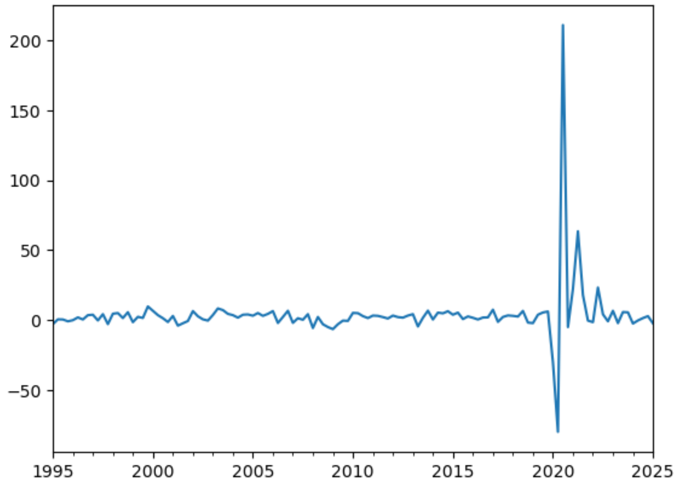


Figure 2.1: US real PCE on food and accommodation services (quarterly log-difference)

to compute an initial least squares estimator and scaled residuals. Observations are then classified into 'good' and 'outliers', depending on whether their residuals exceed a cut-off chosen by the user. The IIS estimator is given by the OLS estimator using the observations classified as 'good'.

As a motivating example, consider US personal consumption expenditure (PCE) on food and accommodation services in Figure 2.1. From 2020 onwards the series contains 'outliers' associated with the shutdown of hospitality services during the COVID-19 pandemic. At the end of sample, the series appears to revert back to its pre-2020 behaviour. It could therefore be of interest to fit an autoregressive equation $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ to the data, while allowing a small number of observations around the pandemic to be outlying. IIS offers a systematic approach to handling this problem.

We study the asymptotic properties of IIS under outlier contamination. This contrasts with existing work on IIS, where all data have been assumed 'clean' (Johansen and Nielsen, 2016; Hendry et al., 2008). We show that IIS has an 'oracle' property: it is asymptotically equivalent to an infeasible least squares estimator that perfectly removes all outliers. We prove this for (iterated) Huber-skip estimators (Huber, 1964; Bickel, 1975), which contain IIS as a special case.

Our conditions for the oracle property are general and allow for a range of data generating processes, such as cross-sections and time series. The main assumptions are that (i) the magnitude of outliers and (ii) the IIS cut-off are growing in the sample size.

Having a growing magnitude for the outliers is inspired by a recent maximum likelihood model for the least trimmed squares (LTS) estimator (Berenguer-Rico et al., 2023). Our theory and simulations suggest that IIS performs best when the rate of growth in outlying errors is sufficiently fast, such that the 'outlying' observations are not too concentrated near the tails of 'good' errors. If there is little separation between the two groups at the tails, larger sample sizes are needed for the oracle approximation to kick in.

Having growing cut-offs avoids nuisance parameters in the asymptotic distribution of IIS, which appear if a fixed cut-off is used (Johansen and Nielsen, 2016). We consider growing cut-offs that are tuned to a reference distribution, such as the standard normal, for the good errors. To guide selection of the cut-off, we find a limiting distribution for the *gauge*, the number of misclassified 'good' observations. Given a cut-off, IIS allows the number of outliers to be estimated simultaneously with the regression parameters, contrasting with other robust methods, such as LTS and M-estimators, where the share of outliers needs to be estimated separately.

Using the oracle property, we write down models where IIS inference is unaffected by the search for outliers and can be based on standard least squares critical values. For stationary autoregressive models, we show that standard inference applies with weaker restrictions on *additive* than *innovative* contamination, where the latter are outliers that induce dynamics while the former do not. For unit root testing with standard Dickey-Fuller critical values, a similar finding is uncovered in the simulations. These findings resonate with Berenguer-Rico and Nielsen (forthcoming), who study LTS estimators in a cointegrated model, but do not make a distinction between additive and innovative contamination.

The paper is structured as follows. Section 2 defines stylised IIS and Huber-skip estimators. Section 3 introduces the oracle property. Section 4 discusses choice of the cut-off for Huber-skip estimators. Section 5 establishes the asymptotic distribution of stylised IIS in cross-sectional and time series models with outliers. Section 7 contains simulations. Section 8 provides an illustration with the data from Figure 2.1. Section 9 concludes.

2.2 IIS and Huber-skip estimators

Consider a linear equation $y_i = x_i'\beta + \sigma\varepsilon_i$ for $i = 1, ..., n$. We describe a stylised IIS algorithm inspired by Hendry (1999). The user chooses a cut-off $c > 0$ and an initial subset $\mathcal{T} \subseteq \{1, ..., n\}$ of observations known to be 'good'. The following steps are then taken.

Algorithm 1 - Stylised IIS

1. Calculate initial least squares estimators $\tilde{\beta}, \tilde{\sigma}$ using observations $\mathcal{T}$.
2. Compute indicators $v_i = \mathbb{I}\{|y_i - x_i'\tilde{\beta}| \leq \tilde{\sigma}c\}$ for $i = 1, ..., n$. Classify observations with $v_i = 1$ as 'good' and the remaining as 'outliers'.
3. IIS estimator is given by the OLS estimator using observations classified as 'good'

$$\hat{\beta}_{IIS} = \arg\min_{\beta} \sum_{i=1}^{n} v_i(y_i - x_i'\beta)^2, \ \hat{h} = \sum_{i=1}^{n} v_i, \ \hat{\sigma}^2_{IIS} = \hat{h}^{-1} \sum_{i=1}^{n} v_i(y_i - x_i'\hat{\beta})^2.$$

The stylised IIS algorithm is intended for data that can be split into a 'clean' part and a 'possibly contaminated' part. In macroeconomic data, outliers are often associated with historical events, such as business cycle fluctuations (Balke and Fomby, 1994), geopolitical events, or natural disasters. Such knowledge can motivate the choice of a 'clean' subsample.

In the empirical illustration using data from Figure 2.1, we treat the subsample ending in 2019Q4 as 'clean' and the data from 2020Q1 onwards as 'possibly contaminated'. If the initial estimators are consistent, they can be used to detect which observations in the possibly contaminated part are outliers. Estimates are then updated using least squares on the set of observations classified as 'good'.

IIS has been extended to setting where a 'clean' part is not known a priori (Hendry et al., 2008). IIS with a more complicated search for outliers has also been incorporated into wider model selection algorithms (Doornik, 2009; Pretis et al., 2018).

Algorithm 1 is a special case of an (iterated) Huber-skip estimator (Huber, 1964; Bickel, 1975; Welsh and Ronchetti, 2002). A Huber-skip estimator starts with the choice of a cut-off $c > 0$ and initial estimators $\tilde{\beta}, \tilde{\sigma}$. For stylised IIS, the initial estimators are the least squares estimators in step 1. A one-step Huber-skip estimator is then computed following steps 2-3 of Algorithm 1. If the initial estimator is itself an $(m-1)$-step Huber-skip estimator, the updated estimator is called an m-step Huber-skip estimator.

2.3 An oracle property for Huber-skip estimators

We show that Huber-skip estimators possess an oracle property: they are asymptotically equivalent to an infeasible least squares estimator that removes all 'outliers'. This result is derived under general conditions that cover different data generating processes.

2.3.1 Sequence of data generating processes

Consider the equation $y_i = x_i'\beta + \sigma\varepsilon_i$ for $i = 1, ..., n$. Assume each observation is either a 'good' or an 'outlier'. Indices of the 'good' observations are given by a deterministic sequence of sets ζ_n with h_n elements. We write h for h_n to simplify notation and suppose $h = \lfloor \lambda n \rfloor$ where $\lambda \in (0, 1]$, allowing for a non-vanishing share of outliers.

We state the assumptions used to prove an oracle property for Huber-skip estimators. To unify the treatment of different models, suppose there exist diagonal $p \times p$ normalisation matrices N with, for example, elements $n^{-1/2}$ corresponding to stationary regressors, n^{-1} for random walks, and $n^{-3/2}$ for linear trends.

Assumption 2.1. *Suppose the following.*

(i) ***Regressors*** *satisfy*

- *(a)* $\max_{1\leq i\leq n} |N'x_i| = \mathcal{O}_p(n^{-\phi})$ *for some* $0 < \phi < 1/2$
- *(b)* $\sum_{i\in\zeta_n} N'x_i\varepsilon_i = \mathcal{O}_p(1)$ *and* $\sum_{i\in\zeta_n} N'x_ix_i'N \to_d \Sigma$ *where* Σ *non-singular a.s.*

(ii) Let $m_n^2 = \min\{(\max_{i\in\zeta_n} \varepsilon_i)^2, (\min_{i\in\zeta_n} \varepsilon_i)^2\}$. *The* ***'good' errors*** *satisfy*

- *(a)* $1/m_n^2 = o_p(1)$
- *(b)* $\max_{i\in\zeta_n} \varepsilon_i^2/m_n^2 = \mathcal{O}_p(1)$
- *(c)* $m_n^2 = \mathcal{O}_p(n^\eta)L(n)$ *for some* $0 \leq \eta < \min\{\phi, 1 - 2\phi\}$ *and slowly varying function* L*, that is* $L(kx)/L(x) \to 1$ *as* $x \to \infty$ *for all* $k > 0$.

(iii) ***Cut-off*** $c_n \to \infty$ *satisfies the following.*

- *(a)* $c_n^2 = O(n^\eta)L(n)$.
- *(b) For any* $A > 0$ *there exists* $0 < \rho < \phi - \eta/2$ *such that*

$$P(\sum_{i\in\zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A\} < n^\rho) \to 1 \tag{2.1}$$

and

$$P(\sum_{j\in\zeta_n^c} \mathbb{I}\{|\varepsilon_j| < c_n + n^{-\phi}A\} < n^\rho) \to 1. \tag{2.2}$$

Assumption 2.1 is inspired by Berenguer-Rico and Nielsen (2025) and Johansen and Nielsen (2016) and allows for general types of 'good' errors that could be serially correlated or heteroskedastic. For specific models, the conditions may be simplified. We discuss this with reference to the following running example.

Running example. *Let* $y_i = \beta y_{i-1} + \varepsilon_i$ *where* $\varepsilon_i \overset{iid}{\sim} N(0,1)$ *for* $i \in \zeta_n$*. Suppose* $\varepsilon_j = \max_{i\in\zeta_n} \varepsilon_i + d + u_j$ *for* $j \in \zeta_n^c$*, where* $d \geq 0$ *is constant and* $u_j \overset{iid}{\sim} \text{Uniform}[0,1]$.

Example 2.3.1. *If* $|\beta| < 1$ *and* $N = n^{-1/2}$ *in the running example then Assumption 2.1(ia) holds* $\forall\phi \in (0, 1/2)$*. To see this, note that* $\max_{i\leq n} |y_{i-1}| = \mathcal{O}(\max_{i\leq n} |\varepsilon_i|)$ *since* $|\beta| < 1$*, and combine this with* $\max_{i\in\zeta_n} \varepsilon_i = O_p(\sqrt{2\log h})$ *from normality of the good innovations.*

Example 2.3.2. *If* $\beta = 1$ *and* $N = n^{-1}$ *in the running example then Assumption 2.1 requires* $\max_{i\leq n} |y_{i-1}| = \mathcal{O}_p(n^{1-\phi})$*. It holds* $y_i = \sum_{s\in\zeta_n:s\leq i} \varepsilon_s + \sum_{t\in\zeta_n^c:t\leq i} \varepsilon_t$*, where the first term is* $\mathcal{O}_p(n^{1/2})$ *and the second term is of order* $(2\log h)^{1/2}(n-h)$*. Thus, we need* $n - h = \mathcal{O}(n^{1-\phi})$ *for some* $0 < \phi < 1/2$*, requiring the share of outliers to be vanishing.*

Assumption 2.1(ii) holds for standard normal good errors, such as in the running example, with $\eta = 0$ and $L(n) = 2\log n$. For t_d distributed good errors it holds with $L(n) = 1$ and $\eta = 2/d$ (Berenguer-Rico and Nielsen, 2025, Appendix B), provided that $2/d < \min\{\phi, 1-2\phi\} \leq 1/3$. The proofs only depend on L satisfying $L(x) = o(x^\alpha)$ for any $\alpha > 0$, which holds if L is slowly varying (Galambos, 1987, Appendix III).

Assumption 2.1(iii) is a general assumption on the cut-off. Section 2.4 discusses how this assumption simplifies under i.i.d. good errors, such as those in the running example.

Assumption 2.1(iii) requires the cut-offs to grow with the sample size. This avoids nuisance parameters in the asymptotic distribution of Huber-skip estimators (cf. Johansen and Nielsen, 2016). Condition (2.2) further requires that the number of 'outlying' errors smaller than $c_n + n^{-\phi}A$ is limited. As $c_n \to \infty$, it follows that the majority of outlying errors are also growing in magnitude.

For the initial estimators, we require consistency at appropriate rates.

Assumption 2.2. *Initial estimators satisfy* $N^{-1}(\tilde{\beta} - \beta), n^{1/2}(\tilde{\sigma} - \sigma) = \mathcal{O}_p(1)$.

Consistent scale estimation in the presence of outliers is challenging. For example, M-estimators for scale are generally inconsistent (Klooster and Nielsen, 2025). For the stylised IIS algorithm, Assumption 2.2 can be viewed as a condition on the initial sets $\mathcal{I}_n$. We provide sufficient conditions on the initial sets for specific models in Section 2.5.

2.3.2 Oracle property

If the above assumptions hold, Huber-skip estimators possess an 'oracle property': they are asymptotically equivalent to an infeasible least squares estimator that perfectly removes outliers.

Theorem 2.1. *Suppose Assumption 2.1, 2.2. Then one-step Huber-skip estimators* $\hat{\beta}_{(1)}, \hat{\sigma}_{(1)}$ *satisfy*

$$N^{-1}(\hat{\beta}_{(1)} - \hat{\beta}_{\zeta_n}),\ n^{1/2}(\hat{\sigma}_{(1)} - \hat{\sigma}_{\zeta_n}) = o_p(1), \tag{2.3}$$

where $\hat{\beta}_{\zeta_n}, \hat{\sigma}_{\zeta_n}$ *are least squares estimators using observations* ζ_n*. If further* $N^{-1}(\hat{\beta}_{\zeta_n} - \beta)$ *and* $n^{1/2}(\hat{\sigma}_{\zeta_n} - \sigma)$ *are* $\mathcal{O}_p(1)$ *then* (2.3) *holds for the* m*-step Huber-skip for any* $m \geq 1$*.*

Theorem 2.1 allows heteroskedastic good errors, as long as they satisfy Assumption 2.1(*ii*). For inference, consistency of an appropriate covariance estimator needs to be established separately.

Exogeneity conditions on the regressors x_{in} are not needed for Theorem 2.1, which only concerns the relationship between Huber-skip and oracle estimators. Exogeneity is typically needed to control the properties of $n^{1/2}(\hat{\beta}_{\zeta_n} - \beta)$, which determines the asymptotic distribution of IIS when the oracle property holds. We give some specific examples in Section 2.5.

heorem 2.1 relies on a high-level condition on the cut-offs, specified in Assumption 2.1(*iii*). We will now simplify these conditions in the case where good errors are i.i.d.

2.4 Choice of the cut-off for Huber-skip estimators

We consider two ways of tuning the cut-off for a Huber-skip estimator when good errors are independent and identically distributed. In the first tuning, cut-offs grow at the rate of maxima of the good errors. Outliers and good observations then need to be 'separated' in order for the oracle property to hold. In the second tuning, cut-offs grow at the rate of intermediate order statistics from a reference distribution. This allows for weaker separation, but simulations (Section 2.6) suggest larger samples sizes are needed for the asymptotics to kick in.

2.4.1 Fast growing cut-offs

We impose more structure on the good errors, and require them to be independent and identically distributed, such as in the running example. This contrasts with Assumption 2.1(ii), which allows for more general good errors.

Assumption 2.3. *Suppose good errors ε_i for $i \in \zeta_n$ are independent and identically distributed with distribution function F and density f supported on $\mathbb{R}$ satisfying*

(i) f is symmetric, continuous, and decreasing in the tails

(ii) $f(c_n)/\{1 - F(c_n)\} = \mathcal{O}(c_n)$

(iii) $f(c_n - n^{-\phi}A)/f(c_n) = \mathcal{O}(1)$ for all $A > 0$ and $0 < \phi < 1/2$ as in Assumption 2.1.

Assumption 2.3 holds for the standard normal distribution if $c_n = \mathcal{O}(n^{\phi})$ and for a t_d-distribution with any $c_n \to \infty$ (Appendix 2.D). Assumption 2.3(ii) restricts the hazard function $f(c_n)/\{1 - F(c_n)\}$ and is implied by the von Mises condition $\lim_{t\to\infty} t\{1 - F(t)\}/f(t) = \alpha$ for a type 2 extreme value distribution. Assumption 2.3(iii) bounds the local rate of change in f.

Assumption 2.4. *Suppose the following.*

(i) Cut-offs satisfy $n\{1 - F(c_n)\} = \tau$ for some $\tau > 0$.

(ii) There exists distribution G and a_n, b_n such that $P(a_n\{\max_{i\in\zeta_n} \varepsilon_i - b_n\} \leq x) \to G(x)\ \forall x \in \mathbb{R}$.

(iii) Define $m_n^2 = \min\{(\max_{i\in\zeta_n} \varepsilon_i)^2, (\min_{i\in\zeta_n} \varepsilon_i)^2\}$ as above. For any $A > 0$ there exists $0 < \rho < \phi - \eta/2$ such that for all $x \geq 0$

$$P(\sum_{j\in\zeta_n^c} \mathbb{I}\{|\varepsilon_j| < m_n + x/a_n + n^{-\phi}A\} < n^{\rho}) \to 1.$$

Assumption 2.4(i) tunes the cut-off to tail probabilities of the good errors, such as the normal distribution in our running example. Johansen and Nielsen (2016) used this tuning to analyse IIS in clean samples. Cut-offs that satisfy Assumption 2.4(i) grow in n.

Assumption 2.4(ii) is adopted for convenience in the proofs and requires F to be in the domain of attraction of an extreme value distribution. This holds for the standard normal distribution with $a_n = (2 \log h)^{1/2}$ and a t_d distribution with $a_n = h^{-1/d} C_d$ for a constant $C_d > 0$.

Assumption 2.4(iii) restricts how many outliers can be smaller than the tails of good observations. This is inspired by a model where the least trimmed squares estimator is maximum likelihood (Berenguer-Rico et al., 2023). For the t_d-distribution, $a_n = h^{-1/d} C_d$ is decreasing so Assumption 2.4(iii) is stronger than for the standard normal case $a_n = (2 \log h)^{1/2}$. We give some examples; proofs are in Appendix 2.D.

Example 2.4.1. *Assumption 2.4(iii) holds if outlying errors are separated from the tails of good errors, that is if $a_n \to \infty$ and $\min_{j \in \zeta_n^c} |\varepsilon_j| \geq m_n + d + o_p(1)$ for some $d > 0$. In the running example, Assumption 2.4(iii) thus holds for any $d > 0$.*

Example 2.4.2. *Assumption 2.4(iii) is violated $\forall F$ if $\varepsilon_j = m_n + u_j$ and $u_j \overset{iid}{\sim}$ Uniform$[0, 1]$ for $j \in \zeta_n^c$. In the running example, Assumption 2.4(iii) is violated if $d = 0$.*

Example 2.4.3. *We illustrate the necessity of Assumption 2.4(iii) in the case when cut-offs are tuned according to Assumption 2.4(i). Consider $y_i = \mu + \varepsilon_i$ and a Huber-skip estimator $\hat{\mu}$ initialised with $\tilde{\mu} = \mu, \tilde{\sigma} = 1$. Suppose $\lambda = 0.99$ and ε_i for $i \in \zeta_n$ are i.i.d. standard normal, while $\varepsilon_j = (\max_{i \in \zeta_n} \varepsilon_i) + u_j$ and $u_j \overset{iid}{\sim}$ Uniform$[0, 1]$ for $j \in \zeta_n^c$. If $n\{1 - \Phi(c_n)\} = \tau$ where Φ is the normal cdf then $\exists \Omega_n : P(\Omega_n) \geq \epsilon > 0$ and $\min_{\omega \in \Omega_n} n^{1/2} |\hat{\mu}(\omega) - \mu| \to \infty$.*

With the above assumptions, cut-offs satisfy conditions needed for the oracle property.

Proposition 2.4.1. *If Assumptions 2.1(iib, iic), 2.3, and 2.4 hold then Assumption 2.1(iii) is satisfied.*

It has been common to choose the cut-off for IIS to control its *gauge*, the expected share of misclassified good observations (Castle et al., 2011). Properties of the gauge have been previously studied in simulations (Hendry and Doornik, 2014) and theoretically in clean samples (Johansen and Nielsen, 2016). We find a limit distribution for the empirical gauge $\hat{\gamma} = h^{-1} \sum_{i \in \zeta_n} \mathbb{I}\{|y_i - x_i' \tilde{\beta}| > \tilde{\sigma} c_n\}$ that is valid under contamination. When cut-offs satisfy $n\{1 - F(c_n)\} = \tau$, we have the following Poisson approximation.

Theorem 2.2. *If Assumptions 2.1, 2.2, 2.3, and 2.4*(i) *hold then* $h\hat{\gamma} \to_d$ *Poisson*$(2\lambda\tau)$*, where* $\lambda \in (0,1]$ *is the share of good observations.*

According to the Theorem 2.2, the choice $n\{1 - F(c_n)\} = \tau$ corresponds to an asymptotic gauge of $2\lambda\tau$. While the share of good observations λ is generally unknown, the gauge can then be bounded from above by 2τ.

2.4.2 Slowly growing cut-offs

With the cut-off tuned as $n\{1 - F(c_n)\} = \tau$, we found that separation between outliers and good observations is needed for the oracle property. We now allow for weaker separation by tuning the cut-off as $n\{1 - F(c_n)\} = n^\rho$, resulting in more slowly-growing cut-offs that grow at the rate of intermediate order statistics from F.

Assumption 2.5.
(i) *The cut-off satisfies* $n\{1 - F(c_n)\} = n^\rho$ *for some* $0 < \rho < \phi - \eta/2$ *and* ρ, η *as in Assumption 2.1.*
(ii) *Let* $|\varepsilon_i|$ *for* $i \in \zeta_n$ *be ordered as* $\psi_1, ..., \psi_h$*. For all* $0 < \rho < 1, \exists C_\rho \in (0,1)$ *such that* $\psi_{h-\lfloor h^\rho \rfloor}/m_n \leq C_\rho + o_p(1)$*.*
(iii) *Outlying errors satisfy* $\min_{j\in\zeta_n^c} \varepsilon_i^2 \geq m_n^2\{1 + o_p(1)\}$*.*

Assumption 2.5(ii) requires separation between extremes values and intermediate order statistics of good errors. This holds for the standard normal, Laplace, and t_d-distributions with $d > 4$ (Berenguer-Rico and Nielsen, 2025). Assumption 2.5(iii) allows weaker separation between good and outlying errors than Assumption 2.4(iii). For instance, the running example with $d = 0$ is now allowed (cf. Example 2.4.2).

The new tuning is checked to satisfy conditions for the oracle property.

Proposition 2.4.2. *If Assumption 2.1*(iib, iic)*, 2.3, and 2.5 hold then Assumption 2.1*(iii) *is satisfied.*

With the new tuning, distribution of the empirical gauge $h\hat{\gamma}$ is approximately normal in large samples. Since Poisson$(2\lambda n^\rho)$ is well approximated by $N(2\lambda n^\rho, 2\lambda n^\rho)$ for large n, we closely mirror the result in Theorem 2.2.

Theorem 2.3. *If Assumptions 2.1, 2.2, 2.3, and 2.5*(i) *hold then* $(2\lambda n)^{-\rho/2}\{h\hat{\gamma} - 2\lambda n^\rho\} \to_d N(0,1)$*, where* $\lambda \in (0,1]$ *is the share of good observations.*

While the new tuning allows for weaker separation, the simulations in Section 2.6 suggest that larger sample sizes are required for the asymptotic approximation to be accurate if outliers lie close to the tails of the good observations. The conditions found in Section 2.4.1 therefore remain useful for understanding small samples properties.

2.5 Asymptotic distribution of IIS under contamination

Using the oracle property, we write down models where classical OLS inference applies to IIS. The main tasks are establishing: (1) consistency of initial estimators required in Assumption 2.2 and (2) distribution of the oracle estimator. Analysis of the latter is more intricate in dynamic models, where outliers propagate across observations.

2.5.1 Cross-section with outliers

Consider $y_i = x_i'\beta + \sigma\varepsilon_i$ for $i = 1, ..., n$, scalar y_i, and vector x_i. We adopt classical OLS assumptions for the 'good' observations.

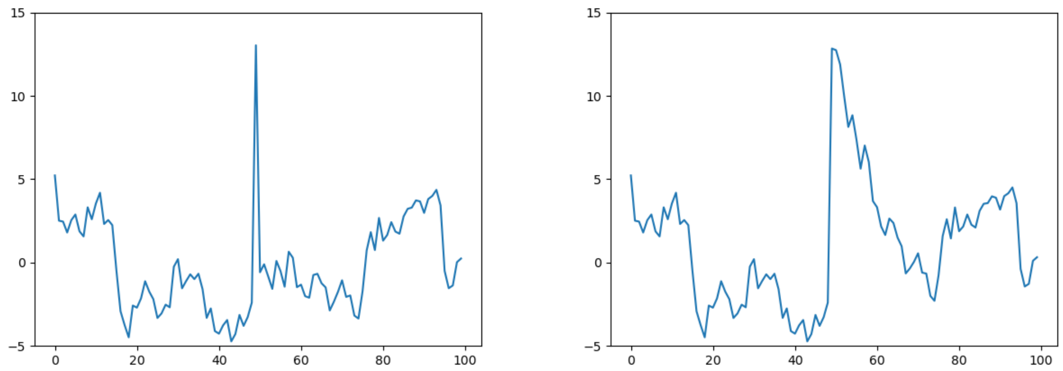
Assumption 2.6.
(i) (y_i, x_i) *for* $i \in \zeta_n$ *are independent and identically distributed*
(ii) $E(\varepsilon_i x_i) = 0$, $E(\varepsilon_i^2|x_i) = 1$, $E(\varepsilon_i^4) < \infty$, *and* $E(x_i x_i')$ *is invertible for* $i \in \zeta_n$
(iii) initial sets $\mathcal{I}_n$ *are independent of* $\{y_i, x_i\}_{i\in\zeta_n}$ *satisfying* $\mathcal{I}_n \subseteq \zeta_n$ *and* $n = \mathcal{O}_p(\#\mathcal{I}_n)$.

Assumptions 2.6(i, ii) are classical homoskedastic conditions for OLS. Here, we only require these for the good observations and allow outliers to violate them. The results could be extended to allow for conditional heteroskedasticity, provided that an appropriate covariance estimator were used.

Assumption 2.6(iii) requires the initial set to only contain good observations, but to be chosen independently of good data. Independence may be violated if the initial set is selected after inspecting the data, but would hold for randomly drawn initial sets, such as in Autometrics (Doornik, 2009), forward search (Atkinson and Riani, 2000), and FAST-LTS (Rousseeuw and Van Driessen, 2006). Future work should study these algorithms in more detail .

Given Assumption 2.6, the oracle estimator has a normal limiting distribution. If the oracle property holds then IIS inherits this limiting distribution and, for example,

Figure 2.2: Time series with additive and innovative outliers



(a) Additive outlier (b) Innovative outlier

Notes: Data are generated from $y_i = 0.9y_{i-1} + \tilde{\varepsilon}_i + \xi_{inn,i} + \xi_{add,i} - 0.9\xi_{add,i-1}$ *for* $i = 1, ..., 100$ *with* $\tilde{\varepsilon}_i \overset{iid}{\sim} N(0,1)$*. We set* $\xi_{add,50} = 15$ *in (a) and* $\xi_{inn,50} = 15$ *in (b).*

t-tests based on IIS can use standard normal critical values. This is summarised in the following result.

Theorem 2.4. *Suppose Assumptions 2.1, 2.6 hold with* $N = n^{-1/2}$*. Then* $(\sum_{i=1}^n v_i x_i x_i')^{-1/2}(\hat{\beta}_{IIS} - \beta)/\hat{\sigma}_{IIS} \to_d N(0, I)$.

2.5.2 Stationary autoregression with outliers

Outliers in time series are often either *additive* or *innovative* (Fox, 1972). Figure 2.2(a) contains an additive outlier at period 50. This corresponds to an extreme observation with no dynamic effects. Figure 2.2(b) contains an innovative outlier at period 50. This induces not only a contemporaneous effect, but also a dynamic adjustment path.

We introduce an autoregressive model with additive and innovative outliers. Let $\tilde{y}_i = \tilde{x}_i'\beta + \tilde{\varepsilon}_i$ be a clean process where $\tilde{x}_i = (1, \tilde{y}_{i-1}, ..., \tilde{y}_{i-q})$ and $\beta = (\beta_0, \beta_1, ..., \beta_q)$. We will assume the innovations $\tilde{\varepsilon}_i$ follow a well-behaved distribution, such as the standard normal. Introducing innovative contamination results in the process $y_i^* = \beta' x_i^* + \sigma(\tilde{\varepsilon}_i + \xi_{inn,i})$ for $i = 1, ..., n$ and $x_i^* = (1, y_{i-1}^*, ..., y_{i-q}^*)$. The process $\xi_{inn,i}$ can induce outliers similar to Figure 2.2(b).

The observed values y_i are further contaminated additively such that

$$y_i = y_i^* + \sigma\xi_{add,i}. \tag{2.4}$$

The process $\xi_{add,i}$ can induce outliers similar to Figure 2.2(a). The estimated equation is

$$y_i = x_i'\beta + \sigma\varepsilon_i \text{ for } i = 1, ..., n, \tag{2.5}$$

relating observed values y_i to $x_i = (1, y_{i-1}, ..., y_{i-q})$ and an initial value x_1. Note that

$$\varepsilon_i = \tilde{\varepsilon}_i + \xi_{inn,i} + \xi_{add,i} - \sum_{k=1}^{q} \beta_k, \xi_{add,i-k}. \tag{2.6}$$

so an additive outlier can equivalently be generated by a sequence of innovative outliers. Since our focus is not on identifying ξ_{inn} and ξ_{add}, this equivalence does not pose an issue. The present framework is simply a convenient setting for writing down different assumptions for additive and innovative contamination.

Example 2.5.1. *We relate the present discussion to the running example in Section 2.3. Suppose $\xi_{inn,i}\xi_{add,i} = 0$, so that each period has only one type of contamination. Then in the running example all outliers must be innovative. This holds since $\varepsilon_j, \varepsilon_{j+1}$ in the running example are independent given the good errors, while additive contamination would introduce dependence. The running example thus corresponds to having $\xi_{inn,j} = -\tilde{\varepsilon}_j + \max_{i\in\zeta_n} \varepsilon_j + d + u_j$ for $j \in \zeta_n^c$.*

The observations unaffected by innovative and additive contamination are $\zeta_{inn,n} = \{i \leq n : \xi_{inn,i} = 0\}$ and $\zeta_{add,n} = \cap\{i \leq n : \xi_{add,i-k} = 0 \text{ for } k = 0, ..., q\}$, respectively. Define the set of good observations as

$$\zeta_n = \zeta_{add,n} \cap \zeta_{inn,n}. \tag{2.7}$$

From (2.6), the good set thus consists of observations with $\varepsilon_i = \tilde{\varepsilon}_i$. A single additive outlier leads to the exclusion of many observations from the good set if the lag order is high. This reflects a broader challenge in detecting 'cellwise' outliers, which IIS is not designed to deal with.

For the clean process, we place classical OLS assumptions for stationary autoregressions. We also require stronger constraints on the innovative outliers because their effect propagates across observations.

Assumption 2.7. *Consider the above data generating process . Suppose the following.*

(i) Good observations satisfy

(a) $\tilde{\varepsilon}_i$ *are i.i.d. with a Lebesgue density,* $E\tilde{\varepsilon}_1 = 0$, $E\tilde{\varepsilon}_1^2 = 1$, *and* $E\tilde{\varepsilon}_1^{4+\delta} < \infty$ *for some* $\delta > 0$

(b) $|x_1^*| = o_p(n^{1/2})$

(c) *roots of* $1 - \sum_{k=1}^{q} \beta_k z^k$ *have modulus greater than one.*

(ii) There exists $0 < \phi < 1/2$ *such that*

(a) $\max_{1 \le i \le n} |\xi_{inn,i}| = \mathcal{O}_p(n^{1/2-\phi})$

(b) $\#\zeta_{inn,n}^c = o(n^{\phi})$.

(iii) Initial sets $\mathcal{I}_n$ *are deterministic,* $\mathcal{I}_n \subseteq \zeta_n$, *and* $n = \mathcal{O}(\#\mathcal{I}_n)$.

Assumption 2.7(i) is standard for analysing least squares in stationary autoregressions. Assumption 2.7(ii) bounds the number and magnitude of innovative outliers. An increasing number of innovative outliers is allowed, but their share needs to be vanishing. Assumption 2.7(iii) requires initial sets to be deterministic. This simplifies proofs, but could potentially be relaxed to independent initial sets such as in Theorem 2.4.

Under the provided conditions, IIS inherits a normal limiting distribution from the oracle.

Theorem 2.5. *Consider the data generating process described above. Suppose Assumptions 2.1 and 2.7 hold with* $N = n^{-1/2}$. *Then* $(\sum_{i=1}^{n} v_i x_i x_i')^{-1/2}(\hat{\beta}_{IIS} - \beta)/\hat{\sigma}_{IIS} \to_d N(0, I)$.

The proof of Theorem 2.5 departs slightly from standard analysis of stationary autoregressions. Firstly, we allow a non-vanishing share of additive outliers, so the 'good' observations may not occur in contiguous blocks of increasing size. Further, the proofs need to control the effect of innovative outliers on future observations.

While Assumption 2.7 does not concern additive outliers, some restrictions enter indirectly through Assumption 2.1. Assumption 2.1(ia) restricts the magnitude of both additive and innovative contamination, as illustrated by the following example.

Example 2.5.2. *We highlight a connection between Assumption 2.7*(ii) *and Assumption 2.1*(ia). *Consider* $y_i = \beta y_{i-1} + \varepsilon_i$, *where* $\varepsilon_i = \tilde{\varepsilon}_i + \xi_{inn,i} + \xi_{add,i} - \beta\ \xi_{add,i-1}$ *and* $|\beta| < 1$. *Suppose* $\tilde{\varepsilon}_i$ *are i.i.d. standard normal and* $\xi_{inn,i}\xi_{add,i} = 0$ *so that each period has only one type of contamination. Arguing as in Example 2.3.1, Assumption 2.1*(ia) *requires* $\max_{i \le n} |\varepsilon_i| = \mathcal{O}_p(n^{1/2-\phi})$ *for some* $\phi \in (0, 1/2)$. *Since* $\tilde{\varepsilon}_i = \mathcal{O}_p(\sqrt{2 \log n})$, *this requirement is equivalent to* $\max_{i \le n} |\xi_{inn,i}|, \max_{i \le n} |\xi_{add,i}| = \mathcal{O}_p(n^{1/2-\phi})$, *imposing the condition in Assumption 2.7*(iia) *on both additive and innovative contamination.*

Both additive and innovative outliers also need to be sufficiently large for Assumption 2.1(iii) to hold. In the case of additive outliers, contamination must be large relative to the autoregressive coefficients so that observations where the lagged dependent variables are contaminated are correctly detected. This is illustrated with the following example.

Example 2.5.3. *We illustrate how the assumption of large outliers applies to additive contamination. Let* $y_i = \beta y_{i-1} + \sigma\varepsilon_i$ *where* $|\beta| \leq 1$ *and* $\varepsilon_i = \tilde{\varepsilon}_i + \xi_{add,i} - \beta\xi_{add,i-1}$*. Define* ζ_n *as in* (2.7)*. Suppose* $\xi_{add,i}\xi_{add,i-1} = 0$ *so contaminations are not adjacent. Then Assumption 2.5(iii) holds if* $\min_{i\in\zeta_n^c} |\beta\xi_{add,i}| \geq 2\max_{i\leq n} |\tilde{\varepsilon}_i|\{1 + o_p(1)\}$*. The details are in Appendix 2.D.*

Theorem 2.5 could be extended to models with different deterministic components, such as linear or quadratic time trends. For models with additional regressors, a stand would also need to be taken on how these explanatory variables are contaminated.

2.5.3 Unit root testing under outlier contamination

The equation of interest is $y_i = \mu + \alpha y_{i-1} + \sigma\varepsilon_i$. The hypothesis $\alpha = 1$ is tested using a t-statistic $t = (\hat{\alpha}_{IIS} - 1)/\hat{\sigma}_\alpha$, where $\hat{\sigma}_\alpha^2$ is the (2,2)-element of $(\sum_{i=1}^n v_i x_i x_i')^{-1}\hat{\sigma}_{IIS}^2$, $x_i' = (1, y_{i-1})$, and v_i is given by Algorithm 1. Let data be generated from $y_i = y_{i-1} + \sigma\varepsilon_i$, where $\varepsilon_i = \tilde{\varepsilon}_i + \xi_i$ and define the good set as $\zeta_n = \{i \leq n : \xi_i = 0\}$. The innovations $\tilde{\varepsilon}_i$ are again assumed to follow a well-behaved distribution, such as the standard normal. The shocks ξ_i can generate both additive and innovative contamination.

Unlike the stationary case, we do not explicitly distinguish between additive and innovative contamination. Preliminary analysis suggests that a non-vanishing share of additive contamination can introduce nuisance parameters in the unit root setting. Instead of characterising this nuisance, we impose sufficient conditions for standard inference. The simulations in Section 2.6 indicate that size distortions from additive contamination are typically smaller than those from innovative contamination, so a distinction between the two types remains relevant for practical purposes even in the unit root case.

Assumption 2.8. *Consider the above data generating process. Suppose the following.*

(i) The clean errors satisfy

- *(a)* $\tilde{\varepsilon}_i$ *are i.i.d. with* $E\tilde{\varepsilon}_1 = 0$, $E\tilde{\varepsilon}_1^2 = 1$, *and* $E\tilde{\varepsilon}_1^{2+\delta} < \infty$ *for some* $\delta > 0$.
- *(b)* $\tilde{\varepsilon}_i$ *is independent of* $\{\xi_j\}_{j<i}$ *and* y_0 *for all* $i \leq n$.
- *(c)* $Ey_0^2 = \mathcal{O}(n)$

(ii) There exists $0 < \phi < 1/2$ *such that*

(a) $\max_{i \le n} E\xi_i^2 = \mathcal{O}(n^{1-2\phi})$

(b) $\#\zeta_n^c = \mathcal{O}(n^{\phi})$.

(iii) Initial sets $\mathcal{I}_n$ *are deterministic and contiguous blocks of observations such that* $\mathcal{I}_n \subseteq \zeta_n$, *and* $n = \mathcal{O}(\#\mathcal{I}_n)$.

Assumption 2.8 differs from Assumption 2.7 in a few respects. First, Assumption 2.8(ib) requires independence between clean errors and past outliers. Second, Assumption 2.8(ii) bounds the number and magnitude of outlier of both additive and innovative outliers. Assumption 2.8(ii) resonates with Assumption 2.1(ia), as seen by comparing the conditions to the one found in Example 2.3.2. Third, the initial sets need to be contiguous blocks such as $\mathcal{I}_n = \{\lfloor n/3 \rfloor, \ldots, \lfloor 2n/3 \rfloor\}$. This corresponds to how IIS is often used (Castle et al., 2012) and simplifies the proofs, but could potentially be relaxed to allow for some 'gaps' in the initial sets.

Under the provided conditions, standard Dickey-Fuller critical values are valid for testing a unit root with the t-statistic based on IIS (cf. Phillips and Perron, 1988).

Theorem 2.6. *Consider the above DGP and t-statistic. Suppose Assumption 2.1, 2.8 with* $x_i = (1, y_{i-1})$ *and* $N = \mathrm{diag}(n^{-1/2}, n^{-1})$. *Then* $t \to_d A_1/A_2^{1/2}$, *where* B *is a standard Brownian motion,* $A_1 = \int_0^1 B_s dB_s - B_1 \int_0^1 B_s ds$, *and* $A_2 = \int_0^1 B_s^2 ds - \big(\int_0^1 B_s ds\big)^2$.

We conjecture that similar results hold for models with more lags and different deterministic components. The results resonate with Berenguer-Rico and Nielsen (forthcoming), who analyse the least trimmed squares estimator in a cointegrated model with outliers.

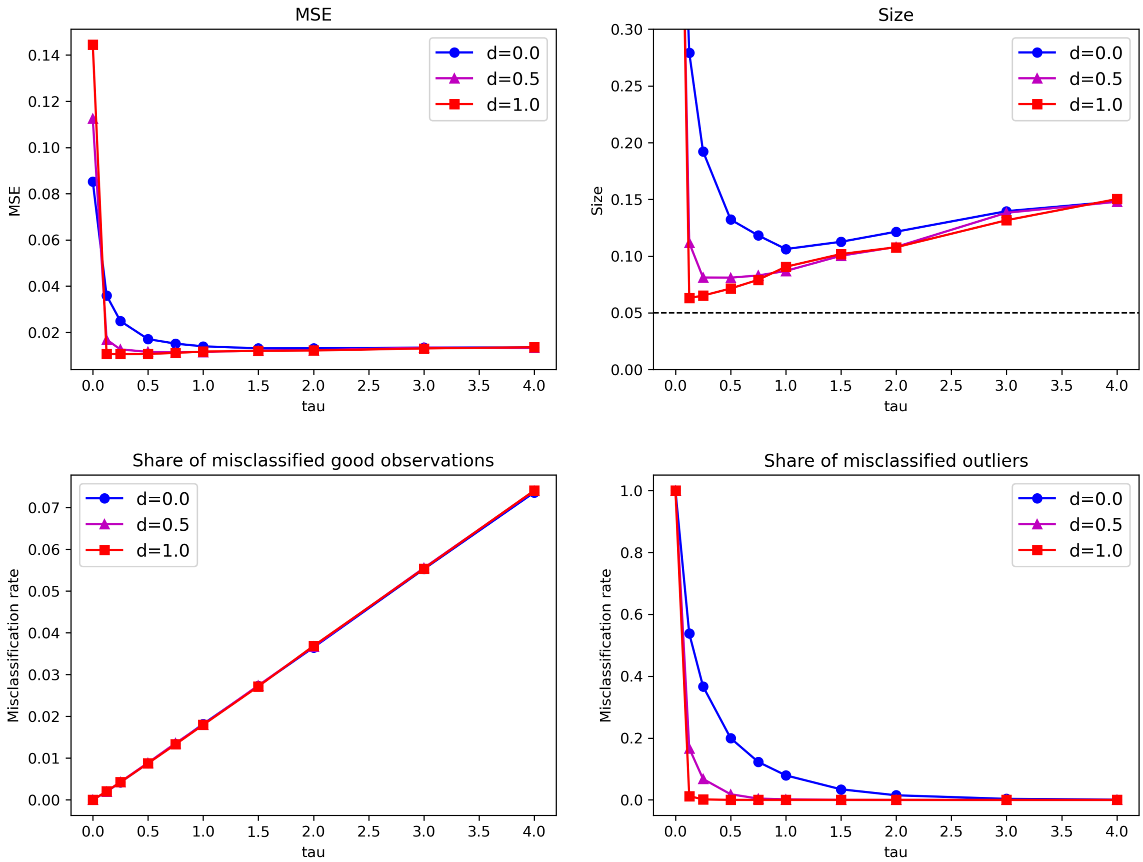
2.6 Simulations

Simulations assess the asymptotic theory relative to finite sample properties. Computations are implemented in Python (3.13.7) and are based on 10^4 repetitions.

2.6.1 Location-scale model

The equation of interest is $y_i = \mu + \sigma\varepsilon_i$ and we generate data with $\mu = 0$ and $\sigma = 1$. Similar to the asymptotic theory, we have an h-set of 'good' observations ζ_n and an $(n-h)$-set of 'outliers' ζ_n^c. We draw ε_i independently from $N(0,1)$ for $i \in \zeta_n$ and let

Figure 2.3: IIS in location-scale simulation ($h = 100$)



$\varepsilon_j = \max_{i \in \zeta_n} \varepsilon_i + \text{Uniform}[0,1] + d$ for $j \in \zeta_n^c$. The parameter d controls separation between good and outlying observations.

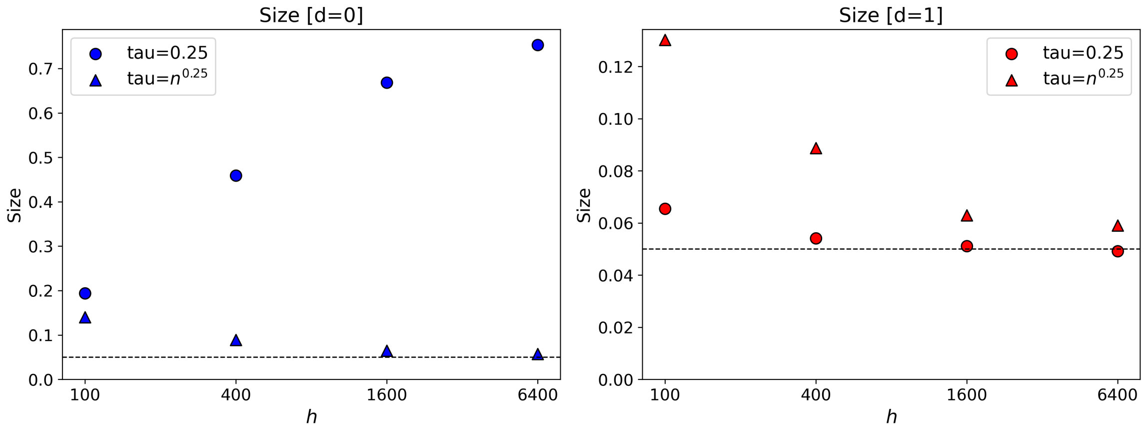
We compute stylised IIS estimators for μ, σ following Algorithm 1. The initial sets are $n_{good}/2$ randomly selected 'good' observations. Cut-offs c_n are determined through $h\{1 - \Phi(c)\} = \tau$, where Φ is the standard normal cdf and τ is varied across simulations. The choice $\tau = 0$ corresponds to $c_n = \infty$, giving full sample OLS.

Results ($h = 100$). We start with a relatively small sample size with $h = 100$ and $n - h = 10$. Figure 2.3 shows results from simulations with $d \in \{0, 0.5, 1\}$ and $\tau \in [0, 4]$.

In the top left panel of Figure 2.3, mean squared errors (MSE) are seen to converge to a value near zero for all values of 'd' as τ is increased. The full sample OLS ($\tau = 0$) suffers from a high MSE.

The top right panel of Figure 2.3 shows the size of a t-test for $\beta = 0$. The t-statistic is computed as $t = \hat{h}^{1/2}\hat{\beta}_{IIS}/\hat{\sigma}_{IIS}$ and the null is rejected if $|t| > 1.96$. If there is some separation between the good and outlying errors ($d = 1$), size is close

Figure 2.4: IIS in location-scale simulation (growing h)



to the nominal level 0.05 for small values of τ. If there is no separation ($d = 0$), tests are oversized across all values of τ.

The bottom panels of Figure 2.3 show shares of misclassified good and outlying observations. The share of misclassified good observations increases linearly in τ, reflecting Theorem 2.2. The rate of outlier misclassification falls steeply in τ when $d = 0.5, 1$ while with $d = 0$ the rate of decrease is slower. For $d = 0.5, 1$, small values of τ have low misclassification rates in both groups, explaining the better size control seen in the top right panel.

Results (increasing n). We generate data with $h \in \{100, 400, 1600, 6400\}$ and $n - h = h/10$. Figure 2.4 shows the size of IIS t-tests with $\tau = 0.25$ and $\tau = n^{0.25}$.

In the left panel of Figure 2.4 there is no separation between good errors and outliers ($d = 0$). In line with Figure 2.3, tests with $\tau = 0.25, n^{0.25}$ are oversized when $h = 100$. With a fast-growing cut-off ($\tau = 0.25$), size does not converge to the nominal level 0.05 as the sample size is increased. This finding resonates with the stronger separation condition in Proposition 2.4.1 and Example 2.4.3. Meanwhile, slowly-growing cut-offs ($\tau = n^{0.25}$) have size converging towards the nominal level 0.05, in line with Proposition 2.4.2. The convergence is relatively slow, and the nominal level seems only to be reached when $h \geq 1600$.

In the right panel of Figure 2.4 there is separation between good errors and outliers ($d = 1$). Both fast- and slowly-growing cut-offs now have sizes converging towards the nominal level, but the choice $\tau = 0.25$ yields faster convergence.

2.6.2 Autoregressive model

DGPs. The equation of interest is $y_i = \mu + \alpha y_{i-1} + \varepsilon_i$. We generate data with either innovative or additive outliers. We draw $\tilde{\varepsilon}_i$ independently from $N(0,1)$ for $i = 1, ..., n$. For innovative outliers, let $\varepsilon_i = \tilde{\varepsilon}_i + \xi_i$. For additive outliers, let $\varepsilon_i = \tilde{\varepsilon}_i + \xi_i - \alpha\xi_{i-1}$. Contamination occurs in periods $\Xi_n = \{i \leq n : \xi_i \neq 0\}$ which are evenly spaced from $\lfloor n/2 \rfloor + 1$ to $n - 1$. The h-set of 'good' observations is $\zeta_n = \{i \leq n : \varepsilon_i = \tilde{\varepsilon}_i\}$. We consider $h \in \{100, 400, 1600, 6400\}$ and vary the number of outliers $n - h$. We generate $\xi_i = -\tilde{\varepsilon}_i + \ell_n + \text{Uniform}[0, 1]$ for $i \in \Xi_n$ and vary ℓ_n to control the outlier magnitude that is central in Assumptions 2.1,2.7,2.8.

Statistics. We compute IIS estimators with initial sets $\{1, ..., \lfloor n/2 \rfloor\}$ and cut-offs c_n satisfying $h\{1 - \Phi(c_n)\} = 1/4$. This calibration performs well in the location-scale simulation when outliers and good observations are separated. Using the observations selected by IIS, we compute t-statistics $t = (\hat{\alpha} - \alpha)/se(\hat{\alpha})$, where se($\hat{\alpha}$) is the homoskedastic standard error. We also report t-statistics from a full sample OLS and an oracle estimator using the true good observations ζ_n.

2.6.2.1 Stable autoregression

We generate data with $\mu = 0, \beta = 0.8$ and test null $\alpha = 0.8$ using the t-statistics described above, rejecting if $|t| > 1.96$. Table 2.1 shows rejection rates (size) in DGPs with innovative and additive outliers. The number and magnitude of outliers varies across the columns while the rows show different estimators and numbers of good observations. For a test of true size 0.05, the Monte Carlo standard error is 0.002.

With innovative outliers, the oracle test has good size control across the DGPs while OLS is erratic. A t-test based on IIS converges to the nominal level 0.05 when h is increased and $n - h$ grows moderately. Good size control with IIS is sometimes attained even if Assumptions 2.1 and 2.7 are violated, such as in Table 2.1a column (7) where $\ell_n = n - h = h^{1/2}$. In Table 2.1a column (8) we have $\ell_n = h^{1/2}, n_{out} = h/5$, violating Assumption 2.1(ia). Consequently, the oracle property of IIS is seen to break down and an IIS t-test has inflated size even in large samples.

With additive outliers the oracle tests are again correctly sized while OLS tests are heavily inflated. IIS also struggles when ℓ_n grows slowly, such as in Table 2.1b columns $(3, 4)$, where the oracle property seems to break. This is explained by IIS having trouble detecting observations where the explanatory variable is contaminated by an additive shock, in reflection of Example 2.5.3. When the magnitude of outliers

Table 2.1: Size of t-tests in stationary AR(1) with outliers

(a) Innovative outliers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ℓ_n	$(2\log h)^{1/2}+1$				$h^{1/2}$			
$n-h$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$
IIS $[h=100]$	0.087	0.091	0.107	0.151	0.139	0.153	0.218	0.346
IIS $[h=400]$	0.060	0.059	0.059	0.059	0.069	0.072	0.091	0.294
IIS $[h=1600]$	0.052	0.049	0.053	0.049	0.054	0.057	0.059	0.270
IIS $[h=6400]$	0.053	0.052	0.047	0.054	0.050	0.050	0.052	0.265
Oracle $[h=100]$	0.060	0.058	0.062	0.062	0.059	0.056	0.057	0.057
Oracle $[h=6400]$	0.053	0.051	0.047	0.053	0.047	0.048	0.049	0.050
OLS $[h=6400]$	0.050	0.050	0.130	1.000	0.000	0.000	1.000	1.000

(b) Additive outliers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ℓ_n	$(2\log h)^{1/2}+1$				h			
$n-h$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$
IIS $[h=100]$	0.207	0.251	0.341	0.501	0.091	0.090	0.087	0.080
IIS $[h=400]$	0.119	0.135	0.305	0.820	0.059	0.061	0.064	0.053
IIS $[h=1600]$	0.083	0.103	0.336	1.000	0.054	0.055	0.053	0.051
IIS $[h=6400]$	0.062	0.085	0.417	1.000	0.051	0.049	0.051	0.054
Oracle $[h=100]$	0.068	0.067	0.064	0.064	0.069	0.070	0.070	0.064
Oracle $[h=6400]$	0.052	0.050	0.049	0.050	0.050	0.048	0.050	0.053
OLS $[h=6400]$	0.130	0.481	1.000	1.000	1.000	1.000	1.000	1.000

grows fast, these problems disappear. Column (8) of Table 2.1b shows IIS being close to nominal size even when ℓ_n and $n-h$ both grow at the rate h, mirroring the absence of quantity and magnitude restrictions on additive outliers in Assumption 2.7.

IIS t-tests are oversized in small samples ($h=100$) across the DGPs. Oracle tests are close to the nominal level even with $h=100$, indicating that issues with IIS size control arise from a failure to correctly separate the good and outlying observations in a small sample. In many DGPs, the asymptotic approximation becomes more reasonable when $h=400$. Iterating the detection of outliers is a potential way of improving small sample performance, but we have not implemented simulations to check this.

2.6.2.2 Unit root tests

We generate data with $\mu = 0, \alpha = 1$. A unit root is tested with the t-statistics described above and rejected if $t < -2.86$ using the critical value from Banerjee et al. (1993, Table 4.2). Table 2.2 shows rejection rates in DGPs with innovative and additive outliers.

Table 2.2: Size of unit root tests under outlier contamination

(a) Innovative outliers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ℓ_n	$(2\log h)^{1/2}+1$				$h^{1/4}$			
$n-h$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$
IIS $[h=100]$	0.192	0.253	0.314	0.287	0.175	0.213	0.287	0.286
IIS $[h=400]$	0.069	0.088	0.328	0.616	0.069	0.092	0.324	0.625
IIS $[h=1600]$	0.041	0.030	0.140	0.817	0.037	0.027	0.221	0.844
IIS $[h=6400]$	0.046	0.032	0.024	0.839	0.042	0.018	0.107	0.902
Oracle $[h=100]$	0.028	0.017	0.008	0.005	0.035	0.019	0.009	0.004
Oracle $[h=6400]$	0.046	0.032	0.005	0.002	0.042	0.017	0.004	0.001
OLS $[h=6400]$	0.044	0.013	0.000	0.000	0.032	0.001	0.000	0.000

(b) Additive outliers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ℓ_n	$(2\log h)^{1/2}+1$				h			
$n-h$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$
IIS $[h=100]$	0.182	0.207	0.246	0.302	0.124	0.121	0.118	0.105
IIS $[h=400]$	0.073	0.076	0.106	0.240	0.066	0.063	0.063	0.054
IIS $[h=1600]$	0.058	0.057	0.067	0.294	0.052	0.052	0.056	0.043
IIS $[h=6400]$	0.050	0.055	0.061	0.445	0.054	0.051	0.052	0.041
Oracle $[h=100]$	0.054	0.056	0.057	0.048	0.059	0.058	0.055	0.053
Oracle $[h=6400]$	0.049	0.052	0.052	0.043	0.054	0.050	0.052	0.041
OLS $[h=6400]$	0.055	0.065	0.107	0.729	1.000	1.000	1.000	1.000

With innovative outliers, IIS has size close to the oracle when the number and magnitude of outliers is moderate (columns 1,2,5,6). Having exact size control seems more elusive than for the stationary case, but this challenge is shared by both the oracle and IIS estimators. When $\ell_n(n-h)$ is moderate (columns 1,2,3,5,6), IIS and oracle tests are both undersized in large samples. When $\ell_n(n-h)$ grows fast (columns

4,7,8), Assumption 2.8 is violated and IIS has an inflated size that deviates from the oracle.

Full sample OLS tests are undersized in large samples when outliers are innovative. When examining power against local alternatives, IIS shows power gains relative to full-sample OLS (Appendix 2.A). Thus, although exact size control is more difficult in unit root testing with innovative outliers, IIS improves over OLS by having bounded size and better power in many of the DGPs.

Full sample OLS tests generally suffer from a bias towards stationarity under additive outlier contamination (Franses and Haldrup, 1994). This is seen in columns 5-8 of Table 2.2b. In contrast, IIS tests have size close to the nominal and oracle levels in large samples. The exception is column (4) where IIS is oversized while the oracle estimator is closer to the nominal level. The breakdown of the oracle property in column (4) again arises from IIS struggling to detect observations with a contaminated regressor when the magnitude of outliers grows slowly.

In Table 2.2b columns 4 and 8 there is a fixed proportion of additive contamination, so Assumption 2.8 is violated. In large samples, the oracle test has size that differs from the nominal level of 0.05, the Monte-Carlo standard error being 0.002. This finding is in line with our preliminary analysis, which indicates that a non-vanishing share of additive outliers introduces nuisance parameters into the oracle distribution. Nonetheless, the size distortions seen in the table are relatively modest.

2.7 Empirical Illustration

We return to the motivating example and fit equation $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ to the data in Figure 2.1 using IIS. The data[1] have $n = 121$ quarterly observations from 1995Q1-2025Q1. Computations are implemented in Python (3.13.7).

Full sample OLS estimates are shown in the first row of Table 2.3. Slope coefficients and standard errors (in parentheses) are reported, along with residual diagnostic tests and p-values (in parentheses) for autocorrelation ($F_{\text{ar}1-5}$), normality (χ^2_{norm}), autoregressive conditional heteroskedasticity ($F_{\text{arch}1-4}$), and heteroskedasticity (F_{hetero}).

OLS estimate of β_1 is negative and statistically significant at conventional levels. Diagnostics suggest severe misspecification: the normality test and residual moments (kurtosis 100.45, skewness -8.89) point to asymmetric outlier contamination.

[1]U.S. Bureau of Economic Analysis, series DFSARL1Q225SBEA, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org.

Table 2.3: AR(1) estimates for real PCE on food and accommodation services

	Parameter estimates		Model diagnostics			
	β_0	β_1	$F_{\text{ar1}-5}$	χ^2_{norm}	$F_{\text{arch1}-4}$	F_{hetero}
Full sample LS	4.82 (1.96)	-0.25 (0.09)	7.13 (0.07)	13756.14 (0.00)	9.99 (0.04)	88.96 (0.00)
Initial estimator	1.76 (0.37)	0.22 (0.10)	1.59 (0.66)	2.85 (0.24)	2.95 (0.56)	1.11 (0.57)
IIS estimator	1.65 (0.33)	0.21 (0.04)	1.40 (0.71)	2.32 (0.31)	3.89 (0.42)	0.87 (0.65)

We re-estimate the equation using stylised IIS (Algorithm 1) to allow for outliers around the COVID-pandemic. We use observations 1995Q1-2019Q4 to compute initial estimates shown in the second row of Table 2.3. The autoregressive coefficient is now positive and significant at 5% level and the diagnostic tests pass at conventional levels.

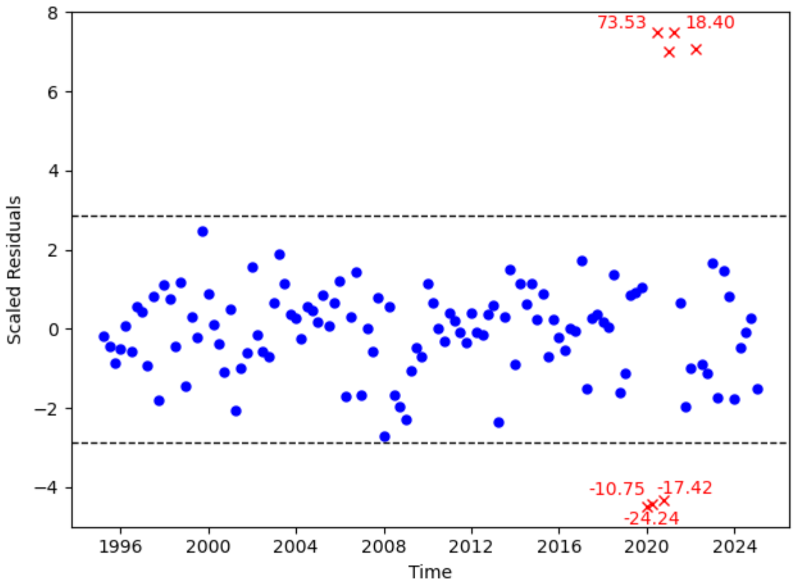
We calibrate the cut-off to a standard normal distribution using $n\{1 - \Phi(c)\} = 0.25$, which yields $c = 2.87$ and upper bounds the gauge at 0.5%. IIS estimates are shown in the third row of Table 2.3. These are similar to those from the initial fit, but with a smaller standard error on the autoregressive coefficient. All the diagnostic tests again pass.

Figure 2.5 plots residuals from the initial fit. Observations exceeding the cut-off are classified as outliers and are shown as crosses; these occur at 2020Q2–2021Q3 and 2022Q2, corresponding to the COVID-19 pandemic and the war in Ukraine. Some further investigation into the observations at 2008Q1 and 1999Q4, corresponding to the financial crisis and the turn of millennium, could be warranted, as they are included in initial set but have residuals close to the cut-off.

2.8 Conclusion

We established an oracle property for a stylised IIS estimator in data generating processes with outliers. A distribution theory for the gauge that is valid under contamination is provided as a guideline for choosing the IIS cut-off. Using the oracle property, we wrote down cross-sectional and autoregressive models with outliers where standard critical values can be used for hypothesis testing with IIS. Our theory highlights how additive and innovative contamination affect inference differently in time series models.

Figure 2.5: Index plot of scaled residuals using initial estimator



Notes: Figure shows scaled residuals from the initial fit. The scale is computed as the residual standard deviation from the initial observations. Residuals outside the plotting range are labeled with their values.

Our theory and simulations highlight some challenges IIS faces when the magnitude of outliers and the sample size is small. It would be useful to develop some additional tools for choosing the IIS cut-off, going beyond the gauge.

Future work should also extend the theory to algorithms without a known 'clean set'. Versions of IIS that do this have been proposed by Hendry et al. (2008) and Doornik (2009). Other related algorithm include forward search (Atkinson and Riani, 2000; Atkinson et al., 2025) and FAST-LTS (Rousseeuw and Van Driessen, 2006), to which the ideas developed here could be applied.

Appendix to Chapter 2

2.A Additional simulations: power of unit root tests

Set-up of the simulations is the same as in Section 2.6. Data are generated from $y_i = (1 - 0.95/h)y_{i-1} + \varepsilon_i$ to assess power of unit root tests against local alternatives.

Table 2.A.1: Rejection rates of t-tests against local alternatives $\beta = 1 - 0.95/h$

(a) Innovative outliers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ℓ_n	$(2\log h)^{1/2}+1$				$h^{1/4}$			
$n-h$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$
IIS $[h=100]$	0.426	0.656	0.930	0.985	0.367	0.525	0.820	0.968
IIS $[h=400]$	0.245	0.339	0.951	0.996	0.245	0.345	0.947	0.997
IIS $[h=1600]$	0.164	0.237	0.974	0.999	0.191	0.325	0.998	1.000
IIS $[h=6400]$	0.140	0.202	0.988	0.998	0.150	0.402	1.000	1.000
Oracle $[h=100]$	0.255	0.469	0.869	1.000	0.210	0.341	0.704	0.998
Oracle $[h=6400]$	0.138	0.200	0.987	1.000	0.149	0.397	1.000	1.000
OLS $[h=6400]$	0.126	0.064	0.000	0.000	0.115	0.024	0.000	0.000

(b) Additive outliers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ℓ_n	$(2\log h)^{1/2}+1$				h			
$n-h$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$	$\log(h)$	$h^{1/3}$	$h^{1/2}$	$h/5$
IIS $[h=100]$	0.308	0.350	0.405	0.502	0.221	0.225	0.218	0.205
IIS $[h=400]$	0.177	0.183	0.242	0.436	0.147	0.149	0.144	0.142
IIS $[h=1600]$	0.138	0.142	0.173	0.459	0.123	0.134	0.130	0.122
IIS $[h=6400]$	0.126	0.137	0.152	0.513	0.132	0.125	0.131	0.125
Oracle $[h=100]$	0.135	0.145	0.141	0.139	0.137	0.141	0.142	0.138
Oracle $[h=6400]$	0.122	0.127	0.125	0.123	0.130	0.122	0.129	0.123
OLS $[h=6400]$	0.139	0.166	0.299	0.995	1.000	1.000	1.000	1.000

2.B Proofs — Oracle property and gauge

2.B.1 Proof of Theorem 2.1

Lemma 2.B.1. *Suppose Assumption 2.1(iic, iib, iii, ia) and 2.2. Then* $\forall\epsilon > 0$, $\exists A_1 > 0$ *and sets* $\mathcal{B}_n$ *with* $P(\mathcal{B}_n) \geq 1-\epsilon$ *such that on* $\mathcal{B}_n$ *and large* n *it holds for all* $1 \leq i \leq n$

$$\mathbb{I}\{|\varepsilon_i| \leq c_n - n^{-\phi}A_1\} \leq \mathbb{I}\{|y_i - x_i'\tilde{\beta}| \leq \tilde{\sigma}c_n\} = v_i \leq \mathbb{I}\{|\varepsilon_i| \leq c_n + n^{-\phi}A_1\}. \quad (2.8)$$

Proof. 1. Construction of $\mathcal{B}_n$. Let $0 < \phi < 1/2$ and $0 < \eta < \min\{\phi, 1-2\phi\}$ be given by Assumption 2.1. By Assumption 2.1(iib, iic, ia), $\exists A_0 > 0$ such that $P(\mathcal{B}_n') \geq 1-\epsilon$

for all n, where

$$\mathcal{B}'_n = \left\{ N^{-1}|\tilde{\beta} - \beta| + n^{1/2}|\tilde{\sigma} - \sigma| + \max_{1\le i\le n} |n^{\phi} N' x_i| + \frac{\max_{i\in\zeta_n} \varepsilon_i^2}{m_n^2} + \frac{m_n^2}{n^{\eta} L(n)} \le A_0 \right\}.$$

2. Deterministic analysis on $\mathcal{B}_n$. Introduce quantity

$$s_i = \tilde{\sigma} c_n - y_i + x_i'\tilde{\beta} + \sigma\varepsilon_i = \sigma c_n + n^{-1/2} n^{1/2}(\tilde{\sigma} - \sigma) c_n + n^{-\phi}(n^{\phi} N' x_i)'\{N^{-1}(\tilde{\beta} - \beta)\}.$$

By Assumption 2.1(*iii*) it holds $c_n = \mathcal{O}(n^{\eta/2})L(n)$. The condition $\eta < 1 - 2\phi$ implies $-1/2 + \eta/2 + \nu = -\phi$ for some $\nu > 0$. Thus, on $\mathcal{B}_n$ it holds for $A_1 = 2A_0^2/\sigma$ and large n

$$s_i \le \sigma c_n + n^{-1/2+\eta/2+\nu} \left(\frac{L(n)}{n^{\nu}}\right) \left(\frac{c_n}{n^{\eta/2} L(n)}\right) A_0 + n^{-\phi} A_0^2 \le \sigma(c_n + n^{-\phi} A_1).$$

$$s_i \ge \sigma c_n - n^{-1/2+\eta/2+\nu} \left(\frac{L(n)}{n^{\nu}}\right) \left(\frac{c_n}{n^{\eta/2} L(n)}\right) A_0 - n^{-\phi} A_0^2 \ge \sigma(c_n - n^{-\phi} A_1).$$

The two inequalities are combined to give

$$\sigma(c_n - n^{-\phi} A_1 - \varepsilon_i) \le \tilde{\sigma} c_n - y_i + x_i'\tilde{\beta} \le \sigma(c_n + n^{-\phi} A_1 - \varepsilon_i)$$

which implies

$$\mathbb{I}\{\varepsilon_i \le c_n - n^{-\phi} A_1\} \le \mathbb{I}\{y_i - x_i'\tilde{\beta} \le \tilde{\sigma} c_n\} \le \mathbb{I}\{\varepsilon_i \le c_n + n^{-\phi} A_1\}.$$

Analogous inequalities hold for $\mathbb{I}\{y_i - x_i'\tilde{\beta} \ge -\tilde{\sigma} c_n\}$, and thus

$$\mathbb{I}\{|\varepsilon_i| \le c_n - n^{-\phi} A_1\} \le \mathbb{I}\{|y_i - x_i'\tilde{\beta}| \le \tilde{\sigma} c_n\} = v_i \le \mathbb{I}\{|\varepsilon_i| \le c_n + n^{-\phi} A_1\}. \qquad \square$$

Lemma 2.B.2. *Suppose Assumption 2.1 and 2.2. Then for $v_i = \mathbb{I}\{|y_i - x_i'\tilde{\beta}| \le \tilde{\sigma} c_n\}$ it holds*

$$\sum_{i\in\zeta_n} (1 - v_i)|g_i|, \sum_{i\in\zeta_n^c} v_i |g_i| = o_p(1), \tag{2.9}$$

for g_i equal to n^{-1}, $n^{-1/2}(\varepsilon_i^2 - \sigma^2)$, $N' x_i x_i' N$, or $N' x_i \varepsilon_i$.

Proof of Lemma 2.B.2. We construct sets $\mathcal{B}_n$ and show the sums in (2.9) are uniformly $o(1)$ on $\mathcal{B}_n$. We then check that $\forall\epsilon > 0$, $\mathcal{B}_n$ can constructed so that $P(\mathcal{B}_n) \geq 1 - \epsilon$ for large n.

1. Construction of $\mathcal{B}_n$. Let $0 < \phi < 1/2$ and $A_1 > 0$ be such that on $\mathcal{B}_n$ and large n

$$\mathbb{I}\{|\varepsilon_i| \leq c_n - n^{-\phi}A_1\} \leq \mathbb{I}\{|y_i - x_i'\tilde{\beta}| \leq \tilde{\sigma}c_n\} = v_i \leq \mathbb{I}\{|\varepsilon_i| \leq c_n + n^{-\phi}A_1\}. \quad (2.10)$$

Further, for $0 < \eta < \min\{\phi, 1 - 2\phi\}$ and $0 < \rho < \phi - \eta$ it holds

$$\sum_{i\in\zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A_1\} \leq n^\rho, \quad \sum_{i\in\zeta_n^c} \mathbb{I}\{|\varepsilon_i| \leq c_n + n^{-\phi}A_1\} \leq n^\rho. \quad (2.11)$$

Finally, there is $A_2 > 0$ such that on $\mathcal{B}_n$

$$\psi_n^2 = \max_{i\in\zeta_n} \varepsilon_i^2 \leq n^\eta L(n)A_2, \quad \max_{i\leq n} |N'x_i| \leq n^{-\phi}A_2. \quad (2.12)$$

2. Deterministic analysis on $\mathcal{B}_n$. Consider $g_i = n^{-1}$. By (2.10), on $\mathcal{B}_n$ it holds

$$(1 - v_i) \leq \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A_1\}, \quad v_i \leq \mathbb{I}\{|\varepsilon_i| \leq c_n + n^{-\phi}A_1\}. \quad (2.13)$$

Combine (2.11), (2.13) to conclude $n^{-1}\sum_{i\in\zeta_n}(1 - v_i), n^{-1}\sum_{i\in\zeta_n^c} v_i \leq n^{\rho-1} \leq o(1)$.

Consider $g_i = n^{-1/2}(\varepsilon_i^2 - \sigma^2)$. For $i \in \zeta_n$, bound $|\varepsilon_i^2 - \sigma^2| \leq \psi_n^2 + \sigma^2$ and use (2.12), (2.13) to see that for any $A_3 > A_2$ and large n it holds

$$(1 - v_i)|\varepsilon_i^2 - \sigma^2| \leq \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A_1\}(\psi_n^2 + \sigma^2) \leq \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A_1\}n^\eta L(n)A_3. \quad (2.14)$$

Sum over $i \in \zeta_n$, use (2.11) and $\eta + \rho < \phi < 1/2$ to conclude

$$n^{-1/2}\sum_{i\in\zeta_n}(1 - v_i)|\varepsilon_i^2 - \sigma^2| \leq n^{-1/2+\rho+\eta}L(n)A_3 = o(1).$$

For $i \in \zeta_n^c$, by (2.13) and $c_n^2 = O(n^\eta)L(n)$ from Assumption 2.1(*iiia*), $\exists A_4 > 0$

such that

$$v_i|\varepsilon_i^2-\sigma^2| \leq \mathbb{I}\{|\varepsilon_i| \leq c_n+n^{-\phi}A_1\}\{(c_n+n^{-\phi}A_1)^2+\sigma^2\} \leq \mathbb{I}\{|\varepsilon_i| \leq c_n+n^{-\phi}A_1\}n^{\eta}L(n)A_4. \tag{2.15}$$

Sum over $i \in \zeta_n^c$ and use $\eta + \rho < \phi < 1/2$ and (2.11) to conclude

$$n^{-1/2}\sum_{i\in\zeta_n^c} v_i|\varepsilon_i^2 - \sigma^2| \leq n^{-1/2+\rho+\eta}L(n)A_4 = o(1).$$

Consider $g_i = N'x_ix_i'N, N'x_i\varepsilon_i$. The spectral norm satisfies $\|N'x_ix_i'N\| = |N'x_i|^2$. Since $\max_{i\leq n}|N'x_i| \leq n^{-\phi}A_2$ by (2.12) then $|N'x_i|^2 < |N'x_i| < 1$ for all i when n is large. It is thus enough to check the claim for $g_i = |N'x_i||\varepsilon_i|^s$ and $s = 0, 1$.

For $i \in \zeta_n$, use (2.13) and $|N'x_i||\varepsilon_i|^s \leq n^{\eta/2-\phi}L(n)A_2^{3/2}$ from (2.12) to bound

$$(1 - v_i)|N'x_i||\varepsilon_i|^s \leq \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A_1\}n^{\eta/2-\phi}L(n)A_2^{3/2}. \tag{2.16}$$

Sum over $i \in \zeta_n$, then use (2.11) and $\rho + \eta/2 < \phi$ to conclude

$$(1 - v_i)|N'x_i||\varepsilon_i|^s \leq \sum_{i\in\zeta_n}(1 - v_i)n^{\rho+\eta/2-\phi}L(n)A^{3/2} = o(1).$$

For $i \in \zeta_n^c$, use (2.13) to bound $v_i|N'x_i||\varepsilon_i|^s \leq \mathbb{I}\{|\varepsilon_i| \leq c_n + n^{-\phi}A_1\}|N'x_i||c_n + n^{-\phi}A_1|^s$. Since $\max_{i\leq n}|N'x_i||c_n + n^{-\phi}A_1|^s \leq n^{\eta/2-\phi}L(n)A_2^2$ it follows (2.12) then $v_i|N'x_i||\varepsilon_i|^s \leq \mathbb{I}\{|\varepsilon_i| \leq c_n + n^{-\phi}A_1\}n^{\eta/2-\phi}L(n)A_2^2$. Sum over $i \in \zeta_n^c$ then use (2.11) and $\rho + \eta/2 < \phi$ to conclude

$$\sum_{i\in\zeta_n^c} v_i|N'x_i||\varepsilon_i|^s \leq n^{\rho+\eta/2-\phi}L(n)A_2^2 = o(1).$$

3. Probability analysis. Since Assumption 2.1(iic, iib, iii, ia) and conditions on the initial estimators hold, by Lemma 2.B.1 we can find $A_1 > 0$ and sets $\mathcal{B}_{1n}$ with $P(\mathcal{B}_{1n}) \geq 1 - \epsilon$ such that (2.10) holds on $\mathcal{B}_{1n}$ for large n. By Assumption 2.1($iiib$), there exist sets $\mathcal{B}_{2n}$ with $P(\mathcal{B}_{2n}) \to 1$ such that (2.11) holds on $\mathcal{B}_{2n}$. By Assumption 2.1($iib, iic, iiia, ia$), $\exists A_2 > 0$ and sets $\mathcal{B}_{3n}$ such that $P(\mathcal{B}_{3n}) \geq 1 - \epsilon$ and (2.12) holds on $\mathcal{B}_{3n}$. Then for $\mathcal{B}_n = \mathcal{B}_{1n} \cap \mathcal{B}_{2n} \cap \mathcal{B}_{3n}$ we have $P(\mathcal{B}_n) \geq 1 - 3\epsilon$ for large n. □

Proof of Theorem 2.1. We show $N^{-1}(\hat{\beta}_{(1)}-\hat{\beta}_{LS,good}), n^{1/2}(\hat{\sigma}_{(1)}-\hat{\sigma}_{LS,good})=o_p(1)$. If further $N^{-1}(\hat{\beta}_{LS,good}-\beta), n^{1/2}(\hat{\sigma}_{LS,good}-\sigma)=\mathcal{O}_p(1)$ then also $N^{-1}(\hat{\beta}_{(1)}-\beta), n^{1/2}(\hat{\sigma}_{(1)}-\sigma)=\mathcal{O}_p(1)$. The claim for $m\geq 2$ follows from the 1-step result with $\tilde{\beta}=\hat{\beta}_{(m-1)}, \tilde{\sigma}=\hat{\sigma}_{(m-1)}$.

By Lemma 2.B.2, for g_i equal to $(N'x_ix_i'N)$, $N'x_i\varepsilon_i$, $n^{1/2}(\varepsilon_i^2-\sigma^2)$, and n^{-1} it holds

$$\sum_{i=1}^{n} v_i g_i=\sum_{i\in\zeta_n} g_i-\sum_{i\in\zeta_n}(1-v_i)g_i+\sum_{i\in\zeta_n^c} v_i g_i=\sum_{i\in\zeta_n} g_i+o_p(1), \tag{2.17}$$

By Assumption 2.1(ib) and (2.17) with $g_i=(N'x_ix_i'N)$, we get $\sum_{i=1}^n v_iN'x_ix_i'N\to_d\Sigma$, where Σ is non-singular a.s.. Thus, $\exists\Omega_n: P(\Omega_n)\to 1$ such that on Ω_n the matrix $\sum_{i=1}^n v_iN'x_ix_i'N$ is invertible. On Ω_n, the least squares solution $N^{-1}(\hat{\beta}_{(1)}-\beta)=(\sum_{i=1}^n v_iN'x_ix_i'N)^{-1}(\sum_{i=1}^n v_iN'x_i\varepsilon_i)$ applies. By (2.17) and Assumption 2.1(ib), conclude $N^{-1}(\hat{\beta}_{(1)}-\beta)=N^{-1}(\hat{\beta}_{LS,good}-\beta)+o_p(1)$. Combined with (2.17), $n^{1/2}(\hat{\sigma}_{(1)}-\sigma)=n^{1/2}(\hat{\sigma}_{LS,good}-\sigma)=o_p(1)$ follows. □

2.B.2 Proofs for Propositions 2.4.1 and 2.4.2

Throughout the subsection, we refer to Leadbetter et al. (1983) as 'LLR'.

Lemma 2.B.3. *Suppose Assumption 2.1(iib, iic), 2.3(i), and $n\{1-F(c_n)\}\to\tau\in(0,\infty]$. Then $c_n^2=O(n^\eta)L(n)$.*

Proof. Let $\psi_n=\max_{i\in\zeta_n}|\varepsilon_i|$ and $\Omega_{1n}=\{\psi_n>c_n\}$. By symmetry, $1-P(|\varepsilon_1|\leq c_n)=2\{1-F(c_n)\}\to 2\tau$. Thus, by LLR (Theorem 1.5.1.) $P(\Omega_{1n})=1-P(\psi_n\leq c_n)\to 1-\exp(-2\tau)>0$.

Let $\epsilon>0$ be such that $1-\exp(-2\tau)>2\epsilon$. By Assumption 2.1(iib, iic), $\psi_n=O_p(n^{\eta/2})L(n)$. Thus, $\exists B>0$ such that $\Omega_{2n}=\{\psi_n\leq n^{\eta/2}L(n)B\}$ satisfies $P(\Omega_{2n})\geq 1-\epsilon\ \forall n$. On $\Omega_{1n}\cap\Omega_{2n}$ it holds $c_n<\psi_n\leq Bn^{\eta/2}L(n)$ and $P(\Omega_{1n}\cap\Omega_{2n})\geq P(\Omega_{1n})-P(\Omega_{2n}^c)\geq 1-\exp(-2\tau)-2\epsilon>0$ for large n . Since c_n is deterministic, conclude $c_n\leq n^{\eta/2}L(n)B$ for large n. □

Lemma 2.B.4. *Let F be a distribution satisfying Assumption 2.3, $\varepsilon_i\sim F$, and $p_n=n\{1-F(c_n)\}$. Then $nP\left(c_n-n^{-\phi}A\leq|\varepsilon_i|\leq c_n+n^{-\phi}A\right)=\mathcal{O}(p_nc_nn^{-\phi})$ for all $A>0$.*

Proof of Lemma 2.B.4. Since f is symmetric then the distribution of $|\varepsilon_i|$ has density $2f$. Thus, $\mathcal{W}_n=nP\left(c_n-n^{-\phi}A\leq|\varepsilon_i|\leq c_n+n^{-\phi}A\right)=n\int_{c_n-n^{-\phi}A}^{c_n+n^{-\phi}A}2f(x)dx$. By the continuity in Assumption 2.3(i) and mean value theorem for definite integrals, $\exists c^* \in (c_n - n^{-\phi}A, c_n + n^{-\phi}A)$ such that $\mathcal{W}_n = 4n^{1-\phi}f(c^*)A$. Multiplying and dividing, rewrite this as

$$\mathcal{W}_n = n\{1 - F(c_n)\}\frac{f(c^*)}{f(c_n - n^{-\phi}A_1)}\frac{f(c_n)}{c_n\{1 - F(c_n)\}}\frac{f(c_n - n^{-\phi}A_1)}{f(c_n)}c_n 4n^{-\phi}A.$$

The first term is p_n. The second, third, and fourth terms are $\mathcal{O}(1)$ by Assumption 2.3(i, ii, iii). Thus, conclude $\mathcal{W}_n = \mathcal{O}(p_n c_n n^{-\phi})$. □

Lemma 2.B.5. *Suppose Assumptions 2.1(iib, iic), 2.3. Let* $p_n = n\{1 - F(c_n)\} \to (0, \infty]$ *and* $p_n = \mathcal{O}(n^\rho)$ *for some* $0 < (\rho + \eta)/2 < \phi$. *Then* $\forall A > 0, \exists \delta^* > 0$ *such that* $\forall 0 < \delta < \delta^*$ *it holds* $P(\sum_{i\in\zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A\} < n^{\rho+\delta}) \to 1$. *If further* $p_n = n^\rho$ *then* $P(\sum_{i\in\zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n + n^{-\phi}A\} > n^{\rho-\delta}) \to 1$.

Proof of Lemma 2.B.5. Let $p_n^* = n\{1 - P(|\varepsilon_i| \le c_n)\}$ and $\delta > 0$ such that $(\rho + \eta + \delta)/2 < \phi$. For constants k_n, u_n such that $\{1+k_n/n\}^{1/2} = o(1)$, adding and subtracting p_n^* gives

$$U_n = \frac{k_n - n\{1 - P(|\varepsilon_i| \le u_n)\}}{k_n^{1/2}(1 - k_n/n)^{1/2}} = \big(\frac{k_n - p_n^*}{k_n^{1/2}} + \frac{p_n^* - n\{1 - P(|\varepsilon_i| \le u_n)\}}{k_n^{1/2}}\big)\{1 + o(1)\}. \tag{2.18}$$

We show $U_n \to \infty$ when $u_n = c_n - n^{-\phi}A$ and $k_n = n^{\rho+\delta}$. The first claim thus follows by LLR (Theorem 2.5.2). Since $p_n^* = 2p_n$ by Assumption 2.3(i) and $p_n = \mathcal{O}(n^\rho)$ then $k_n^{-1/2}(k_n - p_n^*) = k_n^{1/2}\{1 + o(1)\} \to \infty$. We show $r_n = p_n^* - n\{1 - P(|\varepsilon_i| \le u_n)\} = o(k_n^{1/2})$. Since $r_n = -nP(c_n - n^{-\phi}A < |\varepsilon_i| \le c_n)$ then Lemma 2.B.4 shows $r_n = \mathcal{O}(p_n c_n n^{-\phi})$. Since $p_n = \mathcal{O}(n^\rho)$ and $c_n = \mathcal{O}(n^{\eta/2})L(n)$ by Lemma 2.B.3, use $(\rho + \eta)/2 < \phi$ to conclude $r_n = \mathcal{O}(n^{\rho+\eta/2-\phi})L(n) = o(k_n^{1/2})$.

We show $U_n \to -\infty$ when $u_n = c_n + n^{-\phi}A$, $k_n = n^{\rho-\delta}$, and $p_n = n^\rho$. The second claim again follows by LLR (Theorem 2.5.2). Using $p_n^* = 2n^\rho$ get $k_n^{-1/2}(k_n - p_n^*) = n^{(\rho-\delta)/2}\{-2 + o(1)\} \to -\infty$. Arguing as above, $r_n = P(c_n < |\varepsilon_i| \le c_n + n^{-\phi}A) = \mathcal{O}(p_n c_n n^{-\phi}) = o(k_n^{1/2})$.

□

Lemma 2.B.6. *Suppose Assumption 2.1(iic), 2.3, and 2.4. Then* $\forall A > 0, \exists 0 < \rho < \phi - \eta/2$ *such that*

$$P(\sum_{i\in\zeta_n^c} \mathbb{I}\{|\varepsilon_i| < c_n + n^{-\phi}A\} < n^\rho) \to 1. \tag{2.19}$$

Proof. Let $A > 0$ and let $0 < \rho < \phi - \eta/2$ be provided by Assumption 2.4(iii). We construct sets Ω_n and show $\sum_{i \in \zeta_n^c} \mathbb{I}\{|\varepsilon_i| < c_n + n^{-\phi}A\} < n^\rho$ holds on Ω_n. We then show $\forall \delta > 0$, Ω_n can be constructed such that $P(\Omega_n) \geq 1 - \delta$ for all large n.

1.Construction of Ω_n. Let $x_c \in \mathbb{R}$ be such that $G(x_c)^2 > \exp(-2\tau)$ and $x_m < x_c$. Define $\Omega_{1n} = \{m_n > x_m/a_n + b_n\}$, $\Omega_{2n} = \{\sum_{i \in \zeta_n^c} \mathbb{I}\{|\varepsilon_i| < m_n + (x_c - x_m)/a_n + n^{-\phi}A\} < n^\rho\}$, and $\Omega_n = \Omega_{1n} \cap \Omega_{2n}$.

2.Analysis on Ω_n. Assumptions 2.3(i), 2.4(i) imply $n\{1 - P(|\varepsilon_1| \leq c_n)\} = 2n\{1 - F(c_n)\} = 2\tau$. For $\psi_n = \max_{i \in \zeta_n} |\varepsilon_i|$, LLR (Theorem 1.5.1.) then shows $P(\psi_n \leq c_n) \to \exp(-2\tau)$.

By symmetry of f and Assumption 2.4(ii)

$$P(\min_{i \in \zeta_n} \varepsilon_i \geq -x_c/a_n - b_n) = P(\max_{i \in \zeta_n} \varepsilon_i \leq x_c/a_n + b_n) \to G(x_c). \tag{2.20}$$

By LLR (Theorem 1.8.2.) it follows

$$P(\psi_n \leq x_c/a_n + b_n) = P(\max_{i \in \zeta_n} \varepsilon_i \leq x_c/a_n + b_n, \min_{i \in \zeta_n} \varepsilon_i \geq -x_c/a_n - b_n) \to G(x_c)^2. \tag{2.21}$$

Thus, $\exists \epsilon > 0$ such that $P(\psi_n \leq x_c/a_n + b_n) > \exp(-2\tau) + \epsilon > P(\psi_n \leq c_n)$ for large n. Since a distribution function is increasing, conclude that for large n

$$x_c/a_n + b_n \geq c_n. \tag{2.22}$$

By definition of Ω_{1n} and (2.22), $c_n \leq m_n + x/a_n$ holds on Ω_n for large n. Thus, $\mathbb{I}\{|\varepsilon_i| < c_n + n^{-\phi}A\} \leq \mathbb{I}\{|\varepsilon_i| < m_n + x/a_n + n^{-\phi}A\}$. By definition of Ω_{2n}, conclude that on Ω_n and large n

$$\sum_{i \in \zeta_n^c} \mathbb{I}\{|\varepsilon_i| < c_n + n^{-\phi}A\} \leq \sum_{i \in \zeta_n^c} \mathbb{I}\{|\varepsilon_i| < m_n + x/a_n + n^{-\phi}A\} < n^\rho.$$

3. Probability analysis. We have $P(\Omega_{2n}) \to 1$ by Assumption 2.4(iii). We show $\forall \delta > 0$ taking x_m such that $\{1 - G(x_m)\}^2 > 1 - 2\delta$ implies $P(\Omega_{1n}) \geq 1 - \delta$ for large n. By definition,

$$P(\Omega_{1n}) = P(|\min_{i \in \zeta_n} \varepsilon_i| \ > x_m/a_n + b_n, |\max_{i \in \zeta_n} \varepsilon_i| > x_m/a_n + b_m). \tag{2.23}$$

Since f has support $\mathbb{R}$ then $P(\max_{i\in\zeta_n}\varepsilon_i > 0, \min_{i\in\zeta_n}\varepsilon_i < 0) \to 1$ and thus

$$P(\Omega_{1n}) = P(\min_{i\in\zeta_n}\varepsilon_i < -x_m/a_n - b_n, \max_{i\in\zeta_n}\varepsilon_i > x_m/a_n + b_m) + o(1).$$

Using $P(A\cap B) = 1 - P(A^c) - P(B^c) + P(A^c \cap B^c)$ get further

$$P(\Omega_{1n}) = 1 - P(\min_{i\in\zeta_n}\varepsilon_i \geq -\frac{x_m}{a_n} - b_n) - P(\max_{i\in\zeta_n}\varepsilon_i \leq \frac{x_m}{a_n} + b_m) + P(\psi_n \leq \frac{x_m}{a_n} + b_n) + o(1).$$

By (2.20) and (2.21), conclude $P(\Omega_{1n}) = \{1 - G(x_c)\}^2 + o(1) > 1 - \delta$ for large n. □

Proof of Proposition 2.4.1. Assumptions 2.1(iib, iic), 2.3, 2.4(i) and Lemma 2.B.3 check Assumption 2.1($iiia$). We check Assumption 2.1($iiib$).

Let $A > 0$. By Assumption 2.1(iic), 2.3, 2.4 and Lemma 2.B.6, $\exists 0 < \rho < \phi - \eta/2$ such that $P(\sum_{i\in\zeta_n^c}\mathbb{I}\{|\varepsilon_i| < c_n + n^{-\phi}A\} < n^\rho) \to 1$, checking (2.2). By Assumption 2.1(iib, iic), 2.3, 2.4(i) and Lemma 2.B.5, for small $\delta > 0$ it holds $P(\sum_{i\in\zeta_n}\mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A\} < n^{\rho+\delta}) \to 1$, checking (2.1). □

Proof of Proposition 2.4.2. Assumptions 2.1(iib, iic), 2.3, 2.5(i) and Lemma 2.B.3 check Assumption 2.1($iiia$). We check Assumption 2.1($iiib$).

Let $A > 0$. By Assumption 2.1(iib, iic), 2.3, 2.5(i) and Lemma 2.B.5, $\exists \rho + \delta < \phi - \eta/2$ such that $P(\sum_{i\in\zeta_n}\mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi}A\} < n^{\rho+\delta}) \to 1$, checking (2.1).

By Lemma 2.B.5, $\exists \delta > 0$ such that $P(\sum_{i\in\zeta_n}\mathbb{I}\{|\varepsilon_i| > c_n + n^{-\phi}A\} > n^{\rho-\delta}) \to 1$, which is equivalent to $P(\psi_{n-\lfloor n^{\rho-\delta}\rfloor} \geq c_n + n^{-\phi}A) \to 1$. By Assumption 2.5($ii$), for any $C_{\rho-\delta} + \epsilon < 1$ it holds $P(\psi_{n-\lfloor n^{\rho-\delta}\rfloor} \leq \{C_{\rho-\delta} + \epsilon\}m_n) \to 1$. Intersecting the two sets shows $P(\{C_{\rho-\delta} + \epsilon\}m_n \geq c_n + n^{-\phi}A) \to 1$. Since $P(\min_{i\in\zeta_n^c}|\varepsilon_i| \geq \{C_{\rho-\delta} + \epsilon\}m_n) \to 1$ by Assumption 2.5(iii), conclude $P(\min_{i\in\zeta_n^c}|\varepsilon_i| \geq c_n + An^{-\phi}) \to 1$, which implies (2.2). □

2.B.3 Proofs for Theorems 2.2 and 2.3

Proof of Theorem 2.2. Let $\hat{\gamma} = h^{-1}\sum_{i\in\zeta_n}(1 - v_i)$ where $v_i = \mathbb{I}\{|y_i - x_i'\tilde{\beta}| \leq c_n\tilde{\sigma}\}$. By Assumption 2.1, 2.2 and Lemma 2.B.1, $\forall \epsilon > 0$, $\exists A_1 > 0$ and sets $\mathcal{B}_n$ with

$P(\mathcal{B}_{1n}) \geq 1 - \epsilon$ such that

$$S_n^\ell = \sum_{i \in \zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n + n^{-\phi} A_1\} \leq h\hat{\gamma} \leq \sum_{i \in \zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n - n^{-\phi} A_1\} = S_n^u \tag{2.24}$$

on $\mathcal{B}_n$ and large n. We show $S_n^r \to_d \text{Poisson}(2\lambda\tau)$ for $r = \ell, u$, which implies $h\hat{\gamma} \to_d \text{Poisson}(2\lambda\tau)$.

By Assumption 2.3(i), $1 - P\{|\varepsilon_i| \leq c_n\} = 2\{1 - F(c_n)\}$ for $i \in \zeta_n$. Since $n\{1 - F(c_n)\} = \tau$ by Assumption 2.4(i) and $h/n = \lambda + o(1)$ then for $i \in \zeta_n$

$$hp_n^* = h(1 - P\{|\varepsilon_i| \leq c_n\}) = 2n\{1 - F(c_n)\}(h/n) = 2\lambda\tau + o(1). \tag{2.25}$$

For $i \in \zeta_n$, use Assumption 2.3, 2.4(i) and Lemma 2.B.4 to bound

$$\mathcal{W}_n = nP\left(c_n - n^{-\phi} A \leq |\varepsilon_i| \leq c_n + n^{-\phi} A\right) = \mathcal{O}(\tau c_n n^{-\phi})$$

Since $c_n^2 = \mathcal{O}_p(n^\eta)L(n) = o(n^\phi)$ by Lemma 2.B.3, conclude $\mathcal{W}_n = o(1)$. Combine $\mathcal{W}_n = o(1)$ and (2.25) to see that for $i \in \zeta_n$

$$hp_n^u = h(1 - P\{|\varepsilon_i| \leq c_n - n^{-\phi} A_1\}) = hP\{c_n - n^{-\phi} A < |\varepsilon_i| \leq c_n\} + hp_n^* = 2\lambda\tau + o(1). \tag{2.26}$$

and by similar reasoning

$$hp_n^\ell = h\{1 - P(\varepsilon_i| \leq c_n + n^{-\phi} A_1)\} = hp_n^* - hP\{c_n < |\varepsilon_i| \leq c_n + n^{-\phi} A_1\} = 2\lambda\tau + o(1). \tag{2.27}$$

By (2.26), (2.27), and the Poisson limit theorem (LLR, Theorem 2.1.1.), conclude

$$\sum_{i \in \zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n + n^{-\phi} A_1\}, \sum_{i \in \zeta_n} \mathbb{I}\{|\varepsilon_i| > c_n + n^{-\phi} A_1\} \to_d \text{Poisson}(2\lambda\tau). \qquad \square$$

Proof of Theorem 2.3. Consider (2.24). We show $(2\lambda n)^{-\rho/2}\{S_n^r - 2\lambda n^\rho\} \to_d N(0, 1)$ for $r = \ell, u$.

Since $h = \lfloor \lambda n \rfloor$ then $h/n = \lambda + \mathcal{O}(n^{-1})$. Thus, by Assumption 2.3(i), 2.5(i), for

$\varepsilon_i \sim F$ it holds

$$hp_n^* = h\{1 - P(|\varepsilon_i| \leq c_n)\} = 2n\{1 - F(c_n)\}(h/n) = 2n^\rho\{\lambda + \mathcal{O}(n^{-1})\} = 2\lambda n^\rho + o(1). \tag{2.28}$$

Note that $\mathcal{W}_n = nP\left(c_n - n^{-\phi}A \leq |\varepsilon_i| \leq c_n + n^{-\phi}A\right) = \mathcal{O}(n^\rho c_n n^{-\phi})$ by Assumption 2.3, 2.4(i) and Lemma 2.B.4. Since $c_n^2 = \mathcal{O}_p(n^\eta)L(n)$ by Lemma 2.B.3, then using $\rho + \eta/2 < \phi$ conclude $\mathcal{W}_n = o(1)$.

Arguing as for (2.26),(2.27) checks $hp_n^r = h\{1-P(|\varepsilon_i| \leq c_n - n^{-\phi}A_1)\} = 2\lambda n^\rho + o(1)$ for $r = u, \ell$. Thus, $\forall t \in \mathbb{R}$, $k_n = 2\lambda n^\rho + t(2\lambda n)^{\rho/2}$, and $r = u, \ell$

$$\frac{k_n - hp_n^r}{k_n^{1/2}(1 - k_n/h)^{1/2}} = \frac{t(2\lambda n)^{\rho/2} + o(1)}{(2\lambda n)^{\rho/2}\{1 + o(1)\}} = t + o(1).$$

By LLR (Theorem 2.5.1.) it follows $P(S_n^r \leq k_n) \to \Phi(t)$, which is equivalent to $P\{(2\lambda n)^{-\rho/2}\{S_n^r - 2\lambda n^\rho\} \leq t\} \to \Phi(t)$. □

2.C Proofs — Asymptotic distribution of stylised IIS

2.C.1 Proof of Theorem 2.4

Proof of Theorem 2.4. Let P_{good} be the distribution of (x_i, ε_i) for $i \in \zeta_n$. Let $\{\tilde{x}_i, \tilde{\varepsilon}_i\}_{i=1}^\infty$ be a sequence of i.i.d. random variables with $(\tilde{x}_1, \tilde{\varepsilon}_1) \sim P_{good}$.

We show for $\mathcal{S}_n = \mathcal{I}_n, \zeta_n$ and $s_n = \#\mathcal{S}_n$, OLS estimators $\hat{\beta}_{\mathcal{S}_n}, \hat{\sigma}^2_{\mathcal{S}_n}$ using observations $\mathcal{S}_n$ satisfy

$$s_n^{-1} \sum_{i \in \mathcal{S}_n} x_i x_i' \to_p E\{\tilde{x}_1 \tilde{x}_1'\}, \quad s_n^{-1/2} \sum_{i \in \mathcal{S}_n} x_i \varepsilon_i \to_d N(0, E\{\tilde{x}_1 \tilde{x}_1'\}\sigma^2). \tag{2.29}$$

$$s_n^{1/2}(\hat{\beta}_{\mathcal{S}_n} - \beta) \to_d N(0, E\{\tilde{x}_1 \tilde{x}_1'\}^{-1}\sigma^2) \tag{2.30}$$

$$s_n^{1/2}(\hat{\sigma}^2_{\mathcal{S}_n} - \sigma^2) \to_d N(0, \sigma^4 E\{\tilde{\varepsilon}_1^4\} - \sigma^4) \tag{2.31}$$

Note that $\sum_{i \in \mathcal{S}_n} x_i x_i'$, $\sum_{i \in \mathcal{S}_n} x_i \varepsilon_i$ have the same distribution as $\sum_{i=1}^{s_n} \tilde{x}_i x_i'$, $\sum_{i=1}^{s_n} \tilde{x}_i \tilde{\varepsilon}_i$ for $\mathcal{S}_n = \mathcal{I}_n, \zeta_n$ since $\mathcal{I}_n$ is independent of $\{x_i, \varepsilon_i\}_{i \in \zeta_n}$. Since $E(\tilde{x}_1 \tilde{x}_1') < \infty$ by Assumption 2.6(ii) then (2.29) holds by the law of large numbers and central limit theorem, which show $s_n^{-1}\sum_{i=1}^{s_n}\tilde{x}_i\tilde{x}_i' \to_p E\{\tilde{x}_1\tilde{x}_1'\}$ and $s_n^{-1/2}\sum_{i=1}^{s_n}\tilde{x}_i\tilde{\varepsilon}_i \to_d N(0, E\{\tilde{x}_1\tilde{x}_1'\})$.

Using (2.29), (2.30) can be argued as in Hayashi (2000, Proposition 2.1). Using (2.29),(2.30) and $E\{\tilde{\varepsilon}_1^4\} < \infty$ from Assumption 2.6(ii), (2.31) can be argued as in Amemiya (1985, Theorem 3.5.6.).

Take $\mathcal{S}_n = \mathcal{I}_n$ in (2.30),(2.31) and use $n = \mathcal{O}_p(\#\mathcal{I}_n)$ from Assumption 2.6(iii) to show $n^{1/2}(\tilde{\beta}-\beta), n^{1/2}(\tilde{\sigma}^2-\sigma) = \mathcal{O}_p(1)$. Thus, by Theorem 2.1 and (2.31), get $h^{1/2}(\hat{\beta}_{IIS}-\beta) = h^{1/2}(\hat{\beta}_{LS,good}-\beta) + o_p(1)$ and $\hat{\sigma}_{IIS} = \sigma + o_p(1)$. By (2.30) with $\mathcal{S}_n = \zeta_n$ it follows $h^{1/2}(\hat{\beta}_{IIS}-\beta)/\hat{\sigma}_{IIS} \to_d N(0, E\{\tilde{x}_1\tilde{x}_1'\})$. The claim follows by Assumption 2.1, Lemma 2.B.2, and (2.29) with $\mathcal{S}_n = \zeta_n$, which show $(h^{-1}\sum_{i=1}^n v_i x_i x_i') = (h^{-1}\sum_{i\in\zeta_n} x_i x_i') + o_p(1) = E(\tilde{x}_i\tilde{x}_i') + o_p(1)$, □

2.C.2 Proof of Theorem 2.5

Let $\mathcal{F}_{in}$ be a sigma algebra increasing in i and let y_{in} be $\mathcal{F}_{in}$-measurable. The array $\{z_{in}, \mathcal{F}_{in}\}_{i\le s_n, n\in\mathbb{N}}$ is an L1-mixingale if $|E(z_{in}|\mathcal{F}_{i-m,n})| \le c_{in}\psi_m \ \forall i \le s_n, m \ge 0, n \ge 1$ and some constants $c_{in} \ge 0$ and $\lim_{m\to\infty}\psi_m = 0$ (Andrews, 1988, Definition 2). A sequence $\{z_i\}_{i=0}^\infty$ is strong mixing (α-mixing) if $\mathcal{F}_i = \sigma(z_0, ..., z_i)$ and $\mathcal{F}_i^\infty = \sigma(z_i, z_{i+1}, \ldots)$ satisfy $\alpha_m = \sup_i \sup_{A\in\mathcal{F}_i, B\in\mathcal{F}_{i+m}^\infty} |P(A\cap B) - P(A)P(B)| \to 0$.

Lemma 2.C.1. *Let $\mathcal{S}_n \subseteq \{1, ..., n\}$ be deterministic with cardinality $s_n \to \infty$. Suppose $\{z_i\}_{i=0}^\infty$ is strong mixing, $Ez_i = 0$, and $E|z_i|^r \le C < \infty$ for some $r > 1$ and all $i \in \mathbb{N}$. Then $s_n^{-1}\sum_{i\in\mathcal{S}_n} z_i = o_p(1)$.*

Proof. Let $\mathcal{F}_i = \sigma(z_0, ..., z_i)$. Let $\mathcal{S}_n$ have ordered elements $k_{1n}, ..., k_{s_n n}$. Then we can write $\{z_i, \mathcal{F}_i\}_{i\in\mathcal{S}_n, n\in\mathbb{N}} = \{z_{k_{in}}, \mathcal{F}_{k_{in}}\}_{i\le s_n, n\in\mathbb{N}}$. Since $Ez_i = 0$, by Davidson (1994, Theorem 14.2)

$$|E(z_{k_{in}}|\mathcal{F}_{k_{i-m,n}})| \le 6\alpha_{k_{in}-k_{i-m,n}}^{1-1/r}(E|z_i|^r)^{1/r}, \tag{2.32}$$

where α_m is the strong mixing coefficient associated with $\{z_i\}_{i=0}^\infty$. Since $k_{in} - k_{i-m,n} \ge m$ and α_m is decreasing then $\alpha_{k_{in}-k_{i-m,n}} \le \alpha_m$. By (2.32) and the assumption $E|z_i|^r \le C < \infty$ then $|E(z_{k_{in}}|\mathcal{F}_{k_{i-m,n}})| \le c_{in}\psi_m$, where $c_{in} = 6C^{1/r}$ and $\psi_m = \alpha_m^{1-1/r}$. Since $\alpha_m \to 0$ by strong mixing, conclude $\{z_{k_{in}}, \mathcal{F}_{k_{in}}\}_{i\le s_n, n\in\mathbb{N}}$ is an L1-mixingale array.

Since $s_n^{-1}\sum_{i=1}^{s_n} c_{in} = 6C^{1/r}$ and $\{z_{k_{in}}\}_{i\le s_n, n}$ is uniformly integrable by Davidson (1994, Theorem 12.10), Andrews (1988, Theorem 2) shows $s_n^{-1}\sum_{i=1}^{s_n} z_{k_{in}} = o_p(1)$. □

Remark 2.C.1. *We state some properties of autoregressive processes. Consider $\tilde{y}_i = \tilde{x}_i'\beta + \tilde{\varepsilon}_i$ for $i = 1, ..., n$, where $\tilde{x}_i' = (1, \tilde{y}_{i-1}, ..., \tilde{y}_{i-q})$ and Assumption 2.7(ic) holds.*

Then (i) $\tilde{y}_i = \sum_{j=0}^{i-1} \pi_j \tilde{\varepsilon}_{i-j} + \boldsymbol{\pi}_i' \tilde{x}_1$ *for some* $\sum_{i=1}^{\infty} |\pi_i|, \sum_{i=1}^{\infty} |\boldsymbol{\pi}_i| < \infty$. *From (i) and triangle inequality, (ii) if* $E|\tilde{x}_1|^r, E|\tilde{\varepsilon}_i|^r < \infty$ *uniformly in i then* $E|\tilde{y}_i|^r < \infty$ *uniformly in i. Finally, because of the autoregressive structure, (iii) if* $\tilde{\varepsilon}_i$ *is i.i.d. with* $E|\tilde{\varepsilon}_i|^{4+\delta} < \infty$ *then* $\exists \tilde{x}_1$ *independent of* $\tilde{\varepsilon}_i$ *such that* $E|\tilde{x}_1|^{4+\delta} < \infty$ *and* $\tilde{y}_i$ *is stationary.*

Lemma 2.C.2. *Suppose Assumption 2.7(ia, ic). Consider* $\tilde{y}_i = \tilde{x}_i'\beta + \sigma\tilde{\varepsilon}_i$, *where* $\tilde{x}_i = (1, \tilde{y}_{i-1}, ..., \tilde{y}_{i-q})$ *and* $\tilde{x}_1$ *corresponds to the stationary solution from Remark 2.C.1(iii). If* $\mathcal{S}_n \subseteq \{1, ..., n\}$ *are deterministic with cardinality* s_n *then*

$$s_n^{-1} \sum_{i \in \mathcal{S}_n} \tilde{x}_i \tilde{x}_i' = E(\tilde{x}_1 \tilde{x}_1') + o_p(1) \tag{2.33}$$

$$s_n^{-1/2} \sum_{i \in \mathcal{S}_n} \tilde{x}_i \tilde{\varepsilon}_i \to_d N(0, E\{\tilde{x}_1 \tilde{x}_1'\}). \tag{2.34}$$

Proof. Showing (2.33). Let $\Sigma = E(\tilde{x}_1 \tilde{x}_1')$. Since $\tilde{\varepsilon}_i$ has a Lebesgue density [Assumption 2.7(*ia*)] and Assumption 2.7(*ic*) holds, Mokkadem (1988, Theorem 1) shows $\tilde{y}_i$ is strong mixing. Then $\tilde{y}_i \tilde{y}_{i-k}$ is also strong mixing $\forall 0 \leq k \leq q$ by Davidson (1994, Theorem 14.1). Thus, each element of $M_i = \tilde{x}_i \tilde{x}_i' - \Sigma$ is strong mixing. Since $EM_i = 0$ then $s_n^{-1} \sum_{i \in \mathcal{S}_n} M_i = o_p(1)$ by Lemma 2.C.1.

Showing (2.34). We show $s_n^{-1/2} \sum_{i \in \mathcal{S}_n} w_i \to_d N(0, t'\Sigma t)$ for $w_i = t'\tilde{x}_i \tilde{\varepsilon}_i$ and any t. If $\mathcal{F}_i = \sigma(\tilde{y}_0, \ldots, \tilde{y}_i)$ then $\{w_i, \mathcal{F}_i\}_{i \in \mathcal{S}_n, n}$ is a martingale difference array. We check conditions for a martingale CLT (Davidson, 1994, Theorem 24.3): (*a*) $s_n^{-1} \sum_{i \in \mathcal{S}_n} E(w_i^2) = t'\Sigma t$, (*b*) $s_n^{-1} \sum_{i \in \mathcal{S}_n} w_i^2 = t'\Sigma t + o_p(1)$, and (*c*) $s_n^{-1/2} \max_{i \in \mathcal{S}_n} |w_i| = o_p(1)$.

Condition (*a*) holds since $E(w_{in}^2) = t'\Sigma t$ by stationarity and Assumption 2.7(*ia*). To check (*b*), write

$$s_n^{-1} \sum_{i \in \mathcal{S}_n} w_i^2 = s_n^{-1} \sum_{i \in \mathcal{S}_n} (t'\tilde{x}_i)^2 (\tilde{\varepsilon}_i^2 - 1) + s_n^{-1} \sum_{i \in \mathcal{S}_n} (t'\tilde{x}_i)^2 \tag{2.35}$$

The second term (2.35) converges to $t'\Sigma t$ by (2.33). We show that the first term is $o_p(1)$.

Let $z_i = (t'\tilde{x}_i)^2 (\tilde{\varepsilon}_i^2 - 1)$. By stationarity and $E|\tilde{x}_1|^{2+\delta} < \infty$, $\exists \delta > 0 : E|\tilde{x}_i|^{2+\delta} < \infty$ uniformly in i. Then by Cauchy-Schwarz inequality and independence of $\tilde{x}_i$ and $\tilde{\varepsilon}_i$ also $\exists \delta > 0 : E|z_i|^{1+\delta} < \infty$ uniformly in i. Thus, $\{z_i, \mathcal{F}_i\}_{i \in \mathcal{S}_n, n}$ is a uniformly integrable martingale difference array, and $s_n^{-1} \sum_{i \in \mathcal{S}_n} z_i = o_p(1)$ by Davidson (1994, Theorem 19.7).

Arguing as for (b), $\exists \delta > 0 : E|w_i|^{2+\delta} < \infty$ uniformly in i. Thus, $s_n^{-1/2} w_i$ satisfies a Liapunov condition which implies (c) (Davidson, 1994, Theorems 23.11 and 23.16). □

Lemma 2.C.3. *Let $\mathcal{S}_n \subseteq \{1, ..., n\}$. If $\sum_{i \in \mathcal{S}_n} |x_{in} - \tilde{x}_{in}|^2 = o_p(n)$ and $\sum_{i \in \mathcal{S}_n} |\tilde{x}_{in}|^2 = \mathcal{O}_p(n)$ then $\sum_{i \in \mathcal{S}_n} x_{in} x'_{in} = \sum_{i \in \mathcal{S}_n} \tilde{x}_{in} \tilde{x}'_{in} + o_p(n)$.*

Proof. If $v_{in} = x_{in} - \tilde{x}_{in}$ then $x_{in} x'_{in} = \tilde{x}_{in} \tilde{x}'_{in} + v_{in} v'_{in} + \tilde{x}_{in} v'_{in} + v_{in} \tilde{x}'_{in}$ and

$$\| \sum_{i \in \mathcal{S}_n} x_{in} x'_{in} - \sum_{i \in \mathcal{S}_n} \tilde{x}_{in} \tilde{x}'_{in} \| \leq \sum_{i \in \mathcal{S}_n} \| v_{in} v'_{in} \| + 2 \sum_{i \in \mathcal{S}_n} \| \tilde{x}_{in} v'_{in} \| = R_{1n} + 2 R_{2n}.$$

by triangle inequality. By assumption, $R_{1n} = \sum_{i \in \mathcal{S}_n} |v_{in}|^2 = o_p(n)$. Further, by Cauchy-Schwarz $\sum_{i \in \mathcal{S}_n} |\tilde{x}_{in}||v_{in}| \leq (\sum_{i \in \mathcal{S}_n} |\tilde{x}_{in}|^2)^{1/2} (\sum_{i \in \mathcal{S}_n} |v_{in}|^2)^{1/2} = O_p(n)$. Since $\|ab'\| = |a||b|$, conclude $R_{2n} \leq (\sum_{i \in \mathcal{S}_n} |\tilde{x}_{in}|^2)^{1/2} (\sum_{i \in \mathcal{S}_n} |v_{in}|^2)^{1/2} = o_p(n)$. □

Remark 2.C.2. *Let $\tilde{\varepsilon}_i$ be i.i.d and $E|\tilde{\varepsilon}_1|^{4+\delta} < \infty$ for some $\delta > 0$. Then $\max_{i \leq n} |\tilde{\varepsilon}_i| = \mathcal{O}_p(n^{1/4})$. To see this, note $P(\max_{i \leq n} |\tilde{\varepsilon}_i| > c) = P(\cup_{i \leq n} \{|\tilde{\varepsilon}_i| > c\}) \leq \sum_{i \leq n} P(|\tilde{\varepsilon}_i| > c) = nP(|\tilde{\varepsilon}_1| > c)$. Since $nP(|\tilde{\varepsilon}_1| > n^{1/4}) = nP(|\tilde{\varepsilon}_1|^{4+\delta} > n^{1+\delta/4}) \leq n^{-\delta/4} E|\tilde{\varepsilon}_1|^{4+\delta} = o(1)$ by Markov's inequality, the claim follows.*

Lemma 2.C.4. *Let $y_i^* = \beta' x_i^* + \sigma(\tilde{\varepsilon}_i + \xi_{inn,i})$, where $x_i^* = (1, y_{i-1}^*, \ldots, y_{i-q}^*)'$. Define $\tilde{y}_i, \tilde{x}_i$ as in Lemma 2.C.2 and ζ_n as in Section 2.5.2. If Assumption 2.7 holds then for $\mathcal{S}_n \subseteq \zeta_n$*

$$\sum_{i \in \mathcal{S}_n} x_i^* (x_i^*)' = \sum_{i \in \mathcal{S}_n} \tilde{x}_i \tilde{x}'_i + o_p(n) \tag{2.36}$$

$$\sum_{i \in \mathcal{S}_n} x_i^* \tilde{\varepsilon}_i = \sum_{i \in \mathcal{S}_n} \tilde{x}_i \tilde{\varepsilon}_i + o_p(n^{1/2}). \tag{2.37}$$

Proof. Write $\xi_i = \sigma \xi_{inn,i}$ and $\zeta_{inn} = \zeta_{inn,n}$ for ease of notation.

Showing (2.36). We check (a) $\sum_{i \in \mathcal{S}_n} |\tilde{x}_i|^2 = \mathcal{O}_p(n)$ and (b) $\sum_{i \in \mathcal{S}_n} |x_i^* - \tilde{x}_i|^2 = o_p(n)$. Then (2.36) holds by Lemma 2.C.3. By Remark 2.C.1(i), $\exists \sum_{i=0}^{\infty} |\pi_i|, \sum_{i=1}^{\infty} |\boldsymbol{\pi}_i| < \infty$ such that

$$\tilde{y}_i = \sum_{j=0}^{i-1} \pi_j \sigma \tilde{\varepsilon}_{i-j} + \boldsymbol{\pi}'_i \tilde{x}_1, \quad y_i^* = \sum_{j=0}^{i-1} \pi_j \sigma (\tilde{\varepsilon}_{i-j} + \xi_i) + \boldsymbol{\pi}'_i x_1^*. \tag{2.38}$$

We check (a). This holds since $\sum_{i \in \mathcal{S}_n} |\tilde{x}_i|^2 = \sum_{i \in \mathcal{S}_n} \mathrm{tr}(\tilde{x}_i \tilde{x}'_i) = \mathcal{O}_p(n)$ by (2.33).

We check (b). We show $S_k = \sum_{i\in\mathcal{S}_n}(y^*_{i-k} - \tilde{y}_{i-k})^2 = o_p(n)$ for $1 \leq k \leq q$, which implies $\sum_{i\in\mathcal{S}_n}|x^*_i - \tilde{x}_i|^2 = \sum_{k=1}^q S_k = o_p(n)$. From (2.38), since $\xi_i = 0$ for $i \in \zeta_{inn}$ then

$$y^*_{i-k} - \tilde{y}_{i-k} = M_{ik} + N_{ik}, \text{ where } M_{ik} = \sum_{j\leq i-k: j\in\zeta^c_{inn}} \pi_{i-k-j}\xi_j, \tag{2.39}$$

and $N_{ik} = \boldsymbol{\pi}'_{i-k}(x^*_1 - \tilde{x}_1)$ for $i > k$ and $|N_{ik}| \leq |x^*_1 - \tilde{x}_1|$ otherwise. From $(a+b)^2 \leq 2a^2 + 2b^2$ it follows $S_k \leq 2\sum_{i\in\mathcal{S}_n} M^2_{ik} + 2\sum_{i\in\mathcal{S}_n} N^2_{ik}$.

By Cauchy-Schwarz inequality, $M^2_{ik} \leq (\sum_{j\leq i-k:j\in\zeta^c_{inn}} \pi^2_{i-k-j})(\sum_{j\leq i-k:j\in\zeta^c_{inn}} \xi^2_j)$. Using $\sum_{j\leq i-k:j\in\zeta^c_{inn}} \xi^2_j \leq (\max_{i\leq n}\xi^2_i)(\#\zeta^c_{inn})$ thus gives

$$\sum_{i\in\mathcal{S}_n} M^2_{ik} \leq (\max_{i\leq n}\xi^2_j)(\#\zeta^c_{inn}) \sum_{i\in\mathcal{S}_n}\sum_{j\leq i-k:j\in\zeta^c_{inn}} \pi^2_{i-k-j}. \tag{2.40}$$

Swapping order of summations and using $\sum_{i=0}^\infty \pi^2_i = C < \infty$ shows

$$\sum_{i\in\mathcal{S}_n}\sum_{j\leq i-k:j\in\zeta^c_{inn}} \pi^2_{i-k-j} \leq \sum_{j\in\zeta^c_{inn}}\sum_{i\geq k+j} \pi^2_{i-k-j} \leq (\#\zeta^c_{inn})C. \tag{2.41}$$

Insert (2.41) in (2.40) to get $\sum_{i\in\mathcal{S}_n} M^2_{ik} \leq (\max_{i\leq n}\xi^2_i)(\#\zeta_{inn})^2 C$. Combine with $(\max_{i\leq n}\xi^2_i)(\#\zeta_{inn})^2 = o_p(n)$ from Assumption 2.7(ii) to conclude $\sum_{i\in\mathcal{S}_n} M^2_{ik} = o_p(n)$.

By Cauchy-Schwarz, $N^2_{ik} \leq |\boldsymbol{\pi}_{i-k}|^2|x^*_1 - \tilde{x}_1|^2$ when $i > k$ and $N^2_{ik} \leq |x^*_1 - \tilde{x}_1|^2$ otherwise. Thus, $\sum_{i\in\mathcal{S}_n} N^2_{ik} \leq |x^*_1 - \tilde{x}_1|^2(k + \sum_{i=1}^\infty |\boldsymbol{\pi}_i|^2)$. Since $\sum_{i=1}^\infty |\boldsymbol{\pi}_i| < \infty$ by Remark 2.C.1(i), $|x^*_1|^2 = o_p(n)$ by Assumption 2.7(ib), and $|\tilde{x}_1|^2 = \mathcal{O}_p(1)$ by construction, conclude $\sum_{i\in\mathcal{S}_n} N^2_{ik} = o_p(n)$.

Showing (2.37). We show $W_{nk} = \sum_{i\in\mathcal{S}_n}|y^*_{i-k} - \tilde{y}_{i-k}||\tilde{\varepsilon}_i| = o_p(n^{1/2})$ for $k = 1,\ldots,q$, which implies $|\sum_{i\in\mathcal{S}_n}(x^*_i - \tilde{x}_i)\tilde{\varepsilon}_i| \leq \sum_{k=1}^q W_{nk} = o_p(n^{1/2})$.

By (2.39) and triangle inequality $W_{nk} \leq \sum_{i\in\mathcal{S}_n}|M_{ik}||\tilde{\varepsilon}_i| + \sum_{i\in\mathcal{S}_n}|N_{ik}||\tilde{\varepsilon}_i|$. We have

$$\sum_{i\in\mathcal{S}_n}|M_{ik}||\tilde{\varepsilon}_i| \leq (\max_{i\leq n}|\xi_i|)\sum_{i\in\mathcal{S}_n}\sum_{j\leq i-k:j\in\zeta^c_{inn}} |\pi_{i-k-j}||\tilde{\varepsilon}_i| = (\max_{i\leq n}|\xi_i|)M_k. \tag{2.42}$$

Using $E|\tilde{\varepsilon}_i| = E|\tilde{\varepsilon}_1|$, swapping summations, and using $\sum_{i=0}^\infty|\pi_i| = K < \infty$ shows $E|M_k| \leq E|\tilde{\varepsilon}_1|\sum_{j\in\zeta^c_{inn}}\sum_{i\geq k+j}|\pi_{i-k-j}| \leq E|\tilde{\varepsilon}_1|(\#\zeta^c_{inn})K$. By Assumption 2.7$(iib)$, conclude $E|M_k| = o(n^\phi)$ implying $M_k = o_p(n^\phi)$. Insert this in (2.42) and use Assumption 2.7(iia) to conclude $\sum_{i\in\mathcal{S}_n}|M_{ik}||\tilde{\varepsilon}_i| = o_p(n^{1/2})$.

By Cauchy-Schwarz $\sum_{i\in\mathcal{S}_n}|N_{ik}||\tilde{\varepsilon}_i| \leq |x^*_1 - \tilde{x}_1|\sum_{i\in\mathcal{S}_n}|\boldsymbol{\pi}_{i-k}||\tilde{\varepsilon}_i| = |x^*_1 - \tilde{x}_1|N_k$.

Use $\sum_{i=1}^{\infty} |\boldsymbol{\pi}_i| < \infty$ and $E|\tilde{\varepsilon}_i| = E|\tilde{\varepsilon}_1|$ to see $E|N_k| = \mathcal{O}(1)$ which implies $N_k = \mathcal{O}_p(1)$. Since $|x_1^* - \tilde{x}_1| = o_p(n^{1/2})$ by Assumption 2.7(ib), conclude $\sum_{i\in\mathcal{S}_n} |N_{ik}||\tilde{\varepsilon}_i| = o_p(n^{1/2})$. □

Proof of Theorem 2.5. Let $\hat{\beta}_{\mathcal{S}_n}, \hat{\sigma}_{\mathcal{S}_n}$ be OLS estimators using observations $\mathcal{S}_n$. We have $x_i = x_i^*$ and $\varepsilon_i = \tilde{\varepsilon}_i$ for $i \in \zeta_n$ by definition. Thus, for $\mathcal{S}_n = \mathcal{I}_n, \zeta_n$ and $s_n = \#\mathcal{S}_n$ it holds $s_n^{1/2}(\hat{\beta}_{\mathcal{S}_n} - \beta) = \{s_n^{-1}\sum_{i\in\mathcal{S}_n} x_i^*(x_i^*)'\}^{-1}\{s_n^{-1/2}\sum_{i\in\mathcal{S}_n} x_i^*\sigma\tilde{\varepsilon}_i\}$. By Lemmas 2.C.2, 2.C.4 we get $s_n^{1/2}(\hat{\beta}_{\mathcal{S}_n} - \beta) \to_d N(0, \Sigma^{-1}\sigma^2)$ where $\Sigma = E(\tilde{x}_1\tilde{x}_1')$.

Using $\varepsilon_i = \tilde{\varepsilon}_i$ and $x_i = x_i^*$ holding for $i \in \mathcal{S}_n = \mathcal{I}_n, \zeta_n$ we also get

$$n^{1/2}(\hat{\sigma}^2_{\mathcal{S}_n} - \sigma^2) = n^{-1/2}\sum_{i\in\mathcal{S}_n} \sigma^2\{\tilde{\varepsilon}_i^2 - 1\} - \{n^{1/2}(\hat{\beta}_{\mathcal{S}_n} - \beta)\}'\{n^{-1}\sum_{i\in\mathcal{S}_n} x_i^*\sigma\tilde{\varepsilon}_i\} = W_{1n} - W_{2n}.$$

By the central limit theorem, $W_{1n} = \mathcal{O}_p(1)$. Further, $n^{1/2}(\hat{\beta}_{\mathcal{S}_n} - \beta) = \mathcal{O}_p(1)$ since $s_n(\hat{\beta}_{\mathcal{S}_n} - \beta) = \mathcal{O}_p(1)$ from above and $n/s_n = \mathcal{O}(1)$ by Assumption 2.7(iii). Combine with $n^{-1}\sum_{i\in\mathcal{S}_n} x_i^*\tilde{\varepsilon}_i = o_p(1)$ from Lemmas 2.C.2, 2.C.4 to conclude $W_{2n} = o_p(1)$.

Taking $\mathcal{S}_n = \mathcal{I}_n$ in the above checks Assumption 2.2. Thus, $h^{1/2}(\hat{\beta}_{IIS} - \beta) = h^{1/2}(\hat{\beta}_{\zeta_n} - \beta) + o_p(1)$ and $\hat{\sigma}_{IIS} = \hat{\sigma}_{\zeta_n} + o_p(1)$ by Theorem 2.1. The above with $\mathcal{S}_n = \zeta_n$ shows $h^{1/2}(\hat{\beta}_{\zeta_n} - \beta) = N(0, \Sigma^{-1}\sigma^2)$ and $\hat{\sigma}_{\zeta_n} = \sigma + o_p(1)$. Meanwhile, $h^{-1}\sum_{i=1}^n v_i x_i x_i' = h^{-1}\sum_{i\in\zeta_n} x_i x_i' + o_p(1) = h^{-1}\sum_{i\in\zeta_n} x_i^*(x_i^*)' + o_p(1) = \Sigma + o_p(1)$, where the first equality is by Lemma 2.B.2, the second since $x_i = x_i^*$ for $i \in \zeta_n$, and the third by Lemmas 2.C.2,2.C.4. Combining, conclude $(\sum_{i=1}^n v_i x_i x_i')^{1/2}(\hat{\beta}_{IIS} - \beta)/\hat{\sigma}_{IIS} \to_d N(0, I)$. □

2.C.3 Proof of Theorem 2.6

We need to control unit root tests using, say, initial observations $n/2, \ldots, n$. When data are from a random walk, the initial value $y_{n/2}$ in this subsample is of order $n^{1/2}$. Standard references on unit root testing (e.g. Phillips and Perron, 1988) assume a bounded initial value and are not directly applicable. The following deals deals with this. Throughout the subsection, let e_i denote the i-th standard basis vector.

Lemma 2.C.5. *Consider data from $\tilde{y}_i = \tilde{y}_{i-1} + \sigma\tilde{\varepsilon}_i$ for $i = 1, ..., n$ and OLS estimator $\tilde{\beta} = (\tilde{\mu}, \tilde{\alpha})'$ for $\tilde{y}_i = \mu + \beta\tilde{y}_{i-1} + \sigma\tilde{\varepsilon}_i$. Let $\tilde{x}_{in} = (1, n^{-1/2}\tilde{y}_{i-1})'$, $M_{xx} = n^{-1}\sum_{i=1}^n \tilde{x}_{in}\tilde{x}_{in}'$, and $\tilde{\sigma}^2_\alpha = e_2' M_{xx}^{-1} e_2 \sigma^2$. If Assumption 2.8(ia) holds and $\tilde{y}_0 = \mathcal{O}_p(n^{1/2})$ then*
(i) $(n\{\tilde{\alpha} - 1\}, \tilde{\sigma}^2_\alpha) \to_d (A_1/A_2, A_2^{-1})$ *for A_1, A_2 defined as in Section 2.5.3*
(ii) $n^{1/2}\tilde{\mu} = \mathcal{O}_p(1)$
(iii) $M_{xx}^{-1} = \mathcal{O}_p(1)$

(iv) $\{M_{xx} + o_p(1)\}^{-1} = M_{xx}^{-1} + o_p(1)$

Proof. Let $S_i = \sum_{j=1}^{i} \tilde{\varepsilon}_j$ for $i \geq 1$, $S_0 = 0$, and $\bar{y}_{-1} = n^{-1} \sum_{i=1}^{n} \tilde{y}_{i-1}$. Note that $\tilde{y}_{i-1} = S_{i-1} + \tilde{y}_0$ and $\bar{y}_{-1} = \bar{S}_{-1} + \tilde{y}_0$, where $\bar{S}_{-1} = n^{-1} \sum_{i=1}^{n} S_{i-1}$. If $\tilde{x}_i = (1, \tilde{y}_{i-1})'$,

$$R' = \begin{pmatrix} 1 & 0 \\ -\bar{y}_{-1} & 1 \end{pmatrix}, \text{ and } N = \begin{pmatrix} n^{-1/2} & 0 \\ 0 & n^{-1} \end{pmatrix},$$

then $N'R'\tilde{x}_i = (n^{-1/2}, n^{-1}\{S_{i-1} - \bar{S}_{-1}\})'$. Thus, by Chan and Wei (1988) it holds jointly

$$K_n = \sum_{i=1}^{n} N'R'\tilde{x}_i\tilde{x}_i'RN = \begin{pmatrix} 1 & 0 \\ 0 & n^{-2} \sum_{i=1}^{n} (S_{i-1} - \bar{S}_{-1})^2 \end{pmatrix} \to_d \begin{pmatrix} 1 & 0 \\ 0 & A_2\sigma^2 \end{pmatrix}, \quad (2.43)$$

$$L_n = \sum_{i=1}^{n} N'R'\tilde{x}_i\sigma\tilde{\varepsilon}_i = \begin{pmatrix} n^{-1/2} \sum_{i=1}^{n} \sigma\tilde{\varepsilon}_i \\ n^{-1} \sum_{i=1}^{n} (S_{i-1} - \bar{S}_{-1})\sigma\tilde{\varepsilon}_i \end{pmatrix} \to_d \begin{pmatrix} \sigma B(1) \\ A_1\sigma^2 \end{pmatrix}. \quad (2.44)$$

We show (i). We have $K_n^{-1}L_n = N^{-1}R^{-1}(\sum \tilde{x}_i\tilde{x}_i')^{-1}(\sum \tilde{x}_i\sigma\tilde{\varepsilon}_i) = N^{-1}R^{-1}(\tilde{\beta} - \beta)$, where $\beta = (0, 1)'$. Since $R^{-1} = \left(\begin{smallmatrix} 1 & \bar{y}_{-1} \\ 0 & 1 \end{smallmatrix}\right)$ then

$$K_n^{-1}L_n = (n^{1/2}\tilde{\mu} + n^{1/2}\bar{y}_{-1}\{\tilde{\alpha} - 1\}, \, n\{\tilde{\alpha} - 1\})'. \quad (2.45)$$

In particular, $n(\tilde{\alpha} - 1) = e_2'K_n^{-1}L_n$. Meanwhile, $\tilde{\sigma}_{\alpha}^2 = \{n^{-2} \sum_{i=1}^{n} (\tilde{y}_{i-1} - \bar{y}_{-1})^2\}^{-1}\sigma^2 = e_2'K_n^{-1}e_2\sigma^2$. By (2.43),(2.44) conclude $(n\{\tilde{\alpha} - 1\}, \tilde{\sigma}_{\alpha}^2) \to_d (A_1/A_2, A_2^{-1})$.

We show (ii). Since $n^{-1/2}\tilde{y}_0 = \mathcal{O}_p(1)$ by assumption and $n^{-1/2}\bar{S}_{-1} \to_d \int_0^1 B(s)ds$ by Chan and Wei (1988) then $n^{-1/2}\bar{y}_{-1} = \mathcal{O}_p(1)$. Combine with $n(\tilde{\alpha} - 1) = \mathcal{O}_p(1)$ from (i) to see $w_{1n} = n^{1/2}\bar{y}_{-1}(\tilde{\alpha} - 1) = \mathcal{O}_p(1)$. Since also $w_{2n} = e_1'K_n^{-1}L_n = \mathcal{O}_p(1)$ by (2.43),(2.44) then from (2.45) conclude $n^{1/2}\tilde{\mu} = w_{2n} - w_{1n} = \mathcal{O}_p(1)$.

We show (iii). Note $M_{xx}^{-1} = (\sum_{i=1}^{n} N'\tilde{x}_i\tilde{x}_i'N)^{-1} = R_nK_n^{-1}R_n'$ where $R_n = N^{-1}RN$. Thus, by sub-multiplicativity $\|M_{xx}^{-1}\| \leq \|R_n\|^2\|K_n^{-1}\|$. We have

$$R_n = \begin{pmatrix} 1 & -n^{-1/2}\bar{y}_{-1} \\ 0 & 1 \end{pmatrix},$$

which is bounded element-wise and thus also in spectral norm. Since $\|K_n^{-1}\| = \mathcal{O}_p(1)$ by (2.43), conclude $\|M_{xx}^{-1}\| = \mathcal{O}_p(1)$.

By Horn and Johnson (2013, p.381), $\{M_{xx} + o_p(1)\}^{-1} - M_{xx}^{-1} = \|M_{xx}^{-1}\|^2 o_p(1)$, so (iv) follows from (iii). □

Lemma 2.C.6. *Let $y_i = y_{i-1} + \sigma\varepsilon_i$ for $i = 1, ..., n$, where $\varepsilon_i = \tilde{\varepsilon}_i + \xi_i$. Suppose Assumption 2.8. Define $\tilde{y}_i = \tilde{y}_{i-1} + \sigma\tilde{\varepsilon}_i$ for $i = 1, ..., n$, $\tilde{y}_0 = y_0$, $x_{in} = (1, n^{-1/2}y_{i-1})$, and $\tilde{x}_{in} = (1, n^{-1/2}\tilde{y}_{i-1})$. Then for $\mathcal{S}_n = \zeta_n, \mathcal{I}_n$ and $s_n = \#\mathcal{S}_n$ it holds*

$$\max_{i\le n} |\tilde{x}_{in}| = \mathcal{O}_p(1), \quad \max_{i\le n} E|x_{in} - \tilde{x}_{in}|^2 = o(1) \tag{2.46}$$

$$\sum_{i\in\mathcal{S}_n} x_{in}x_{in}' = \sum_{i=\min(\mathcal{S}_n)}^{\max(\mathcal{S}_n)} \tilde{x}_{in}\tilde{x}_{in}' + o_p(n) \tag{2.47}$$

$$\sum_{i\in\mathcal{S}_n} x_{in}\varepsilon_i = \sum_{i=\min(\mathcal{S}_n)}^{\max(\mathcal{S}_n)} \tilde{x}_{in}\tilde{\varepsilon}_i + o_p(n^{1/2}), \tag{2.48}$$

$$\sum_{i\in\mathcal{S}_n} \tilde{x}_{in}\tilde{x}_{in}' = \mathcal{O}_p(n), \quad \sum_{i\in\mathcal{S}_n} \tilde{x}_{in}\tilde{\varepsilon}_i = \mathcal{O}_p(n^{1/2}). \tag{2.49}$$

Proof. Let $i_{\min} = \min(\mathcal{S}_n)$, $i_{\max} = \max(\mathcal{S}_n)$, $\Delta_n = (i_{\max} - i_{\min} + 1)$, and $\mathcal{S}_n^c = \{i_{\min} \le i \le i_{\max} : i \notin \mathcal{S}_n\}$. Since $\mathcal{S}_n^c \subseteq \zeta_n^c$ or $\mathcal{S}_n = \varnothing$ by Assumption 2.8(iii) then by Assumption 2.8(iib)

$$\#\mathcal{S}_n^c = o(n). \tag{2.50}$$

Showing (2.46). Since $\tilde{x}_{in} = (1, n^{-1/2}\tilde{y}_{i-1})$ then $\max_{i\le n} |\tilde{x}_{in}| \le 1 + n^{-1/2}\max_{i\le n} |\tilde{y}_i|$. Since $n^{-1/2}\sum_{j=1}^{\lfloor rn \rfloor} \tilde{\varepsilon}_j$ converges to a Brownian motion, its maximum is $\mathcal{O}_p(1)$. As $y_0 = \mathcal{O}_p(n^{1/2})$ by Assumption 2.8(ic) then $n^{-1/2}\max_{i\le n} |\tilde{y}_i| \le \sigma \max_{i\le n} n^{-1/2}|\sum_{j=1}^{i} \tilde{\varepsilon}_j| + n^{-1/2}|y_0| = \mathcal{O}_p(1)$, showing $\max_{i\le n} |\tilde{x}_{in}| = \mathcal{O}_p(1)$.

Since $v_{in} = x_{in} - \tilde{x}_{in} = (0, n^{-1/2}\sum_{j\le i-1: j\in\zeta_n^c} \sigma\xi_i)$ then by Cauchy-Schwarz $|v_{in}|^2 = n^{-1}(\sum_{j\le i-1:j\in\zeta_n^c} \sigma\xi_i)^2 \le n^{-1}\sigma^2(\#\zeta_n^c)(\sum_{i\in\zeta_n^c} \xi_i^2)$. By Assumption 2.8$(iia, iib)$, conclude $\max_{i\le n} E|v_{in}|^2 \le n^{-1}\sigma^2(\#\zeta_n^c)^2(\max_{i\le n} E\xi_i^2) = o(1)$.

Showing (2.47). By triangle inequality $\|\sum_{i\in\mathcal{S}_n} x_{in}x_{in}' - \sum_{i=i_{\min}}^{i_{\max}} \tilde{x}_{in}\tilde{x}_{in}'\| \le R_{1n} + R_{2n}$, where $R_{1n} = \sum_{i\in\mathcal{S}_n^c} \|\tilde{x}_{in}\tilde{x}_{in}'\|$ and $R_{2n} = \|\sum_{i\in\mathcal{S}_n}(x_{in}x_{in}' - \tilde{x}_{in}\tilde{x}_{in}')\|$. It holds $R_{1n} = \sum_{i\in\mathcal{S}_n^c} |\tilde{x}_{in}|^2 \le (\#\mathcal{S}_n^c)(\max_{i\le n} |\tilde{x}_{in}|^2) = o_p(n)$ by (2.46),(2.50).

We check (a) $\sum_{i\in\mathcal{S}_n} |\tilde{x}_{in}|^2 = \mathcal{O}_p(n)$ and (b) $\sum_{i\in\mathcal{S}_n} |x_{in} - \tilde{x}_{in}|^2 = o_p(n)$. Then $R_{2n} = o_p(n)$ follows by Lemma 2.C.3. Since $\max_{i\le n} |\tilde{x}_{in}| = \mathcal{O}_p(1)$ by (2.46), (a) holds. Further, by (2.46), $E\sum_{i\in\mathcal{S}_n} |x_{in} - \tilde{x}_{in}|^2 \le (\#\mathcal{S}_n)\max_{i\le n} E|x_{in} - \tilde{x}_{in}|^2 = o(n)$, implying $\sum_{i\in\mathcal{S}_n} |x_{in} - \tilde{x}_{in}|^2 = o_p(n)$, checking (b).

Showing (2.48). Since $\varepsilon_i = \tilde{\varepsilon}_i$ for $i \in \mathcal{S}_n = \zeta_n, \mathcal{I}_n$ then

$$\sum_{i\in\mathcal{S}_n} x_{in}\varepsilon_i - \sum_{i=\min(\mathcal{S}_n)}^{\max(\mathcal{S}_n)} \tilde{x}_{in}\tilde{\varepsilon}_i = \sum_{i\in\mathcal{S}_n}(x_{in}-\tilde{x}_{in})\tilde{\varepsilon}_i - \sum_{i\in\mathcal{S}_n^c}\tilde{x}_{in}\tilde{\varepsilon}_i = S_{1n} - S_{2n}. \tag{2.51}$$

If $w_{in} = n^{-1/2}(y_{i-1} - \tilde{y}_{i-1})$ then $S_{1n} = (0, \sum_{i\in\mathcal{S}_n} w_{in}\tilde{\varepsilon}_i)'$ and it follows $|S_{1n}|^2 = (\sum_{i\in\mathcal{S}_n} w_{in}\tilde{\varepsilon}_i)^2$. By Assumption 2.8($i, ic$), $w_{in}\tilde{\varepsilon}_i = (n^{-1/2}\sum_{j\le i-1}\sigma\xi_j)\tilde{\varepsilon}_i$ is a martingale difference. By independence of w_{in} and $\tilde{\varepsilon}_i$ then $E|S_{1n}|^2 = \sum_{i\in\mathcal{S}_n} E\{w_{in}^2\tilde{\varepsilon}_i^2\} = \sum_{i\in\mathcal{S}_n} Ew_{in}^2$. Since $\max_{i\le n} E|w_{in}|^2 = o(1)$ holds by (2.46), conclude $E|S_{1n}|^2 \le (\#\mathcal{S}_n)\max_{i\le n} E\xi_i^2 = o(n)$, which implies $ES_{1n} = o_p(n^{1/2})$.

We have $|S_{2n}|^2 = (\sum_{i\in\mathcal{S}_n^c}\tilde{\varepsilon}_i)^2 + (n^{-1/2}\sum_{i\in\mathcal{S}_n^c}\tilde{y}_{i-1}\tilde{\varepsilon}_i)^2 = S_{2n}' + S_{2n}''$. By Assumption 2.8($ia$) and (2.50), $ES_{2n}' = (\#\mathcal{S}_n^c) = o(n)$. By Assumption 2.8($ia, ic$), $\tilde{y}_{i-1}\tilde{\varepsilon}_i$ is a martingale difference and thus $ES_{2n}'' = n^{-1}\sum_{i\in\mathcal{S}_n^c} E\tilde{y}_{i-1}^2$. Using (2.50) further $n^{-1}\sum_{i\in\mathcal{S}_n^c} E\tilde{y}_{i-1}^2 \le o(1)\max_{i\le n} E\tilde{y}_{i-1}^2$. Since $\tilde{y}_{i-1} = \sum_{j=1}^{i-1}\sigma\tilde{\varepsilon}_i + y_0$ then $\max_{i\le n} E\tilde{y}_{i-1}^2 = \sigma^2(n-1) + Ey_0^2 = \mathcal{O}(n)$ by Assumption 2.8(i, ic). Conclude $E|S_{2n}|^2 = o(n)$, which implies $S_{2n} = o_p(n^{1/2})$

Showing (2.49). By (2.46), $\max_{i\le n}|\tilde{x}_{in}|^2 = \mathcal{O}_p(1)$. By triangle inequality and $\|\tilde{x}_{in}\tilde{x}_{in}'\| = |\tilde{x}_{in}|^2$ it follows $\|\sum_{i\in\mathcal{S}_n}\tilde{x}_{in}\tilde{x}_{in}'\| \le \sum_{i\in\mathcal{S}_n}|\tilde{x}_{in}|^2 = \mathcal{O}_p(n)$.

We check $E|T_n|^2 = \mathcal{O}(n)$ for $T_n = \sum_{i\in\mathcal{S}_n}\tilde{x}_{in}\tilde{\varepsilon}_i$, implying $T_n = \mathcal{O}_p(n^{1/2})$. We have $|T_n|^2 = T_{1n} + T_{2n}$ where $T_{1n} = (\sum_{i\in\mathcal{S}_n}\tilde{\varepsilon}_i)^2$ and $T_{2n} = (n^{-1/2}\sum_{i\in\mathcal{S}_n}\tilde{y}_{i-1}\tilde{\varepsilon}_i)^2$. Arguing as above, we have $ET_{1n} = (\#\mathcal{S}_n) = \mathcal{O}(n)$ and $ET_{2n} \le n^{-1}(\#\mathcal{S}_n)\max_{i\le n} E\tilde{y}_{i-1}^2 = \mathcal{O}(n)$. □

Proof of Theorem 2.6. Let $\hat{\beta}_{\mathcal{S}_n}, \hat{\sigma}_{\mathcal{S}_n}$ be OLS estimators of $y_i = x_i'\beta + \varepsilon_i$ using observations $\mathcal{S}_n$. By Lemmas 2.C.6, 2.C.5(iii, iv), for $\mathcal{S}_n = \zeta_n, \mathcal{I}_n$

$$\begin{aligned} N^{-1}(\hat{\beta}_{\mathcal{S}_n} - \beta) &= (n^{-1}\sum_{i\in\mathcal{S}_n} x_{in}x_{in}')^{-1}(n^{-1/2}\sum_{i\in\mathcal{S}_n} x_{in}\sigma\varepsilon_i) \\ &= (n^{-1}\sum_{\min(\mathcal{S}_n)}^{\max(\mathcal{S}_n)}\tilde{x}_{in}\tilde{x}_{in}')^{-1}(n^{-1/2}\sum_{\min(\mathcal{S}_n)}^{\max(\mathcal{S}_n)}\tilde{x}_{in}\tilde{\varepsilon}_i) + o_p(1). \end{aligned} \tag{2.52}$$

Re-indexing the observations $\min(\mathcal{S}_n), ..., \max(\mathcal{S}_n)$ as '$1, ..., n$', it follows $N^{-1}(\hat{\beta}_{\mathcal{S}_n} - \beta) = \mathcal{O}_p(1)$ by Lemma 2.C.5(i, ii).

Since $\varepsilon_i = \tilde{\varepsilon}_i$ for $i \in \mathcal{S}_n = \zeta_n, \mathcal{I}_n$ we have $n^{1/2}(\hat{\sigma}_{\mathcal{S}_n}^2 - \sigma^2) = W_{1n} - W_{2n}$, where $W_{1n} = n^{-1/2}\sum_{i\in\mathcal{S}_n}\sigma^2\{\tilde{\varepsilon}_i^2 - 1\}$ and $W_{2n} = \{N^{-1}(\hat{\beta}_{\mathcal{S}_n} - \beta)\}'\{n^{-1}\sum_{i\in\mathcal{S}_n} x_{in}\sigma\tilde{\varepsilon}_i\}$. It holds $W_{1n} = \mathcal{O}_p(1)$ by the central limit theorem. Meanwhile, $W_{2n} = o_p(1)$ since

$N^{-1}(\hat{\beta}_{\mathcal{S}_n} - \beta) = \mathcal{O}_p(1)$ from above and $\sum_{i\in\mathcal{S}_n} x_{in}\tilde{\varepsilon}_i = \mathcal{O}_p(n^{1/2})$ by (2.48, 2.49). It follows $n^{1/2}(\hat{\sigma}_{\mathcal{S}_n} - \sigma^2) = \mathcal{O}_p(1)$.

Let $h_\circ = \max(\zeta_n) - \min(\zeta_n) + 1$, $x_{ih} = (1, h_\circ^{-1/2} y_{i-1})$, and $\tilde{x}_{ih} = (1, h_\circ^{-1/2}\tilde{y}_{i-1})$. Consider $t = h_\circ(\hat{\alpha}_{IIS} - 1)/\hat{\sigma}_\alpha$, where $\hat{\sigma}^2_\alpha = e_2'(h_\circ^{-1}\sum_{i=1}^n v_i x_{ih} x_{ih}')^{-1} e_2 \hat{\sigma}^2_{IIS}$. Let $\tilde{t} = h_\circ(\tilde{\alpha} - 1)/\tilde{\sigma}_\alpha$ where $\tilde{\alpha}$ is the OLS estimator for $\tilde{y}_i = \mu + \alpha\tilde{y}_{i-1} + \sigma\varepsilon_i$ using $i = \min(\zeta_n), \dots, \max(\zeta_n)$, and $\tilde{\sigma}^2_\alpha = e_2'(h_\circ^{-1}\sum_{i=\min(\zeta_n)}^{\max(\zeta_n)} \tilde{x}_{ih}\tilde{x}_{ih}')^{-1} e_2\sigma^2$.

We show $h_\circ(\hat{\alpha}_{IIS} - 1) = h_\circ(\tilde{\alpha} - 1) + o_p(1)$. Taking $\mathcal{S}_n = \mathcal{I}_n$ in the above checks Assumption 2.2. By Theorem 2.1, it follows $h_\circ(\hat{\alpha}_{IIS} - 1) = h_\circ(\hat{\alpha}_{\zeta_n} - 1) + o_p(1)$. The above with $\mathcal{S}_n = \zeta_n$ further shows $h_\circ(\hat{\alpha}_{\zeta_n} - 1) = h_\circ(\tilde{\alpha} - 1) + o_p(1)$.

We show $\hat{\sigma}^2_\alpha = \tilde{\sigma}^2_\alpha + o_p(1)$. Theorem 2.1 and the above with $\mathcal{S}_n = \zeta_n$ show $\hat{\sigma}_{IIS} = \hat{\sigma}_{\zeta_n} + o_p(1) = \sigma + o_p(1)$. Lemma 2.B.2 shows $h_\circ^{-1}\sum_{i=1}^n v_i x_{ih}x_{ih}' = h_\circ^{-1}\sum_{i\in\zeta_n} x_{ih}x_{ih}' + o_p(1)$ while (2.47) gives $h_\circ^{-1}\sum_{i\in\zeta_n} x_{ih}x_{ih}' = h_\circ^{-1}\sum_{i=\min(\zeta_n)}^{\max(\zeta_n)} \tilde{x}_{ih}\tilde{x}_{ih}' + o_p(1) = M_{xx} + o_p(1)$. By Lemma 2.C.5($iv$), conclude $\hat{\sigma}^2_\alpha = e_2' M_{xx}^{-1} e_2\sigma^2 + o_p(1)$.

Using $t = \{h_\circ(\tilde{\alpha} - 1) + o_p(1)\}/\{\tilde{\sigma}^2_\alpha + o_p(1)\}^{1/2}$ and Lemma 2.C.5, conclude $t \to A_1/A_2^{1/2}$. □

2.D Proofs for examples

2.D.1 Examples on Assumption 2.3

Example 2.D.1. *If F is standard normal then Assumption 2.3 holds for $c_n = \mathcal{O}(n^\phi)$ and Assumption 2.3(ii) holds for any $c_n \to \infty$. This is checked with the inequality $\left\{(4 + c^2)^{1/2} - c\right\}/2 < (1 - \Phi(c))/\phi(c)$ (Sampford, 1953). Assumption 2.1(iii) holds for $c_n = \mathcal{O}(n^\phi)$ since $2\log\{\phi(c_n - n^{-\phi})/\phi(c_n)\} = c_n^2 - (c_n - n^{-\phi}A)^2 = 2c_n n^{-\phi}A - n^{-2\phi}A^2$.*

Example 2.D.2. *If F is a t_d-distribution then Assumption 2.3 holds $\forall c_n \to \infty$. To check Assumption 2.3(ii), combine $1 - F(c_n) \sim B_d\, c_n^{-d}$ for some $B_d > 0$ (Soms, 1976) together with $f(c_n) = \mathcal{O}(c_n^{-1-d})$, which shows $f(c_n)/[c_n\{1 - F(c_n)\}] = \mathcal{O}(c_n^{-2})$. For Assumption 2.3(iii), note $f(c_n - n^{-\phi}A)/f(c_n) = \{1 + (c_n - n^{-\phi}A)^2\}^{-(d+1)/2}/\{1 + c_n^2\}^{-(d+1)/2)} = 1 + o(1)$.*

Remark 2.D.1. *If $\ell(x) = \log f(x)$ is uniformly continuous, such as the Laplace distribution where $f(x) \propto \exp(-|x|)$, then Assumption 2.3(iii) holds. This is checked since by uniform continuity $\exists\delta > 0$ such that $\log\{f(c_n - An^{-\phi})/f(c_n)\} = \ell(c_n - A_n^{-\phi}) - \ell(c_n) \le \delta$ for all c_n and large n.*

2.D.2 Further examples

Proof of Example 2.4.2. Define $S_n = \sum_{j \in \zeta_n^c} \mathbb{I}\{-u_j > -x/a_n - n^{-\phi}A\}$. Assumption 2.4(*iii*) is equivalent to: $\forall A > 0$, $\exists 0 < \rho < \phi - \eta/2 : \forall x \geq 0$ it holds $P(S_n < n^\rho) = P(S_n \leq \lceil n^\rho \rceil - 1) \to 1$.

Let $k_n = \lceil n^\rho \rceil - 1$ and $p_n = 1 - P(-u_j \leq -x/a_n - n^{-\phi}A)$. By Leadbetter et al. (1983, Theorem 2.5.1), $P(S_n < n^\rho) \to 1$ holds if and only if

$$W_n = \frac{k_n - (n-h)p_n}{k_n^{1/2}\{1 - k_n/(n-h)\}^{1/2}} = \frac{k_n - np_n(1 - \lambda + o(1))}{k_n^{1/2}\{1 + o(1)\}} \to \infty.$$

When u_j is uniform, $p_n = x/a_n + n^{-\phi}A \geq n^{-\phi}A$. Then $k_n - np_n(1 - \lambda + o(1)) \leq n^\rho - n^{1-\phi}\mathcal{O}(1) \to -\infty$ $\forall 0 < \rho < \phi < 1/2$, implying that $k_n - np_n(1 - \lambda + o(1)) \leq 0$. As this holds for all A, a_n, x, Assumption 2.4(*iii*) is violated for any F. □

Proof of Example 2.4.3. *1. Construction of* Ω_n. Let $v_i = \mathbb{I}\{|y_i - \mu| \leq c_n\}$ and $\hat{h} = \sum_{i=1}^n v_i$. By normality, if $\Omega_{1n} = \{\max_{i \in \zeta_n} \varepsilon_i \geq 1\}$ then $P(\Omega_{1n}) \to 1$. Let Ω_{2n} be such that $c_n - \max_{i \in \zeta_n} \varepsilon_i \geq x/a_n$ on Ω_{2n} for some $x > 0$ and $a_n \sim (2 \log h)^{1/2}$. Let Ω_{3n} be such that $\sum_{j \in \zeta_n^c} \mathbb{I}\{u_j \leq x/a_n\} \geq n^{3/4}$ on Ω_{3n}. Define $\Omega_n = \Omega_{1n} \cap \Omega_{2n} \cap \Omega_{3n}$.

2. Analysis on Ω_n. Note $n^{1/2}|\hat{\mu} - \mu| = n^{1/2}\hat{h}^{-1}\,|\sum_{i=1}^n v_i \varepsilon_i| \geq n^{-1/2}\left|\sum_{i \in \zeta_n^c} v_j \varepsilon_j\right| = \mathcal{W}_n$. Since $\varepsilon_j \geq \max_{i \in \zeta_n} \varepsilon_i \geq 1$ for $j \in \zeta_n^c$ on Ω_{1n} then $\mathcal{W}_n \geq n^{-1/2}\sum_{i \in \zeta_n^c} v_j$. Since $v_j = \mathbb{I}\{\max_{i \in \zeta_n} \varepsilon_i + u_j \leq c_n\} \geq \mathbb{I}\{u_j \leq x/a_n\}$ on Ω_{2n} then $\sum_{j \in \zeta_n^c} v_j \geq n^{3/4}$ on $\Omega_{2n} \cap \Omega_{3n}$. Conclude $\mathcal{W}_n \geq n^{1/4} \to \infty$ on Ω_n.

3. Probability analysis. We show $\exists x > 0 : P(\Omega_{2n}) \to \epsilon > 0$ by showing $\exists x_c \in \mathbb{R}, a_n, x > 0$ such that (*i*) $\max_{i \in \zeta_n} \varepsilon_i \leq (x_c - x)/a_n + b_n$ on sets Ω_{1n} with $P(\Omega_{1n}) \to \epsilon$ and (*ii*) $c_n \geq x_c/a_n + b_n$ for large n. By normality and LLR (Theorem 1.5.3.), $P(\max_{i \in \zeta_n} \varepsilon_i \leq \{x_c - x\}/a_n + b_n) \to \exp(-\exp\{x_c - x\}) > 0$ for some $a_n \sim b_n \sim (2 \log n)^{1/2}$ and all x_c, x, checking (*i*).

We check (*ii*). Since $n\{1 - \Phi(c_n)\} = \tau$ then by $P(\max_{i \in \zeta_n} \varepsilon_i \leq c_n) \to exp(-\tau)$ by LLR (Theorem 1.5.1.) while $P(\max_{i \in \zeta_n} \varepsilon_i \leq x/a_n + b_m) \to \exp(-\exp\{-x\})$ for any x_c by LLR (Theorem 1.5.3.). Then $\forall x_c : \exp(x_c) > \tau$ it holds $P(\max_{i \in \zeta_n} \varepsilon_i \leq x_c/a_n + b_n) < P(\max_{i \in \zeta_n} \varepsilon_i \leq c_n)$ for large n, and (*ii*) follows by monotonicity of a distribution function.

We show $P(\Omega_{3n}) \to 1$. Let $n_{out} = n - h$, $p_n = 1 - P(u_j \leq x/a_n) = 1 - x/a_n$, and $k_n = n_{out} - n^{3/4}$. Since $n_{out} - n = 0.01 + o(1)$ and $a_n \sim (2 \log h)^{1/2}$ then

$k_n - n_{out}p_n = a_n^{-1}n_{out}\{x + o(1)\}$ and $k_n\{1 - k_n/n_{out}\} = n^{3/4}\{1 + o(1)\}$. Thus,

$$\frac{k_n - (n-h)p_n}{[k_n\{1 - k_n/(n-h)\}]^{1/2}} = \frac{n_{out}\{x + o(1)\}}{a_n n^{3/8}\{1 + o(1)\}} \to \infty, \tag{2.53}$$

By LLR (Theorem 2.5.2.) it follows $P(\sum_{j\in\zeta_n^c} \mathbb{I}\{u_j > x/a_n\} \le n_{out} - n^{3/4}) \to 1$, which is equivalent to $P(\Omega_{3n}) \to 1$. □

Proof of Example 2.5.3. By definition, $\zeta_n = \{1 \le i \le n : \xi_{add,i} = 0$ and $\xi_{add,i-1} = 0\}$. Since $\xi_{add,i}\xi_{add,i-1} = 0$ for all i then $\zeta_n^c = \{i \le n : \xi_{add,i} = 0$ and $\xi_{add,i-1} \neq 0\} \cup \{i \le n : \xi_{add,i} \neq 0$ and $\xi_{add,i-1} = 0\}$.

For $i \in \{i \le n : \xi_{add,i} = 0$ and $\xi_{add,i-1} \neq 0\}$ reverse triangle inequality gives $|\varepsilon_i| \ge |\beta\xi_{add,i-1}| - |\tilde{\varepsilon}_i| \ge |\beta\xi_{add,i-1}| - \max_{i\le n}|\tilde{\varepsilon}_i|$. If $\min_{i\in\zeta_n}|\xi_{add,i}| \ge 2|\beta^{-1}|\max_{i\le n}|\tilde{\varepsilon}_i|\{1 + o_p(1)\}$ then conclude $|\varepsilon_i| \ge \max_{i\le n}|\tilde{\varepsilon}_i|\{1 + o_p(1)\} \ge m_n\{1 + o_p(1)\}$.

For $i \in \{i \le n : \xi_{add,i} \neq 0$ and $\xi_{add,i-1} = 0\}$ reverse triangle inequality gives $|\tilde{\varepsilon}_i| \ge |\xi_{add,i}| - |\tilde{\varepsilon}_i| \ge |\xi_{add,i}| - \max_{i\le n}|\tilde{\varepsilon}_i|$. If $\min_{i\in\zeta_n}|\xi_{add,i}| \ge 2|\beta^{-1}|\max_{i\le n}|\tilde{\varepsilon}_i|\{1 + o_p(1)\}$ then using $|\beta| \le 1$ conclude $|\varepsilon_i| \ge \max_{i\le n}|\tilde{\varepsilon}_i|\{1 + o_p(1)\} \ge m_n\{1 + o_p(1)\}$. □

References for Chapter 2

Amemiya, T. (1985). Advanced Econometrics. Cambridge, Mass: Harvard University Press.

Andrews, D. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* 4(3), 458–467.

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

Atkinson, A. C., M. Riani, A. Corbellini, D. Perrotta, and V. Todorov (2025). Robust Statistics Through the Monitoring Approach : Applications in Regression. 1st ed. Springer Series in Statistics. Springer Nature Switzerland.

Balke, N. S. and T. B. Fomby (1994). Large shocks, small shocks, and economic fluctuations: Outliers in macroeconomic time series. *Journal of Applied Econometrics* 9(2), 181–200.

Banerjee, A., J. J. Dolado, J. W. Galbraith, and D. F. Hendry (1993). Co-integration, Error Correction, and the Econometric Analysis of Non-Stationary Data. Advanced texts in econometrics. New York: Oxford University Press.

Berenguer-Rico, V., S. Johansen, and B. Nielsen (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 B. Nielsen (2025). Least Trimmed Squares: Nuisance parameter free asymptotics. *Econometric Theory*, 1–39.
— (forthcoming). Least Trimmed Squares: Cointegration and outliers. *Oxford Bulletin of Economics and Statistics.*
Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association* 70(350), 428–434.
Castle, J. L., J. Doornik, and D. F. Hendry (2012). Model selection when there are multiple breaks. *Journal of Econometrics* 169(2), 239–246.
— (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics* 3(1).
Chan, N. H. and C. Z. Wei (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* 16(1), 367–401.
Davidson, J. (1994). Stochastic Limit Theory : An Introduction for Econometricians. Oxford: Oxford University Press.
Doornik, J. (2009). Autometrics. *The Methodology and Practice of Econometrics.* Ed. by N. Shephard and J. L. Castle. Oxford, UK: Oxford University Press, 88–121.
Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society, Series B* 34(3), 350–363.
Franses, P. H. and N. Haldrup (1994). The effects of additive outliers on tests for unit roots and cointegration. *Journal of Business & Economic Statistics* 12(4), 471–478.
Galambos, J. (1987). The Asymptotic Theory of Extreme Order Statistics. 2nd ed. Malabar, FL: Robert E. Krieger Publishing Company.
Hayashi, F. (2000). Econometrics. Princeton, N.J.: Princeton University Press.
Hendry, D. F. (1999). An econometric analysis of US food expenditure 1931–1989. *Methodology & Tacit Knowledge: Two Experiments in Econometrics.* Ed. by J. Magnus and M. Morgan. Chichester, UK: John Wiley & Sons, 41–362.
Hendry, D. F. and J. Doornik (2014). Empirical Model Discovery and Theory Evaluation. Cambridge, Massachusetts: The MIT Press.
Hendry, D. F., S. Johansen, and C. Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 23, 317–335.
Horn, R. A. and C. R. Johnson (2013). Matrix analysis. 2nd ed. Cambridge, UK: Cambridge University Press.
Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35(1), 73–101.
Johansen, S. and B. Nielsen (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics* 43(2), 321–348.
Klooster, J. and B. Nielsen (2025). On (in)consistency of M-estimators under contamination. arXiv: 2502.09145.
Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). Extremes and related properties of random sequences and processes. 1st ed. Springer-Verlag New York.
Mokkadem, A. (1988). Mixing properties of ARMA processes. *Stochastic processes and their applications* 29(2), 309–315.

Phillips, P. C. and P. Perron (1988). Testing for a unit root in time series regression. *Biometrika* 75(2), 335–346.

Pretis, F., J. J. Reade, and G. Sucarrat (2018). Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software* 86, 1–44.

Rousseeuw, P. J. and K. Van Driessen (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery* 12(1), 29–45.

Sampford, M. R. (1953). Some inequalities on Mill's ratio and related functions. *Annals of Mathematical Statistics* 24(1), 130–132.

Soms, A. P. (1976). An asymptotic expansion for the tail area of the t-distribution. *Journal of the American Statistical Association* 71(355), 728–730.

Welsh, A. H. and E. Ronchetti (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* 103(1-2), 287–310.

Chapter 3

Demand vs. supply decomposition of inflation: Cross-country evidence with applications

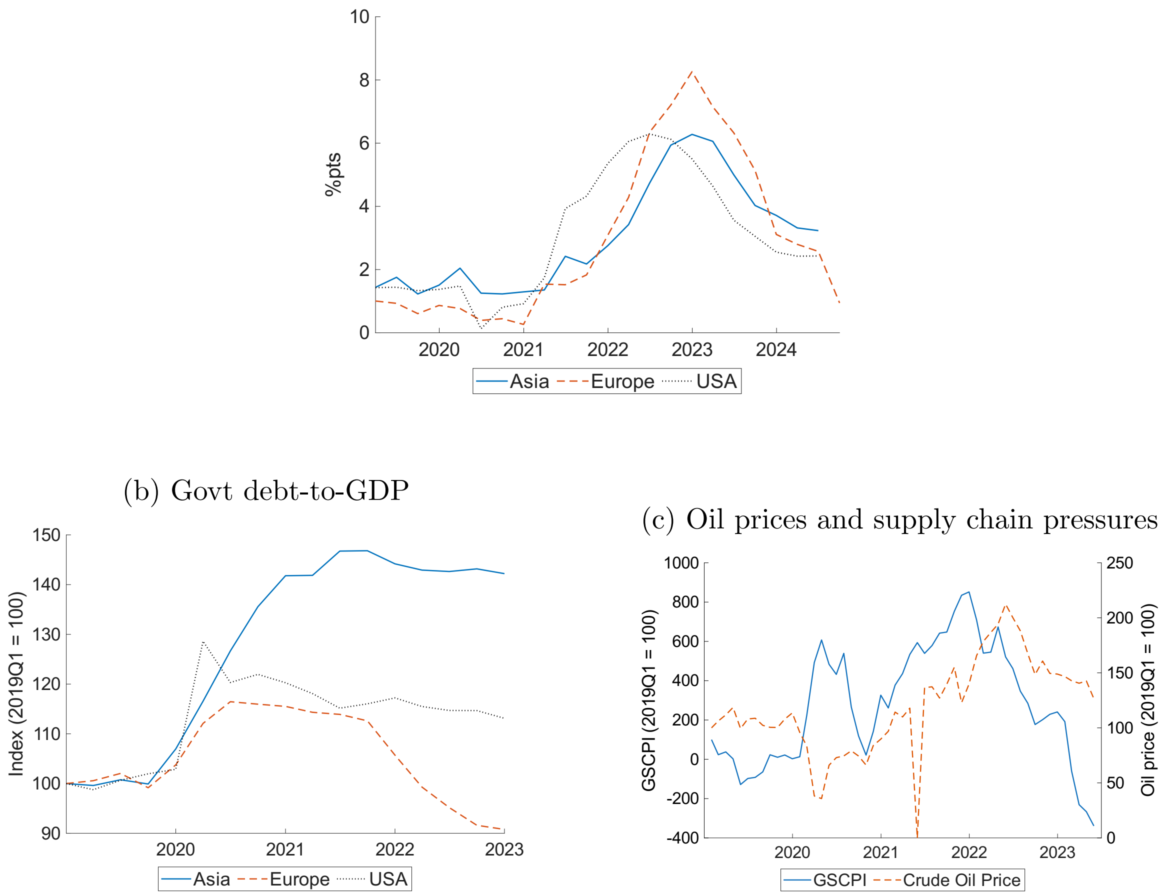
Abstract: *We decompose inflation into demand- and supply-driven components in 32 countries using disaggregated personal consumption expenditures data. The decompositions are updated frequently and can be used for inflation monitoring. We validate the decompositions by assessing their responses to external measures of demand and supply shocks. We then use the decompositions in two applications. First, we examine the roles demand- and supply-side factors played in the surge in inflation following the COVID-19 pandemic. Second, we study whether the empirical Phillips curve relationship differs depending on whether inflation is demand- or supply-driven.*

3.1 Introduction

The global economy experienced a historic surge in inflation beginning in 2021. Inflation in advanced economies rose to levels not seen since the 1970s, while emerging market economies saw multi-decade highs (Figure 3.1a). This rise in inflation has been attributed to both demand- and supply-side factors (e.g. Furman, 2022; Di Giovanni et al., 2022; Carrière-Swallow et al., 2023). Demand pressures on inflation

Figure 3.1: Post-pandemic inflation and its potential drivers

(a) year-on-year PCE inflation



Notes: Europe includes GBR, DEU, FRA, ITA, SWE, DNK. Asia includes AUS, NZL, THA, PHL, KOR, IDN. Median across countries is displayed.

arose, in part, from accommodative monetary policies and fiscal stimulus measures (Figure 3.1b). Supply-side explanations have included supply chain disruptions and increases in commodity prices (Figure 3.1c).

Policymakers need information on the relative contributions of demand and supply factors to inflation in order to design an appropriate policy response. For example, an established view in monetary policy is that central banks should 'look through' inflation if it is caused by temporary supply shocks. To improve policy monitoring, it is valuable to have a tool that can be updated frequently and compared across countries. The aim of this paper is to address these needs.

We decompose inflation into demand- and supply-driven components across 32 countries, including both advanced economies and emerging market economies, over the last three decades. The analysis uses a methodology by Shapiro (2026), who applied it to study inflation in the US. We extend the implementation and validation of the Shapiro (2026) methodology to a multi-country setting, and use it in applications across economies.

The methodology uses disaggregated personal consumption expenditure (PCE) data. The decompositions are based on bivariate vector autoregressive (VAR) equations for prices and quantities fit at item level. An item of the PCE basket is classified as demand (supply) 'affected' if residuals from the VAR have the same (opposite) signs. The 'demand-driven' component of headline PCE inflation for a given country is then constructed as a weighted sum of all demand-affected items in that period. The methodology is motivated by the idea that demand shocks should move prices and quantities in the same direction, whereas supply shocks should move them in opposite directions.

We validate the decomposed inflation series by examining their relationship to external measures of demand and supply shocks, extending the validation by Shapiro (2026) for the US to a panel of countries. We confirm that a country-level monetary policy tightening shock (Deb et al., 2023) leads to a persistent decline in demand-driven prices and a muted and statistically insignificant effect on supply-driven prices. Meanwhile, supply chain pressures (Benigno et al., 2022) and oil price shocks (Baumeister and Hamilton, 2019) result in a substantial increase in supply-driven prices and a more limited and transitory effect on demand-driven prices.

We use the decomposed inflation series in two applications. As a first application, we study the roles demand- and supply-side factors played in the surge in inflation following the COVID-19 pandemic. We find that demand factors were the primary reason for the initial inflation uptick in 2021, and that supply-driven emerged later but were the main explanation for the inflation peak in 2022–23.

As a second application, we examine whether the empirical Phillips curve relationship differs between demand- and supply-driven inflation. Recent literature has documented an apparent flattening of the Phillips curve. The various explanations for this flattening include changes in monetary policy credibility (Ball and Mazumder, 2011), globalization (Auer et al., 2017; Aquilante et al., 2024), market concentration (Heise et al., 2022), and production networks (Rubbo, 2023). We take a different approach and examine whether the empirical Phillips curve relationship varies depending on the shocks driving inflation.

We find that across countries a statistically significant relationship holds between demand-driven inflation and output gaps, while the same relation is insignificant for supply-driven inflation. These results resonate with findings in Bergholt et al. (2023)

for the US, and suggest that a flattening of the Phillips curve may be explained by supply-driven inflation becoming more prominent.

The paper is structured as follows. Section 2 describes the data and methodology for inflation decompositions. Section 3 validates the decompositions using external measures of demand and supply shocks. Section 4 uses the decompositions to study inflation dynamics after 2020. Section 5 examines the differences between Phillips curves for demand- and supply-driven inflation. Section 6 concludes.

3.2 Data and methodology

We employ the methodology proposed by Shapiro (2026) to decompose inflation into demand- and supply-driven parts. The methodology uses data on prices and quantities of personal consumption expenditure (PCE) at item level. For estimation, we use an outlier robust method that automatically detects outlying observations around the COVID-19 pandemic.

3.2.1 Data

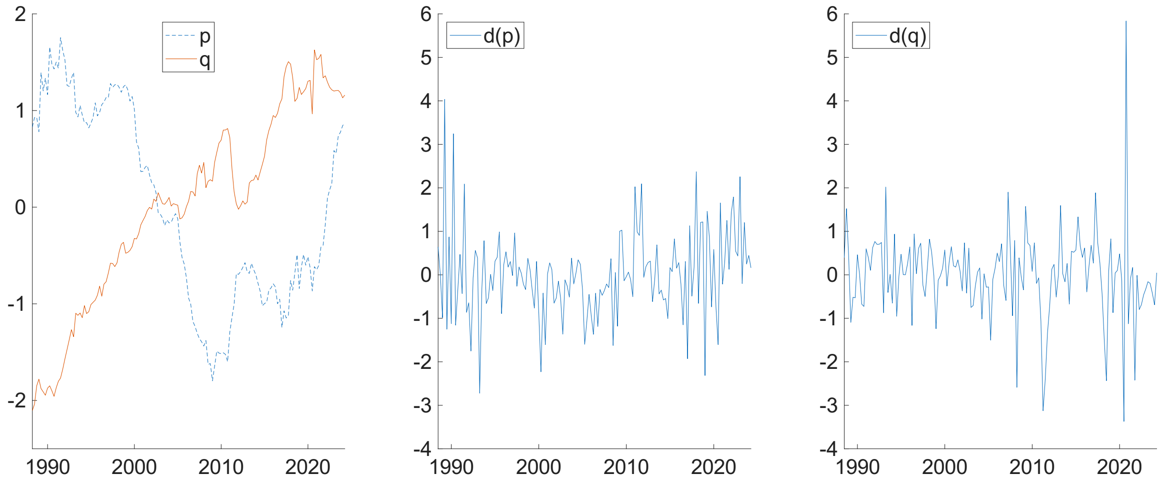
We use data on quarterly personal consumption expenditure (PCE) by expenditure item, retrieved from Haver Analytics and Eurostat. For each expenditure item, quantity is measured by the real series and price by the corresponding deflator. The data have 450 price-quantity pairs across 32 countries, including 23 advanced and 9 emerging economies. Time-coverage varies from the 1970s up to 2024. The granularity of expenditure items also differs by country, ranging from 99 items in the US to four in some European countries. Details on time and item coverage are in Table 3.A.1.

PCE data is used instead of consumer price indices (CPI) because our methodology requires both prices and quantities, and CPI subcategories do not have matching measures of quantity. Correlation between PCE and CPI inflation is generally high (Table 3.A.2). We use seasonally pre-adjusted series and construct them with X13-ARIMA-SEATS[1] when unavailable in the data source. All analyses use log-levels of the series.

Figure 3.2 shows a typical price-quantity pair in the data. Log-levels of the series are trending, while log differences look more stationary. The quantity series contains some clear outliers, most notably during the COVID-19 pandemic in 2020.

[1] Yvan Lengwiler: 'X-13 Toolbox for Matlab' (1.51), Mathworks File Exchange, 2014-2021.

Figure 3.2: Consumption expenditure in the UK on household textiles



Notes: The first panel shows centered and standardised log-levels of the series.

3.2.2 Inflation decomposition methodology

We use the methodology by Shapiro (2026) to decompose PCE inflation into demand- and supply-driven components. The equation of interest is

$$y_{cit} = \mu_{ci} + \sum_{k=1}^{K} B_{cik} y_{ci,t-k} + u_{cit}, \tag{3.1}$$

where $y_{cit} = (p_{cit}, q_{cit})'$ has price p_{cit} and quantity q_{cit} of country c, expenditure item i, and period t. After fitting (3.1), an expenditure item is categorised as 'demand affected' (supply affected) if residuals $\hat{u}^p_{cit}, \hat{u}^q_{cit}$ have the same (different) signs. The set of demand-affected items D_{ct} and supply-affected items S_{ct} are thus

$$D_{ct} = \{i : \hat{u}^p_{cit}\hat{u}^q_{cit} \geq 0\},\ S_{ct} = \{i : \hat{u}^p_{cit}\hat{u}^q_{cit} < 0\}. \tag{3.2}$$

Define headline PCE inflation as $\pi_{ct} = \sum_{i=1}^{I_c} \omega_{cit}\pi_{cit}$, where $\pi_{cit} = p_{cit} - p_{ci,t-1}$, I_c is the number of items in country c, and ω_{cit} is the year-to-date nominal expenditure share of item i in quarter $t-1$. Discrepancies between π_{ct} and the rates reported by statistical agencies may arise from the log approximation, differences in item weights, and aggregation issues introduced by the series being seasonally adjusted individually. In our data, π_{ct} is highly correlated with the rates reported by statistical agencies (Table 3.A.2).

Using (3.2), the headline rate can be decomposed as $\pi_{ct} = \pi^d_{ct} + \pi^s_{ct}$, where $\pi^d_{ct} =$

$\sum_{i=1}^{I_c} \mathbf{1}_{i \in D_{ct}} \omega_{cit} \pi_{cit}$ is the 'demand-driven' and $\pi_{ct}^s = \sum_{i=1}^{I_c} \mathbf{1}_{i \in S_{ct}} \omega_{cit} \pi_{cit}$ the 'supply-driven' component of inflation. Table 3.A.3 shows cumulative expenditure shares of the four largest expenditure items by country. Inflation decompositions may be sensitive to the classification of a small number of items when these shares are large.

The classification in (3.2) is binary: an item in a given period is classified as either demand- or supply-affected. An item classified as demand (supply) affected indicates the presence of *at least* one demand (supply) shock (Shapiro, 2026). This differs from structural vector autoregressive (SVAR) models, which aim to distinguish among multiple shocks occurring within the same period.

Shapiro (2026) notes that certain item-level characteristics, such as relative elasticities of demand and supply, may bias the classification of an item towards either demand or supply. If these biases are time-invarying, *changes* in the classification of an item may still carry useful information about the underlying composition of shocks. Our validation (Section 3.3) shows that the supply- and demand-driven components respond to external measures of shocks in line with economic theory.

3.2.3 Estimation

We implement unit root and cointegration tests. The tests are calibrated to a 5% significance level without adjustment for multiple testing, so a positive share rejections is expected even if all null hypotheses were true. The null of a unit root in levels is rejected against an alternative with a linear trend in 20% of the 900 PCE series. With a few exceptions, stationarity in first differences is accepted. Evidence of cointegration is found in 25% of the 450 price–quantity pairs. Overall, the diagnostics point to heterogeneity in integration orders and cointegration. Details on the tests and country-level results are in Appendix 3.B.2.

We estimate (3.1) in vector error correction (VEC) form $\Delta y_{cit} = \mu_{ci} + \Pi_{ci} y_{ci,t-1} + \sum_{k=1}^{K} \Gamma_{cik} \Delta y_{ci,t-k} + u_{cit}$. Since our interest is in the innovations, we remain agnostic about integration and cointegration and estimate the VEC without restrictions on Π. When estimated by OLS, residuals from the VEC coincide with those from (3.1) estimated in levels, which is the approach taken by Shapiro (2026). For each quantity-price pair, we compute OLS estimators with lag lengths $1, ..., 8$ in the subsample ending in 2019Q4, and select K as the minimiser of the Akaike information criterion.

Full sample OLS estimation displays signs of misspecification (Table 3.B.2). Normality of the residuals is rejected in most price-quantity pairs. Visual inspection of the residuals aligns with Figure 3.2, pointing to outliers particularly around the COVID-19 pandemic.

Baseline. Given the sensitivity of least squares to individual observations, we employ an outlier robust estimator. The estimator is inspired by Impulse Indicator Saturation (IIS) (Hendry et al., 2008) and automatically detects the timing of outliers for each series, which is essential given that the large number of series makes manual handling of the outliers infeasible. Properties of IIS are discussed in Johansen and Nielsen (2016) for clean data and in Chapter 2 of this thesis for contaminated data.

The IIS algorithm first fits the VEC equations by OLS using data up to 2019Q4, thereby excluding extreme COVID-19 observations. Based on this initial fit, we classify observations in the whole sample as 'clean' or 'outliers' according to the size of their residuals. We then re-estimate OLS using only the 'clean' observations and update the classification using the new residuals. Our final estimator is the OLS estimator using the observations classified as 'clean' in the second step. Details on the algorithm are in Appendix 3.B.2.

The procedure relies critically on the initial OLS estimator using data up to 2019Q4. If this sample contains bad leverage points, the initial estimator may be distorted, leading to misclassification of outliers in subsequent steps. Two factors mitigate this concern. First, the most extreme observations occur during the pandemic period, which is excluded from the initial estimation. Second, we iterate the detection of outliers, which may reduce influence of the initial estimator, although this claim has not been established formally. Specification tests indicate that the procedure succeeds at detecting outlying observations (Table 3.B.2).

Robustness checks. As alternative specifications, we implement estimation of the VEC equations by *i*) full sample OLS, *ii*) OLS in the subsample ending in 2019Q4, *iii*) IIS with lag length 8, and *iv*) 10-year rolling-window OLS with residuals computed as one-step-ahead forecast errors. Table 3.B.3 reports correlations between baseline and alternative decompositions by country. The average correlation across countries exceeds 0.9 for most specifications.

To address concerns about estimation errors and sensitivity to small innovations, we adopt an approach used by Shapiro (2026). Namely, we classify items where either the price or quantity residual is smaller than 0.1 standard deviations as ambiguous, and report robustness checks where only the non-ambiguously classified items are included in the demand- and supply-driven components.

3.3 Validation with external demand and supply shocks

We validate the inflation decompositions by assessing their relationship to external measures of demand and supply shocks. On the demand side, we look at the effects of monetary policy shocks. On the supply side, we examine the effects of supply chain pressure shocks and oil shocks. We find that the responses of decomposed inflation series to external shocks align with standard economic theory. Namely, demand-driven inflation is more responsive to a monetary policy shock, while supply-driven inflation is more responsive to supply chain pressures and oil shocks.

3.3.1 Monetary policy shocks

Macroeconomic theory posits that monetary policy controls inflation mainly through the demand-side (Smets and Wouters, 2003; Christiano et al., 2005). We check whether this prediction is confirmed by the inflation decompositions. We use country-level monetary policy (MP) shocks from Deb et al. (2023), who construct the shocks using the C. Romer and D. Romer (2004) methodology. MP shocks are matched for 22 countries over the period 1990Q1-2019Q4.

We estimate impulse response functions of demand- and supply-driven inflation to MP shocks using local projections (Jordà, 2005). We fit equations

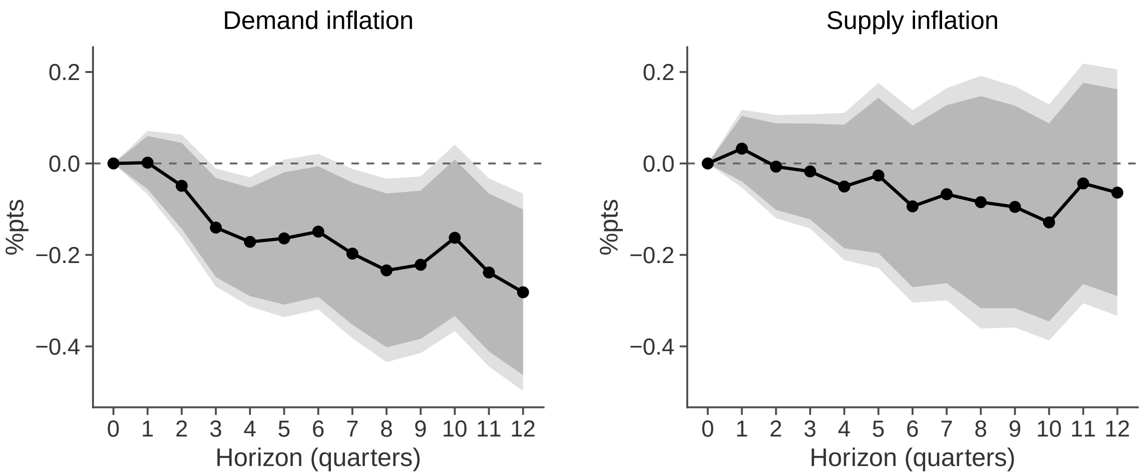
$$\pi^{j}_{c,t+h:t} = \beta^{jh} Shock^{MP}_{ct} + z'_{ct}\theta^{jh} + \alpha^{jh}_{c} + \gamma^{jh}_{t} + \varepsilon^{jh}_{ct} \tag{3.3}$$

where $\pi^{j}_{c,t+h:t} = \sum_{s=1}^{h} \pi^{j}_{c,t+s}$ for $j \in \{demand, supply\}$ is interpreted as the cumulative price change from period t to $t+h$ due to demand- or supply factors. The parameters of interest are β^{j}_{h} which capture the effect at horizon h. The vector z_{ct} includes four lags of the dependent variable and shocks. Country fixed effects control for characteristics such as monetary policy credibility, financial development, and exchange rate regime. Time fixed effects account for common time-varying developments such as global financial conditions and oil price fluctuations.

Figure 3.3 shows responses of demand-driven (left) and supply-driven (right) inflation to a 100 basis point MP tightening shock. The parameters are estimated by OLS and shaded areas correspond to pointwise confidence intervals with 90% (in dark) and

95% (in light) coverage computed with standard errors clustered by country. Following a MP shock, demand-driven inflation declines gradually over a three-year period. The effects are persistent and significant across most horizons. Effects on supply-driven inflation are more muted, less persistent, and not significant over a 3-year horizon. These results align with standard economic theory, and imply that effectiveness of monetary policy is more limited when inflation is mainly supply driven. The results are robust to controlling for nominal effective exchange rate (exchange rate pass-through), cyclically adjusted primary balance (fiscal policy) and real GDP growth (see Figure 3.C.4).

Figure 3.3: Monetary policy transmission to demand- and supply-driven inflation



3.3.2 Supply shocks

We estimate (3.3) with the Global Supply Chain Pressures Index (GSCPI) (Benigno et al., 2022) as the shock variable on the right-hand-side. The GSCPI is constructed from a factor model of 27 supply chain variables and is not a structurally identified shock. The results can thus be viewed as a reduced-form exercise. As GSCPI is common across countries, we drop time fixed effects from (3.3), but include controls for cyclically-adjusted primary balance, nominal effective exchange rate, and real GDP growth.

Figure 3.4 displays the responses of demand-driven (left) and supply-driven (right) inflation to an adverse one standard deviation GSCPI shock. The shock is associated with a significant and persistent increase in supply-driven inflation, while the effects on demand-driven inflation are muted and not significant across most horizons.

Figure 3.4: GSPCI transmission to demand and supply-driven inflation

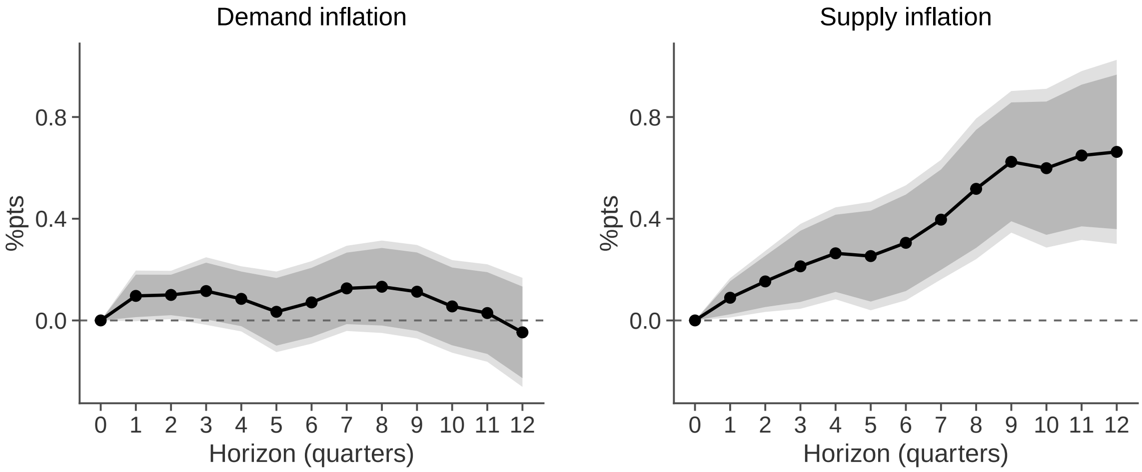


Table 3.C.5 in Appendix 3.C shows results from an additional supply-side validation exercise using oil supply shocks from Baumeister and Hamilton (2019). Both demand- and supply driven inflation rise in response to a negative oil supply shock. The effect on supply driven inflation is persistent while the cumulative price change through demand factors reverts back to zero at higher horizons.

3.4 Application 1: Inflation dynamics after 2020

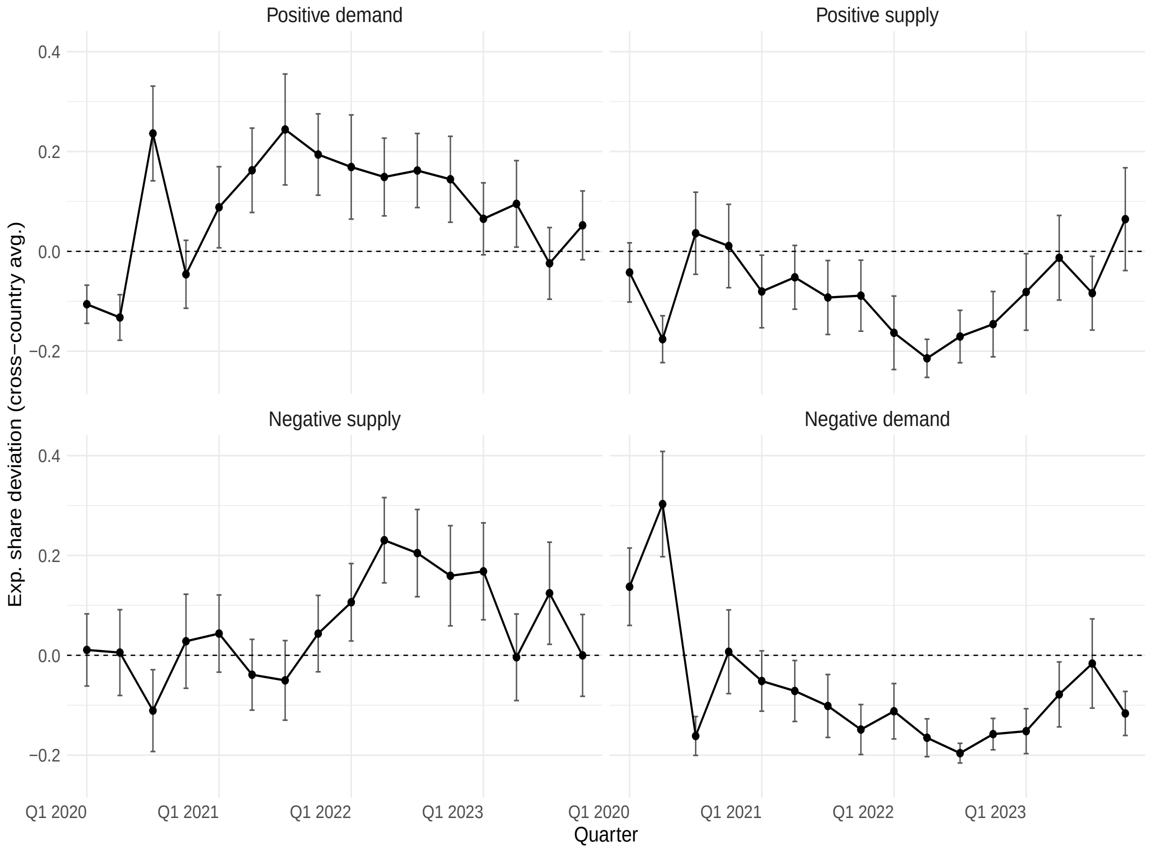
Timing of the shocks. We examine how the share of items affected by demand and supply shocks behaved around the inflationary episode starting in 2020.

We refine the classification of items in (3.2) further by sign, and define $D^+_{ct} = \{i \in D_{ct} : \hat{u}^q_{cit} \geq 0\}$, $D^-_{ct} = \{i \in D_{ct} : \hat{u}^q_{cit} < 0\}$ as items affected by positive and negative demand shocks, and $S^+_{ct} = \{i \in S_{ct} : \hat{u}^q_{cit} \geq 0\}$, $\quad S^-_{ct} = \{i \in S_{ct} : \hat{u}^q_{cit} < 0\}$ as those affected by signed supply shocks.

We fit equations $\gamma_{ct} = \sum_{s \in \mathcal{S}} \theta_s \mathbf{1}\{t = s\} + \alpha_c + \varepsilon_{ct}$. The dependent variable measures the share of items affected by a given signed shock, defined as $\gamma_{ct} = \sum_{i=1}^{\mathcal{I}_c} \omega_{ict} \mathbf{1}\{i \in A\}$, where ω_{ict} are expenditure weights and A is one of $D^+_{ct}, D^-_{ct}, S^+_{ct}, S^-_{ct}$. We include dummies for periods $\mathcal{S}$ corresponding to the 16 quarters from 2020Q1 to 2023Q4. Country fixed effects isolate changes relative to a long-run mean, which can account for time-invarying biases in the classification of items (see Section 3.2.2).

The coefficient θ_s captures the average deviation across countries in the share of items experiencing a given shock relative to country the mean. The parameters are

Figure 3.5: Changes in the share of shocks after 2020

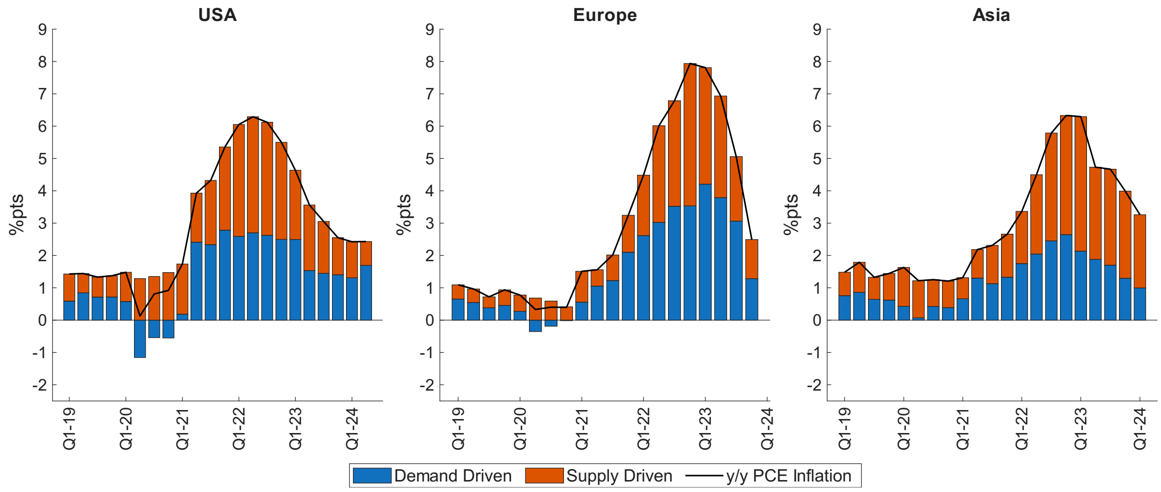


estimated by OLS for each of the four dependent variables using observations from 2010Q1 to the latest available data.

Figure 3.5 shows estimates of θ_s across the four dependent variables with 95% confidence intervals allowing for heteroskedasticity by country. The standard errors do not account for estimation errors in the VEC stage, but the results are robust to the removal of ambiguously classified items (Figure 3.C.1).

The start of the COVID-19 pandemic and the associated lockdowns coincide with an increased prevalence of large negative demand shocks in 2020:Q1-Q2, as seen in the bottom right panel of Figure 3.5. The short lifting of lockdowns in Summer 2020 shows up as an increased prevalence of positive demand in 2020Q3. The global surge in inflation starts in 2021. The results in Figure 3.5 suggest that the initial inflationary push came mainly from positive demand shocks, which reached their peak in mid-2021. The share of supply shocks only significantly exceeds its long-run level in 2022Q1, coinciding with rise in commodity prices after the military escalation in Ukraine.

Figure 3.6: Decomposition of post-2020 inflation



Notes: 'Europe' shows the median of DEU, DNK, FRA, GBR, ITA, and SWE. 'Asia' shows the median of AUS, IDN, KOR, NZL, PHL, THA.

Decomposition of inflation after 2020. Figure 3.6 shows decompositions of year-on-year headline PCE inflation to demand- and supply-driven components for 2019Q1-2024Q1. The qualitative features are robust to excluding ambiguously classified items (Figure 3.C.2).

Figure 3.6 confirms the findings in Figure 3.5 and gives some additional regional insights. Negative demand shocks associated with the COVID-19 lockdowns pushed prices downwards in 2020 while supply-driven inflation remained largely unaffected. The initial inflationary push starting in 2021 was mainly associated with a resurgence in demand-driven inflation. In the US, the pick-up in demand-driven inflation is particularly quick, corresponding to a stronger fiscal support through cash handouts and a swifter labor market recovery (Bernanke and Blanchard, 2025, Ball et al., 2022).

Supply-side inflation plays a central role in generating the peak of inflation across the regions. In the US, supply-side inflation accelerates in late 2021 while demand-driven inflation remains stable, leading to a peak in early 2022. Meanwhile, in Europe and Asia, supply-side inflation emerges later starting from 2022Q1. These findings align with a narrative that US supply-side inflation came more from supply chain distortions and domestic economic conditions, while in Europe and Asia supply-side inflation was predominantly driven by commodity prices.

3.5 Application 2: Phillips curve

The standard New Keynesian Phillips curve suggests a positive relationship between inflation (π_t) and output gap ($\hat{y}_t$) after controlling for inflation expectations ($E_t\pi_{t+1}$) and cost-push shocks (u_t), given by

$$\pi_t = \kappa\hat{y}_t + \beta E_t\pi_{t+1} + u_t, \tag{3.4}$$

where κ is the theoretical Phillips curve coefficient. Since supply shocks move prices and output in opposite directions, they can bias estimates of the Phillips curve coefficient downwards. In practice, controlling for supply shocks in a reduced-form specification is difficult because many supply-side disturbances are unobserved. We address this challenge by estimating separate Phillips curves for demand-driven and supply-driven inflation.

We estimate hybrid Phillips curves

$$\pi^j_{ct} = \kappa^j\hat{y}_{ct} + \beta^j\pi^E_{ct} + \theta^j\Delta\pi^m_{ct} + \Sigma^4_{k=1}\gamma^j_k\pi^j_{c,t-k} + \alpha^j_c + \delta^j_t + \epsilon^j_{ct}, \tag{3.5}$$

where π^j_{ct} for $j \in \{demand, supply, aggregate\}$ is either demand-driven, supply-driven, or aggregate year-on-year (yoy) inflation in percentage points. Output gap ($\hat{y}_{ct}$) is measured by the percentage point deviation of country i's real GDP from its HP-filtered trend. We control for one-year-ahead inflation expectations (π^E_{ct}) measured by Consensus Forecasts and yoy changes in import prices ($\Delta\pi^m_{ct}$). We include country and time fixed effects as well as four lags of the dependent variable.

Columns 2-4 of Table 3.1 show estimates of selected parameters from (3.5) for the three dependent variables, with standard errors clustered by country in parentheses. Demand-driven inflation has a positive and statistically significant relationship with output gap. By contrast, the Phillips curve coefficient for supply-driven inflation is negative and not statistically significant. For aggregate inflation, the relationship with output gap is weaker than for demand-driven inflation and not statistically significant. The results are robust to controlling for changes in input costs (Δppi_{ct}), measured by trade-weighted partner producer price indices (columns 5-7).

The results in Table 3.1 support the argument that supply shocks generate a downward bias in the Phillips curve coefficient, and that a strong Phillips curve relationship holds when supply-side disturbances are excluded from aggregate inflation.

Table 3.1: Slope of the Phillips curve with demand- vs. supply-driven inflation

	Demand	Supply	Agg.	Demand	Supply	Agg.
$\hat{y}_{ct}$	0.0361***	-0.0150	0.0262	0.0359***	-0.0151	0.0259
	(0.0110)	(0.0146)	(0.0166)	(0.0109)	(0.0145)	(0.0166)
π^E_{ct}	0.1209***	0.3287***	0.4948***	0.1172***	0.3238***	0.4845***
	(0.0194)	(0.0312)	(0.0706)	(0.0200)	(0.0299)	(0.0694)
$\Delta\pi^m_{ct}$	0.3171	0.3288	0.6248	0.3156	0.3228	0.6231
	(0.2770)	(0.2989)	(0.5282)	(0.2714)	(0.2871)	(0.5055)
Δppi_{ct}				0.0050	0.0069	0.0126**
				(0.0056)	(0.0051)	(0.0049)
Observations	2,366	2,366	2,366	2,366	2,366	2,366
Countries	27	27	27	27	27	27

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A potential implication of these results is that the empirical Phillips curve relationship would weaken during periods where supply-side shocks are more pronounced.

We have assessed robustness of the results to computing output gaps using the Hamilton (2018) regression filter (Table 3.C.1) and to using decompositions that only include non-ambiguously classified items (Table 3.C.2). We also allow the Phillips curve coefficient in (3.5) to differ across advanced and emerging economies (Table 3.C.3). The robustness checks support the presence of a significant Phillips curve relationship between demand-driven inflation and the absence of such a relationship between supply-driven inflation.

We have also examined time variation in the Phillips curve relationship by estimating (3.5) in 10-year rolling windows. The sensitivity of demand- and supply-driven inflation to economic activity is more stable than that of aggregate inflation. This is consistent with the view that the variation in the slope of aggregate Phillips curve may be a result of changes in the composition of shocks driving inflation.

3.6 Conclusion

We have implemented inflation decompositions to demand- and supply-driven components in 32 countries using the Shapiro (2026) methodology. The series are updated quarterly and suitable for policy monitoring across countries.

We validated the decompositions by looking at their relationship to external measures of demand and supply shocks. The validations align with economic theory:

demand-driven inflation reacts more strongly to monetary policy shocks, while supply-driven inflation responds more strongly to supply chain pressures and oil shocks.

In an application to studying inflation dynamics after 2020, we provided insight into the timing and relative contributions of supply and demand shocks. On average across countries, the initial uptick in inflation in 2021 is associated with an increased prevalence of demand shocks. By contrast, the peak of inflation in 2022-23 is mainly explained by increased supply-driven inflation. In the US, supply shocks emerge earlier than in Europe and Asia, suggesting differences in the underlying sources of supply shocks.

We also found evidence of a steeper and more significant Phillips curve for demand-driven inflation compared to supply-driven and aggregate inflation. These results suggest that a flattening of the Phillips curve may be explained by supply-driven inflation becoming more prominent. Policymakers may therefore have weaker control over prices when supply-side shocks are more pronounced.

Appendix to Chapter 3

3.A Appendix: Descriptive statistics

Table 3.A.1: Data Description

ISO3	Source	SA	#items	Sample	ISO3	Source	SA	#items	Sample
AUS	haver	author	26	85Q3-24Q2	KOR	haver	source	12	00Q1-24Q2
AUT	eurostat	source	4	95Q1-23Q1	LUX	haver	source	4	95Q1-24Q2
CAN	haver	source	96	92Q1-24Q2	LVA	haver	source	4	96Q1-24Q2
CYP	haver	author	4	96Q1-24Q2	MEX	haver	author	8	93Q1-23Q1
CZE	haver	source	4	95Q1-24Q2	MLT	haver	source	4	00Q1-24Q1
DEU	haver	source	8	95Q1-24Q1	NLD	haver	author	4	00Q1-24Q3
DNK	haver	author	11	91Q1-24Q1	NOR	haver	source	4	95Q1-24Q2
EST	haver	source	4	95Q1-24Q2	NZL	haver	source	10	87Q2-24Q1
FIN	haver	source	4	90Q1-24Q2	PHL	haver	author	12	98Q1-24Q2
FRA	haver	source	17	90Q1-24Q2	ROU	haver	author	4	95Q1-24Q2
GBR	haver	author	41	88Q1-24Q1	SVK	eurostat	source	4	95Q1-23Q1
HUN	haver	author	4	95Q1-24Q2	SWE	haver	author	9	00Q1-24Q3
IDN	haver	author	7	08Q1-24Q2	THA	haver	source	32	93Q1-24Q2
IRL	haver	author	4	95Q1-24Q2	TWN	haver	author	12	81Q1-24Q2
ITA	haver	source	12	96Q1-23Q4	USA	haver	source	99	88Q1-24Q2
JPN	haver	author	13	94Q1-23Q1	ZAF	haver	source	4	70Q1-24Q2

Notes: Columns 'SA' denote seasonal adjustment.

Table 3.A.2: Correlations between constructed PCE and aggregate CPI and PCE

ISO3	CPI	Agg.PCE	ISO3	CPI	Agg.PCE	ISO3	CPI	Agg.PCE
AUS	93.4	99.9	AUT	92.3	100	CAN	88.6	97.3
CYP	86.1	99.5	CZE	96.6	100	DEU	93.3	99.6
DNK	92.7	93.3	EST	95.9	99.9	FIN	94.7	99.7
FRA	88.7	99.8	GBR	93.8	98.5	HUN	97.5	99.6
IDN	77.1	97.8	IRL	83.4	99.5	ITA	95.9	99.9
JPN	92.4	99.9	KOR	92.6	98.5	LUX	86.6	95.3
LVA	87.7	99.7	MEX	98.8	99.8	MLT	75.5	98.1
NLD	81.3	94.6	NOR	85.7	94.7	NZL	92.0	99.8
PHL	89.7	97.5	ROU	95.3	98.2	SVK	93.4	98.6
SWE	97.1	93.6	THA	92.7	99.4	TWN	85.7	98.2
USA	95.7	99.7	ZAF	90.7	97.8			

Mean: CPI = 90.716 Agg.PCE = 98.366

Notes: Correlations in %pts between annual constructed PCE inflation and aggregate PCE and CPI inflation reported by statistical agencies.

Table 3.A.3: Cumulative expenditure share of four largest items

ISO3	1	2	3	4	ISO3	1	2	3	4
AUS	0.18	0.29	0.37	0.43	KOR	0.20	0.33	0.45	0.55
AUT	0.51	0.79	0.90	1	LUX	0.48	0.82	0.92	1
CAN	0.15	0.24	0.29	0.34	LVA	0.48	0.86	0.95	1
CYP	0.53	0.84	0.93	1	MEX	0.45	0.80	0.86	0.91
CZE	0.45	0.85	0.93	1	MLT	0.51	0.82	0.91	1
DEU	0.25	0.42	0.58	0.73	NLD	0.27	0.53	0.77	1
DNK	0.26	0.38	0.50	0.59	NOR	0.48	0.77	0.90	1
EST	0.43	0.82	0.93	1	NZL	0.27	0.41	0.55	0.66
FIN	0.50	0.82	0.91	1	PHL	0.34	0.47	0.59	0.69
FRA	0.19	0.34	0.47	0.54	ROU	0.50	0.84	0.93	1
GBR	0.18	0.27	0.34	0.40	SVK	0.44	0.85	0.93	1
HUN	0.44	0.84	0.93	1	SWE	0.15	0.27	0.39	0.50
IDN	0.39	0.62	0.75	0.85	THA	0.14	0.21	0.27	0.33
IRL	0.50	0.84	0.93	1	TWN	0.19	0.36	0.49	0.61
ITA	0.21	0.36	0.49	0.59	USA	0.12	0.19	0.26	0.31
JPN	0.24	0.39	0.49	0.58	ZAF	0.44	0.76	0.88	1

3.B IIS estimation and specification tests

3.B.1 IIS estimation

We describe an IIS algorithm for a generic equation $y_i = x_i'\beta + \sigma\varepsilon_i$ with observations $i = 1, ..., n$, scalar outcome y_i, and vector x_i. The user chooses an initial set $\mathcal{T}$, cut-off $c > 0$, and the number of iterations $m \geq 1$. The following steps are then taken.
(1) Calculate initial least squares estimators $\hat{\beta}_{(0)}, \hat{\sigma}_{(0)}$ using observations $\mathcal{T}$.
(2) For $\ell = 1, ..., m$ repeat the following steps, yielding the final estimator $\hat{\beta}_{(m)}$.
(3) Compute indicators $v_i = \mathbf{1}\{|y_i - x_i'\hat{\beta}_{(\ell-1)}| \leq \hat{\sigma}_{(\ell-1)}c\}$ for $i = 1, \ldots, n$. Classify observations with $v_i = 1$ as 'good' and the remaining as 'outliers'.
(4) Compute the OLS estimator using 'good' observations in the current step: $\hat{\beta}_{(\ell)} = \arg\min_\beta \sum_{i=1}^n v_i(y_i - x_i'\beta)^2$, $\quad \hat{h} = \sum_{i=1}^n v_i$, $\quad \hat{\sigma}^2_{(\ell)} = \hat{h}^{-1}\sum_{i=1}^n v_i(y_i - x_i'\hat{\beta}_{(\ell)})^2$.

We estimate the bivariate VECM $\Delta y_{cit} = \mu_{ci} + \Pi_{ci} y_{ci,t-1} + \sum_{k=1}^K \Gamma_{cik}\Delta y_{ci,t-k} + u_{cit}$ equation-by-equation using the above IIS algorithm with $m = 2$ iterations. The initial set $\mathcal{T}$ for each item is the subsample ending in 2019Q4. The cut-off c is chosen such that $n_c\{1 - \Phi(c)\} = 1/4$, where Φ is the standard normal cdf and n_c is the number of observations in country c. Chapter 2 of this thesis shows that this calibration of the cut-off bounds the expected share of misclassified good observations by 0.5% if the true 'good' innovations are standard normal.

3.B.2 Specification tests

Table 3.B.1 shows rejection rates of unit root and cointegration tests for the PCE series using standard Matlab (2025b) implementations. We report a test for unit roots (Phillips and Perron, 1988) in levels against an alternative of linear trend stationarity [pp(level)] and in first differences against an alternative with a constant [pp(diff)]. The column 'jcitest' shows a maximum eigenvalue test (Johansen, 1995) for the null of Π having rank 0 against an alternative of positive rank, allowing a constant and linear trends in the levels and cointegrating relation. We include 8 lags of the dependent variables, calibrate the tests to 5% size, and implement the tests in subsamples ending in 2019Q4. The subsample corresponds to the initial set used for IIS.

Table 3.B.1: Unit root and cointegration tests — Rejection rates

ISO3	pp(level)	pp(diff)	jcitest	ISO3	pp(level)	pp(diff)	jcitest
AUS	0.231	0.962	0.269	KOR	0.125	1	0.167
AUT	0.125	1	0.5	LUX	0.25	1	0
CAN	0.141	0.995	0.24	LVA	0.125	1	0
CYP	0	1	0	MEX	0.063	1	0.625
CZE	0.25	1	0	MLT	0.625	1	0.5
DEU	0.125	1	0	NLD	0.375	1	0.25
DNK	0.227	1	0.364	NOR	0.25	1	0
EST	0.375	1	0	NZL	0.3	1	0.5
FIN	0.375	1	0.25	PHL	0.125	1	0.333
FRA	0.118	0.971	0.118	ROU	0.5	1	0.25
GBR	0.159	0.988	0.366	SVK	0	1	0
HUN	0.375	1	0.25	SWE	0.222	1	0.333
IDN	0.214	0.857	0.857	THA	0.156	1	0.344
IRL	0	1	0.25	TWN	0	1	0.25
ITA	0	0.958	0.417	USA	0.081	0.975	0.131
JPN	0.154	1	0.231	ZAF	0	1	0

Cross-country average: pp(level) = 0.19, pp(diff) = 0.991, jcitest = 0.244.

Table 3.B.2 shows rejection rates of tests for autocorrelation (LB), autoregressive conditional heteroskedasticity (arch), and normality (norm) of the residuals. The LB and arch tests check dependence up to eight lags.

Implementation of the diagnostic tests with IIS requires special care. The normality test is implemented on the set of selected good observations, similar to Berenguer-Rico and Nielsen (2026). The autocorrelation test (LB) treats the outlying residuals as missing at random. The arch-test is implemented by replacing outlying residuals with zeroes. These choices for the LB and arch tests are influenced by the standard Matlab implementation. The appropriate way of testing for autocorrelation and conditional heteroskedasticity after outlier removal remains an open problem.

Table 3.B.2: Diagnostics tests — Rejection rates

	OLS (full sample)			OLS (pre-2020)			IIS		
	arch	LB	norm	arch	LB	norm	arch	LB	norm
AUS	0.327	0.154	0.788	0.231	0.077	0.596	0.192	0.173	0.058
AUT	0.75	0.625	0.875	0.125	0.25	0.375	0.125	0.375	0.125
CAN	0.427	0.135	0.807	0.255	0.036	0.573	0.188	0.109	0.083
CYP	0.25	0.25	0.875	0.375	0.125	0.875	0	0.125	0.25
CZE	0.25	0	1	0	0	0.75	0	0.125	0
DEU	0.375	0.125	0.813	0.125	0	0.438	0	0.063	0.063
DNK	0.364	0.136	0.818	0.045	0.045	0.409	0.091	0.136	0
EST	0.375	0.125	0.75	0.125	0.125	0.5	0.25	0	0
FIN	0.125	0	0.75	0.375	0	0.75	0.125	0.25	0
FRA	0.412	0.118	0.824	0.206	0.059	0.735	0.265	0.118	0.059
GBR	0.427	0.207	0.768	0.341	0.098	0.573	0.28	0.159	0.024
HUN	0.25	0.125	0.25	0	0.125	0.25	0.125	0.125	0
IDN	0.143	0.071	0.571	0.143	0.071	0.571	0.071	0.214	0.071
IRL	0.5	0.125	0.75	0.375	0.125	0.75	0.125	0.125	0
ITA	0.458	0.208	0.958	0.125	0.042	0.625	0.208	0.208	0.125
JPN	0.308	0.077	0.962	0.231	0.038	0.808	0.115	0.269	0.231
KOR	0.125	0.083	0.458	0.042	0	0.167	0.083	0.083	0.042
LUX	0.25	0	0.75	0.25	0	0.625	0	0	0
LVA	0.375	0.125	1	0.25	0	1	0.125	0.375	0.125
MEX	0.063	0.063	0.813	0.125	0.125	0.688	0	0.125	0
MLT	0.125	0	0.75	0.125	0	0.125	0.125	0	0
NLD	0.25	0.125	0.375	0	0	0.5	0	0.25	0
NOR	0.375	0.375	0.5	0	0	0.5	0.25	0	0
NZL	0.55	0.15	1	0.25	0.05	0.8	0.15	0.05	0.05
PHL	0.25	0.125	0.75	0.292	0.042	0.625	0.167	0.208	0.042
ROU	0.375	0	0.375	0.125	0	0.25	0.25	0	0
SVK	0.125	0.125	0.875	0	0	0.875	0.125	0.5	0.125
SWE	0.222	0.056	0.667	0	0	0.444	0.056	0.056	0
THA	0.422	0.031	0.875	0.344	0.016	0.75	0.281	0.172	0.125
TWN	0.458	0.125	0.625	0.458	0.083	0.5	0.417	0.167	0.083
USA	0.399	0.086	0.874	0.242	0.04	0.611	0.227	0.106	0.106
ZAF	0	0	1	0.25	0	1	0.75	0.25	0.125
Average	0.316	0.123	0.758	0.182	0.049	0.595	0.161	0.154	0.06

Table 3.B.3: Correlations between baseline and alternative decompositions

(a) Demand-driven inflation (year-on-year)

ISO3	(1)	(2)	(3)	(4)	ISO3	(1)	(2)	(3)	(4)
AUS	93.2	88.2	73.3	94.1	KOR	94.9	93.5	85.9	93.1
AUT	98.7	82.3	88.8	91.8	LUX	76.5	65.7	68.1	59.7
CAN	96.3	86	88.7	95.6	LVA	95.5	96.6	82.5	97.1
CYP	94.5	89.2	90.9	99.4	MEX	74.1	86.1	49.3	84.2
CZE	81.4	78.8	64.7	77.3	MLT	85.3	32.9	-0.7	-7.7
DEU	95.8	93.4	87.3	97.8	NLD	96.9	98.2	96.9	97.3
DNK	96.5	96	74.7	68.6	NOR	91.5	68.4	27.9	78.9
EST	91	86.6	58.8	89.3	NZL	91.4	93.3	81	96.8
FIN	84	86.7	68.6	88.7	PHL	91.9	92.5	93.9	92.3
FRA	95.1	94.6	87.7	96.8	ROU	89.6	90.4	85.8	75.7
GBR	95.1	95.4	95.8	96.8	SVK	97.7	94.9	89.5	94.3
HUN	92.8	88.5	62.2	93.4	SWE	97.6	93.4	91.8	94.9
IDN	75.3	64.1	69.4	63.7	THA	92.6	97.2	87.2	91.5
IRL	94.6	96.3	92.2	95.2	TWN	98.1	98.1	88.9	96
ITA	98.8	94.9	96.5	97	USA	98.7	94.1	94.7	91
JPN	98.6	98.5	96.3	97.4	ZAF	77.8	71.9	55.9	90.1
Mean: (1) = 91.6, (2) = 87.1, (3) = 77.3, (4) = 86.5									

(b) Supply-driven inflation (year-on-year)

ISO3	(1)	(2)	(3)	(4)	ISO3	(1)	(2)	(3)	(4)
AUS	97.9	96.8	81.4	97.9	KOR	96.6	95.7	82.8	95.3
AUT	98.5	74.8	80.6	91.6	LUX	83.9	63.5	77.6	63.8
CAN	97.2	89.4	93.6	96.1	LVA	97.6	97.8	94	98
CYP	88.1	67.6	78	98.8	MEX	93.3	86.2	13.3	93.7
CZE	98	94.1	81.8	97.2	MLT	90.6	42.9	60.5	51.9
DEU	98.3	98	96.4	99.3	NLD	87.9	96	94.4	93.7
DNK	96.6	96.6	89.3	85.5	NOR	96.5	88.4	49.4	90.9
EST	97.7	96.6	58.7	96.2	NZL	95.3	96.4	90.3	98.2
FIN	91.2	92.9	79.6	94.4	PHL	96.8	96.9	96.6	96.9
FRA	97.2	95.1	90.6	97.7	ROU	97	97.4	85.6	88.3
GBR	97.3	97.3	93.7	98.3	SVK	98.2	96.5	88.8	95.8
HUN	97.4	96.4	88.4	97	SWE	94.2	91.8	91.1	91.6
IDN	85.5	78	73.2	78.2	THA	94.6	98.1	87.1	90.4
IRL	91.2	95.4	89.4	92.7	TWN	98.4	98.6	92.9	96.4
ITA	96.6	86.6	77.4	93.3	USA	99.4	96.8	96.9	95.7
JPN	98.8	98.6	96.2	97.6	ZAF	95.4	93.2	80.2	97.5
Mean: (1) = 95.1, (2) = 90.3, (3) = 82.2, (4) = 92.2									

Notes: Columns correspond to (1): OLS (pre-2020), (2): OLS (full sample), (3): Rolling window (80 quarters), (4): IIS (8 lags). Correlations multiplied by 100.

3.C Additional results and robustness checks

Figure 3.C.1: Share of shocks by type during the pandemic – Ambiguous shocks excluded

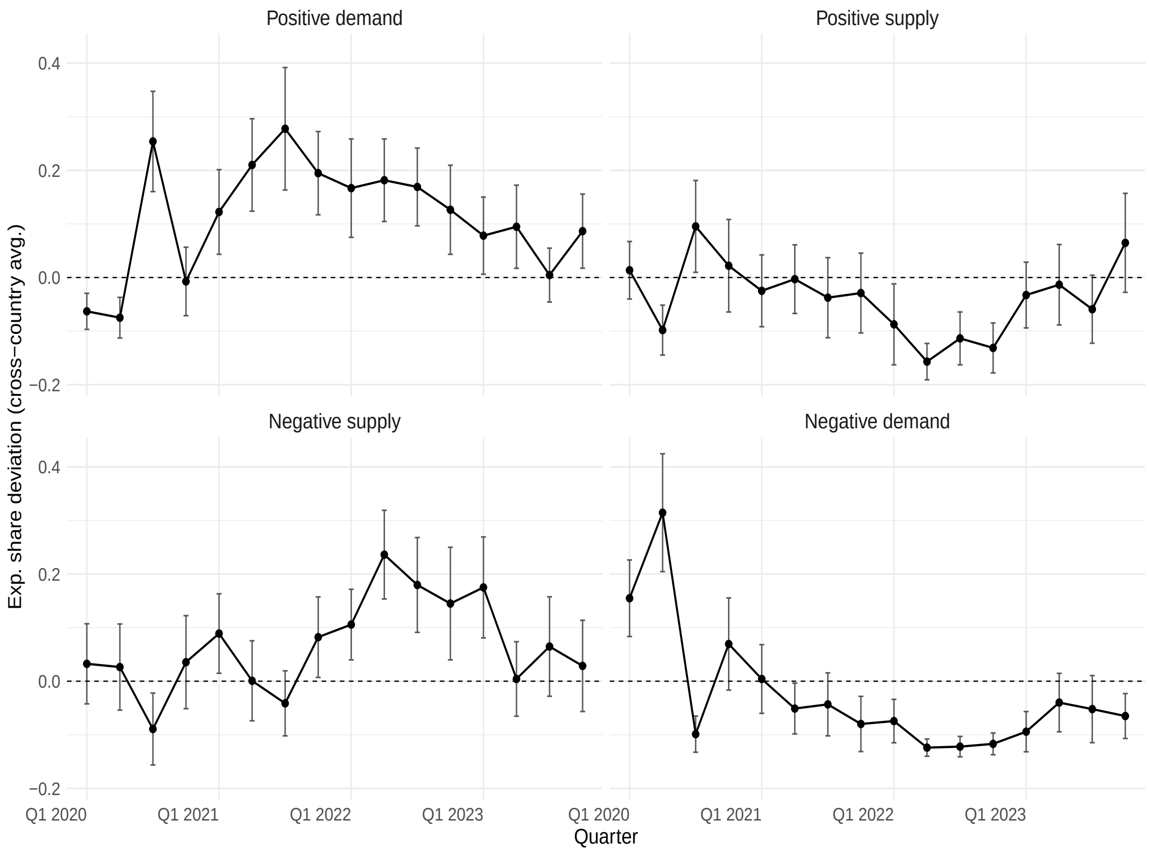


Figure 3.C.2: Decomposition of post-2020 inflation allowing for ambiguous shocks

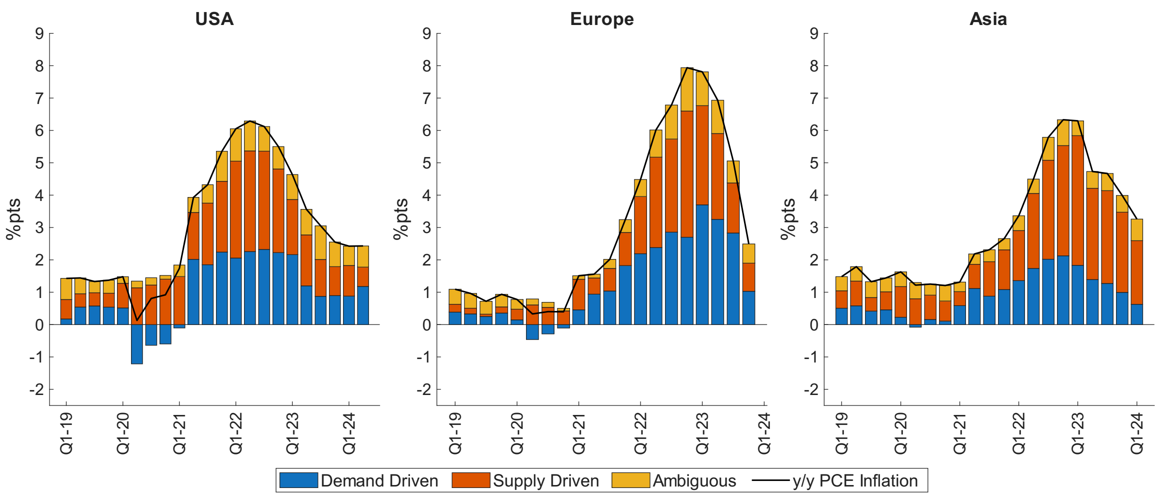


Table 3.C.1: Phillips curves — Output gaps computed with Hamilton regression filter

	Demand	Supply	Agg.	Demand	Supply	Agg.
$\hat{y}_{ct}$	0.0202***	0.0026	0.0269***	0.0201***	0.0024	0.0265***
	(0.0036)	(0.0069)	(0.0058)	(0.0035)	(0.0070)	(0.0059)
π^E_{ct}	0.1225***	0.3238***	0.5001***	0.1192***	0.3190***	0.4903***
	(0.0182)	(0.0317)	(0.0643)	(0.0192)	(0.0304)	(0.0631)
$\Delta\pi^m_{ct}$	0.2565	0.3936	0.6182	0.2552	0.3874	0.6165
	(0.2707)	(0.2948)	(0.5261)	(0.2650)	(0.2834)	(0.5036)
Δppi_{ct}				0.0044	0.0068	0.0120**
				(0.0054)	(0.0050)	(0.0045)
Observations	2,366	2,366	2,366	2,366	2,366	2,366
Countries	27	27	27	27	27	27

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.C.2: Phillips curves — Ambiguous items excluded

	Demand	Supply	Agg.	Demand	Supply	Agg.
$\hat{y}_{ct}$	0.0304***	-0.0255*	0.0262	0.0302***	-0.0260*	0.0259
	(0.0097)	(0.0139)	(0.0166)	(0.0095)	(0.0138)	(0.0166)
π^E_{ct}	0.0582***	0.2068***	0.4948***	0.0533***	0.2009***	0.4845***
	(0.0173)	(0.0334)	(0.0706)	(0.0172)	(0.0311)	(0.0694)
$\Delta\pi^m_{ct}$	0.4468	0.4778	0.6248	0.4427	0.4646	0.6231
	(0.3052)	(0.2980)	(0.5282)	(0.2977)	(0.2798)	(0.5055)
Δppi_{ct}				0.0076	0.0112**	0.0126**
				(0.0049)	(0.0050)	(0.0049)
Observations	2,366	2,366	2,366	2,366	2,366	2,366
Countries	27	27	27	27	27	27

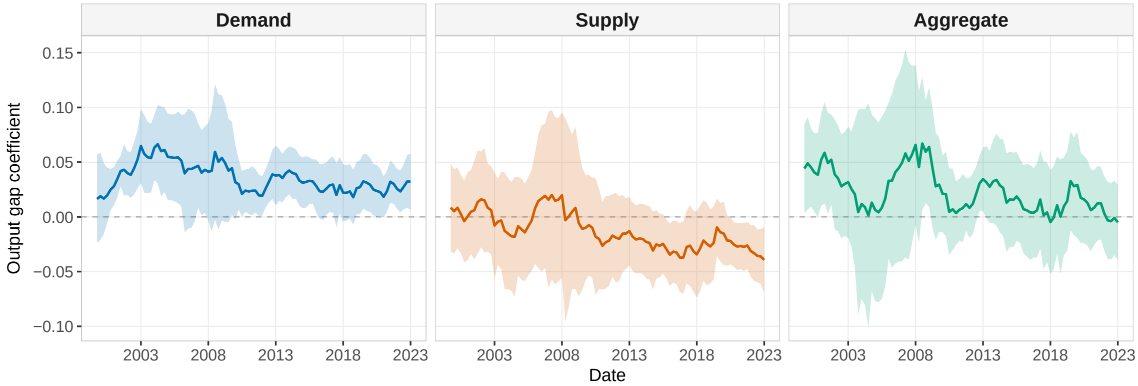
Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.C.3: Phillips Curve coefficients — Advanced vs. Emerging Economies

	Advanced Economies			Emerging Economies		
	Demand	Supply	Agg.	Demand	Supply	Agg.
PC slope	0.0162***	0.0064	0.0262***	0.0282***	−0.0060	0.0271**
	(0.0036)	(0.0059)	(0.0078)	(0.0044)	(0.0145)	(0.0111)

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Country classifications into Advanced and Emerging corresponds to 2023 IMF World Economic Outlook.

Figure 3.C.3: Slope of the Phillips Curve — 40-quarter rolling window estimation



Notes: Date on x-axis corresponds to end of window. Confidence intervals are at 90%.

Figure 3.C.4: Monetary policy transmission – Robustness to additional controls

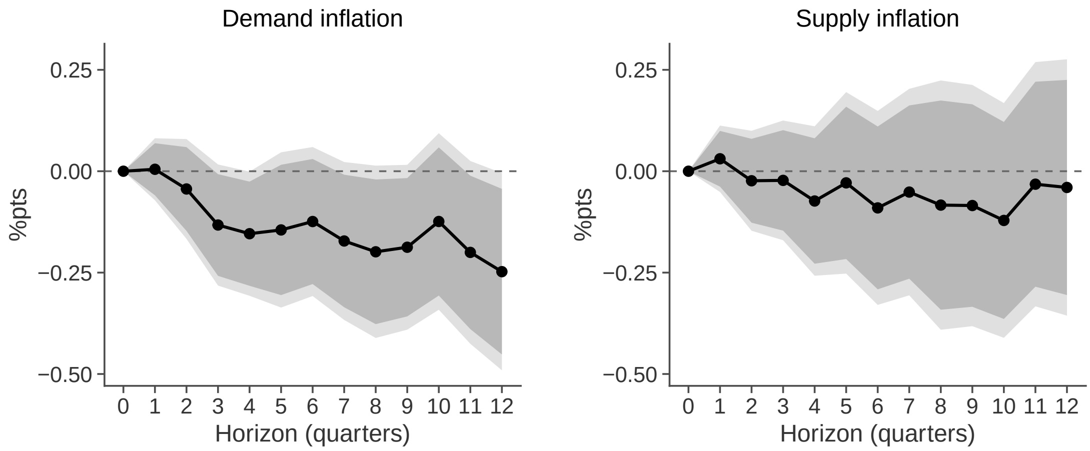
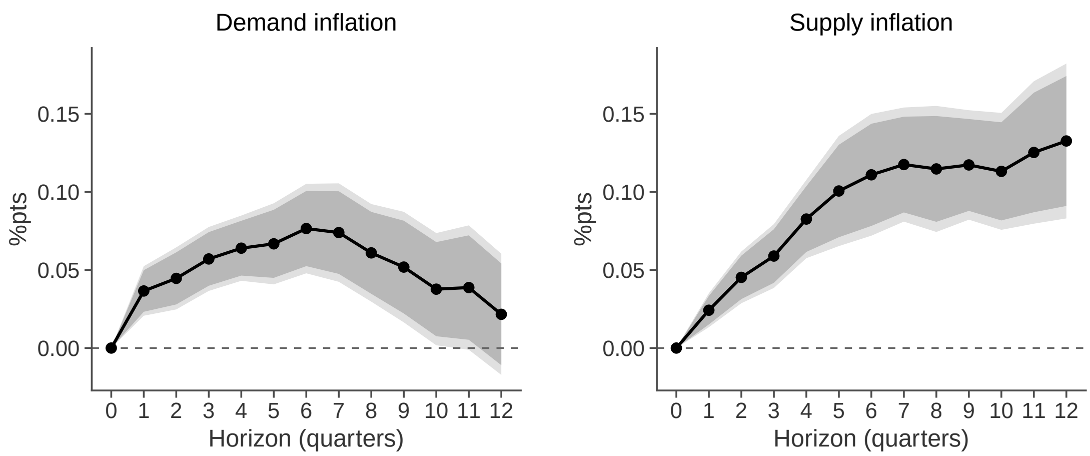


Figure 3.C.5: Oil supply shock transmission to demand and supply-driven inflation



References for Chapter 3

Aquilante, T., A. Dogan, M. Firat, and A. Soenarjo (2024). Global value chains and the dynamics of UK inflation. Staff Working Paper No. 1,060. Bank of England.

Auer, R., C. E. Borio, and A. J. Filardo (2017). The globalisation of inflation: the growing importance of global value chains. BIS Working Paper No 60. Bank for International Settlements.

Ball, L. M., D. Leigh, and P. Mishra (2022). Understanding us inflation during the covid era. NBER Working Paper 30613. National Bureau of Economic Research.

Ball, L. M. and S. Mazumder (2011). Inflation dynamics and the great recession. NBER Working Paper 17044. National Bureau of Economic Research.

Baumeister, C. and J. D. Hamilton (2019). Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *American Economic Review* 109(5), 1873–1910.

Benigno, G., J. Di Giovanni, J. J. Groen, and A. I. Noble (2022). The GSCPI: A new barometer of global supply chain pressures. Federal Reserve Bank of New York Staff Report 1017.

Berenguer-Rico, V. and B. Nielsen (2026). Normality testing after outlier removal. *Econometrics and Statistics* 38, 74–96.

Bergholt, D., F. Furlanetto, and E. Vaccaro-Grange (2023). Did monetary policy kill the Phillips curve? Some simple arithmetics. Working Paper 2/2023. Norges Bank.

Bernanke, B. S. and O. Blanchard (2025). What caused the US pandemic-era inflation? *American Economic Journal: Macroeconomics* 17(3).

Carrière-Swallow, Y., P. Deb, D. Furceri, D. Jiménez, and J. D. Ostry (2023). Shipping costs and inflation. *Journal of International Money and Finance* 130, 102–771.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113(1), 1–45.

Deb, P., J. Estefania-Flores, M. Firat, D. Furceri, and S. Kothari (2023). Monetary Policy Transmission Heterogeneity: Cross-Country Evidence. WP 23/204. International Monetary Fund.

Di Giovanni, J., S. Kalemli-Özcan, A. Silva, and M. A. Yildirim (2022). Global supply chain pressures, international trade, and inflation. NBER Working Paper 30240. National Bureau of Economic Research.

Furman, J. (Apr. 20, 2022). This inflation Is demand-driven and persistent. *Project Syndicate*. URL: `https://www.project-syndicate.org/commentary/inflation-demand-driven-persistent-by-jason-furman-2022-04` (visited on 09/12/2023).

Hamilton, J. D. (2018). Why you should never use the Hodrick-Prescott filter. *Review of Economics and Statistics* 100(5), 831–843.

Heise, S., F. Karahan, and A. Şahin (2022). The missing inflation puzzle: The role of the wage-price pass-through. *Journal of Money, Credit and Banking* 54(S1), 7–51.

Hendry, D. F., S. Johansen, and C. Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics* 23, 317–335.

Johansen, S. (1995). Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Advanced texts in econometrics. Oxford, UK: Oxford University Press.

Johansen, S. and B. Nielsen (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics* 43(2), 321–348.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review* 95(1), 161–182.

Phillips, P. C. and P. Perron (1988). Testing for a unit root in time series regression. *Biometrika* 75(2), 335–346.

Romer, C. and D. Romer (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review* 94(4), 1055–1084.

Rubbo, E. (2023). Networks, Phillips curves, and monetary policy. *Econometrica* 91(4), 1417–1455.

Shapiro, A. H. (2026). Decomposing supply and demand driven inflation. *Journal of Money, Credit and Banking* 58(2), 365–388.

Smets, F. and R. Wouters (2003). An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association* 1(5), 1123–1175.

Conclusion

This thesis examined the properties of outlier-robust statistical methods under contamination.

Chapter 1 studied outlier robustness of the Least Trimmed Squares estimator in linear models with categorical covariates. We found uniform boundedness guarantess and showed that LTS is robust in a wider range of settings than suggested by existing boundedness and breakdown point results. We proposed a data-driven approach to choosing an initial LTS tuning parameter.

Chapter 2 developed asymptotic theory for Impulse Indicator Saturation (IIS) under contamination. We showed the asymptotic equivalence of IIS to an infeasible least squares estimator and used this to derive the distribution of IIS in cross-sectional and time series models with outliers. We further found a limit theory for the number of misclassified 'good' observations to guide tuning parameter selection for IIS.

Chapter 3 decomposed inflation to demand- and supply-driven components in a panel of 32 countries. The estimations used IIS to guard against outliers occuring around the COVID-19 pandemic. We validated the decomposed series by examining their relationship to external measures of demand and supply shocks. The decompositions were used for applications to inflation dynamics after 2020 and Phillips curves.