Occasionally Misspecified

Jean-Jacques Forneron*

November 10, 2023

Abstract

When fitting a particular Economic model on a sample of data, the model may turn out to be heavily misspecified for some observations. This can happen because of unmodelled idiosyncratic events, such as an abrupt but short-lived change in policy. These outliers can significantly alter estimates and inferences. A robust estimation is desirable to limit their influence. For skewed data, this induces another bias which can also invalidate the estimation and inferences. This paper proposes a robust GMM estimator with a simple bias correction that does not degrade robustness significantly. The paper provides finite-sample robustness bounds, and asymptotic uniform equivalence with an oracle that discards all outliers. Consistency and asymptotic normality ensue from that result. An application to the "Price-Puzzle," which finds inflation increases when monetary policy tightens, illustrates the concerns and the method. The proposed estimator finds the intuitive result: tighter monetary policy leads to a decline in inflation.

JEL Classification: C11, C12, C13, C32, C36.
Keywords: Leveraged outliers, Structural Vector-Autoregression, Instrumental Variables.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA.
Email: jjmf@bu.edu, Website: http://jjforneron.com.
I would like to thank Richard Crump, Claire Labonne, participants at the Wisconsin Econometrics seminar, and the BC/BU Greenline workshop for useful comments and suggestions.

1 Introduction

Empirical data is routinely used to fit and test Economic models or predictions. Although the model may explain much of the variation in the data, it may also turn out to be particularly misspecified for some observations. This can result from sudden, yet temporary, changes in policy. To illustrate: monetary policy is often measured via changes in interest rates. Between 1979 and 1982, the Federal Reserve no longer fixed the Federal Funds Rate as a policy tool, targetting monetary aggregates instead (Coibion, 2012, p3). Sharp changes in interest rates during that period generate significant identifying power on the effects of monetary policy. Yet, misspecification threatens the validity of the resulting estimates and inferences. Other factors that can cause occasional misspecification include imperfect data matching, or some rare - but significant - prediction errors when generating regressors.

A robust estimator is desirable in these scenarios: being less sensitive to influential outliers. However, robust estimates can be biased and inconsistent when the underlying data is asymmetric. To illustrate: the sample median is more robust than the mean; however, it estimates a different quantity when the data is skewed. This is relevant as many economic variables – income, prices, and quantity, to name a few – tend to be skewed. When symmetric data is contaminated asymmetrically, both the mean and median are biased. Further, in a linear regression context, Hamilton (1992) stresses that robust M-estimators are "designed for protection against wild errors or y-outliers. x-outliers are its Achilles' heel." Leverage characterizes x-outliers, which is bounded for ordinary least-squares. In the example above: sharp changes in interest rates imply high leverage around 1979-1982. The issue is even more pronounced in non-linear regressions where leverage is not necessarily bounded (St Laurent and Cook, 1992). This superleverage can further exacerbate the influence of outliers.

This paper proposes a robust Generalized Method of Moments (GMM) estimator with a simple bias-correction step. Building on Ronchetti and Trojani (2001), the sample moments are estimated robustly; here using a penalized student log-likelihood criterion. The particular choice of criterion makes the asymptotic asymmetry bias tractable. A linear combination, known as Richardson extrapolation, of two robust moment estimates is asymptotically unbiased. The bias, which depends on higher-order moments, is not estimated. The correction does not degrade robustness significantly. Also, in linear regressions, robust GMM estimates are robust against x-outlier, unlike M-estimates which only screen for large residuals. Given these moment estimates, the model is estimated in the same fashion as a standard GMM.

Finite and large sample results describe the properties of the method against adversarial contamination. First, uniform finite-sample exponential bounds, for cross-sections and mixing time-series, measure how robust moment estimates deviate from their biased target. This provides a worst-case global robustness guarantee for a given level of data contamination. The combination of the student likelihood, which is neither convex nor bounded but has a bounded influence function, with the particular choice of penalty is key for this result.

The large-sample results require the number of outliers to increase more slowly than the sample size. Their influence can grow rapidly: non-robust estimates may be inconsistent, or diverge. This captures the finite-sample setting where a few observations overwhelm the estimation. The bias-corrected robust moment and parameter estimates are shown to be first-order equivalent to an oracle which discards all outliers. Asymptotic normality follows from standard regularity conditions on the oracle. For linear models, the robust GMM estimates can be expressed as weighted least-squares or weighted two-stage least-squares. The weights are easy to compute and report, highlighting which observations were downweighted in the process. This should reduce concerns about black-box results.

Simulations illustrate the small sample properties of the proposed estimator in the presence of x-outliers, which have high leverage. OLS is very sensitive. A robust M-estimator packaged in R is biased and sensitive. Without correction, the procedure is more robust but biased. Bias correction reduces estimation error and improves coverage of t-tests. As the proportion of outliers increases, its performance degrades but remains better than the benchmarks. Undersmoothing, sometimes suggested in the literature, is also less robust than bias correction. Three empirical applications illustrate the relevance of the procedure.

The first estimates the effect of a monetary policy shock on inflation using a structural Vector Autoregressive (VAR) model as in Stock and Watson (2001). OLS estimates a "Price-Puzzle:" predicting an inflation increase when monetary policy tightens. Two historical sub-periods of unusual monetary policy – including 1979-1982 – significantly influence this result. The proposed estimates find the intuitive result: a negative impact on inflation. The weights reveal that the two historical subperiods are downweighted to get this result. Robust estimates overweight some observations. Bias correction re-adjusts towards equal weighting.

Recently, Young (2022) found that many instrumental variable (IV) results involve highly leveraged regressions, and are very sensitive to outliers. Two applications illustrate the methodology in this setting. The first considers the relationship between trade openness and inflation (Romer, 1993). The second is about the effect of segregation on the quality of government (Alesina and Zhuravskaya, 2011). Both regressions are highly influenced by a few observations. Robust estimates have significantly smaller standard errors, producing more precise inferences. Bias correction reveals non-negligible bias in robust estimates.

Structure of the paper. Section 2 motivates the paper with the Price Puzzle example. Section 3 surveys the existing literature. Section 4 introduces the setting, sampling assumptions, and the estimator. Derivations for a simplified estimator give insights for the finite and large sample results. Section 5 provides finite-sample bounds and asymptotic results. Simulated and empirical applications are in Section 6. Appendices A, B give the proofs for the main results and preliminary ones. Supplemental Appendices C, D, E, F, G, H provide proofs for the preliminary results, simple derivations with leveraged outliers, derivations for influence and leverage in IV regressions, additional simulation and empirical results, and detailed numerical Algorithms to perform the estimation.

2 Motivating Example: the Price Puzzle

To illustrate the issues considered in this paper, consider estimating the impact of monetary policy with a recursive vector autoregressive (VAR) model as in Stock and Watson (2001). There are three variables: inflation (π_t), unemployment rate (u_t), and the federal funds rate (R_t). The VAR is estimated by OLS with four lags on U.S. data from 1960Q1 to 2000Q4.

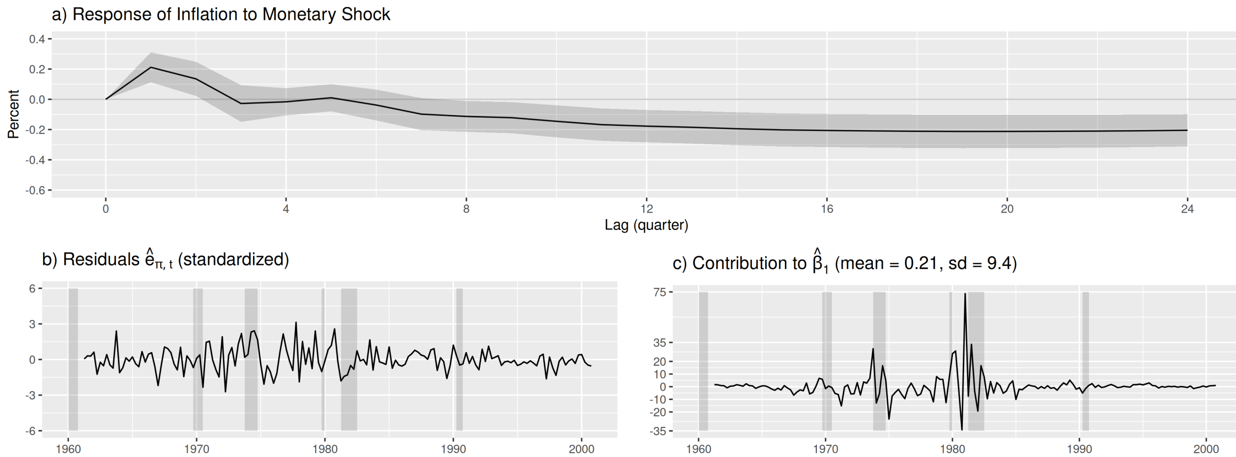
Panel a) in Figure 1 plots the estimated response of inflation to a unit increase in R_t. It shows a positive and significant increase in inflation for nearly four consecutive quarters. This was first observed by Sims (1992) and immediately coined as a 'Price Puzzle' by Eichenbaum (1992). It has since been studied extensively. Rusnák et al. (2013) performed a meta-analysis of 1000 estimates and put forward several potential forms of model misspecification to explain the puzzle. The number of specifications they explore is several times greater than the sample size so there should be some concerns about overfitting, however.

The following presents some simple diagnostics that indicate two time periods strongly influence the estimates. The puzzle begins with a positive and significant initial impact. It is measured by β_1 in the regression:

$$\pi_t = \beta_0 + \beta_1 R_{t-1} + \beta_2 u_{t-1} + \beta_3 \pi_{t-1} + \cdots + \beta_{10} R_{t-4} + \beta_{11} u_{t-4} + \beta_{12} \pi_{t-4} + e_{\pi,t}. \quad (1)$$

Figure 1 investigates this regression more closely. Panel a) plots the residuals $\hat{e}_{\pi,t}$ over time. Besides some increased volatility between 1970-1982, there are no obvious outliers in the series. In fact, the skewness and kurtosis are 0.36 and 3.78, respectively, not far from a normal distribution. Panel c) approximates the contribution of each t to $\hat{\beta}_1$. Since $\hat{\beta}_n = \sum_{t=1}^n (X'X/n)^{-1} x_t y_t / n$ is a sample mean, $(X'X/n)^{-1} x_t y_t$ approximates the contribution of each t to the mean. Some observations stand out: for instance, 1981Q1 alone positively contributes $\approx 75/n = 0.47$ to $\hat{\beta}_1 = 0.21$, about 3.5 standard errors.[1]

Figure 1: Recursive VAR: Impulse Response, Diagnostics



Note: a) Estimated response of inflation π to a unit increase in interest rate R, shaded = estimates $\pm$ one standard error, b) Standardized Residuals = $\hat{e}_{\pi,t}/\hat{\sigma}_{\hat{e}_\pi}$, c) Contribution of observation t to $\hat{\beta}_n$ measured by $(X'X/n)^{-1}x_t\pi_t$, x_t is the vector of regressors. b,c) Shaded vertical bars = NBER recession dates.

Panels b,c) show that, although none of the residuals $\hat{e}_{\pi,t}$ are particularly large, two time periods, around 1974-1975 and 1979-1982, have a disproportionate influence on the results. The latter has historical significance: the Federal Reserve changed to non-borrowed reserves targeting where the interest rate R_t was no longer a fixed policy instrument, as discussed in the introduction.Richmond FED President, Robert P. Black, summarized the tactical change during the October 1979 FOMC meeting as follows:

> "I often think of our position as being analogous to that of a monopolist in the sense that we control the money supply. A monopolist has a choice of controlling either price or quantity but he can never control both. I believe we've been trying to control the quantity of money by setting the price and we have misjudged. We've jiggled the price, in terms of the federal funds rate, one way or the other, and we‘ve usually met with less than complete success in judging what quantity of money will be forthcoming from that." (FOMC, 1979, p23)

This has several implications for the VAR estimates. First, R_t was no longer a direct measure of monetary policy: the recursive VAR may not correctly identify monetary shocks

[1]Most coefficients in (1) are strongly influenced by a few observations as shown in Table G9. The contribution reported here is related to Cook's distance which measures changes in predicted values $\hat{y}_t$ when observation t is excluded in the estimation (Cook, 1977). Here, the effect of observation t on the estimated regression coefficients is the object of interest – this will be referred to as contribution.

during that time period. Importantly, this goes beyond parameter instability. Time-varying parameters, regime-switching, or structural break models would still require R_t to provide a measure of monetary policy shocks. As emphasized by Robert Black, monetary policy was conducted on monetary aggregates at that time, not interest rates. Second, interest rates were significantly more volatile with the policy change;[2] producing significant regression leverage. This, as highlighted in Figure 1, gives excess influence to these observations.

Misspecification arises because the central bank relies on multiple policy instruments, the VAR only uses R_t. Friedman and Schwartz (1963) argued that well-known historical events clearly identify large monetary shocks. This narrative approach was popularized by Romer and Romer (1989), Romer and Romer (2004). Narrative and VAR estimates can differ when the central bank relies on different instruments throughout the sample (Coibion, 2012; Monnet, 2014). Narrative estimates, however, aggregate multiple types of monetary policies; results cannot be interpreted as e.g. an interest rate shock.

To identify the effect of an interest rate shock, a robust estimation is desirable. However, as noted in the introduction robust M-estimates may be biased and may not be robust to these x-outliers. Because residuals are small, robust M-estimates with Huber loss and high-breakdown MM estimates (*rlm*, *lmRob* in R) are nearly identical to Figure 1 (not reported).

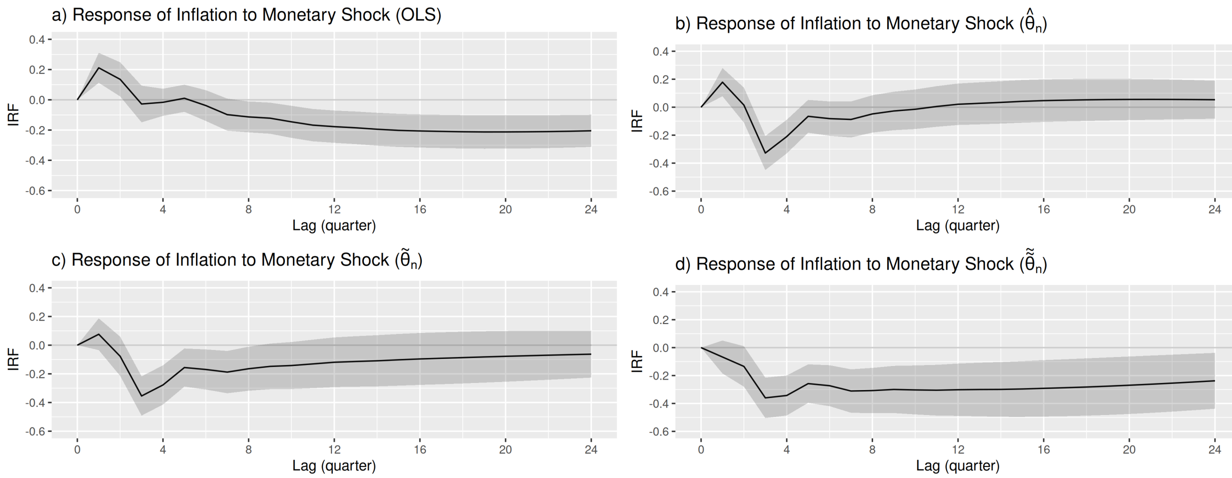
Diagnostics, as presented above, are useful to assess whether the estimation might present some irregularities. A robust estimation, presented below, is meant to reduce the influence of abnormal observations. The two are complementary, see Huber and Ronchetti (2011, Ch1.2.4) for further discussion.

Figure 2 re-estimates the effect on the same data, with the same model specification: using OLS (panel a), the proposed robust estimator without bias correction (panel b), with bias correction (panel c), with bias correction and a small sample correction (panel d). Without bias correction, the price puzzle remains – but does not last 4 quarters anymore. With bias correction, the price puzzle disappears; the initial effect is not significant. With the additional adjustment, the effect is qualitatively larger and negative.

As discussed above, the estimates can be seen as weighted least-squares. Figure 3 compares the weights, for each time period, used by each method on a regular and a log-scale (resp. top, bottom). OLS uses equal weighting (black/dashed). Without bias correction, robust estimates downweigh the leveraged outliers, especially 1979-1982, but overweigh other periods (black/solid). Bias correction re-adjusts towards equal weighting (blue/dot). The small sample adjustment further re-adjusts in that direction (purple/triangle).

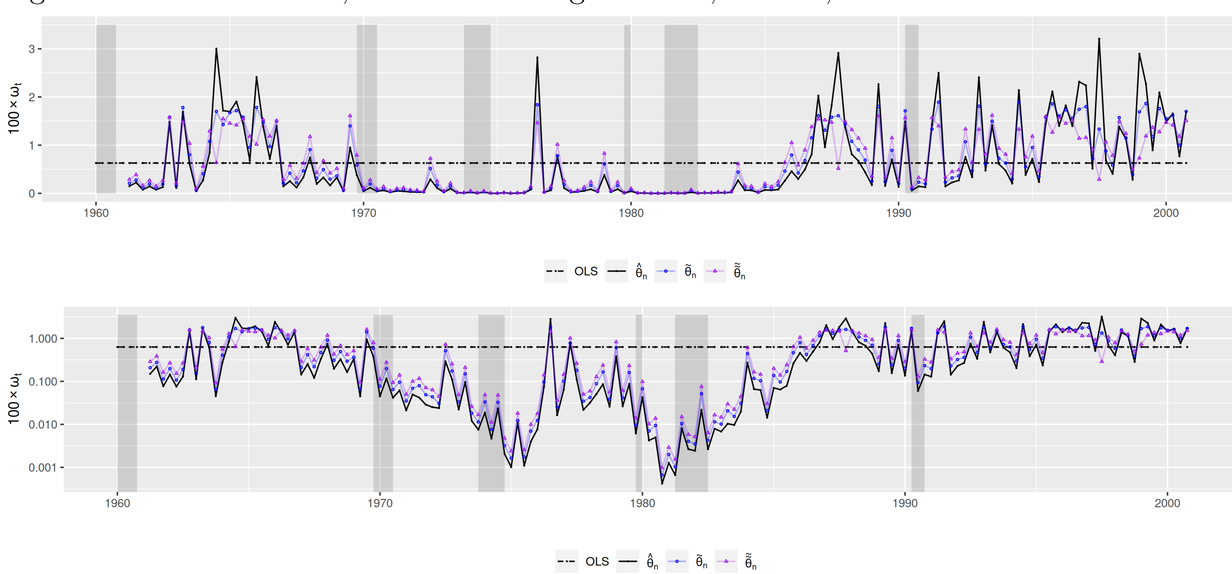
[2]This was anticipated and monitored by board members as shown by FOMC Transcripts of 1979-1982.

Figure 2: Recursive VAR, IRF: OLS, Robust and Bias-Corrected Estimates



Note: a) OLS estimates, b) $\hat{\theta}_n$ robust estimates without bias correction, c) $\tilde{\theta}_n$ robust estimates with bias correction, d) $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction. b,c,d) Estimates computed with tuning parameter $\hat{\nu}_n = 8.99$. Results for other ν in Appendix G. Bands: estimates $\pm$ one standard error.

Figure 3: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates



Note: Top and bottom panels: levels and log scale, respectively. Estimation weights ω_t, implicitly used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

3 Related Literature

The paper is mainly related to the literature on robust estimation, mostly developed in statistics. Textbook references such as Huber and Ronchetti (2011) and Maronna et al. (2019) survey a wide range of estimators and their properties. To focus the discussion, consider a linear regression: $y_t = x_t'\theta + e_t$. Robust M-estimators minimize the loss $\sum_{t=1}^{n} \psi(y_t - x_t'\theta)$ over θ. While OLS uses a quadratic ψ, least-absolute deviation (LAD), and the Huber (1964) loss are non-quadratic. They increase linearly with large residuals $|y_t - x_t'\theta|$. This reduces the influence of y-outliers. Winsorizing and trimming are popular alternatives. Huber (1964, p80) notes that trimming can be sensitive around the cutoffs. The first-order condition implies the solution $\hat{\theta}_n$ satifies $\sum_{t=1}^{n} x_t\psi'(y_t - x_t'\hat{\theta}_n)$. Large residuals $\hat{e}_t = y_t - x_t'\hat{\theta}_n$ are handled by ψ'. However, x-outliers with a large x_t, are not screened by ψ'. When the distribution of e_t is symmetric and the sample is contaminated symmetrically, robust estimates are consistent and asymptotically normal under regularity conditions. Symmetry is critical. Jaeckel (1971) derived, for estimating a location parameter, with asymmetric contamination of symmetric data, an asymptotic bias of order $n^{-1/2}$ when the proportion of outliers is $O(n^{-1/2})$ – i.e. $n_o = O(n^{1/2})$. n_o is the number of outliers.

For asymmetric data, the estimator is generally not consistent, see Carroll and Welsh (1988) for linear regressions. Quasi-Maximum Likelihood estimation, with a student distribution for the errors, is commonly used to estimate volatility models. Newey and Steigerwald (1997) show that the estimates may not be consistent without symmetry conditions. In a parametric setup, Cantoni and Ronchetti (2001) provide analytical bias formulas for generalized linear models, used to correct the first-order condition of the M-estimation. Here, parametric assumptions are not required. Zhou et al. (2018) derive bias bounds and exponential inequalities for linear regressions with the Huber loss when e_t has finite variance. They do not consider sample contamination and require sub-gaussian regressors - i.e. no x-outliers. These two issues are particularly relevant for the Price Puzzle. Another approach bounds the asymptotic bias in a local neighborhood of the model using the influence curve (IC) of Hampel (1974), see e.g. Huber and Ronchetti (2011, Ch4.9). Andrews (1986) relates the IC to the stability of estimators. Recently, several papers have used the IC to study and bound local misspecification bias for GMM, e.g. Andrews et al. (2017), Armstrong and Kolesár (2021), Bonhomme and Weidner (2022). Under these local asymptotics, the estimator remains consistent and asymptotically normal with a bias proportional to sampling uncertainty. In this paper, the model is grossly misspecified, but only for $1 \leq n_o \ll n$ outliers. Non-robust estimates can be inconsistent, or diverge: a robust estimation is required.

Christensen and Connault (2023) propose global sensitivity analyses on distributional assumptions, the model is otherwise correctly specified. It is common in Economics to apply more robust testing to non-robust estimates, assuming consistency, asymptotic normality – unlike here. One can adjust standard errors (e.g. MacKinnon, 2012), critical values (e.g. Müller, 2020; Pötscher and Preinerstorfer, 2023), or both. Sasaki and Wang (2023) propose a test for finite moments at a point, as required for consistency and central limit theory. Cowell and Victoria-Feser (1996) and Cowell and Flachaire (2007) consider the robustness properties of inequality measures, e.g. Gini coefficient. Surveying a large number of empirical results, Young (2022) finds that many IV regressions are highly leveraged and sensitive to a few observations, or clusters of observations.

Ronchetti and Trojani (2001) proposed a robust GMM estimator that is locally asymptotically robust, using the IC criteria. Hill and Renault (2010), Čížek (2016) consider trimming in GMM estimation. Rohatgi and Syrgkanis (2022) use a FILTER algorithm to screen out outliers in GMM estimation. The median-of-means is popular in prediction problems, which could also be considered here: the dataset is split into $K \geq 2$ subsamples of $m = n/K$ observations. K sample means are computed. The median of the K means is the estimator. The estimate is robust for up to $n_o \leq K/2 - 1$ outliers, see e.g. Lecué and Lerasle (2020), Laforgue et al. (2021). To accommodate an increasing n_o, having $K \to \infty$ as $n \to \infty$ is necessary. This introduces a bias, bounded above by $\sigma/\sqrt{m} = \sigma\sqrt{K/n}$.[3] Even for K fixed, an asymptotic bias can arise. Without a tractable expression for the bias, it is not clear how one would correct the asymptotic bias. Here, the choice of loss function makes the asymptotic bias tractable. An alternative is undersmoothing where the tuning parameter diverges fast enough that the bias is asymptotically negligible. It only requires to bound the asymptotic bias. Section 6.1 illustrates that it is less robust than bias-correction.

4 Models, Sample, Estimator

This paper considers estimations from unconditional moment restrictions:

$$\mathbb{E}_P \left[g\left(z_t; \theta \right) \right] = 0 \Leftrightarrow \theta = \theta_0, \tag{2}$$

where $z_t \overset{d}{\sim} P$ and the solution $\theta_0 \in \Theta$, a compact subset of $\mathbb{R}^k$. OLS regressions correspond to $g(z_t;\theta) = x_t(y_t - x_t'\theta)$ where $z_t = (y_t, x_t)$ collects the dependent variable and the regressors. For instrumental variable regressions, take $g(z_t;\theta) = w_t(y_t - x_t'\theta)$ where $z_t = (y_t, x_t, w_t)$

[3]For any distribution, the median and the mean differ by at most: $|\text{median}(X) - \mathbb{E}(X)| \leq \sigma(X)$.

collects the dependent variable, the regressors and the instruments. Non-linear estimations also fit into this framework. Concave Likelihood maximization, such as Probit or Logit, would set (2) to be the first-order condition. The main examples are linear.

The dataset consists of n observations but $z_t \sim P$ may not hold for all $t = 1, \dots, n$. This is presented in the following Assumption.

Assumption 1 (Sample)**.** *There are $n = n_P + n_o$ observations such that*

i) for $t \in \{1, \dots, n_P\}$, $z_t \sim P$ for which (2) holds, are either iid or strictly stationary, β-mixing with rate $\beta_m \leq a \exp(-bm)$ for $0 < a, b < \infty$;

ii) for $t \in \{n_P + 1, \dots, n\}$ and $0 < A, \alpha < \infty$:

$$z_t \in \mathcal{O}_n := \{z \text{ } s.t. \sup_{\theta \in \Theta} \|g(z; \theta)\|^2 \leq A n^\alpha\}. \tag{3}$$

The first n_P observations are such that (2) holds. However, the last n_o observations, or outliers, can be arbitrary in $\mathcal{O}_n$. The ordering between observations simplifies notation and, for time-series, preserves the dependence structure of the good n_P observations. The mixing condition typically holds for stationary VAR models, as in the motivating example. In practice, the user does not know which observations are drawn from P and those that are not. The n_o outliers could be allocated anywhere within the sample. The outliers will be chosen in an adversarial fashion, looking at the least-favorable collection $(z_{n_P+1}, \dots, z_n) \in \mathcal{O}_n$ for each θ, without restrictions on dependence.

The goal here is to derive finite-sample robustness properties against the *worst-case realization* of the n_o outliers. Ex-ante, if the n_o outliers are randomly distributed, such that $\mathbb{P}(\sup_{\theta \in \Theta} \|g(z_t; \theta)\| > t) \leq t^{-\varepsilon}$ for some $\varepsilon > 0$. Then $\mathbb{P}(z_t \in \mathcal{O}_n$ for each $t = n_P + 1, \dots, n) \geq 1 - A^{-\varepsilon} n_o n^{-\alpha\varepsilon}$, can be made arbitrarily close to 1 setting α large enough. In practice, the user does not specify (A, α). For random data contamination, Assumption 1 can be interpreted as conditioning on a realization with n_o outliers in the set $\mathcal{O}_n$ which has arbitrarily high-probability given an appropriate choice of n_o, A and α.[4]

Outliers can take many forms in (3). Figure 1 illustrates that residuals $y_t - x_t'\theta$ are not the only source of influence, captured here by $x_t(y_t - x_t'\theta)$. High leverage observations are only influential if $|y_t - x_t'\theta| \gg 0$. Likewise, $x_t(y_t - x_t'\theta)$ can be large when neither $y_t - x_t'\theta$ nor x_t are individually large but their product is non-negligible. This implies that screening residuals and regressors separately, as suggested in Hamilton (1992), can be insufficient. The influence

[4]See also Remark 1 in Laforgue et al. (2021).

of a single observation can also vary depending on the model specification: a regression that is linear in x_t is typically less leveraged than in a quadratic specification with (x_t, x_t^2) as regressors. Collinearity also plays a role on influence, as $(X'X/n)^{-1}x_t y_t$, reported in Figure 1, can be greatly inflated by the collinearity factor $(X'X/n)^{-1}$. In the motivating example, the regressors are lagged variables which are autocorrelated, i.e. collinear. A rotation invariance property is important to ensure robustness when there are multiple regressors. For instrumental variable regressions, the relevant quantity $w_t(y_t - x_t'\theta)$ involves the instruments w_t and the residual. In the context of time-series, one concern would be innovation outliers associated with a large shock $y_t - x_t'\theta$. Another, similar to the description in the motivating example, would be additive outliers. Here the effect is isolated, as in a different regime that occurs only once within the sample.

The main concern here is that the sample mean $\overline{g}_n(\theta) = 1/n \sum_{t=1}^n g(z_t;\theta)$ is not a consistent estimator for $\mathbb{E}_P[g(z_t;\theta)]$ when $(n_o n^\alpha)/n \not\to 0$. This allows to capture the concern that a minority of observations has significant influence, even as the sample size n increases. For $n_o = 1$, $\alpha = 1/2$ the estimates are consistent but asymptotically biased, standard error estimates are also affected.[5] For $n_o = 1$, $\alpha = 1$ estimates are inconsistent. They diverge when $\alpha > 1$. Mild outliers are also problematic: for $n_o = n^{1/4}$ and $\alpha = 1/4$ estimates are asymptotically biased.

To handle contaminated samples, Ronchetti and Trojani (2001) showed that a robust estimate of $\mathbb{E}_P[g(z_t;\theta)]$ is required. The following first computes a robust estimate of $\mu(\theta) = \mathbb{E}_P[g(z_t;\theta)]$, then corrects the first-order asymptotic bias, and finally solves for $\mu(\theta) = 0$.

Step 1. For each $\theta \in \Theta$, find $\hat{\psi}_n(\theta;\nu)$ which minimizes the sample criterion:

$$Q_n(\psi;\theta) = \frac{\nu + p}{n} \sum_{t=1}^n \log\left(1 + \frac{\|g(z_t;\theta) - \mu\|^2_{\Sigma^{-1}}}{\nu}\right) + \log|\Sigma| + \frac{\kappa_1}{\nu}\|\mu\|^2_{\Sigma^{-1}} + \frac{\kappa_2}{\nu}\text{trace}(\Sigma), \quad (4)$$

where $\psi = (\mu, \Sigma)$ and $p = \dim(g(z_t;\theta))$. The location and scale parameters are estimated jointly to ensure the first is invariant to rotation and less sensitive to re-scaling. The loss Q_n consists of a student quasi-likelihood plus two penalization terms. The tuning parameter $\nu > 0$ controls the robustness of the estimates. Here, it is not estimated and acts as a critical value. For observations such that $\|g(z_t;\theta) - \mu\|^2_\Sigma \ll \nu$, the loss is approximately quadratic $\frac{\nu+p}{n}\sum_{t=1}^n \log\left(1 + \|g(z_t;\theta) - \mu\|^2_{\Sigma^{-1}}/\nu\right) \simeq \|g(z_t;\theta) - \mu\|^2_{\Sigma^{-1}}/n$ which approximates the Gaussian log-likelihood. In contrast, for observations such that $\|g(z_t;\theta) - \mu\|^2_{\Sigma^{-1}} \gg \nu$

[5] This is illustrated in Appendix D.

the loss is logarithmic $\frac{\nu+p}{n}\sum_{t=1}^{n}\log\left(1+\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}/\nu\right)\simeq \nu/n[2\log(\|g(z_t;\theta)-\mu\|_{\Sigma^{-1}})-\log(\nu)]$. Large discrepancies, as measured by $\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}$, have less impact than using a Gaussian log-likelihood. To fully capture robustness, the parameter space Ψ for ψ is unbounded:

$$\Psi=\{(\mu,\Sigma),\ \mu\in\mathbb{R}^p, 0<s_0\leq\lambda_{\min}(\Sigma)\leq\lambda_{\max}(\Sigma)\leq+\infty\},$$

where s_0 is such that $s_0\leq\lambda_{\min}(\text{var}_P[g(z_t;\theta)])<+\infty$ for all $\theta\in\Theta$. In the presence of outliers, the main concern is in estimating a large $\hat{\mu}_n$ and/or $\hat{\Sigma}_n$. Here, setting $s_0>0$ simplifies some derivations to focus on finite-sample upper bounds.

Without regularization, setting $\kappa_1=\kappa_2=0$, the derivative $\partial_\mu Q_n(\psi;\theta)=0$ for $\|\mu\|=+\infty$ and any Σ. This can cause numerical instability when fitting data using a student log-likelihood. For non-zero κ_1,κ_2, $\partial_\mu Q_n(\psi;\theta)\to\infty$ when $\|\mu\|\to\infty$, the solution is bounded as shown in the next Section. The self-normalization $\|\mu\|_{\Sigma^{-1}}$ is invariant to rotations of the moments and less sensitive to scale. At the solution $\theta=\theta_0$, $\mu(\theta_0)=\mathbb{E}_P[g(z_t;\theta_0)]=0$ holds. This motivates penalizing towards zero in this particular setting.

Simultaneously estimating the location and scale parameters can seem problematic. A large $\hat{\Sigma}_n$ is effectively similar to using a large ν, leading to less robust location estimates $\hat{\mu}_n$. The second penalty trace(Σ) is important in that regard, as it ensures $\hat{\Sigma}_n$ cannot be too large in finite samples. This is shown in the next Section.

Step 2. For each $\theta\in\Theta$, compute:

$$\tilde{\mu}_n(\theta;\nu)=2\hat{\mu}_n(\theta;\nu)-\hat{\mu}_n(\theta;\nu/2). \tag{5}$$

This type of adjustment is known as Richardson extrapolation in numerical analysis. Unlike the sample mean, the estimator $\hat{\mu}_n(\theta;\nu)$ is typically biased for $\nu<+\infty$. Taking $\nu\to\infty$ with $n\to\infty$ at an appropriate rate, the adjustment $2\hat{\mu}_n(\theta;\nu)-\hat{\mu}_n(\theta;\nu/2)$ corrects the first-order asymptotic bias. The bias depends on higher order moments (see below). Estimating it is not straightforward: estimating the first moment is already a challenge in this setting. The correction (5) does not compute the bias, is simple to implement and widely applicable.

Step 3. Find $\tilde{\theta}_n$ such that:

$$\|\tilde{\mu}_n(\tilde{\theta}_n;\nu)\|^2_{W_n}\leq\inf_{\theta\in\Theta}\|\tilde{\mu}_n(\theta;\nu)\|^2_{W_n}+o_p(n^{-1}). \tag{6}$$

The estimated $\tilde{\theta}_n$ inherits the asymptotic bias properties of the corrected $\tilde{\mu}_n$.

Step 1. continuously updates both μ and Σ with θ. The scaling $\hat{\Sigma}_n(\theta;\nu)$ used to normalize the estimation of $\hat{\mu}_n(\theta;\nu)$ adapts to the value of θ. Appendix H gives generic Algorithms 1, 2 used to compute $\hat{\psi}_n$, $\tilde{\theta}_n$ in the applications. $\tilde{\mu}_n(\theta;\nu)$ is as smooth as $g(z_t;\theta)$ – cf. implicit function Theorem. Gradient-based optimizers, e.g. gradient-descent or Gauss-Newton, can be used. They are globally convergent under rank conditions (Forneron and Zhong, 2023, Th1,2). Unlike trimmed moments, the estimated $\tilde{\mu}_n(\theta;\nu)$ varies continuously with ν. This implies that the estimates $\tilde{\theta}_n$ can be less sensitive to small changes in tuning parameters. Figures G5-G7 reproduce Figure 2 with larger values of ν, illustrating that the estimated impulse response function changes gradually with ν.

Numerical software typically proceeds iteratively, see e.g. Huber and Ronchetti (2011, Ch7.8). Fix a tuning parameter and fit an initial regression $\hat{\theta}_n^1$. Then, update the scale parameter - here $\hat{\Sigma}_n^1$, re-estimate the regression $\hat{\theta}_n^2$, re-estimate the scale parameter, and repeat until convergence. The same scaling is applied for all θ at each stage. For least-squares, *rreg* in Stata and *rlm* in R proceed this way. Stata's *rreg* is initialized with a non-robust OLS estimate. The properties of the estimates after many iterations are not easy to derive, especially as scale estimates are less robust than those of location. Here, uniform-in-θ non-asymptotic concentration inequalities for the joint parameter $\hat{\psi}_n(\theta;\nu)$ are derived. This gives some finite-sample guarantees for step 1. above.

Intuition for the results. To better understand the role of the tuning parameter ν and the bias-correction step, consider estimating a scalar parameter $\theta_0 = \mathbb{E}_P(z_t)$ using:

$$\hat{\mu}_n(\nu) = \frac{1}{n}\sum_{t=1}^{n} \frac{z_t}{1+|z_t|^2/\nu},$$

which simplifies the first-order condition of Q_n with respect to μ.[6] For any z, $\frac{|z|}{1+|z|^2/\nu} \leq \frac{\sqrt{\nu}}{2}$ bounds the influence of a single observation. Let $\mu(\nu) = \mathbb{E}_P\left(z_t/(1+|z_t|^2/\nu)\right)$. If z_t are iid for $t \in \{1,\dots,n_P\}$, regardless of the remaining n_o observations:[7]

$$\mathbb{P}\left(\sup_{z_t \in \mathcal{O}_n, t>n_P} |\hat{\mu}_n(\nu) - \mu(\nu)| \geq \frac{\sqrt{\nu} n_o}{n} + \frac{n_P}{n}\frac{x}{\sqrt{n_P}}\right) \leq 2\exp\left(-\frac{x^2}{2\sigma_\nu^2 + \frac{2}{3}\sqrt{\frac{\nu}{n_P}}x}\right),$$

[6] The first-order condition $\partial_\mu Q_n = 0$ reads $\frac{\nu+p}{\nu n}\sum_{t=1}^n \frac{z_t-\mu}{1+\|z_t-\mu\|^2_{\Sigma^{-1}}/\nu} + \frac{\kappa_1 \mu}{\nu} = 0$.

[7] This inequality implies $\mathbb{P}(\sup_{z_t\in\mathcal{O}_n, t>n_P} |\hat{\mu}_n(\nu) - \mu(\nu)| \geq \frac{\sqrt{\nu}n_o}{n} + C\frac{n_P}{n}[\sqrt{\frac{x}{n_P}} + \frac{x}{n_P}]) \leq 2\exp(-x)$ for some constant C. This is the form used in a later Theorem.

using Bernstein's inequality, with $\sigma_\nu^2 = \text{var}_P\left(\frac{z_t}{1+|z_t|^2/\nu}\right) \to \text{var}_P(z_t)$ as $\nu \to \infty$. The right-hand-side is approximately sub-Gaussian for $x \ll \sqrt{n_P/\nu}$ and sub-exponential for $x \gg \sqrt{n_P/\nu}$. The factor $\sqrt{\nu/n_P}$ indicates the rate at which the estimator becomes sub-Gaussian.

As expected, outliers introduce a bias. The worst-case bias is at most $\sqrt{\nu}n_o/n$. Consistency of $\hat{\mu}_n$ requires $(\sqrt{\nu}/n)n_o = o(1)$ and asymptotic normality $(\sqrt{\nu/n})n_o = o(1)$. More contamination n_o requires a smaller ν to compensate. The same ν introduces another bias:

$$\mu(\nu) = \theta_0 - \frac{1}{\nu}\mathbb{E}_P\left(\frac{z_t^3}{1+z_t^2/\nu}\right),$$

as measured by the last term. It is typically non-zero when the distribution is not symmetric around 0. The bias is at most $\mathbb{E}_P(|z_t|^3)/\nu$ or $\mathbb{E}_P(|z_t|^2)/(2\sqrt{\nu})$ if, respectively, the third or second moment is finite. Consistency requires $\nu \to \infty$ and asymptotic normality $\sqrt{n}/\nu = o(1)$. There is some tradeoff between the outlier bias $\sqrt{\nu}n_o/n$, which mandates a smaller ν, and this robustness bias, which compels using a larger ν. A bias reduction that does not significantly degrade robustness can be achieved using $\tilde{\mu}_n = 2\hat{\mu}_n(\nu) - \hat{\mu}_n(\nu/2)$, since:

$$\tilde{\mu}(\nu) = 2\mu(\nu) - \mu(\nu/2) = \theta_0 - \frac{1}{\nu^2}\mathbb{E}_P\left(\frac{z_t^5}{(1+z_t^2/\nu)(1+2z_t^2/\nu)}\right).$$

Now the bias is at most $\mathbb{E}_P(|z_t|^5)/\nu^2$ or $\mathbb{E}_P(|z_t|^4)/\nu^{3/2}$ if, respectively, the fifth or fourth moment is finite. For the former, asymptotic normality only requires $\sqrt{n}/\nu^2 = o(1)$. The effect of a single observation on the estimate $\tilde{\mu}_n$ is no more than $\sqrt{2\nu} + \sqrt{\nu}/2$, compared to $\sqrt{\nu/2}$ for the non-corrected $\hat{\mu}_n$. The bias correction does require more regularity from the uncontaminated data in terms of moments - 5 instead of 3 finite ones.

Higher-order Richardson extrapolation could further reduce the order of the asymptotic bias. Simulations suggest the following can give better results in small samples. Applying the correction once more using $\tilde{\tilde{\mu}}_n = 2\tilde{\mu}_n(\nu) - \tilde{\mu}_n(\nu/2)$ flips the sign of the asymptotic bias and can have some small sample effects:

$$\tilde{\tilde{\mu}}(\nu) = \theta_0 + \frac{2}{\nu^2}\mathbb{E}_P\left(\frac{z_t^5(1-4z_t^4/\nu^2)}{(1+z_t^2/\nu)(1+2z_t^2/\nu)(1+2z_t^2/\nu)(1+4z_t^2/\nu)}\right).$$

To illustrate, take $z_t = \theta_0$ constant. Then $\tilde{\mu}(\nu) = \theta_0$ if, and only if, $\theta_0 = 0$ whereas $\tilde{\tilde{\mu}}(\nu) = \theta_0$ if $\theta_0 \in \{\theta_0, -\sqrt{\nu/2}, \sqrt{\nu/2}\}$. For finite ν, the bias of $\tilde{\tilde{\mu}}$ has two additional roots. Simulations in Section 6.1 indicate small-sample improvements for estimation and inference.[8]

[8]Note that averaging $2/3\tilde{\mu}(\nu) + 1/3\tilde{\tilde{\mu}}(\nu) = \theta_0 + o(\nu^{-2})$ can reduce the asymptotic bias, by the dominated convergence Theorem. This is not pursued here.

5 Properties of the Estimator

5.1 Finite Sample Bounds

The following Lemma shows the importance of the penalization κ_1, κ_2 in (4) which effectively bounds the parameter space Ψ.

Lemma 1. *For any $\theta \in \Theta$ and $\nu > 0$, the minimizer $\hat{\psi}_n = (\hat{\mu}_n, \hat{\Sigma}_n)$ of (4) over Ψ satisfies:*

$$\|\hat{\Sigma}_n^{-1/2}\hat{\mu}_n\| \leq \frac{\nu^{3/2}(1+p/\nu)}{2\kappa_1}, \quad trace(\hat{\Sigma}_n) \leq \frac{\nu^2(1+p/\nu)}{\kappa_2} + \frac{\nu^4(1+p/\nu)^2}{4\kappa_1\kappa_2} + \frac{p\nu}{\kappa_2}. \tag{7}$$

The dependence of $\hat{\psi}_n$ on θ, ν is omitted to simplify notation. Lemma 1 implies $\|\hat{\mu}_n\| \leq \nu^{7/2}$ and $\hat{\Sigma}_n \leq \nu^4$, up to constants. Although Ψ is unbounded, the estimates are bounded with probability 1. In the following, Ψ will be replaced with:

$$\Psi_n = \{(\mu, \Sigma) \in \Psi \text{ s.t. (7) holds}\},$$

without loss of generality. The upper bounds increase rapidly. With Lemma A1, they imply an envelope function of size ν^{17} which diverges too quickly to directly apply standard empirical process results, e.g. van der Vaart and Wellner (1996, Th2.14.1). Instead, the results directly rely on the functional form of (4) and the following assumption to derive exponential inequalities under cross-sectional and time-series dependence (Lemma A2).

Assumption 2. *$z_t \sim P$, a distribution such that for two $0 \leq M_2, M_4 < \infty$:*
i. $\sup_{\theta\in\Theta} \mathbb{E}_P(\|g(z_t;\theta)\|^2) \leq M_2$, ii. for all $(\theta_1, \theta_2) \in \Theta$, $\|g(z_t;\theta_1) - g(z_t;\theta_2)\| \leq G_t\|\theta_1 - \theta_2\|$ with $\mathbb{E}_P(\|G_t\|^2) \leq M_2$, iii. $\sup_{\theta\in\Theta} \mathbb{E}_P(\|g(z_t;\theta)\|^4) \leq M_4$. In ii. $G_t = G(z_t)$ is either iid or strictly stationary and mixing with rate β_m found in Assumption 1 i.

Let $Q_\nu = \mathbb{E}_P(Q_n)$ be the population analog of Q_n without any contamination:

$$Q_\nu(\psi;\theta) = \mathbb{E}_P\left[(\nu+p)\log\left(1 + \|g(z_t;\theta) - \mu\|^2_{\Sigma^{-1}}/\nu\right)\right] + \log|\Sigma| + \frac{\kappa_1}{\nu}\|\mu\|^2_{\Sigma^{-1}} + \frac{\kappa_2}{\nu}\text{trace}(\Sigma).$$

Proposition 1. *Take $x \geq 0$ and $1 \leq \nu \leq n$, suppose Assumptions 1 and 2 i-ii hold with z_t iid for $t \in \{1, \dots, n_P\}$. For each $\theta \in \Theta$, let $\hat{\psi}_n(\theta;\nu)$ be the minimizer of (4) and $\psi(\theta;\nu)$ the minimizer of Q_ν on Ψ. Set $C_n = 1 + (k+2p^2)[\log(p) + \log(\nu) + \log(n_P)]$, with $p = dim(g)$*

and $k = dim(\theta)$ then:

$$\mathbb{P}\Bigg(\sup_{\theta \in \Theta} \sup_{z_t \in \mathcal{O}_n, t > n_P} \Big\{ Q_\nu(\hat{\psi}_n(\theta;\nu);\theta) - Q_\nu(\psi(\theta;\nu);\theta) \Big\} \geq C_{\mathcal{O}} \frac{n_o(\nu+p)}{n}[1+\log(n)] + L \frac{n_P}{n}(\nu+p)\log(1+\nu p) \left[\sqrt{\frac{x}{n_P}} + \frac{x}{n_P} + \sqrt{\frac{C_n}{n_P}} + \frac{C_n}{n_P} \right] \Bigg) \leq 4\exp(-x),$$

for a constant L which depends on $s_0, \kappa_1, \kappa_2, M_2$ and $C_{\mathcal{O}}$ depends on $s_0, M_2, \kappa_1, A, \alpha$. If z_t is strictly stationary and β-mixing for $t \in \{1, \dots, n_P\}$, then:

$$\mathbb{P}\Bigg(\sup_{\theta \in \Theta} \sup_{z_t \in \mathcal{O}_n, t > n_P} \Big\{ Q_\nu(\hat{\psi}_n(\theta;\nu);\theta) - Q_\nu(\psi(\theta;\nu);\theta) \Big\} \geq C_{\mathcal{O}} \frac{n_o(\nu+p)}{n}[1+\log(n)] + \tilde{L} \frac{n_P}{n}(\nu+p)\log(1+\nu p) \left[\sqrt{\frac{(x+C_n)x}{n_P}} + \frac{(x+C_n)x}{n_P} + \sqrt{\frac{C_n}{n_P}} + \frac{C_n}{n_P} \right] \Bigg) \leq 12\exp(-x),$$

for $\tilde{L}$ which additionally depends on the mixing coefficients a, b.

Because $\psi(\theta;\nu)$ is a minimizer, $Q_\nu(\hat{\psi}_n(\theta;\nu);\theta) - Q_\nu(\psi(\theta;\nu);\theta) \geq 0$ always holds. Proposition 1 gives exponential inequalities for deviations from the biased solutions $\psi(\theta;\nu)$, uniformly in both parameters θ and outliers $z_t \in \mathcal{O}_n$, with respect to the loss Q_ν. The bounds only require finite second moments, allowing for heavy tails under P. This is important in macroeconomic and financial applications since P typically does not have sub-exponential, or Gaussian, tails.[9] The worst-case contamination bias is of order $n_0(\nu+p)/n[1+\log(n)]$ which depends on the proportion of outliers n_o/n and the tuning parameter ν. It differs from the $(\sqrt{\nu}/n)n_o$ term for the simple estimator above. The proofs indicate that $(\nu/n)n_o$ corresponds to the influence of outliers when estimating Σ.

For iid data, similar to Bernstein's inequality, the tails are thin: approximately sub-Gaussian for small $x \ll \sqrt{n_P}$ and sub-exponential for large $x \gg \sqrt{n_P}$.[10] For time-series data, the tails are thicker: approximately sub-Gaussian for $x \ll C_n$, sub-exponential for $C_n \ll x \ll \sqrt{n_P}$ and sub-Weibull for $x \gg \sqrt{n_P}$ with tail parameter $1/2$ (Vladimirova et al., 2020). This is comparable to Bernstein inequalities for sample means of bounded β-mixing processes in Doukhan (1994).

[9] Heavy-tailed distributions, unlike the exponential and Gaussian distributions, may not have all finite moments. Student and Pareto are both heavy-tailed distributions.

[10] The inequality $\mathbb{P}(Z \geq \sqrt{x/n} + x/n + a_n) \leq 4\exp(-x)$ implies $\mathbb{P}(Z \geq u/\sqrt{n} + a_n) \leq 4\min[\exp(-u^2), \exp(-\sqrt{n}u)]$ which is sub-Gaussian for $u \ll \sqrt{n}$ and sub-exponential for $u \gg \sqrt{n}$.

Estimating both μ and Σ consistently requires $(\nu/n)\log(n)n_o \to 0$. This is more restrictive than $(\sqrt{\nu}/n)n_o \to 0$ which appears under local asymptotics for μ. This is related to the discussion above on iterative procedures and joint estimation of ψ. The dependence on the number of moment conditions p is made explicit to show how it affects the bounds. The p^2 term in C_n comes from estimating $p(p+1)/2$ coefficients in Σ. For the large sample results below, the number of parameters k and moments p will be assumed to be fixed and finite.

5.2 Asymptotic Properties

The following builds on Proposition 1 to derive uniform consistency and then oracle equivalence results which involve the amount of contamination n_o and the bias. The large-sample results can be used to compute standard errors and compute confidence intervals the usual way (i.e. reporting $\tilde{\theta}_n \pm 1.96\text{se}(\tilde{\theta}_n)$).

Corollary 1. *Suppose the conditions for Proposition 1, Assumption 2 iii hold, and:*

$$n_o = o\left(\frac{n}{\nu\log(n)}\right),\ \nu\log(\nu) = o\left(\sqrt{\frac{n}{\log(n)}}\right).$$

Let $\psi(\theta;\infty)$ *denote the pair* $\mu(\theta;\infty) = \mathbb{E}_P[g(z_t;\theta)]$, $\Sigma(\theta;\infty) = var_P[g(z_t;\theta)]$, *then:*

$$\sup_{\theta\in\Theta}\left(\sup_{z_t\in\mathcal{O}_n, t>n_P}\|\hat{\psi}_n(\theta;\nu) - \psi(\theta;\infty)\|\right) = o_p(1).$$

Proposition 1 and the following two bounds: $|Q_\nu(\psi;\theta) - Q_\infty(\psi;\theta)| \leq O(\nu^{-1})$ and $\|\psi(\theta;\nu) - \psi(\theta;\infty)\| \leq O(\nu^{-1})$, uniformly in θ, imply the uniform consistency result above. Taking the supremum over $\mathcal{O}_n$ ensures the result is robust against the least favorable outliers.

Proposition 2. *Suppose the conditions of Corollary 1 hold. Let* $\max[\mathbb{E}_P(\|g(z_t;\theta_0)\|^{r+\delta}), \mathbb{E}_P(|G_t|^{r+\delta})] := M_{r,\delta}$ *for* $r \geq 1$ *and* $\delta > 0$. *Let* $\overline{g}_{n_P}(\theta) = \frac{1}{n_P}\sum_{t=1}^{n_P} g(z_t;\theta)$, *if* $M_{3,\delta}$ *is finite for some* $\delta > 0$*:*

$$\sup_{\theta\in\Theta}\left(\sup_{z_t\in\mathcal{O}_n, t>n_P}\|\hat{\mu}_n(\theta;\nu) - \overline{g}_{n_P}(\theta)\|\right) = O_p\left(\max\left[\frac{1}{\nu}, \frac{\sqrt{\nu}n_o}{n}\right]\right)$$

If, in addition $M_{5,\delta}$ *is finite for some* $\delta > 0$*:*

$$\sup_{\theta\in\Theta}\left(\sup_{z_t\in\mathcal{O}_n, t>n_P}\|\tilde{\mu}_n(\theta;\nu) - \overline{g}_{n_P}(\theta)\|\right) = O_p\left(\max\left[\frac{1}{\nu^2}, \frac{\sqrt{\nu}n_o}{n}\right]\right).$$

Using the same two inequalities, and a bound on the score, Proposition 2 shows that the robust and bias-corrected estimates are uniformly close to an oracle that computes the sample mean using only the good n_P observations. An empirical researcher might want to trim out outliers without altering, as much as possible, the rest of the sample. This oracle result precisely states this property. In that sense, it gives a more desirable characterization than limit theorems for $\hat{\mu}_n(\theta;\nu) - \mu(\theta;\infty)$ and $\tilde{\mu}_n(\theta;\nu) - \mu(\theta;\infty)$.

Similar to non-parametric regressions which derive bias from smoothness, stronger moment conditions are needed to derive faster rates of convergence. Without outliers, OLS estimates are asymptotically normal for iid data when $\mathbb{E}_P(\|x_t y_t\|^2), \mathbb{E}_P(\|x_t\|^4) < \infty$. Here the condition is more restrictive, it reads $\mathbb{E}_P(\|x_t y_t\|^{5+\delta}), \mathbb{E}_P(\|x_t\|^{10+2\delta}) < \infty$.

The worst-case impact of outliers is of order $(\sqrt{\nu}/n)n_o$, with and without bias correction. Note that the estimator $\hat{\mu}_n$ is "redescending." The maximal influence of a single observation z given by $(\sqrt{\nu/2})/n$, is attained at $\|g(z;\theta) - \mu\|_{\Sigma^{-1}} = \sqrt{\nu}$ and then monotonically declines to zero as $\|g(z;\theta) - \mu\|_{\Sigma^{-1}}$ increases.[11] The result requires $\hat{\Sigma}_n(\theta;\nu)$ uniformly convergent. Importantly, the influence function is not redescending for Σ: it is strictly increasing and bounded above by $\nu > \sqrt{\nu}$. Hence, consistency of $\hat{\Sigma}_n$ is more restrictive: $(\nu/n)n_o \to 0$.

Assumption 3. *i.* $\mathbb{E}_P[g(z_t;\cdot)]$ *is continuously differentiable in* $\theta \in \Theta$, *ii.* $\mathbb{E}_P[g(z_t;\theta)] = 0$ *if, and only if,* $\theta = \theta_0 \in int(\Theta)$, *iii.* $G(\theta_0) := \partial_\theta \mathbb{E}_P[g(z_t;\theta_0)]$ *has full rank, iv. for any* $\delta_{n_P} \to 0$, $\sup_{\|\theta-\theta_0\| \le \delta_{n_P}} \sqrt{n_P}\|\overline{g}_{n_P}(\theta) - \overline{g}_{n_P}(\theta_0) - \partial_\theta \mathbb{E}_P[g(z_t;\theta_0)](\theta-\theta_0)\|/[1+\sqrt{n_P}\|\theta-\theta_0\|] = o_p(1)$, *v.* $\sqrt{n_P}\overline{g}_{n_P}(\theta_0) \overset{d}{\to} \mathcal{N}(0,\Sigma_0)$, *vi.* $W_n \overset{p}{\to} W$ *positive definite.*

Assumption 2 repeats conditions from Newey and McFadden (1994), only for the good n_P observations. They imply consistency and asymptotic normality of $\hat{\theta}_{n_P}$, an oracle estimator.

Theorem 1. *Suppose Assumption 3 and the conditions of Proposition 2 hold with* $M_{5,\delta}$ *finite for some* $\delta > 0$. *Suppose* n_o *and* ν *are such that:*

$$\frac{\sqrt{n}}{\nu^2} = o(1), \text{ and } \sqrt{\frac{\nu}{n}} n_o = o(1).$$

Let $\hat{\theta}_{n_P} = argmin_{\theta \in \Theta} \|\overline{g}_{n_P}(\theta)\|_{W_n}$, *the estimator* $\tilde{\theta}_n$ *satisfies:*

$$\sup_{z_t \in \mathcal{O}_n, t > n_P} \|\sqrt{n_P}(\tilde{\theta}_n - \hat{\theta}_{n_P})\| = o_p(1), \quad \text{and} \quad \sqrt{n_P}(\tilde{\theta}_n - \theta_0) \overset{d}{\to} \mathcal{N}(0,V)$$

for any sequence $z_t \in \mathcal{O}_n$, $t = n_P+1, \dots, n$, *where* $V = (G'WG)^{-1}G'W\Sigma_0 WG(G'WG)^{-1}$, $G = \partial_\theta \mathbb{E}_P[g(z_t;\theta_0)]$.

[11] This is also discussed in McDonald and Newey (1988, p432), Huber and Ronchetti (2011, Ch4.8).

Theorem 1 presents the main result: the bias-corrected estimates are asymptotically equivalent to the oracle $\hat{\theta}_{n_P}$. They inherit its asymptotic properties. The supremum over $\mathcal{O}_n$ ensures robustness against least favorable outliers. The bias is asymptotically negligible if $\nu^2 = o(\sqrt{n})$. If n_o were known, setting $\nu \asymp (n/n_o)^{2/5}$ would achieve the optimal rate in Proposition 2. For this choice of ν, the condition $\sqrt{n}/\nu^2 = o(1)$ reads $n_o = o(n^{3/8})$. Setting $\nu = O(n^{1/4}\log(n))$ is nearly optimal when n_o becomes arbitrarily close to this bound as it requires implies $n_o = o(n^{3/8}/\sqrt{\log(n)})$. A data-driven rule is given below to select ν in practice while enforcing this rate. The large sample properties of $\tilde{\tilde{\mu}}_n(\theta;\nu) = 2\tilde{\mu}_n(\theta;\nu) - \tilde{\mu}_n(\theta;\nu/2)$, from Section 4, and the resulting $\tilde{\tilde{\theta}}_n$ follow from those of $\tilde{\mu}(\theta;\nu)$, $\tilde{\theta}_n$.

With the oracle result (Proposition 2) and the regularity conditions (Assumption 3), further results could be derived. One could consider two-step GMM with robust weighting $W_n = \hat{\Sigma}_n(\tilde{\theta}_n;\nu)^{-1}$ in the second step, robust overidentifying restrictions, quasi-Likelihood Ratio and Lagrange multiplier tests, etc. This is not pursued here.

Proposition 3. *Suppose the assumptions for Theorem 1 hold. For each $\theta \in \Theta$ and $\nu > 0$, the estimates $\hat{\mu}_n(\theta;\nu)$, $\tilde{\mu}_n(\theta;\nu)$ satisfy:*

$$\hat{\mu}_n(\theta;\nu) = \sum_{t=1}^{n} \omega_t(\theta;\nu) g(z_t;\theta), \quad \tilde{\mu}_n(\theta;\nu) = \sum_{t=1}^{n} \tilde{\omega}_t(\theta;\nu) g(z_t;\theta)$$

where the weights are given by $\omega_t(\theta;\nu) = \frac{(1+p/\nu)/n[1+q_t(\theta;\nu)/\nu]^{-1}}{(1+p/\nu)/n\sum_{t=1}^{n}[1+q_t(\theta;\nu)/\nu]^{-1}+\kappa_1/\nu}$ and $\tilde{\omega}_t(\theta;\nu) = 2\omega_t(\theta;\nu) - \omega_t(\theta;\nu/2)$ using $q_t(\theta;\nu) = \|g(z_t;\theta) - \hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}$.

Let $\hat{\varepsilon}_t(\theta) = g(z_t;\theta) - \hat{\mu}_n(\theta;\nu)$, $\tilde{\varepsilon}_t(\theta) = g(z_t;\theta) - \tilde{\mu}_n(\theta;\nu)$. The following weighted variance estimators are consistent:

$$\hat{\Sigma}_{n,\omega}(\theta) = \sum_{t=1}^{n} \omega_t(\theta;\nu)\hat{\varepsilon}_t(\theta)\hat{\varepsilon}_t(\theta)' \xrightarrow{p} \Sigma(\theta), \quad \tilde{\Sigma}_{n,\omega}(\theta) = \sum_{t=1}^{n} \tilde{\omega}_t(\theta;\nu)\tilde{\varepsilon}_t(\theta)\tilde{\varepsilon}_t(\theta)' \xrightarrow{p} \Sigma(\theta),$$

where $\Sigma(\theta) = var_P(g(z_t;\theta))$ here denotes the short-run variance under P.

For cross-sections and serially uncorrelated moments, $\hat{\Sigma}_n(\hat{\theta}_n,\nu)$ can be used to estimate Σ_0 in Theorem 1. Because of the penalty $\kappa_2 > 0$ on Σ, it tends to be downward biased. An alternative is to use the same weights as $\tilde{\mu}_n$ to match the properties of the estimator more closely. Proposition 3 above shows that such an estimator is also consistent for the short-run variance. Long-run variance estimates, required for serially correlated moments, are not considered here.

The weighted average representation further implies, for linear models, that $\tilde{\theta}_n$ are weighted least-squares estimates since $\tilde{\mu}_n(\tilde{\theta}_n;\nu) = \sum_{t=1}^n \tilde{\omega}(\tilde{\theta}_n;\nu)x_t(y_t - x_t'\tilde{\theta}_n) = 0 \Rightarrow \tilde{\theta}_n = (\sum_{t=1}^n \tilde{\omega}(\tilde{\theta}_n;\nu)x_t x_t')^{-1}\sum_{t=1}^n \tilde{\omega}(\tilde{\theta}_n;\nu)x_t y_t$. The weighting can be used to interpret the results.

Data-driven choice of tuning parameter ν. The following describes a data-driven procedure to select the tuning parameter ν. Take $0 < a_0 < a_1 < \cdots < a_J$ and $\nu_j = a_j n^s$ with $1/4 < s < 1/2$ or $\nu_j = a_j n^s \log(n)$ with $1/4 \leq s < 1/2$. The simulated and empirical examples use $s = 1/4$ and $0.5 = \log(a_0) < \cdots < \log(a_J) = 35$ so that each $\nu_j = O(n^{1/4}\log(n))$ satisfies the requirements for Theorem 1.

Using $\nu = \nu_0$ as a baseline, compute a preliminary estimate $\hat{\theta}_n$ and the corresponding moment estimates $\hat{\psi}_n(\hat{\theta}_n;\nu_0)$. In the absence of outliers, it can be shown that $|Q_n(\hat{\psi}_n(\hat{\theta}_n;\nu_0);\nu_j) - Q_n(\hat{\psi}_n(\hat{\theta}_n;\nu_0);\infty)| = O_p(\nu_j^{-1})$. This implies that, in the absence of outliers, the fit should be comparable accross different values of ν: $|Q_n(\hat{\psi}_n(\hat{\theta}_n;\nu_0);\nu_j) - Q_n(\hat{\psi}_n(\hat{\theta}_n;\nu_0);\nu_0)| \leq O_p(\nu_0^{-1})$. The selection rule picks the largest value of ν_j such that the fit remains comparable:

$$\hat{\nu}_n = \max\left\{\nu_j, \text{ s.t. } |Q_n(\hat{\psi}_n(\hat{\theta}_n;\nu_0);\nu_j) - Q_n(\hat{\psi}_n(\hat{\theta}_n;\nu_0);\nu_0)| \leq \frac{1+\log(n)}{\nu_0}\right\}.$$

The parameters and the moments are only estimated once, to reduce computation, at the smallest ν_0 which produces the most robust estimate of the grid $\nu_0, \dots, \nu_J$.

By design, $\hat{\nu}_n$ has rate $O(n^s)$ or $O(n^s\log(n))$ which satisfies the conditions of Theorem 1 given restrictions on n_o. The following heuristic motivates the choice of criteria. As discussed above, the outliers have an asymptotic impact on non-robust estimates if $n_o n^\alpha/n \not\to 0$, and the estimator is robust as long as $n_o = o(\sqrt{n/\nu})$. Set $n_o = c\sqrt{n/\nu_0}$ then the sum over outliers in $Q_n(\cdot;\nu)$ increases proportionally to $\nu c\sqrt{\nu_0/n}\log(1+n^{2\alpha}/\nu) \sim c\nu\sqrt{\nu_0/n}\log(n)$. The change over the n_P terms is a $O_p(\nu_0^{-1})$. For ν_0 relatively small, the upper bound in the criteria above conservatively minors the sum of these two bounds.

6 Simulated and Empirical Applications

All the estimations below use the same $\kappa_1 = \kappa_2 = 10^{-2}$, giving wide bounds in Lemma 1. With the data-driven choice of $\hat{\nu}_n$, the results are not too sensitive to this choice of penalty.

6.1 Simulated Example

To illustrate the finite sample properties of the procedure, consider a linear regression $y_t = x_t'\theta_0 + e_t$. There are three regressors $x_t = (1, x_{1t}, x_{2t}, x_{3t})$, each x_{jt} and e_t is drawn from

$(\chi^2_5 - 5)/\sqrt{10}$ has mean zero and unit variance, $\theta_0 = (0, 1, 1, 1)$. Sample size is $n = 150$, several $n_o = 0, 1, 5, 10$ are reported where each outlier has $x_{jt} = \sqrt{n}$ and $y_t = x_t' \theta_\dagger$, $\theta_\dagger = (0, 1/2, 1/2, 1/2)$. In this example, outliers are leveraged to mimic the motivating example.

The simulations compares full sample $\hat{\theta}^{ols}_n$, an oracle which discards outliers $\hat{\theta}^{ols}_{nP}$, R's robust regression estimates $\hat{\theta}^{rlm}_n$ with $\hat{\theta}_n$, $\tilde{\theta}_n$, $\tilde{\tilde{\theta}}_n$ computed using $\hat{\nu}_n$ as described above. A further $\hat{\theta}^{un}_n$ is computed using $\hat{\nu}^2_n$ to illustrate undersmoothing as opposed to bias correction used in this paper. $\tilde{\tilde{\theta}}_n$ applies the correction step twice as discussed at the end of Section 4.

Table 1: Small sample properties of the estimators ($n = 150$)

	100 × RMSE							Rejection Rate						
	$n_o = 0$													
	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$
θ_0	8.05	8.05	12.00	11.84	9.31	8.11	7.94	0.04	0.04	0.24	0.29	0.14	0.05	0.06
θ_1	8.00	8.00	7.15	7.97	7.79	7.78	7.92	0.06	0.06	0.06	0.11	0.08	0.07	0.06
θ_2	8.10	8.10	7.46	8.45	8.21	8.11	8.06	0.04	0.04	0.05	0.10	0.06	0.05	0.05
θ_3	8.19	8.19	7.43	8.55	8.30	8.16	8.14	0.06	0.06	0.06	0.10	0.07	0.06	0.06
	$n_o = 1$													
	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$
θ_0	10.71	8.04	13.01	14.18	10.97	8.52	10.32	0.03	0.04	0.20	0.46	0.23	0.08	0.08
θ_1	38.57	8.07	15.23	8.27	7.97	7.87	32.24	0.00	0.06	0.01	0.14	0.10	0.07	0.40
θ_2	38.39	8.11	15.09	8.73	8.36	8.14	32.08	0.01	0.04	0.01	0.12	0.06	0.06	0.38
θ_3	39.94	8.20	15.75	8.83	8.49	8.27	33.47	0.00	0.06	0.00	0.12	0.09	0.07	0.39
	$n_o = 5$													
	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$
θ_0	11.98	8.14	16.57	16.98	13.38	9.82	13.45	0.10	0.04	0.24	0.59	0.38	0.13	0.16
θ_1	47.57	8.40	47.17	9.02	8.62	8.40	46.72	0.99	0.06	0.99	0.12	0.08	0.06	0.99
θ_2	47.48	8.26	48.25	9.28	8.80	8.53	47.14	0.99	0.04	1.00	0.12	0.05	0.03	1.00
θ_3	49.17	8.28	49.48	9.33	8.94	8.72	48.65	0.98	0.06	0.98	0.10	0.08	0.04	0.98
	$n_o = 10$													
	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$	$\hat{\theta}^{ols}_n$	$\hat{\theta}^{ols}_{nP}$	$\hat{\theta}^{rlm}_n$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}^{un}_n$
θ_0	12.21	8.21	17.33	16.78	13.27	10.35	14.13	0.09	0.04	0.23	0.47	0.22	0.07	0.17
θ_1	49.14	8.54	48.38	10.22	11.68	19.76	48.65	0.99	0.04	0.99	0.01	0.01	0.09	1.00
θ_2	49.05	8.31	49.67	10.76	12.40	20.28	48.92	0.99	0.04	0.99	0.01	0.01	0.09	1.00
θ_3	50.52	8.51	50.70	11.04	13.00	20.96	50.19	0.98	0.06	0.98	0.00	0.01	0.09	0.99

Legend: $\hat{\theta}^{ols}_n$ full sample OLS, $\hat{\theta}^{ols}_{nP}$ oracle OLS, $\hat{\theta}^{rlm}_n$ robust M-estimator, $\hat{\theta}_n$ robust estimates without bias correction, $\tilde{\theta}_n$ robust estimates with bias correction, $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction, $\hat{\theta}^{un}_n$ undersmoothed robust estimates with $\hat{\nu}^2_n$. 200 Monte-Carlo replications. n_o = number of outliers. Rejection rate for t-test at the 5% significance level. Average $\hat{\nu}_n$: 35.85, 16.00, 11.00, 10.71 for n_0 = 0, 1, 5, 10 respectively. Each $\hat{\nu}_n$ is selected on a grid $[\nu_0, \dots, \nu_J]$ where $\nu_0 = 8.77$, $\nu_J = 584.69$.

Table 1 shows that without outliers ($n_o = 0$) the performance of bias-corrected and undersmoothed estimates is comparable to full sample OLS. The robust M-estimates of the intercept θ_0 are biased, because the errors are skewed. The performance of OLS degrades as soon as $n_o = 1$, as expected. The undersmoothed and *rlm* estimates are also less accurate. The non-corrected estimates $\hat{\theta}_n$ are more robust but biased. Bias correction, $\tilde{\theta}_n$ and $\tilde{\tilde{\theta}}_n$, improves accuracy and rejection rates. The estimators still perform well for $n_o = 5$. Performance degrades for $n_o = 10$. This is perhaps not too surprising since $\log(n_o)/\log(n) \simeq 0.45 > 3/8$ for $n_o = 10$. Additional results for $n = 500$ are reported in Table F8, Appendix F. Tables F6, F7 has results with $\nu = O(n^{1/3})$ in the same Appendix.

6.2 Empirical Applications

Two empirical applications further illustrate robust-GMM estimator in instrumental variable regression settings.

6.2.1 Trade Openness and Inflation

The second empirical application is also inflation-related. Romer (1993) estimates the relationship between trade openness and inflation using country time averages between 1973 and 1993. Trade openness, measured by the share of imports to GDP, can be considered as endogenous given that monetary policy affects both inflation and exchange rates. He considers the following specification:

$$\pi_t = \theta_0 + \theta_1 \text{open}_t + \theta_2 \log(\text{pcinc})_t + e_t,$$

where π measures inflation, pcinc is per-capita income in 1980, assumed exogenous. Romer (1993) further adds dummies in some specifications, these are not included here. The instrument for openness is log(land) measuring the log of the square-mile surface of the country. The idea is that smaller land area economics should be more open to imports. Romer (1993) notes that "A few countries in the sample have extremely high average inflation rates." and is concerned that "the parameter estimates from a linear regression would be determined almost entirely by a handful of observations." As a remedy, he estimates the regression using the log of average inflation $\log(\pi/100)$. The influence of outliers in linear IV regressions is not intuitive because leverage can be either positive or negative (Lemma E3). As a result, unlike OLS, the influence may not have the same sign as the residual: the impact of an outlier is less predictable than with OLS.

Similar to the motivating example, Table 2 provides diagnostics for both specifications. For $y = \pi/100$, the greatest contributors tend to be severely indebted countries that were particularly affected by the 1980s debt crisis. Terra (1998) argues that these countries

Table 2: Romer (1993): 10 Largest Contributors to $\hat{\theta}^{IV}_{1n}$, Sample Moments

Dependent variable: $y = \log(\frac{\pi}{100})$				Dependent variable: $y = \frac{\pi}{100}$			
Country	Contr.	$\log(\frac{\pi}{100})$	Open.	Country	Contr.	$\frac{\pi}{100}$	Open.
Malta	-60.75	-3.17	0.92	Bolivia	-11.27	2.07	0.23
Singapore	-56.77	-3.32	1.64	Argentina	-11.01	1.17	0.09
Bahrain	-49.65	-3.04	0.91	Brazil	-9.40	0.74	0.07
Barbados	-40.74	-2.23	0.73	Israel	4.28	0.75	0.57
United States	39.32	-2.78	0.09	Peru	-3.18	0.49	0.20
Canada	38.08	-2.65	0.25	Chile	-3.15	0.59	0.23
Hong Kong	-37.30	-2.49	0.82	Mexico	-2.73	0.33	0.11
Luxembourg	-32.86	-2.80	0.76	Zaire	-2.57	0.43	0.40
Australia	31.24	-2.35	0.17	Barbados	1.95	0.11	0.73
Mauritius	-29.07	-2.02	0.57	Mauritius	1.92	0.13	0.57
	Sample Moments				Sample Moments		
Mean	-1.25	-2.10	0.37	Mean	-0.34	0.17	0.37
Stdev	15.57	0.71	0.24	Stdev	1.93	0.24	0.24
Skewness	-1.12	1.25	2.09	Skewness	-3.91	5.34	2.09
Kurtosis	6.32	5.38	9.89	Kurtosis	22.22	38.10	9.89

Note: Contr.: Contribution = $(Z'X/n)^{-1}z_i y_i$ to coefficient $\hat{\theta}^{IV}_{1n}$. Open.: Openness. π = average inflation. Sample size $n = 114$. Countries sorted in decreasing order of contribution, in absolute values.

overborrowed in the 1980s and had "less pre-commitment in monetary policy" resulting in higher inflation during the debt crisis.[12] In contrast, for $y = \log(\pi/100)$, the greatest contributors are less indebted and other countries Terra (1998, p647) which have low average inflation.[13] The log increases the influence of low-inflation countries, as one might expect.

Table 3: Romer (1993): IV, Robust and Bias-Corrected Estimates

	Dependent variable: $y = \log(\frac{\pi}{100})$											
	$\hat{\theta}^{\text{IV}}_{0n}$	$\hat{\theta}^{\text{IV}}_{1n}$	$\hat{\theta}^{\text{IV}}_{2n}$	$\hat{\theta}_{0n}$	$\hat{\theta}_{1n}$	$\hat{\theta}_{2n}$	$\tilde{\theta}_{0n}$	$\tilde{\theta}_{1n}$	$\tilde{\theta}_{2n}$	$\tilde{\tilde{\theta}}_{0n}$	$\tilde{\tilde{\theta}}_{1n}$	$\tilde{\tilde{\theta}}_{2n}$
est	-1.21	-1.25	-5.64	-1.19	-1.13	-6.82	-1.18	-1.21	-6.42	-1.19	-1.29	-5.70
se	0.42	0.40	5.60	0.37	0.36	5.01	0.40	0.38	5.41	0.43	0.41	5.70
	Dependent variable: $y = \frac{\pi}{100}$											
	$\hat{\theta}^{\text{IV}}_{0n}$	$\hat{\theta}^{\text{IV}}_{1n}$	$\hat{\theta}^{\text{IV}}_{2n}$	$\hat{\theta}_{0n}$	$\hat{\theta}_{1n}$	$\hat{\theta}_{2n}$	$\tilde{\theta}_{0n}$	$\tilde{\theta}_{1n}$	$\tilde{\theta}_{2n}$	$\tilde{\tilde{\theta}}_{0n}$	$\tilde{\tilde{\theta}}_{1n}$	$\tilde{\tilde{\theta}}_{2n}$
est	0.27	-0.34	0.38	0.21	-0.08	-0.74	0.22	-0.10	-0.75	0.23	-0.13	-0.63
se	0.11	0.16	1.36	0.04	0.04	0.53	0.05	0.05	0.65	0.06	0.06	0.81

Note: $\hat{\theta}^{IV}_n$: IV estimates, $\hat{\theta}_n$: robust estimates, $\tilde{\theta}_n$: bias-corrected robust estimates, $\tilde{\tilde{\theta}}_n$: repeated bias-corrected robust estimates. $\hat{\nu}_n = 38.33$, 14.10 for $y = \log(\pi/100)$ and $\pi/100$, respectively. Estimates for θ_2 reported using log(pcinc)/100 as a regressor. Sample size $n = 114$.

The kurtosis indicates the log-transformed regression is less prone to outliers, the standard deviation suggests the estimates will be significantly less accurate. This reflects the

[12] Terra (1998, p647) classifies Argentina, Bolivia, Brazil, Peru, Mexico, Zaire as severely indebted.

[13] Singapore is the country with the lowest average inflation in the sample.

larger volatility of log-inflation compared to inflation. Also, the log transformation changes the interpretation of the coefficient θ_1 which may not be desirable. The following replicates the original results and estimates the regression in levels, as in Wooldridge (2002, Ch16), to get the desired coefficient interpretation.

Table 3 confirms that the log-transformed regression is less prone to outliers as the IV and robust estimates are very similar after bias-correction.[14] The non-transformed regression is, as Romer (1993) suspected, sensitive to some datapoints. Robust and bias-corrected estimates indicate IV overestimates the relationship between trade openness and inflation. Standard errors indicate the bias-corrected estimates are more accurate than the IV ones. The estimated effect is about one-third of the non-robust one. The bias correction adjusts the estimates by half to a full standard error. The full dataset of weights used to compute the estimates when $y = \pi/100$ are reported in Tables G10, G11, Appendix G.

6.2.2 Segregation and the Quality of Government

The third application considers the relationship between racial and religious discrimination and the quality of government. Alesina and Zhuravskaya (2011) constructed a new dataset on ethnic, linguistic, and religious segregation and fractionalization for a large number of countries. Mobility within a country, which determines segregation, can be endogenous to government quality. To address this particular issue, the authors predict segregation from neighboring country data. The main idea is that when a sub-population is at the border of the neighboring country, the same sub-group is more likely to be located near that border (see Alesina and Zhuravskaya, 2011, Figure 1, p1980). They illustrate using Switzerland as an example: most French speakers live near the French border, and Protestants are more commonly found near the German border. This is one of the papers surveyed in Young (2022), which finds that published IV regressions tend to be highly leveraged and sensitive to a few observations. The following revisits some of the main results in the original paper. The regression specification is given by:

$$\text{Rule of law}_i = \theta_0 + \theta_1 \text{Segregation}_i + \theta_2 \text{Fractionalization}_i + \text{Controls} + u_i,$$

where Segregation and Fractionalization are measured with respect to one of Ethnicity, Language, or Religion leading to three separate IV regressions. The controls are the same as in Table 6, Column 2 of Alesina and Zhuravskaya (2011, p1897). Fractionalization controls

[14]Estimates using a smaller $\nu = 12$ are nearly identical for the log regression (not reported here).

for group heterogeneity in each dimension (ethnicity, language, and religion) as measured by a Herfindahl index. If there is only one group in the population, the index is zero. If there are many equal-sized groups, the measure is closer to 1. See Alesina and Zhuravskaya (2011, pp1779-1780) for further details.

Table 4: Alesina and Zhuravskaya (2011): 10 Largest Contributors to $\hat{\theta}_{1n}$, $\hat{\theta}_{2n}$, Sample Moments, for each Measure of Segregation (ranked on coefficient θ_1)

Ethnicity			Language			Religion		
Country	θ_1	θ_2	Country	θ_1	θ_2	Country	θ_1	θ_2
Zimbabwe	-99.76	19.98	USA	142.04	-22.93	Kazakhstan	-139.56	-2.67
Israel	78.21	-9.10	Zimbabwe	-110.37	20.86	Uzbekistan	76.49	-1.62
Belgium	61.05	6.76	Austria	106.92	-7.21	Cambodia	53.48	7.31
Cote d'Ivoire	-53.93	1.56	Belgium	76.80	-0.48	Indonesia	-51.20	6.27
Guatemala	-32.69	3.02	Canada	-58.79	17.34	Switzerland	46.39	-4.31
Ecuador	-27.86	1.57	New Zealand	-58.21	-0.22	Netherlands	45.01	-1.37
UK	-27.12	-2.17	Togo	-44.16	6.20	CAR	-43.58	11.92
Tajikistan	-26.56	6.03	UK	-40.78	-6.84	Canada	-41.44	12.29
France	-25.16	1.47	Kyrgyzstan	-38.44	2.41	Kenya	-41.28	4.92
Spain	-25.12	9.11	Rwanda	-36.45	11.63	Israel	41.10	-4.39
Sample Moments			Sample Moments			Sample Moments		
Mean	-2.47	0.18	Mean	-1.80	0.31	Mean	-0.87	0.40
Stdev	19.11	5.08	Stdev	28.82	6.06	Stdev	30.53	5.85
Skewness	-0.52	0.74	Skewness	1.31	0.67	Skewness	-1.00	0.39
Kurtosis	12.63	5.11	Kurtosis	12.58	7.46	Kurtosis	7.17	4.21

Note: CAR = Central African Republic.

Table 4 shows the 10 highest contributors for the coefficient θ_1 in each regression as well as the sample moments of coefficient contribution. The regressions for ethnicity and language display somewhat heavy tails, as measured by the kurtosis. This indicates that the baseline results – estimated coefficients, standard errors, or both – may be sensitive to a few observations. Table 5 reports standard IV and robust estimates with(out) bias correction. Robust estimates tend to produce more precise inferences, as measured by standard errors. The baseline results indicate that both ethnic and language segregation have a significant, negative impact on the rule of law in a given country.

Robust results indicate that ethnic segregation if the only significant determinant of the rule of law. Unlike the previous example, the estimate implies a larger effect than standard IV. Notice that because the controls are correlated with the instrument, the direction of the change from standard to robust estimates does not necessarily coincide with the contribution to θ_1 in Table 4.[15] Diagnostics can inform if the results are sensitive to some observations but, as explained in Huber and Ronchetti (2011), are not a substitute for robust estimation. Although non-significant, the coefficient θ_2 for fractionalization does change from positive to negative in the first two regressions. Again, the full dataset of weights used in the three regressions is reported in Tables G12, G13, G14 of Appendix G.

Table 5: Alesina and Zhuravskaya (2011): IV, Robust and Bias-Corrected Estimates

	Ethnicity							
	$\hat{\theta}^{\text{IV}}_{1n}$	$\hat{\theta}^{\text{IV}}_{2}$	$\hat{\theta}_{1n}$	$\hat{\theta}_{2n}$	$\tilde{\theta}_{1n}$	$\tilde{\theta}_{2n}$	$\tilde{\tilde{\theta}}_{1n}$	$\tilde{\tilde{\theta}}_{2n}$
est	-2.47	0.18	-3.18	-0.13	-3.19	-0.13	-2.75	-0.09
se	0.60	0.24	0.33	0.10	0.39	0.14	0.38	0.17
	Language							
	$\hat{\theta}^{\text{IV}}_{1n}$	$\hat{\theta}^{\text{IV}}_{2}$	$\hat{\theta}_{1n}$	$\hat{\theta}_{2n}$	$\tilde{\theta}_{1n}$	$\tilde{\theta}_{2n}$	$\tilde{\tilde{\theta}}_{1n}$	$\tilde{\tilde{\theta}}_{2n}$
est	-1.80	0.31	-0.65	-0.21	-0.65	-0.22	-0.59	-0.20
se	0.80	0.24	0.23	0.06	0.35	0.08	0.54	0.11
	Religion							
	$\hat{\theta}^{\text{IV}}_{1n}$	$\hat{\theta}^{\text{IV}}_{2}$	$\hat{\theta}_{1n}$	$\hat{\theta}_{2n}$	$\tilde{\theta}_{1n}$	$\tilde{\theta}_{2n}$	$\tilde{\tilde{\theta}}_{1n}$	$\tilde{\tilde{\theta}}_{2n}$
est	-0.87	0.40	-0.04	-0.02	0.09	0.10	0.34	0.16
se	1.82	0.23	0.52	0.12	0.69	0.16	0.86	0.16

Note: $\hat{\theta}^{IV}_n$: IV estimates, $\hat{\theta}_n$: robust estimates, $\tilde{\theta}_n$: bias-corrected robust estimates, $\tilde{\tilde{\theta}}_n$: repeated bias-corrected robust estimates. $\hat{\nu}_n = 10.71, 8.55, 11.80$ and sample size $n = 97, 92, 78$ for Ethnicity, Language and Religion, respectively. Sample sizes vary because of missing values.

7 Conclusion

It is important to assess the robustness of empirical findings. Without symmetry restrictions, large differences between robust and non-robust estimates could be attributed to 1) improved resilience, or 2) significant asymmetry bias (or a combination of the two). This paper proposes a procedure with a simple asymptotic bias correction so that 2) is less likely. Reporting the implicit estimation weights makes the final results transparent and interpretable. This is illustrated in three empirical applications.

[15] If an outlier affects a coefficient on the controls and there is collinearity with the instrument, then robust estimates of θ_1 will change with the coefficients on the controls, as they are correlated. Diagnostics may not fully reflect the multivariate effect of the outliers. In addition, when there are multiple outliers, the direction of change depends on the combined effect of the outliers.

References

ALESINA, A. AND E. ZHURAVSKAYA (2011): "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, 101, 1872–1911.

ANDREWS, D. W. (1986): "Stability comparison of estimators," *Econometrica: Journal of the Econometric Society*, 1207–1235.

ANDREWS, I., M. GENTZKOW, AND J. M. SHAPIRO (2017): "Measuring the sensitivity of parameter estimates to estimation moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

ARMSTRONG, T. B. AND M. KOLESÁR (2021): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

BONHOMME, S. AND M. WEIDNER (2022): "Minimizing sensitivity to model misspecification," *Quantitative Economics*, 13, 907–954.

BOSQ, D. (1991): "Inégalité de Bernstein pour un processus mélangeant à temps discret ou continu," *Rapport technique 126, Université Pierre-et-Marie-Curie Paris.*

——— (1998): *Nonparametric statistics for stochastic processes: estimation and prediction*, vol. 110, Springer Science & Business Media.

CANTONI, E. AND E. RONCHETTI (2001): "Robust inference for generalized linear models," *Journal of the American Statistical Association*, 96, 1022–1030.

CARROLL, R. J. AND A. H. WELSH (1988): "A note on asymmetry and robustness in linear regression," *The American Statistician*, 42, 285–287.

CHRISTENSEN, T. AND B. CONNAULT (2023): "Counterfactual sensitivity and robustness," *Econometrica*, 91, 263–298.

ČÍŽEK, P. (2016): "Generalized method of trimmed moments," *Journal of Statistical Planning and Inference*, 171, 63–78.

COIBION, O. (2012): "Are the effects of monetary policy shocks big or small?" *American Economic Journal: Macroeconomics*, 4, 1–32.

COOK, R. D. (1977): "Detection of influential observation in linear regression," *Technometrics*, 19, 15–18.

COWELL, F. A. AND E. FLACHAIRE (2007): "Income distribution and inequality measurement: The problem of extreme values," *Journal of Econometrics*, 141, 1044–1072.

COWELL, F. A. AND M.-P. VICTORIA-FESER (1996): "Robustness properties of inequality measures," *Econometrica: journal of the Econometric Society*, 77–101.

DOUKHAN, P. (1994): *Mixing: Properties and Examples*, vol. 85, Springer.

EICHENBAUM, M. (1992): "Comment on 'Interpreting the macroeconomic time series facts:

The effects of monetary policy': by Christopher Sims," *European Economic Review*, 36, 1001–1011.

FOMC (1979): "TRANSCRIPT: FEDERAL OPEN MARKET COMMITTEE MEETING. October 6, 1979," .

Forneron, J.-J. and L. Zhong (2023): "Convexity Not Required: Estimation of Smooth Moment Condition Models," *arXiv preprint arXiv:2304.14386*.

Friedman, M. and A. J. Schwartz (1963): *A monetary history of the United States, 1867-1960*, vol. 9, Princeton University Press.

Hamilton, L. C. (1992): "How robust is robust regression?" *Stata Technical Bulletin*, 1.

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

Hill, J. B. and E. Renault (2010): "Generalized method of moments with tail trimming," *Dept. of Economics, University of North Carolina-Chapel Hill*.

Huber, P. and E. Ronchetti (2011): *Robust Statistics*, Wiley Series in Probability and Statistics, Wiley.

Huber, P. J. (1964): "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73 – 101.

Jaeckel, L. A. (1971): "Robust estimates of location: Symmetry and asymmetric contamination," *The Annals of Mathematical Statistics*, 42, 1020–1034.

Laforgue, P., G. Staerman, and S. Clémençon (2021): "Generalization bounds in the presence of outliers: a median-of-means study," in *International Conference on Machine Learning*, PMLR, 5937–5947.

Lecué, G. and M. Lerasle (2020): "Robust machine learning by median-of-means: Theory and practice," *The Annals of Statistics*, 48, 906 – 931.

MacKinnon, J. G. (2012): "Thirty years of heteroskedasticity-robust inference," in *Recent advances and future directions in causality, prediction, and specification analysis: Essays in honor of Halbert L. White Jr*, Springer, 437–461.

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera (2019): *Robust statistics: theory and methods (with R)*, John Wiley & Sons.

McDonald, J. B. and W. K. Newey (1988): "Partially adaptive estimation of regression models via the generalized t distribution," *Econometric theory*, 4, 428–457.

Monnet, E. (2014): "Monetary policy without interest rates: Evidence from France's Golden Age (1948 to 1973) using a narrative approach," *American Economic Journal: Macroeconomics*, 6, 137–169.

Müller, U. K. (2020): "A more robust t-test," *The Review of Economics and Statistics*, 1–46.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Newey, W. K. and D. G. Steigerwald (1997): "Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models," *Econometrica: Journal of the Econometric Society*, 587–599.

Pötscher, B. M. and D. Preinerstorfer (2023): "VALID HETEROSKEDASTICITY ROBUST TESTING," *Econometric Theory*, 1–53.

Rio, E. (1999): *Théorie asymptotique des processus aléatoires faiblement dépendants*, vol. 31, Springer Science & Business Media.

Rohatgi, D. and V. Syrgkanis (2022): "Robust Generalized Method of Moments: A Finite Sample Viewpoint," *Advances in Neural Information Processing Systems*, 35, 15970–15981.

Romer, C. D. and D. H. Romer (1989): "Does monetary policy matter? A new test in the spirit of Friedman and Schwartz," *NBER macroeconomics annual*, 4, 121–170.

——— (2004): "A new measure of monetary shocks: Derivation and implications," *American economic review*, 94, 1055–1084.

Romer, D. (1993): "Openness and inflation: theory and evidence," *The quarterly journal of economics*, 108, 869–903.

Ronchetti, E. and F. Trojani (2001): "Robust inference with GMM estimators," *Journal of econometrics*, 101, 37–69.

Rusnák, M., T. Havranek, and R. Horváth (2013): "How to solve the price puzzle? A meta-analysis," *Journal of Money, Credit and Banking*, 45, 37–70.

Sasaki, Y. and Y. Wang (2023): "Diagnostic testing of finite moment conditions for the consistency and root-N asymptotic normality of the GMM and M estimators," *Journal of Business & Economic Statistics*, 41, 339–348.

Sims, C. A. (1992): "Interpreting the macroeconomic time series facts: The effects of monetary policy," *European Economic Review*, 36, 975–1000.

St Laurent, R. T. and R. D. Cook (1992): "Leverage and superleverage in nonlinear regression," *Journal of the American Statistical Association*, 87, 985–990.

Stock, J. H. and M. W. Watson (2001): "Vector autoregressions," *Journal of Economic perspectives*, 15, 101–115.

Terra, C. T. (1998): "Openness and inflation: a new assessment," *The Quarterly Journal* *of Economics*, 113, 641–648.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, vol. 126, Springer New York.

Vershynin, R. (2018): *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press.

Vladimirova, M., S. Girard, H. Nguyen, and J. Arbel (2020): "Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions," *Stat*, 9, e318.

White, H. (2001): *Asymptotic Theory for Econometricians*, Economic Theory, Econometrics, Emerald Group Publishing Limited.

Wooldridge, J. M. (2002): *Introductory econometrics: A modern approach*, Cengage learning.

Young, A. (2022): "Consistency without inference: Instrumental variables in practical application," *European Economic Review*, 147, 104112.

Zhou, W.-X., K. Bose, J. Fan, and H. Liu (2018): "A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing," *Annals of statistics*, 46, 1904.

Appendix A Preliminary Results

Lemma A1. *Let* $q_t(\psi;\theta) = (\nu+p)\log(1+\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}/\nu)$*. For all* $\theta \in \Theta$*:* $\sup_{\psi\in\Psi_n}\|\partial_\mu q_t(\psi;\theta)\| \leq s_0^{-1/2}(1+p/\nu)\nu^{1/2}$, $\sup_{\psi\in\Psi_n}\|\partial_\Sigma q_t(\psi;\theta)\| \leq \nu\left(\frac{\nu^2(1+p/\nu)}{\kappa_2}+\frac{\nu^4(1+p/\nu)^2}{4\kappa_1\kappa_2}+\frac{p\nu}{\kappa_2}\right)^3$.

Lemma A2. *Suppose* $z_t \sim P$ *satisfying Assumption 2, for* $t \in \{1,\dots,n\}$*, take* $1 \leq \nu \leq n$*. Let:*

$$\overline{\Delta}_n(\psi;\theta) = \frac{1}{n}\sum_{t=1}^n\left(\log\left(1+\frac{\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}}{\nu}\right)-\mathbb{E}_P\left[\log\left(1+\frac{\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}}{\nu}\right)\right]\right)$$

for any $\theta,\psi \in \Theta\times\Psi_n$*.*

1) If z_t *are iid, then there exists a constant* $L>0$ *which depends on* $s_0,\kappa_1,\kappa_2,M_2,M_4$ *such that for all* $t \geq 0$*:*

$$\mathbb{P}\left(\sup_{\theta\in\Theta,\psi\in\Psi_n}|\overline{\Delta}_n(\psi,\theta)| \geq L\log(1+p\nu)\left[\sqrt{\frac{t}{n}}+\frac{t}{n}+\sqrt{\frac{C_n}{n}}+\frac{C_n}{n}\right]\right) \leq 4\exp(-t), \quad \text{(A.1)}$$

where $C_n = 1+(k+2p^2)[\log(p)+\log(\nu)+\log(n)]$*.*

2) If z_t *is strictly stationary with mixing coefficient* $\beta_m \leq a\exp(-bm)$ *for* $a,b>0$*, then for another constant* $\tilde{L}>0$ *which further depends on* a,b *such that:*

$$\mathbb{P}\left(\sup_{\theta\in\Theta,\psi\in\Psi_n}|\overline{\Delta}_n(\psi,\theta)| \geq \tilde{L}\log(1+p\nu)\left[\sqrt{\frac{(t+C_n)t}{n}}+\frac{(t+C_n)t}{n}+\sqrt{\frac{C_n}{n}}+\frac{C_n}{n}\right]\right) \quad \text{(A.1')}$$
$$\leq 12\exp(-t),$$

for the same C_n *as 1).*

Appendix B Proofs for the Main Results

Proof of Lemma 1. Note that $Q_n(\psi) \to +\infty$ when $\text{trace}(\Sigma) \to +\infty$ so the solution is s.t. $\text{trace}(\hat{\Sigma}_n) < +\infty$, likewise $\|\hat{\mu}_n\| < \infty$. The first-order condition (foc) wrt μ implies:

$$-\frac{\nu+p}{\nu n}\sum_{t=1}^n\frac{\hat{\Sigma}_n^{-1}(g(Z_t;\theta)-\hat{\mu}_n)}{1+\|g(Z_t;\theta)-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu}+\frac{\kappa_1}{\nu}\hat{\Sigma}_n^{-1}\hat{\mu}_n = 0.$$

Pre-multiply by $\Sigma_n^{1/2}$ and re-arrange terms to find:

$$\|\hat{\Sigma}_n^{-1/2}\hat{\mu}_n\| \leq \frac{\nu}{\kappa_1}(1+p/\nu)\max_t \frac{\|g(Z_t;\theta)-\hat{\mu}_n\|_{\hat{\Sigma}_n^{-1}}}{1+\|g(Z_t;\theta)-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu},$$

where $\max_{x\geq 0}\frac{x}{1+x^2/\nu}=\sqrt{\nu}/2$ yields the desired inequality. Take the foc wrt to Σ^{-1}:

$$\frac{\nu+p}{\nu n}\sum_{t=1}^n \frac{(g(Z_t;\theta)-\hat{\mu}_n)(g(Z_t;\theta)-\hat{\mu}_n)'}{1+\|g(Z_t;\theta)-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu}+\frac{\kappa_1}{\nu}\hat{\mu}_n\hat{\mu}_n'-\frac{\kappa_2}{\nu}\Sigma_n^2-\Sigma_n=0.$$

Pre and post-multiply by $\Sigma_n^{-1/2}$, re-arrange terms and compute the trace to find:

$$\text{trace}(\hat{\Sigma}_n)\leq \frac{\nu}{\kappa_2}\left((1+p/\nu)\max_t \frac{\|g(Z_t;\theta)-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}}{1+\|g(Z_t;\theta)-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu}+\frac{\kappa_1}{\nu}\|\hat{\Sigma}_n^{-1/2}\hat{\mu}_n\|^2+p\right).$$

The max is bounded above by $\sup_{x\geq 0}\frac{x^2}{1+x^2/\nu}=\nu$. Plug-in the bound for $\|\hat{\Sigma}_n^{-1/2}\hat{\mu}_n\|$ to get the desired inequality. □

Proof of Proposition 1. First, note that $\psi(\theta;\nu)\in\Psi_n$ for all $\theta\in\Theta$. By minimization, we have for all $\theta\in\Theta$:

$$\begin{aligned}0\leq Q_\nu(\hat{\psi}_n(\theta;\nu);\theta)-Q_\nu(\psi(\theta;\nu);\theta) &= \underbrace{Q_n(\hat{\psi}_n(\theta;\nu);\theta)-Q_n(\psi(\theta;\nu);\theta)}_{\leq 0}\\ &\quad+(Q_\nu-Q_n)(\hat{\psi}_n(\theta;\nu);\theta)-(Q_\nu-Q_n)(\psi(\theta;\nu);\theta)\\ &\leq 2\sup_{\theta\in\Theta,\psi\in\Psi_n}|(Q_\nu-Q_n)(\psi;\theta)|,\end{aligned}$$

where $Q_n-Q_\nu=(\nu+p)\overline{\Delta}_n$ used in Lemma A2. There are two bounds to derive: one for the n_o outliers and another for the remaining n_P observations. For any $z\in\mathcal{O}_n$, $\psi\in\Psi_n$, $1\leq\nu\leq n$:

$$0\leq\log(1+\|g(z;\theta)-\mu\|^2_{\Sigma^{-1}}/\nu)\leq\log(1+3s_0^{-1}A^2n^{2\alpha}/\nu)+\log(1+3/2\kappa_1^{-1}\nu^{1/2}).$$

We also have $Q_\nu=n_o/nQ_\nu+n_P/nQ_\nu$, the second is the centering term for well-behaved observations. We need to bound the first:

$$0\leq(\nu+p)\mathbb{E}_P[\log(1+\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}/\nu)]\leq 3(1+p/\nu)s_0^{-1}M_2+(\nu+p)\log(1+3/2\kappa_1^{-1}\nu^{1/2}),$$

for any $(\theta, \psi) \in \Theta \times \Psi_n$, using $\log(1+x) \leq x$ for $x \geq 0$, Assumption 2 and Lemma 1. Combine the two bounds to find:

$$2\Big|\frac{\nu+p}{n}\sum_{t=n_P+1}^{n}\log(1+\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}/\nu) - \frac{n_o}{n}Q_\nu(\psi;\theta)\Big| \leq C_{\mathcal{O}}\frac{n_o(\nu+p)}{n}[1+\log(n)],$$

where $C_{\mathcal{O}}$ only depends on $s_0, M_2, \kappa_1, A, \alpha$. Define Q_{n_P} to be the sample average over the n_P uncontaminated observations, $Q_{n_P} - Q_\nu = (\nu+p)\overline{\Delta}_{n_P}$ which satisfies the conditions of Lemma A2. Pre-multiply by n_P/n to get the uncontaminated part of $Q_n - Q_\nu$ and multiply by 2. Replace $L, \tilde{L}$ from Lemma A2 with $2L, 2\tilde{L}$ to get the desired result. □

Proof of Corollary 1. Proceed in several steps: 1) show uniform convergence under the pseudo-distance Q_ν and that it implies some compactness restrictions, 2) derive a norm equivalence on compact sets, 3) combine these two steps with a uniform convergence for $\|\psi(\theta;\nu) - \psi(\theta;\infty)\|$ as $\nu \to \infty$.

Step 1. Uniform convergence is implied by Proposition 1 and the rate conditions. The following shows that this implies: $\sup_{\theta\in\Theta}\|\hat{\psi}_n(\theta;\nu)\| \leq K$ with probability approaching 1 (wpa1), for some constant $K > 0$. Then, all pairs $(\hat{\psi}_n(\theta;\nu), \psi(\theta;\nu))_{\theta\in\Theta}$ will be in a bounded compact subset of Ψ wpa1. First, note that for $1 \leq \nu$:

$$Q_\nu(\psi;\theta) \leq (1+p)\mathbb{E}_P\left[\|g(z_t;\theta)\|^2_{\Sigma^{-1}})\right] + \log|\Sigma| + \kappa_1\|\mu\|^2_{\Sigma^{-1}} + \kappa_2\text{trace}(\Sigma),$$

which implies that $\sup_{\theta\in\Theta,\nu\geq 1}(\inf_{\psi\in\Psi} Q_\nu(\psi;\theta)) \leq K_1$ for some constant K_1 which is less or equal to the largest (over θ) minimal (over ψ) value of the upper bound which is finite by compactness, continuity and strict convexity, wrt ψ, of the upper bound.

$$Q_\nu(\psi;\theta) \geq \mathbb{E}_P\left[(\nu+p)\log(1+\|g(z_t;\theta)\|^2_{\Sigma^{-1}}/\nu)\right] + \log|\Sigma| \geq \log(\lambda_{\max}(\Sigma)) \geq 2K_1,$$

for any θ, ν, μ as soon as $\lambda_{\max}(\Sigma) \geq \exp(2K_1) := s_1$. Assumption 2 ii and compactness of Θ implies that:

$$\|\mu(\theta;\infty)\| = \|\mathbb{E}_P[g(z_t;\theta)]\| \leq K_2,$$

for some constant K_2 which depends on M_2, M_4 and diam(Θ). In addition, for any $M > 0$, Chebychev's inequality implies:

$$\sup_{\theta\in\Theta}\mathbb{P}\left(\|g(z_t;\theta) - \mu(\theta;\infty)\| \geq M\right) \leq M_2/M := \varepsilon > 0.$$

For $\lambda_{\max}(\Sigma) \leq s_1$ above, this implies for any $\theta \in \Theta$ and all $\|\mu\| \geq 2M + K_2$:

$$Q_\nu(\psi;\theta) \geq (\nu+p)(1-\varepsilon)\log(1+s_1^{-1}M/\nu)+p\log(s_0) \geq (1+p)\frac{(1-\varepsilon)s_1^{-1}M}{1+s_1^{-1}M/\nu}+p\log(s_0) \geq 2K_1,$$

for M and $\nu \geq \underline{\nu} \geq 1$ sufficiently large.

The uniform convergence then implies that $\sup_{\theta\in\Theta}\|Q_\nu(\hat{\psi}_n(\theta;\nu);\theta)\| \leq \sup_{\theta\in\Theta}\|Q_\nu(\psi(\theta;\nu);\theta)\|+o_p(1) \leq 2K_1$, wpa1. This implies that $\sup_{\theta\in\Theta}\|\hat{\mu}_n(\theta;\nu)\| \leq K_2+2M$ and $\sup_{\theta\in\Theta}\|\lambda_{\max}(\hat{\Sigma}_n(\theta;\nu))\| \leq \exp(2K_1)$ wpa1, which implies the desired result. The same holds for $\psi(\theta;\nu)$.

Step 2. First, for any $x \geq 0$ we have $\frac{x}{1+x} \leq \log(1+x) \leq x$ which implies $|\log(1+x)-x| \leq \frac{x^2}{1+x}$. Take $(\theta,\psi) \in \Theta\times\Psi$, this implies:

$$\begin{aligned}&\left|\mathbb{E}_p\left[(\nu+p)\log(1+\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}/\nu)-\frac{\nu+p}{\nu}\|g(z_t;\theta)-\mu\|^2_{\Sigma^{-1}}\right]\right|\\&\leq \frac{\nu+p}{\nu^2}\mathbb{E}_P\left[\|g(z_t;\theta)-\mu\|^4_{\Sigma^{-1}}\right] \leq 9s_0^{-2}\frac{\nu+p}{\nu^2}\left[M_4+\|\mu\|^4\right],\end{aligned}$$

using Assumption 2 iii. to bound the 4th moment. This implies that $\sup_{\theta\in\Theta}|Q_\nu(\psi;\theta)-Q_\infty(\psi;\theta)| \leq O(\nu^{-1})$ with respect to ψ on bounded compact sets.

Step 3. Given that $\sup_{\theta\in\Theta}(\|\hat{\psi}_n(\theta;\nu)\|+\|\psi(\theta;\nu)\|) \leq 2K$ from Step 1, Step 2 and the triangular inequality imply:

$$\sup_{\theta\in\Theta}|Q_\infty(\hat{\psi}_n(\theta;\nu);\theta)-Q_\infty(\psi(\theta;\nu);\theta)| = o_p(1).$$

Note that Q_∞ is the Gaussian negative log-likelihood which is strictly convex for each $\theta \in \Theta$, so this also implies $\|\hat{\psi}_n(\theta;\nu)-\psi(\theta;\nu)\| = o_p(1)$ uniformly in θ. Since we are actually interested in $\psi(\theta;\infty)$:

$$\begin{aligned}0 &\leq \sup_{\theta\in\Theta}\{Q_\infty(\psi(\theta;\nu);\theta)-Q_\infty(\psi(\theta;\infty);\theta)\} \leq \sup_{\theta\in\Theta}\{\underbrace{Q_\nu(\psi(\theta;\nu);\theta)-Q_\nu(\psi(\theta;\infty);\theta)}_{\leq 0}\}\\&+\sup_{\theta\in\Theta}[Q_\infty(\psi(\theta;\nu);\theta)-Q_\nu(\psi(\theta;\nu);\theta)-Q_\infty(\psi(\theta;\infty);\theta)+Q_\nu(\psi(\theta;\infty);\theta)]\\&\leq O(\nu^{-1}),\end{aligned}$$

using Step 2 and the compactness from Step 1. This implies the uniform convergence result $\|\hat{\psi}_n(\theta;\nu)-\psi(\theta;\infty)\| = o_p(1)$. □

Proof of Proposition 2. The foc wrt $\hat{\mu}_n(\theta;\nu)$ reads (the dependence on θ, ν is omitted for brievety):

$$\frac{1}{n}\sum_{t=1}^{n} \frac{x_{t,\theta} - \hat{\mu}_n}{1 + \|x_{t,\theta} - \hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu} + \kappa_1 \frac{\hat{\mu}_n}{\nu} = 0,$$

where $x_{t,\theta} = g(z_t;\theta)$ as in the proof of Lemma A2. Re-arrange terms to find:

$$\hat{\mu}_n = \underbrace{\frac{1}{n_P}\sum_{t=1}^{n_P} x_{t,\theta}}_{(A)} - \underbrace{\frac{1}{\nu n_P}\sum_{t=1}^{n_P} \frac{(x_{t,\theta} - \hat{\mu}_n)\|x_{t,\theta} - \hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}}{1 + \|x_{t,\theta} - \hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu}}_{(B)} + \underbrace{\frac{\kappa_1 n}{n_P}\frac{\hat{\mu}_n}{\nu}}_{(C)} + \underbrace{\frac{1}{n_P}\sum_{t>n_P} \frac{\Sigma_n^{1/2}\Sigma_n^{-1/2}(x_{t,\theta} - \hat{\mu}_n)}{1 + \|x_{t,\theta} - \hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu}}_{(D)},$$

where $(A) = \overline{g}_{n_P}(\theta)$ and $\|(C)\| = O_p(\nu^{-1})$ uniformly in θ when $n_P/n \to 1$ using Corollary 1. Then, we have:

$$\begin{aligned}
\sup_{\theta\in\Theta}\|(B)\| &\leq (\sup_{\theta\in\Theta}\lambda_{\max}(\Sigma(\theta;\infty)) + o_p(1))^{-2}\frac{1}{\nu n_P}\sum_{t=1}^{n_P} 8\left(\sup_{\theta\in\Theta}\|x_{t,\theta}\|^3 + \sup_{\theta\in\Theta}\|\hat{\mu}_n\|^3\right)\\
&\leq (\sup_{\theta\in\Theta}\lambda_{\max}(\Sigma(\theta;\infty)) + o_p(1))^{-2}\frac{1}{\nu n_P}\sum_{t=1}^{n_P}\left(64\|x_{t,\theta_0}\|^3 + 64\text{diam}(\Theta)^3 G_t^3 + 8\sup_{\theta\in\Theta}\|\hat{\mu}_n\|^3\right)\\
&= O_p(\nu^{-1}),
\end{aligned}$$

by uniform consistency of $\hat{\mu}_n$ and a strong law of large numbers applied to the sample mean of $\|x_{t\theta_0}\|^3 + G_t^3$ (White, 2001, Cor3.48). We also have:

$$\sup_{\theta\in\Theta}\|(D)\| \leq \left[\sup_{\theta\in\Theta}\lambda_{\max}(\Sigma(\theta;\infty)) + o_p(1)\right]^{1/2}\frac{\sqrt{\nu}n_o}{2n_P} = o(n^{-1/2}),$$

if $n_o = o(\sqrt{\nu/n})$. Corollary 1 required $\nu = o(\sqrt{n})$, this yields the first result:

$$\sup_{\theta\in\Theta}\|\hat{\mu}_n(\theta;\nu) - \overline{g}_{n_P}(\theta)\| = O_p\left(\max\left[\nu^{-1}, \frac{\sqrt{\nu}n_o}{n}\right]\right).$$

To derive results for the bias-corrected estimates, we additionally need convergence rates for $\hat{\Sigma}_n$, take the foc wrt Σ^{-1} and re-arrange terms:

$$\begin{aligned}
\hat{\Sigma}_n = \frac{1}{n}\sum_{t=1}^{n_P}&(x_{t,\theta}-\hat{\mu}_n)(x_{t,\theta}-\hat{\mu}_n)' && \text{(A)}\\
&-\frac{p}{\nu n}\sum_{t=1}^{n_P}(x_{t,\theta}-\hat{\mu}_n)(x_{t,\theta}-\hat{\mu}_n)' && \text{(B)}\\
&-\frac{\nu+p}{\nu^2 n}\sum_{t=1}^{n_P}\frac{(x_{t,\theta}-\hat{\mu}_n)(x_{t,\theta}-\hat{\mu}_n)'\|x_{t,\theta}-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}}{1+\|x_{t,\theta}-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu} && \text{(C)}\\
&+\frac{\nu+p}{\nu n}\sum_{t>n_P}\frac{\hat{\Sigma}_n^{1/2}\hat{\Sigma}_n^{-1/2}(x_{t,\theta}-\hat{\mu}_n)(x_{t,\theta}-\hat{\mu}_n)'\hat{\Sigma}_n^{-1/2}\hat{\Sigma}_n^{1/2}}{1+\|x_{t,\theta}-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu} && \text{(D)}\\
&+\kappa_1\frac{\hat{\mu}_n\hat{\mu}_n'}{\nu}-\kappa_2\frac{\hat{\Sigma}_n^2}{\nu}, && \text{(E)}
\end{aligned}$$

where $\sup_{\theta\in\Theta}\|(E)\| = O_p(\nu^{-1})$ by uniform convergence. $\sup_{\theta\in\Theta}\|(B)\| = O_p(\nu^{-1})$ by applying a uniform law of large numbers to $x_{t,\theta}, x_{t,\theta}^2$ and uniform convergence of $\hat{\mu}_n$. Then, we have:

$$\sup_{\theta\in\Theta}\|(C)\| \leq (\sup_{\theta\in\Theta}\lambda_{\max}(\Sigma(\theta;\infty))+o_p(1))^{-2}\frac{1+p}{\nu n_P}\sum_{t=1}^{n_P}16(\|x_{t,\theta}\|^4+\|\hat{\mu}_n\|^4) = O_p(\nu^{-1}),$$

using a strong law of large numbers for $\|x_{t,\theta_0}\|^4, G_t^4$, as in the bound on (B) for $\hat{\mu}_n$ above. Finally, $\sup_{\theta\in\Theta}\|(D)\| \leq (\sup_{\theta\in\Theta}\lambda_{\max}(\Sigma(\theta;\infty))+o_p(1))\nu(1+p)\frac{n_o}{n} = O_p\left(\frac{\nu n_o}{n}\right)$. Importantly, we also have:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n_P}\left[(x_{t,\theta}-\hat{\mu}_n(\theta;\nu))(x_{t,\theta}-\hat{\mu}_n(\theta;\nu))' - (x_{t,\theta}-\hat{\mu}_n(\theta;\nu/2))(x_{t,\theta}-\hat{\mu}_n(\theta;\nu/2))'\right]\\
&= O_p\left(\max\left[\nu^{-1}, \frac{\sqrt{\nu}n_o}{n}\right]\right),
\end{aligned}$$

since $\hat{\mu}_n(\theta,\nu)-\hat{\mu}_n(\theta,\nu/2) = O_p(\max\left[\nu^{-1}, \frac{\sqrt{\nu}n_o}{n}\right])$ uniformly in θ. This implies that $\hat{\Sigma}_n(\theta;\nu)-\hat{\Sigma}_n(\theta;\nu/2) = O_p(\max[\nu^{-1}, \frac{\nu n_o}{n}])$ uniformly in θ. We now have all the ingredients to expand the bias-corrected estimates $\tilde{\mu}_n(\theta;\nu) = 2\hat{\mu}_n(\theta;\nu) - \hat{\mu}_n(\theta;\nu/2)$, omit their dependence on θ:

$$\begin{aligned}
&\tilde{\mu}_n(\nu) \\
&= \frac{2}{n_P}\sum_{t=1}^{n_P} x_{t,\theta} - \frac{1}{n_P}\sum_{t=1}^{n_P} x_{t,\theta} && \text{(A)} \\
&- \frac{2}{\nu n_P}\sum_{t=1}^{n_P}\left[\frac{(x_{t,\theta}-\hat{\mu}_n(\nu))\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}}{1+\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}/\nu} - \frac{(x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2}))\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}}{1+2\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}/\nu}\right] && \text{(B)} \\
&+ 2\kappa_1\frac{\hat{\mu}_n(\nu)-\hat{\mu}_n(\frac{\nu}{2})}{\nu} && \text{(C)} \\
&+ \frac{2}{n_P}\sum_{t>n_P}\frac{\Sigma_n^{1/2}(\nu)\Sigma_n^{-1/2}(\nu)(x_{t,\theta}-\hat{\mu}_n(\nu))}{1+\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}/\nu} - \frac{1}{n_P}\sum_{t>n_P}\frac{\Sigma_n^{1/2}(\frac{\nu}{2})\Sigma_n^{-1/2}(\frac{\nu}{2})(x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2}))}{1+2\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}/\nu}. && \text{(D)}
\end{aligned}$$

Clearly $(A) = \overline{g}_{n_P}(\theta)$ and $\|(D)\| \leq O_p(\frac{\sqrt{\nu}n_o}{n})$ uniformly in $\theta \in \Theta$ as previously shown. Likewise, $\|(C)\| \leq O_p(\max[\nu^{-2}, \frac{n_o}{\sqrt{\nu}n}]) \leq O_p(\max[\nu^{-2}, \frac{\sqrt{\nu}n_o}{n}])$, uniformly.

Remains to bound the longer term:

$$\begin{aligned}
(B) &= \frac{-2}{\nu n_p}\sum_{t=1}^{n_P}\frac{(x_{t,\theta}-\hat{\mu}_n(\nu))\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)} - (x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2}))\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}}{(1+\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}/\nu)(1+2\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}/\nu)} && \text{(B1)} \\
&+ \frac{2}{\nu^2 n_p}\sum_{t=1}^{n_P}\frac{(x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2}))\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}}{(1+\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}/\nu)(1+2\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}/\nu)} && \text{(B2)} \\
&- \frac{2}{\nu^2 n_p}\sum_{t=1}^{n_P}\frac{(x_{t,\theta}-\hat{\mu}_n(\nu))\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}}{(1+\|x_{t,\theta}-\hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)}/\nu)(1+2\|x_{t,\theta}-\hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}/\nu)}, && \text{(B3)}
\end{aligned}$$

where $\|(B2),(B3)\| = O_p(\nu^{-2})$ using a uniform of large numbers for $\|x_{t,\theta}\|^5$ and uniform convergence of $\hat{\mu}_n(\nu), \hat{\mu}_n(\nu/2)$. The last step is to show that the numerator in $(B1)$ is a

$O_p(\max[\nu^{-1}, \frac{\sqrt{\nu}n_o}{n}])$, let $\delta_n = \nu^{-1} + \frac{\nu n_o}{n}$:

$$
\begin{aligned}
&\|(x_{t,\theta} - \hat{\mu}_n(\nu))\|x_{t,\theta} - \hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)} - (x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2}))\|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}\| \\
&\leq O_p(\delta_n)\|x_{t,\theta} - \hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)} + \|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\| \left[\|x_{t,\theta} - \hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)} - \|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|^2_{\hat{\Sigma}_n^{-1}(\frac{\nu}{2})}\right] \\
&\leq O_p(\delta_n)\|x_{t,\theta} - \hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)} + O_p(\delta_n)s_0^{-2}\|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|^3 \\
&+ \|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|s_0^{-2}\left[\|\hat{\mu}_n(\nu) - \hat{\mu}_n(\frac{\nu}{2})\| \times \|2x_{t,\theta} - \hat{\mu}_n(\nu) - \hat{\mu}_n(\frac{\nu}{2})\|\right] \\
&\leq O_p(\delta_n)\Bigg(\|x_{t,\theta} - \hat{\mu}_n(\nu)\|^2_{\hat{\Sigma}_n^{-1}(\nu)} + s_0^{-2}\|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|^3 \\
&+ s_0^{-2}\|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|(\|x_{t,\theta} - \hat{\mu}_n(\nu)\| + \|x_{t,\theta} - \hat{\mu}_n(\frac{\nu}{2})\|)\Bigg).
\end{aligned}
$$

Apply a uniform law of large numbers to $\|x_{t,\theta}\|^2, \|x_{t,\theta}\|^3$, and invoke uniform convergence of $\hat{\mu}_n(\nu), \hat{\mu}_n(\nu/2)$ to get since the denominator in $(B1)$ is less or equal than 1: $\sup_{\theta\in\Theta}\|(B1)\| \leq O_p(\nu^{-1}\delta_n) = O_p(\max[\nu^{-2}, \frac{\sqrt{\nu}n_o}{n}])$ as desired. Putting everything together, we get the desired result:

$$
\sup_{\theta\in\Theta}\|\tilde{\mu}_n(\theta;\nu) - \overline{g}_{n_P}(\theta)\| \leq O_p\left(\max\left[\nu^{-2}, \frac{\sqrt{\nu}n_o}{n}\right]\right).
$$

□

Proof of Theorem 1. By definition: $\|\tilde{\mu}_n(\tilde{\theta}_n)\|^2_{W_n} \leq \inf_{\theta\in\Theta}\|\tilde{\mu}_n(\theta)\|^2_{W_n} + o_p(n^{-1})$. Proposition 2 implies that, uniformly in $\theta \in \Theta$:

$$
\|\overline{g}_{n_P}(\theta)\|_{W_n} - o_p(n^{-1/2}) \leq \|\tilde{\mu}_n(\theta)\|_{W_n} \leq \|\overline{g}_{n_P}(\theta)\|_{W_n} + o_p(n^{-1/2}).
$$

In particular the asymptotic equivalence and approximate minimization properties imply:

$$
\|\overline{g}_{n_P}(\tilde{\theta}_n)\|_{W_n} \leq \|\tilde{\mu}_n(\tilde{\theta}_n)\|_{W_n} + o_p(n^{-1/2}) \leq \|\tilde{\mu}_n(\hat{\theta}_n)\|_{W_n} + o_p(n^{-1/2}) \leq \|\overline{g}_{n_P}(\hat{\theta}_{n_P})\|_{W_n} + o_p(n^{-1/2}),
$$

which implies that $\tilde{\theta}_n$ is an approximate minimizer of $\|\overline{g}_{n_P}(\cdot)\|_{W_n}$. Assumption 3 then implies continuity and asymptotic normality for both $\tilde{\theta}_n$ and $\hat{\theta}_{n_P}$, e.g. Newey and McFadden (1994, Th2.6, Th7.2) in the iid setting. The results then follow from a first-order expansion of the two estimators, e.g.: $\sqrt{n_P}(\tilde{\theta}_n - \theta_0) = -(G'WG)^{-1}G'W\overline{g}_{n_P}(\theta_0) + o_p(1)$. □

Proof of Proposition 3. The weighted average representation follows from the first-order condition $\partial_\mu Q_n(\hat{\psi}_n;\theta)=0$, which can be re-written as:

$$0=\frac{1+p/\nu}{n}\sum_{t=1}^{n}\frac{\hat{\mu}_n-g(z_t;\theta)}{1+\|g(z_t;\theta)-\hat{\mu}_n\|^2_{\hat{\Sigma}_n^{-1}}/\nu}+\frac{\kappa_1}{\nu}\hat{\mu}_n.$$

Re-arrange terms to find $\hat{\mu}_n=\sum_{t=1}^n \omega_t(\theta;\nu)g(z_t;\theta)$ as in the Proposition. Since $\tilde{\mu}_n(\theta;\nu)=2\hat{\mu}_n(\theta;\nu)-\hat{\mu}_n(\theta;\nu/2)$ we have also have $\tilde{\mu}_n(\theta;\nu)=\sum_{t=1}^n[2\omega_t(\theta;\nu)-\omega_t(\theta;\nu/2)]g(z_t;\theta)$.

Let $\bar{\omega}_n(\theta;\nu)=(1+p/\nu)/n\sum_{t=1}^n[1+q_t/\nu]^{-1}+\kappa_1/\nu$, we have:

$$\begin{aligned}\bar{\omega}_n(\theta;\nu)\hat{\Sigma}_{n,\omega}(\theta)&=\sum_{t=1}^{n}\frac{1+p/\nu}{n}\frac{\hat{\varepsilon}_t\hat{\varepsilon}_t'}{1+\|\hat{\varepsilon}_t\|^2_{\hat{\Sigma}_n^{-1}}/\nu}\\&=\underbrace{\sum_{t=1}^{n_P}\frac{1+p/\nu}{n}\hat{\varepsilon}_t\hat{\varepsilon}_t'}_{(A)}-\underbrace{\sum_{t=1}^{n_P}\frac{1+p/\nu}{n\nu}\frac{\hat{\varepsilon}_t\hat{\varepsilon}_t'\|\hat{\varepsilon}_t\|^2_{\hat{\Sigma}_n^{-1}}}{1+\|\hat{\varepsilon}_t\|^2_{\hat{\Sigma}_n^{-1}}/\nu}}_{(B)}+\underbrace{\sum_{t=n_P+1}^{n}\frac{1+p/\nu}{n}\frac{\hat{\varepsilon}_t\hat{\varepsilon}_t'}{1+\|\hat{\varepsilon}_t\|^2_{\hat{\Sigma}_n^{-1}}/\nu}}_{(C)}\\&\overset{p}{\to}\Sigma(\theta).\end{aligned}$$

To get the result, note that $\|(C)\|\leq\lambda_{\max}(\hat{\Sigma}_n)(1+p/\nu)n_o\nu/n=o_p(1)$. Likewise, $\|(B)\|\leq\nu^{-1}(1+p/\nu)s_0^{-1}[1/n\sum_{t=1}^{n_P}\|\hat{\varepsilon}_t\|^4]=O_p(\nu^{-1})$ using $\hat{\mu}_n\overset{p}{\to}\mathbb{E}_P(g(z_t;\theta))$ and a law of large numbers for $\|g(z_t;\theta)\|^4$. Similarly, a law of large numbers implies $(A)\overset{p}{\to}\Sigma(\theta)$. Using $q_t\geq 0$, we have $\bar{\omega}_n(\theta;\nu)=(1+p/\nu)/n\sum_{t=1}^{n_P}1-(1+p/\nu)/[n\nu]\sum_{t=1}^{n_P}q_t/[1+q_t/\nu]+(1+p/\nu)/n\sum_{t=n_P+1}^{n}[1+q_t/\nu]^{-1}$. The first term converges to 1, the second term is a $O_p(\nu^{-1})$ using a law of large numbers, and the third term is less or equal than $n_o/n(1+p/\nu)=o(1)$. This implies $\bar{\omega}_n(\theta;\nu)\overset{p}{\to}1$. Combine the results to find $\hat{\Sigma}_{n,\omega}(\theta)\overset{p}{\to}\Sigma(\theta)$.

To prove consistency for $\tilde{\Sigma}_{n,\omega}(\theta)$, we will first prove consistency for $\sum_{t=1}^n\omega_t(\theta;\nu)\tilde{\varepsilon}_t(\theta)\tilde{\varepsilon}_t(\theta)'$ and $\sum_{t=1}^n\omega_t(\theta;\nu/2)\tilde{\varepsilon}_t(\theta)\tilde{\varepsilon}_t(\theta)'$. Since $\tilde{\Sigma}_{n,\omega}(\theta)$ equals two times the first minus the second, consistency follows. First, note that $\tilde{\varepsilon}_t=\hat{\varepsilon}_t+\hat{\mu}_n-\tilde{\mu}_n$, where $\hat{\mu}_n-\tilde{\mu}_n=o_p(1)$ by Corollary 1.

$$\begin{aligned}\sum_{t=1}^n\omega_t(\theta;\nu)\tilde{\varepsilon}_t(\theta)\tilde{\varepsilon}_t(\theta)'&=\sum_{t=1}^n\omega_t(\theta;\nu)\hat{\varepsilon}_t(\theta)\hat{\varepsilon}_t(\theta)'\\&\quad+2\sum_{t=1}^n\omega_t(\theta;\nu)\hat{\varepsilon}_t(\theta)(\tilde{\mu}_n-\hat{\mu}_n)+\bar{\omega}_n(\theta;\nu)(\tilde{\mu}_n-\hat{\mu}_n)(\tilde{\mu}_n-\hat{\mu}_n)'\\&\overset{p}{\to}\Sigma_0(\theta),\end{aligned}$$

because the first term is consistent for $\Sigma_0(\theta)$ from the previous result. The last term is a $o_p(1)$ since $\bar{\omega}_n(\theta) = 1 + o_p(1)$ is multiplied by a $o_p(1)$. The second term is equal to $2\hat{\mu}_n o_p(1) - 2\bar{\omega}_n(\theta)\hat{\mu}_n o_p(1) = o_p(1)$. Follow the same steps for $\sum_{t=1}^{n} \omega_t(\theta;\nu/2)\tilde{\varepsilon}_t(\theta)\tilde{\varepsilon}_t(\theta)'$ using $\hat{\mu}_n(\theta;\nu/2)$ instead of $\hat{\mu}_n(\theta;\nu)$ to derive the result and conclude the proof. $\square$

Appendix C Proofs for the Preliminary Results

Proof of Lemma A1. Take derivates wrt μ:

$$\partial_\mu q_t(\psi) = -2\Sigma^{-1/2}\frac{\nu+p}{\nu}\frac{\Sigma^{-1/2}(x_t-\mu)}{1+\|x_t-\mu\|^2_{\Sigma^{-1}}/\nu},$$

where $\lambda_{\max}(\Sigma^{-1/2}) \le s_0^{-1/2}$. Use $\|\Sigma^{-1/2}(x_t-\mu)\|/(1+\|x_t-\mu\|^2_{\Sigma^{-1}}/\nu) \le \sqrt{\nu}/2$ to get the first inequality. Take derivates wrt Σ:

$$\partial_\Sigma q_t(\psi) = -\Sigma^{3/2}\frac{\nu+p}{\nu}\frac{\Sigma^{-1/2}(x_t-\mu)(x_t-\mu)'\Sigma^{-1/2}}{1+\|x_t-\mu\|^2_{\Sigma^{-1}}/\nu}\Sigma^{3/2}.$$

This implies $\|\partial_\Sigma q_t(\psi)\| \le \lambda_{\max}(\Sigma)^3(1+p/\nu)\nu$ where $\lambda_{\max}(\Sigma) \le \text{trace}(\Sigma)$, bounded in (7). $\square$

Proof of Lemma A2 - 1) IID Setting. Let $x_{t,\theta} = g(z_t;\theta)$ and $\Delta_t(\psi;\theta) = \log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu) - \mathbb{E}_P[\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu)]$, $\overline{\Delta}_n(\psi;\theta) = 1/n\sum_{t=1}^n \Delta_t(\psi,\theta)$. For any pair (ψ_j,θ_j), we have:

$$|\overline{\Delta}_n(\psi;\theta)| \le \underbrace{|\overline{\Delta}_n(\psi;\theta)-\overline{\Delta}_n(\psi_j;\theta)|}_{(A)} + \underbrace{|\overline{\Delta}_n(\psi_j;\theta)-\overline{\Delta}_n(\psi_j;\theta_j)|}_{(B)} + \underbrace{|\overline{\Delta}_n(\psi_j;\theta_j)|}_{(C)}.$$

The following bounds each one of (A), (B), and (C), either deterministically or in probability.

1. Bound for (A). Lemma A1 implies that for any $\psi=(\mu,\Sigma), \psi_j=(\mu_j,\Sigma_j)$ in Ψ_n:

$$|\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu) - \log(1+\|x_{t,\theta}-\mu_j\|^2_{\Sigma_j^{-1}}/\nu)| \le p^3\nu^{12}L_1\|\psi-\psi_j\|,$$

where L_1 depends on s_0,κ_1,κ_2. Taking either sample averages or expectations, yields:

$$(A) \le 2p^3\nu^{12}L_1\|\psi-\psi_j\|, \tag{C.2}$$

since the bound is deterministic.

2. Bound for (B)**.** Suppose, without loss of generality that $\|x_{t,\theta}-\mu\|_{\Sigma^{-1}} \geq \|x_{t,\theta_j}-\mu\|_{\Sigma^{-1}}$, then:[16]

$$\begin{aligned} 0 &\leq \log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu) - \log(1+\|x_{t,\theta_j}-\mu\|^2_{\Sigma^{-1}}/\nu) \\ &\leq \log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu - \|x_{t,\theta_j}-\mu\|^2_{\Sigma^{-1}}/\nu). \end{aligned}$$

Using properties of inner-products: $0 \leq \|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu - \|x_{t,\theta_j}-\mu\|^2_{\Sigma^{-1}}/\nu \leq \|x_{t,\theta} - x_{t,\theta_j}\|_{\Sigma^{-1}}\|x_{t,\theta}+x_{t,\theta_j}-2\mu\|_{\Sigma^{-1}}/\nu$.[17] Assumption (2) implies $\|x_{t,\theta}-x_{t,\theta_j}\|_{\Sigma^{-1}} \leq s_0^{-1/2}G_t\|\theta-\theta_j\|$ and $\|x_{t,\theta}+x_{t,\theta_j}\|_{\Sigma^{-1}} \leq 2s_0^{-1/2}G_t\mathrm{diam}(\Theta)$. Also $\psi \in \Psi_n$ implies $\|2\mu\|_{\Sigma^{-1}} \leq \nu^{3/2}(1+p/\nu)\kappa_1^{-1}$. Hence, for some constant L_2 which depends on s_0, κ_1 and $\mathrm{diam}(\Theta)$:

$$|\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu) - \log(1+\|x_{t,\theta_j}-\mu\|^2_{\Sigma^{-1}}/\nu)| \leq \log(1+\nu p L_2(1+G_t)^2\|\theta-\theta_j\|),$$

and then taking expectations and using $\log(1+x) \leq x$ for $x \geq 0$:

$$\mathbb{E}_P|\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu) - \log(1+\|x_{t,\theta_j}-\mu\|^2_{\Sigma^{-1}}/\nu)| \leq 3\nu p L_2(1+M_2)\|\theta-\theta_j\|,$$

by taking expectations over $(1+G_t)^2 \leq 3(1+G_t^2)$. Take $\varepsilon > 0$ and $\|\theta-\theta_j\| \leq \varepsilon$, denote $\ell_{t,\varepsilon} = \log(1+\nu p L_2(1+G_t)^2\varepsilon)$, then:

$$\sup_{\psi\in\Psi_n, \|\theta-\theta_j\|\leq\varepsilon} \underbrace{|\overline{\Delta}_n(\psi_j;\theta) - \overline{\Delta}_n(\psi_j;\theta_j)|}_{(B)} \leq |\bar{\ell}_{n,\varepsilon} - \mathbb{E}_P(\ell_{t,\varepsilon})| + 6\nu p L_2(1+M_2)\varepsilon.$$

Take $u_1 \geq 1$, we have:

$$\mathbb{E}_P(\exp[\ell_{t,\varepsilon}/u_1]) \leq \mathbb{E}_P([1+\nu\varepsilon p L_2(1+G_t)^2]^{1/u_1}) \leq [1+3\nu\varepsilon p L_2(1+M_2)]^{1/u_1} \leq 2,$$

if $u_1 = \max\Big(1, \log(1+\nu\varepsilon p(1+M_4^{1/2})^2)\Big)$, using $\mathbb{E}(X^{1/u_1}) \leq \mathbb{E}(X)^{1/u_1}$ for $u_1 \geq 1$ and $X \geq 0$. This implies that the sub-exponential norm of $\ell_{t,\varepsilon}$ is at most u_1. Because centering preserves sub-exponentiality, Bernstein's inequality (Vershynin, 2018, Cor2.8.3) implies:

$$\mathbb{P}\left(|\bar{\ell}_{n,\varepsilon} - \mathbb{E}_P(\ell_{t,\varepsilon})| \geq u_1\sqrt{\frac{t}{n}} + u_1\frac{t}{n}\right) \leq 2\exp(-Ct),$$

[16] For any $x \geq y \geq 0$, $0 \leq \log(1+x) - \log(1+y) = \log(1+(1+x)/(1+y)-1) = \log(1+(x-y)/(1+y)) \leq \log(1+x-y)$.

[17] For any two vectors a, b, we have $\langle a,a\rangle - \langle b,b\rangle = \langle a-b, a+b\rangle \leq \|a-b\| \times \|a+b\|$.

for some universal constant $C > 0$. From this we deduce that:

$$\mathbb{P}\left(\sup_{\psi\in\Psi_n,\|\theta-\theta_j\|\leq\varepsilon}|\overline{\Delta}_n(\psi_j;\theta)-\overline{\Delta}_n(\psi_j;\theta_j)|\geq u_1\sqrt{\frac{t}{n}}+u_1\frac{t}{n}+6\nu pL_2(1+M_2)\varepsilon\right)$$
$$\leq 2\exp(-Ct). \tag{C.3}$$

3. Bound for (C)**.** The first step is to show that (C) is a sample average over a centered sub-exponential random variable. By Assumption 2, $\sup_{\theta\in\Theta}\mathbb{E}_P(\|x_{t,\theta}\|^2)\leq M_2<\infty$. For any $\theta\in\Theta,\psi\in\Psi_n$: $0\leq\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu)\leq\log(1+3s_0^{-1}\|x_{t,\theta}\|^2/\nu)+\log(1+3/2\kappa_1^{-1}\nu(1+p/\nu))$. This inequality implies that for any $u_2\geq 1$:

$$\mathbb{E}_P(\exp[\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu)/u_2])\leq\mathbb{E}_P[1+3s_0^{-1}\|x_{t,\theta}\|^2/\nu]^{1/u_2}\exp[\log(1+3/2\kappa_1^{-1}\nu(1+p/\nu))/u_2].$$

Take $u_2=\max\left(1,\frac{\log(1+3/2\kappa_1^{-1}\nu(1+p/\nu))}{1/2\log(2)},\frac{3M_2}{1/2\log(2)s_0\nu}\right)$. We have $\mathbb{E}_P([1+3s_0^{-1}\|x_t\|^2/\nu]^{1/u_2})\leq(\mathbb{E}_P[1+3s_0^{-1}\|x_t\|^2/\nu])^{1/u_2}\leq\sqrt{2}$ and $\exp(\log[1+3/2\kappa_1^{-1}\nu(1+p/\nu)]/u)\leq\sqrt{2}$, making the product less than 2. This implies that the sub-exponential norm of $\log(1+\|x_{t,\theta}-\mu\|^2_{\Sigma^{-1}}/\nu)$ is at most u_2 for any ψ,θ. Apply Bernstein's inequality to find:

$$\mathbb{P}\left(|\overline{\Delta}_n(\psi,\theta)|\geq u_2\sqrt{\frac{t}{n}}+u_2\frac{t}{n}\right)\leq 2\exp(-Ct), \tag{C.4}$$

for the same universal constant $C>0$ as above, and for any $(\psi,\theta)\in\Psi_n\times\Theta$.

4. Overall Bound. Take $\varepsilon>0$ and $N(\varepsilon)$ denote the smallest $N\geq 1$ such that there exists $(\psi_j,\theta_j)\in\Psi_n\times\Theta$ such that $\sup_{\psi,\theta\in\Psi_n\times\Theta}(\inf_{j=1,\dots,N}[\|\psi-\psi_j\|+\|\theta-\theta_j\|])\leq\varepsilon$. Using this cover and a union bound, we have:

$$\mathbb{P}\left(\sup_{j=1,\dots,N(\varepsilon)}|\overline{\Delta}_n(\psi_j,\theta_j)|\geq u_2\sqrt{\frac{t+\log[N(\varepsilon)]}{Cn}}+u_2\frac{t+\log[N(\varepsilon)]}{Cn}\right)\leq 2\exp(-t). \tag{C.4'}$$

Take $u=u_1+u_2$ and combine the bounds to find:

$$\mathbb{P}\left(\sup_{\theta\in\Theta,\psi\in\Psi_n}|\overline{\Delta}_n(\psi,\theta)|\geq 2u\sqrt{\frac{t}{Cn}}+u\frac{t}{Cn}+u\left[\sqrt{\frac{\log[N(\varepsilon)]}{Cn}}+\frac{\log[N(\varepsilon)]}{Cn}\right]+L_3\nu^{12}p^3\varepsilon\right) \tag{C.5}$$
$$\leq 4\exp(-t).$$

Let $k = \dim(\theta)$ and $p = \dim(\mu)$. Lemma 1 implies that for some $L_4 > 0$ which depends on κ_1, κ_2, we have for any $(\mu, \Sigma) \in \Psi_n$ that $\|\psi\| = \|\mu\| + \|\Sigma\| \leq L_4 p^2 \nu^4$. This yields the following bound $\log[N(\varepsilon)] \leq k \log(3\text{diam}(\Theta)/\varepsilon) + 2p^2 \log(3L_4 p^2 \nu^4/\varepsilon)$. Pick $\varepsilon = \nu^{-12} p^{-2} n^{-1/2}$, then for some constant $L_5 > 0$ which depends on L_4 and $\text{diam}(\Theta)$: $\log[N(\varepsilon)] \leq L_5(k + 2p^2)[\log(p) + \log(\nu) + \log(n)]$. For the same choice of ε, we have $u \leq \log(1 + \nu p)$, up to a constant that depends on $\kappa_1, \kappa_2, M_2, s_0$. This implies for some constant $L > 0$:

$$\mathbb{P}\left(\sup_{\theta \in \Theta, \psi \in \Psi_n} |\overline{\Delta}_n(\psi, \theta)| \geq L \log(1 + p\nu) \left[\sqrt{\frac{t}{n}} + \frac{t}{n} + \sqrt{\frac{C_n}{n}} + \frac{C_n}{n} \right] \right) \leq 4 \exp(-t), \quad \text{(A.1)}$$

where $C_n = 1 + (k + 2p^2)[\log(p) + \log(\nu) + \log(n)]$. □

Proof of Lemma A2 - 2) Dependent Setting. The core of the proof is similar to the iid setting, the main differences occur in the sub-exponential inequalities for (B)-(C) in the inequality:

$$|\overline{\Delta}_n(\psi; \theta)| \leq \underbrace{|\overline{\Delta}_n(\psi; \theta) - \overline{\Delta}_n(\psi_j; \theta)|}_{(A)} + \underbrace{|\overline{\Delta}_n(\psi_j; \theta) - \overline{\Delta}_n(\psi_j; \theta_j)|}_{(B)} + \underbrace{|\overline{\Delta}_n(\psi_j; \theta_j)|}_{(C)}.$$

1. Bound for (A)**.** Same as iid setting.

2. Bound for (B)**.** The following relies on a proof reduction technique by Bosq (1991).[18] Take an integer $q \geq 1$ and a real number $m \in (0, n)$ such that $m = \frac{n}{2q}$. Take $\varepsilon > 0$, $\ell_{t,\varepsilon} = \log(1 + \nu p L_2 (1 + G_t)^2 \varepsilon)$ from the iid setting, and, for $t \in [0, n]$, let $\mathcal{L}_{t,\varepsilon} = \ell_{[t+1],\varepsilon}$ be its continuous-time extension. By design, $\overline{\ell}_{n,\varepsilon} = \frac{1}{n} \int_0^n \mathcal{L}_{v,\varepsilon} dv$. Let $\mathcal{U}_i = \int_{2(i-1)m}^{(2i-1)m} \mathcal{L}_{v,\varepsilon} dv$, $\mathcal{V}_i = \int_{(2i-1)m}^{2im} \mathcal{L}_{v,\varepsilon} dv$ befine non-overlapping blocks; each contains m consecutive discrete-time observations. By construction, $\overline{\ell}_{n,\varepsilon} = \frac{1}{n} \sum_{i=1}^q (\mathcal{U}_i + \mathcal{V}_i)$.

Both $\mathcal{U}_i$ and $\mathcal{V}_i$ are strictly stationary and β-mixing. Berbee's Lemma (Bosq, 1998, Lem1.1) implies that there exists $(\mathcal{U}_i^*, \mathcal{V}_i^*)_{i=1,\dots,q}$ iid such that $(\mathcal{U}_i^*, \mathcal{V}_i^*) \overset{d}{=} (\mathcal{U}_i, \mathcal{V}_i)$ and $\mathbb{P}(\mathcal{U}_i \neq \mathcal{U}_i^*) \leq \beta_{[m]}$ (likewise for $\mathcal{V}_i, \mathcal{V}_i^*$). The next step is to compute the sub-exponential norm of $\mathcal{U}_i$, $\mathcal{V}_i$. For any $i \in \{1, \dots, q\}$ and $\tilde{u}_1 \geq m \geq 1$, Jensen's inequality and Fubini's Theorem imply:

$$\mathbb{E}_P \left(\exp \left[\int_{2(i-1)m}^{2im} \mathcal{L}_{v,\varepsilon} dv / \tilde{u}_1 \right] \right) \leq \left[\int_{2(i-1)m}^{2im} \mathbb{E}_P \left(\exp \left[\mathcal{L}_{v,\varepsilon} m / \tilde{u}_1 \right] \right) dv \right] / m,$$

[18] See also Doukhan (1994), Bosq (1998).

which is less than 2 if the integrand itself is less than 2 for all v. Following the proof in the iid setting, this is true whenever $\tilde{u}_1 \geq m \max(1, \log[1 + 3\nu\varepsilon p(1 + M_2)])$. Take $u_1 = \tilde{u}_1/m$ After recentering, Bernstein's inequality applied to the iid sequence $\mathcal{U}_i^*$ yields for the same choice of u_1 as the iid setting:

$$\mathbb{P}\left(|\overline{\mathcal{U}}_i^* - \mathbb{E}_P(\mathcal{U}_i^*)| \geq m u_1 \sqrt{\frac{t}{q}} + m u_1 \frac{t}{q}\right) \leq 2\exp(-Ct),$$

for the same universal constant $C > 0$ used in the iid setting, the same holds for $\mathcal{V}_i^*$. To get the bound for (B), we need a tail inequality for $\overline{\ell}_{n,\varepsilon} - \mathbb{E}_P(\ell_{t,\varepsilon}) = \frac{1}{n}\sum_{i=1}^q (\mathcal{U}_i + \mathcal{V}_i - \mathbb{E}_P(\mathcal{U}_i) - \mathbb{E}_P(\mathcal{V}_i))$:[19]

$$\begin{aligned}\mathbb{P}\left(\frac{n}{2q}|\overline{\ell}_n, \varepsilon - \mathbb{E}_P(\ell_{t,\varepsilon})| \geq m u_1 \sqrt{\frac{t}{q}} + m u_1 \frac{t}{q}\right) &\leq 2\mathbb{P}\left(|\overline{\mathcal{U}}_i - \mathbb{E}_P(\mathcal{U}_i)| \geq m u_1 \sqrt{\frac{t}{q}} + m u_1 \frac{t}{q}\right)\\ &\leq 2\mathbb{P}\left(|\overline{\mathcal{U}}_i^* - \mathbb{E}_P(\mathcal{U}_i^*)| \geq m u_1 \sqrt{\frac{t}{q}} + m u_1 \frac{t}{q}\right) + 2q\beta_{[m]}.\end{aligned}$$

The mixing condition and the definition of m imply that $2q\beta_{[m]} \leq \frac{na}{m}\exp(-b[m])$. Then, we can re-write for $m \geq 1$:

$$\mathbb{P}\left(|\overline{\ell}_{n,\varepsilon} - \mathbb{E}_P(\ell_{t,\varepsilon})| \geq 2u_1\sqrt{\frac{mt}{n}} + 2u_1\frac{mt}{n}\right) \leq 4\exp(-Ct) + \frac{na}{\exp(b)}\exp(-bm) = 6\exp(-Ct),$$

for $m = 1 + [Ct + \log(an)]/b$. Note that the effect of m on the tail inequality is comparable to the bounded case found in e.g. Doukhan (1994, Ch1.4), Rio (1999, Ch6). Going back to (B) itself, following the same steps from the above inequality to the result yields:

$$\begin{aligned}&\mathbb{P}\left(\sup_{\psi\in\Psi_n, \|\theta-\theta_j\|\leq\varepsilon} |\overline{\Delta}_n(\psi_j;\theta) - \overline{\Delta}_n(\psi_j;\theta_j)| \geq u_1\sqrt{\frac{mt}{n}} + u_1\frac{mt}{n} + 6\nu p L_2(1+M_2)\varepsilon\right)\\ &\leq 6\exp(-Ct),\end{aligned} \tag{C.3}$$

where m depends on t and n as stated above.

[19] The derivation relies on the inequality: $\mathbb{P}(\sum_{i=1}^q(\mathcal{U}_i + \mathcal{V}_i - \mathbb{E}_P(\mathcal{U}_i) - \mathbb{E}_P(\mathcal{V}_i)) \geq 2t) \leq \mathbb{P}(\sum_{i=1}^q(\mathcal{U}_i - \mathbb{E}_P(\mathcal{U}_i) \geq t) + \mathbb{P}(\sum_{i=1}^q(\mathcal{V}_i - \mathbb{E}_P(\mathcal{U}_i)) \geq t) = 2\mathbb{P}(\sum_{i=1}^q(\mathcal{U}_i - \mathbb{E}_P(\mathcal{U}_i) \geq t)$.

3. Bound for (C)**.** Using the same steps as above, we can take $m = 1 + [Ct + \log(an)]/b$ and the same u_2 found in the iid setting to get the inequality:

$$\mathbb{P}\left(|\overline{\Delta}_n(\psi;\theta)| \geq u_2\sqrt{\frac{mt}{n}} + u_2\frac{mt}{n}\right) \leq 6\exp(-Ct), \tag{C.4}$$

for the same universal constant $C > 0$ and for any $(\psi, \theta) \in \Psi_n \times \Theta$.

4. Overall Bound. Using the same collection $(\psi_j, \theta_j) \in \Psi_n \times \Theta$ as in the iid case, we have:

$$\mathbb{P}\left(\sup_{j=1,\dots,N(\varepsilon)} |\overline{\Delta}_n(\psi_j;\theta_j)| \geq u_2\sqrt{\frac{m_2 t + m_2\log[N(\varepsilon)]}{Cn}} + u_2\frac{m_2 t + m_2 \log[N(\varepsilon)]}{Cn}\right)$$
$$\leq 6\exp(-t), \tag{C.4'}$$

using $m_2 = 1 + [t + \log[N(\varepsilon)] + \log(an)]/b = m + \log[N(\varepsilon)]/b$.

Take $u = u_1 + u_2$ and combine these bounds:

$$\mathbb{P}\Bigg(\sup_{\theta\in\Theta,\psi\in\Psi_n} |\overline{\Delta}_n(\psi,\theta)| \geq 2u\sqrt{\frac{(m+m_2)t}{Cn}} + u\frac{(m+m_2)t}{Cn} + u\left[\sqrt{\frac{\log[N(\varepsilon)]}{Cn}} + \frac{\log[N(\varepsilon)]}{Cn}\right] \tag{C.5}$$
$$+ L_3\nu^{12}p^3\varepsilon\Bigg) \leq 12\exp(-t).$$

Take $\varepsilon = \nu^{-12}p^{-2}n^{-1/2}$ as in the iid case so that $\log[N(\varepsilon)] \leq L_5(k + 2p^2)[\log(p) + \log(\nu) + \log(n)]$. This implies that $t \leq (m+m_2)t = t + t^2/b + \log[N(\varepsilon)]t/b + \log(an)t/b \leq t\tilde{L}_5\,(t + C_n)$, for some constant $\tilde{L}_5$ which depends on a, b and L_5. As in the iid setting $u \leq \log(1 + \nu p)$, up to a constant and for some constant $\tilde{L} > 0$:

$$\mathbb{P}\left(\sup_{\theta\in\Theta,\psi\in\Psi_n} |\overline{\Delta}_n(\psi,\theta)| \geq \tilde{L}\log(1+p\nu)\left[\sqrt{\frac{(t+C_n)t}{n}} + \frac{(t+C_n)t}{n} + \sqrt{\frac{C_n}{n}} + \frac{C_n}{n}\right]\right) \tag{A.1'}$$
$$\leq 12\exp(-t), \tag{C.6}$$

where $C_n = 1 + (k + 2p^2)[\log(p) + \log(\nu) + \log(n)]$. □

Appendix D Leveraged outliers: an illustration

Before introducing the estimator, the following illustrates the asymptotic effect of excess leverage. Consider a single regressor linear model:

$$y_t = \beta_0 + \beta_1 x_t + e_t,$$

for $t = 1, \ldots, n-1$ where $x_t \sim (0, \sigma_x^2), e_t \sim (0, \sigma_e^2)$ are iid with finite fourth moment. The last observation is $y_n = \beta_0 + (\beta_1 + c)x_n$. Here c measures misspecification, and x_n is such that $x_n^2 = \sqrt{n}\sigma_x^2$. Because of leverage, (y_n, x_n) has some influence asymptotically, $\frac{x_n(y_n - \bar{y}_n)}{\sum_{t=1}^{n}(x_t - \bar{x}_n)^2} \approx \frac{x_n^2(\beta_1 + c)}{n\sigma_x^2} = \frac{\beta_1 + c}{\sqrt{n}}$, so that the estimator is asympotically biased:

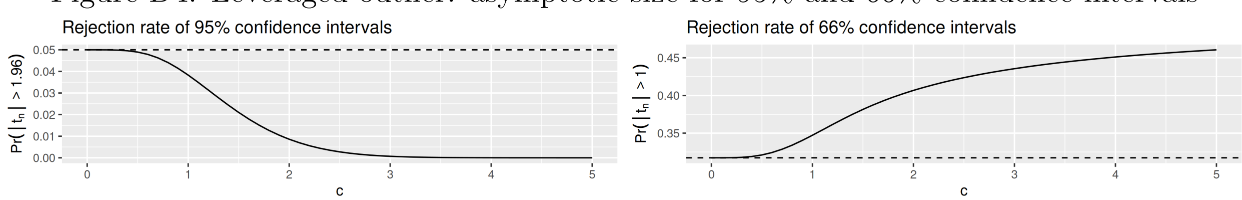
$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}(c, \sigma_e^2/\sigma_x^2),$$

with homoskedastic errors. The outlier further inflates heteroskedasticity-robust standard errors: $\hat{V}_{\hat{\beta}_1} \xrightarrow{p} c^2 + \sigma_e^2/\sigma_x^2$. The misspecification c affects the t-statistic t_n through both estimates and standard errors:

$$t_n = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \xrightarrow{d} \mathcal{N}\left(\frac{c}{\sqrt{c^2 + \sigma_e^2/\sigma_x^2}}, \frac{1}{\sqrt{c^2\sigma_x^2/\sigma_e^2 + 1}}\right).$$

Figure D4 shows the coverage of 95% and 66% confidence intervals when c increases.

Figure D4: Leveraged outlier: asymptotic size for 95% and 66% confidence intervals



Note: Solid line: rejection probability, dashed line: nominal size.

Appendix E Leverage in IV Regressions

The following Lemma gives a measure of influence and leverage in just-identified linear instrumental variable regressions. The model is $y_t = x_t'\theta + e_t$, let $\hat{\theta}_n$ be the IV estimates, $\hat{y}_t = x_t'\hat{\theta}_n$ the predicted value and $\tilde{y}_t = x_t'\hat{\theta}_{-t}$ the leave-one-out predicted value. Using standard notation, Z, X and y refer the to matrix of instruments, regressors and the vector of outcomes.

Lemma E3. *For each t, the difference between the full sample and the leave-one-out predicted value is:*

$$\hat{y}_t - \tilde{y}_t = x_t'(Z'X)^{-1} z_t \tilde{e}_t,$$

where $\tilde{e}_t = y_t - \tilde{y}_t$*. Using the terminology from OLS, leverage is given by* $h_t = x_t'(Z'X)^{-1} z_t$ *and influence is* $h_t \tilde{e}_t$*. Leverage can be positive or negative. Unlike OLS, the sign of influence may not coincide with the sign of the residual* $\tilde{e}_t$*.*

Proof of Lemma E3. The derivations are similar to OLS. The full sample $\hat{\theta}_n = (Z'X)^{-1} Z'y$, the leave-one-out $\hat{\theta}_{-t} = (Z'X - z_t x_t')^{-1}(Z'y - z_t y_t)$. Pre-multiply the latter by $(Z'X)^{-1}(Z'X - z_t x_t')$ to find:

$$\hat{\theta}_{-t} - (Z'X)^{-1} z_t \tilde{y}_t = \hat{\theta}_n - (Z'X)^{-1} z_t y_t.$$

Re-arrange terms and pre-multiply by x_t' to find:

$$\hat{y}_t - \tilde{y}_t = \underbrace{x_t'(Z'X)^{-1} z_t \tilde{e}_t}_{\text{Influence}}.$$

For OLS, $h_t = x_t'(X'X)^{-1} x_t \geq 0$, here $h_t = x_t'(Z'X)^{-1} z_t < 0$ can occur. $\square$

Appendix F Additional Simulation Results

Table F6: Small sample properties of the estimators ($n = 150$) – $\nu = O(n^{1/3})$

	100 × RMSE							Rejection Rate						
	$n_o = 0$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	8.05	8.05	12.00	11.84	9.31	8.11	7.94	0.04	0.04	0.24	0.29	0.13	0.05	0.06
θ_1	8.00	8.00	7.15	7.96	7.78	7.78	7.92	0.06	0.06	0.06	0.11	0.08	0.07	0.06
θ_2	8.10	8.10	7.46	8.44	8.20	8.10	8.06	0.04	0.04	0.05	0.10	0.06	0.05	0.05
θ_3	8.19	8.19	7.43	8.55	8.30	8.15	8.15	0.06	0.06	0.06	0.10	0.07	0.06	0.06
	$n_o = 1$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	10.71	8.04	13.01	14.17	10.95	8.52	10.32	0.03	0.04	0.20	0.46	0.23	0.08	0.08
θ_1	38.57	8.07	15.23	8.27	7.97	7.87	32.28	0.00	0.06	0.01	0.14	0.10	0.07	0.39
θ_2	38.39	8.11	15.09	8.73	8.36	8.13	32.12	0.01	0.04	0.01	0.12	0.06	0.06	0.37
θ_3	39.94	8.20	15.75	8.82	8.49	8.26	33.52	0.00	0.06	0.00	0.12	0.09	0.07	0.39
	$n_o = 5$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	11.98	8.14	16.57	16.96	13.36	9.79	13.44	0.10	0.04	0.24	0.59	0.38	0.13	0.16
θ_1	47.57	8.40	47.17	9.03	8.63	8.41	46.72	0.99	0.06	0.99	0.12	0.08	0.06	0.99
θ_2	47.48	8.26	48.25	9.26	8.78	8.51	47.14	0.99	0.04	1.00	0.11	0.04	0.03	1.00
θ_3	49.17	8.28	49.48	9.34	8.95	8.72	48.64	0.98	0.06	0.98	0.10	0.08	0.04	0.98
	$n_o = 10$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	12.21	8.21	17.33	16.84	13.37	10.98	14.11	0.09	0.04	0.23	0.47	0.22	0.07	0.17
θ_1	49.14	8.54	48.38	10.45	12.31	23.20	48.65	0.99	0.04	0.99	0.01	0.01	0.16	1.00
θ_2	49.05	8.31	49.67	11.02	13.09	24.58	48.92	0.99	0.04	0.99	0.01	0.01	0.16	1.00
θ_3	50.52	8.51	50.70	11.32	13.68	24.91	50.19	0.98	0.06	0.98	0.00	0.01	0.16	0.99

Legend: $\hat{\theta}_n^{ols}$ full sample OLS, $\hat{\theta}_{n_P}^{ols}$ oracle OLS, $\hat{\theta}_n^{rlm}$ robust M-estimator, $\hat{\theta}_n$ robust estimates without bias correction, $\tilde{\theta}_n$ robust estimates with bias correction, $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction, $\hat{\theta}_n^{un}$ undersmoothed robust estimates with $\hat{\nu}_n^2$. 200 Monte-Carlo replications. n_o = number of outliers. Rejection rate for t-test at the 5% significance level. Average $\hat{\nu}_n$: 32.8, 18.0, 11.0, 10.8, 10.8 for $n_0 = 0$, 1, 5, 10, 20 respectively. Each $\hat{\nu}_n$ is selected on a grid $[\nu_0, \ldots, \nu_J]$ where $\nu_0 = 8.82$, $\nu_J = 177.16$.

Table F7: Small sample properties of the estimators ($n = 500$) – $\nu = O(n^{1/3})$

	100 × RMSE							Rejection Rate						
	$n_o = 0$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	4.59	4.59	10.67	9.29	6.29	4.70	4.56	0.07	0.07	0.65	0.51	0.21	0.07	0.07
θ_1	4.21	4.21	3.93	4.57	4.50	4.44	4.21	0.04	0.04	0.05	0.09	0.07	0.07	0.04
θ_2	4.76	4.76	4.21	4.65	4.61	4.60	4.72	0.06	0.06	0.07	0.09	0.09	0.07	0.07
θ_3	4.51	4.51	4.09	4.66	4.56	4.52	4.48	0.09	0.09	0.07	0.13	0.10	0.09	0.09
	$n_o = 1$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	5.40	4.58	10.87	10.92	7.41	4.87	4.88	0.03	0.07	0.64	0.67	0.30	0.09	0.04
θ_1	38.16	4.22	7.98	4.67	4.60	4.56	22.53	0.00	0.04	0.01	0.10	0.07	0.07	0.02
θ_2	38.00	4.77	7.95	4.74	4.68	4.67	22.66	0.00	0.07	0.00	0.10	0.09	0.09	0.04
θ_3	37.38	4.50	7.42	4.73	4.59	4.52	21.92	0.00	0.08	0.01	0.13	0.09	0.08	0.03
	$n_o = 5$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	5.90	4.60	11.52	13.86	9.89	5.89	6.60	0.07	0.06	0.30	0.91	0.56	0.17	0.09
θ_1	47.49	4.20	45.53	4.84	4.77	4.80	46.01	1.00	0.04	0.47	0.11	0.09	0.07	1.00
θ_2	47.41	4.82	45.67	4.93	4.84	4.84	45.96	1.00	0.07	0.46	0.12	0.10	0.08	1.00
θ_3	46.66	4.51	44.65	4.94	4.75	4.67	45.21	1.00	0.07	0.46	0.15	0.10	0.08	1.00
	$n_o = 10$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	5.95	4.60	11.69	15.11	11.06	6.62	7.43	0.07	0.05	0.42	0.94	0.71	0.23	0.13
θ_1	48.95	4.19	49.02	4.89	4.81	4.86	48.44	1.00	0.03	1.00	0.12	0.07	0.04	1.00
θ_2	48.98	4.85	49.27	5.05	4.95	4.96	48.58	1.00	0.07	1.00	0.11	0.10	0.08	1.00
θ_3	48.15	4.56	48.22	5.12	4.89	4.80	47.69	1.00	0.07	1.00	0.16	0.10	0.07	1.00
	$n_o = 20$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{nP}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	6.06	4.61	12.26	16.15	12.03	7.30	15.63	0.06	0.05	0.45	0.95	0.78	0.27	0.79
θ_1	49.71	4.24	49.63	4.97	4.88	4.97	49.20	1.00	0.04	1.00	0.10	0.05	0.04	1.00
θ_2	49.92	4.96	50.07	5.25	5.13	5.15	49.69	1.00	0.07	1.00	0.10	0.04	0.02	1.00
θ_3	48.85	4.56	48.78	5.17	4.89	4.79	48.55	1.00	0.06	1.00	0.14	0.06	0.02	1.00

Legend: $\hat{\theta}_n^{ols}$ full sample OLS, $\hat{\theta}_{nP}^{ols}$ oracle OLS, $\hat{\theta}_n^{rlm}$ robust M-estimator, $\hat{\theta}_n$ robust estimates without bias correction, $\tilde{\theta}_n$ robust estimates with bias correction, $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction, $\hat{\theta}_n^{un}$ undersmoothed robust estimates with $\hat{\nu}_n^2$. 200 Monte-Carlo replications. n_o = number of outliers. Rejection rate for t-test at the 5% significance level. Average $\hat{\nu}_n$: 37.46, 26.39, 16.09, 13.18, 10.79 for $n_0 = 0$, 1, 5, 10, 20 respectively. Each $\hat{\nu}_n$ is selected on a grid $[\nu_0, \ldots, \nu_J]$ where $\nu_0 = 8.83$, $\nu_J = 264.64$.

Table F8: Small sample properties of the estimators ($n = 500$), with $\nu = O(n^{1/4}\log(n))$

	100 × RMSE							Rejection Rate						
	$n_o = 0$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	4.59	4.59	10.67	8.53	5.83	4.67	4.57	0.07	0.07	0.65	0.42	0.18	0.07	0.07
θ_1	4.21	4.21	3.93	4.52	4.45	4.39	4.21	0.04	0.04	0.05	0.08	0.07	0.07	0.04
θ_2	4.76	4.76	4.21	4.63	4.61	4.62	4.73	0.06	0.06	0.07	0.09	0.08	0.07	0.07
θ_3	4.51	4.51	4.09	4.61	4.52	4.49	4.48	0.09	0.09	0.07	0.12	0.10	0.09	0.09
	$n_o = 1$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	5.40	4.58	10.87	10.30	6.93	4.75	5.05	0.03	0.07	0.64	0.62	0.27	0.07	0.04
θ_1	38.16	4.22	7.98	4.64	4.58	4.52	27.54	0.00	0.04	0.01	0.09	0.07	0.06	0.14
θ_2	38.00	4.77	7.95	4.70	4.65	4.64	27.44	0.00	0.07	0.00	0.09	0.09	0.07	0.17
θ_3	37.38	4.50	7.42	4.69	4.57	4.51	26.86	0.00	0.08	0.01	0.12	0.09	0.07	0.14
	$n_o = 5$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	5.90	4.60	11.52	12.93	9.02	5.41	6.44	0.07	0.06	0.30	0.89	0.49	0.14	0.09
θ_1	47.49	4.20	45.53	4.78	4.72	4.72	46.45	1.00	0.04	0.47	0.10	0.09	0.06	1.00
θ_2	47.41	4.82	45.67	4.89	4.81	4.80	46.41	1.00	0.07	0.46	0.10	0.09	0.06	1.00
θ_3	46.66	4.51	44.65	4.87	4.70	4.62	45.64	1.00	0.07	0.46	0.14	0.10	0.07	1.00
	$n_o = 10$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	5.95	4.60	11.69	12.57	8.59	5.10	7.01	0.07	0.05	0.42	0.86	0.43	0.09	0.10
θ_1	48.95	4.19	49.02	4.74	4.68	4.69	48.66	1.00	0.03	1.00	0.07	0.04	0.03	1.00
θ_2	48.98	4.85	49.27	4.90	4.83	4.84	48.81	1.00	0.07	1.00	0.09	0.07	0.04	1.00
θ_3	48.15	4.56	48.22	4.91	4.73	4.64	47.91	1.00	0.07	1.00	0.12	0.07	0.03	1.00
	$n_o = 20$													
	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$	$\hat{\theta}_n^{ols}$	$\hat{\theta}_{n_P}^{ols}$	$\hat{\theta}_n^{rlm}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	$\hat{\theta}_n^{un}$
θ_0	6.06	4.61	12.26	13.32	9.14	5.22	15.91	0.06	0.05	0.45	0.89	0.47	0.10	0.78
θ_1	49.71	4.24	49.63	4.80	4.77	4.86	49.23	1.00	0.04	1.00	0.03	0.02	0.01	1.00
θ_2	49.92	4.96	50.07	5.09	5.02	5.09	49.71	1.00	0.07	1.00	0.03	0.02	0.00	1.00
θ_3	48.85	4.56	48.78	4.91	4.71	4.67	48.58	1.00	0.06	1.00	0.03	0.01	0.00	1.00

Legend: $\hat{\theta}_n^{ols}$ full sample OLS, $\hat{\theta}_{n_P}^{ols}$ oracle OLS, $\hat{\theta}_n^{rlm}$ robust M-estimator, $\hat{\theta}_n$ robust estimates without bias correction, $\tilde{\theta}_n$ robust estimates with bias correction, $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction, $\hat{\theta}_n^{un}$ undersmoothed robust estimates with $\hat{\nu}_n^2$. 200 Monte-Carlo replications. n_o = number of outliers. Rejection rate for t-test at the 5% significance level. Average $\hat{\nu}_n$: 48.78, 28.88, 18.34, 17.95, 14.69 for $n_0 = 0$, 1, 5, 10, 20 respectively. Each $\hat{\nu}_n$ is selected on a grid $[\nu_0, \dots, \nu_J]$ where $\nu_0 = 14.69$, $\nu_J = 979.86$.

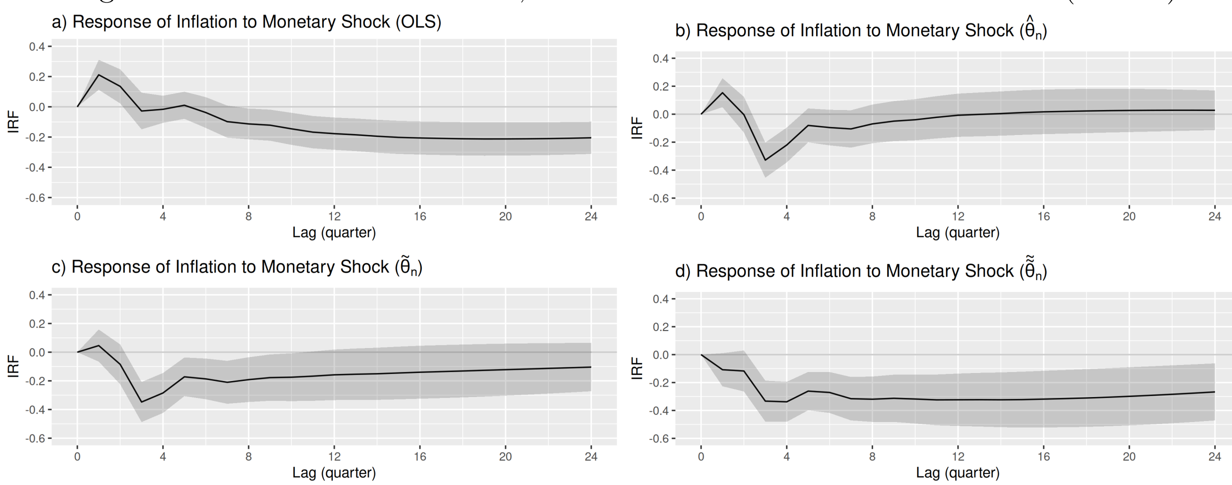
Appendix G Additional Empirical Results

G.1 Additional Results for the Price Puzzle

Table G9: Regression (1): contribution to each coefficient (moments)

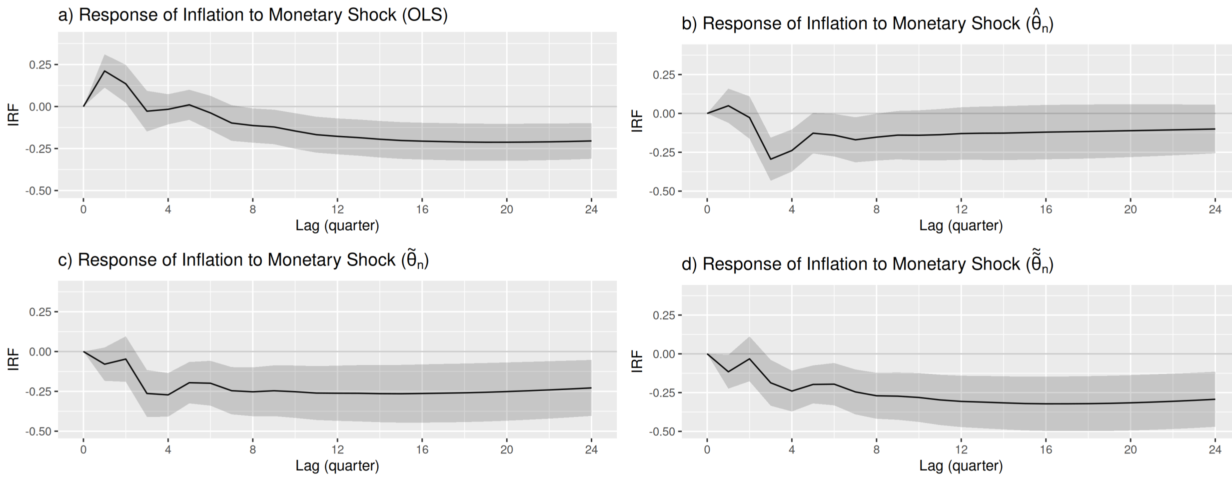
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$
skewness	-0.56	3.24	-0.30	1.30	-2.99	0.95	0.74	-1.48	-0.78	0.52	-2.93	-0.24	0.54
kurtosis	4.42	27.81	8.98	7.88	36.70	9.77	8.41	27.78	6.95	7.76	32.66	9.29	7.24

Figure G5: Recursive VAR: OLS, Robust and Bias-Corrected Estimates ($\nu = 10$)



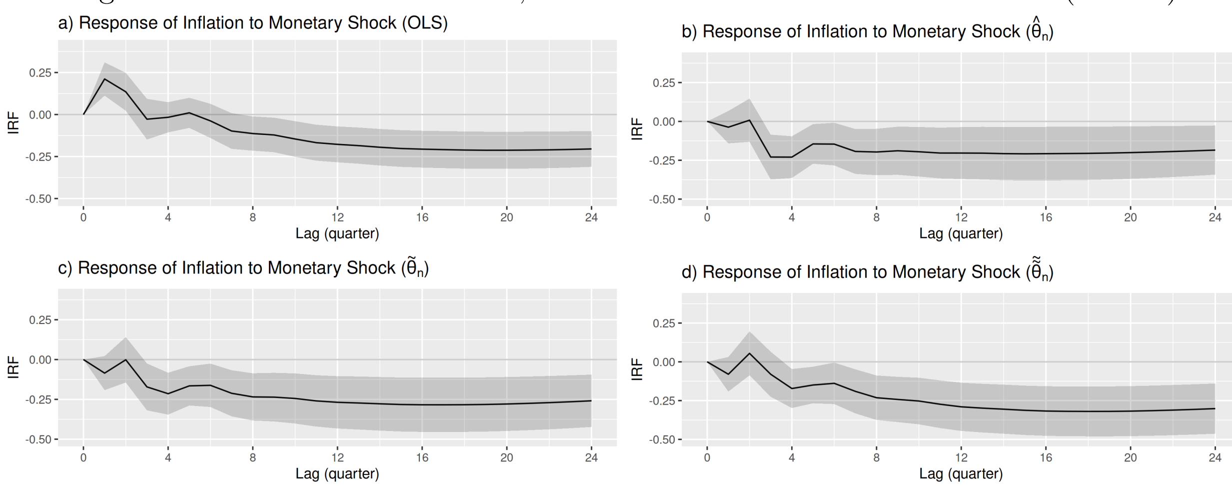
Note: a) OLS estimates, b) $\hat{\theta}_n$ robust estimates without bias correction, c) $\tilde{\theta}_n$ robust estimates with bias correction, d) $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction. Bands: estimates $\pm$ one standard error.

Figure G6: Recursive VAR: OLS, Robust and Bias-Corrected Estimates ($\nu = 15$)



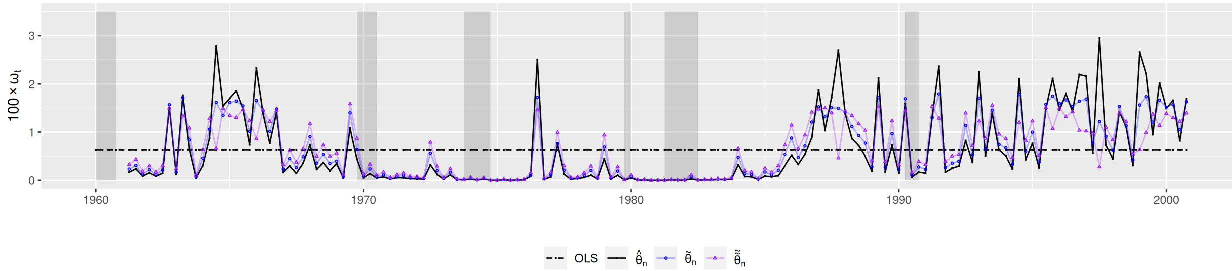
Note: a) OLS estimates, b) $\hat{\theta}_n$ robust estimates without bias correction, c) $\tilde{\theta}_n$ robust estimates with bias correction, d) $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction. Bands: estimates $\pm$ one standard error.

Figure G7: Recursive VAR: OLS, Robust and Bias-Corrected Estimates ($\nu = 20$)



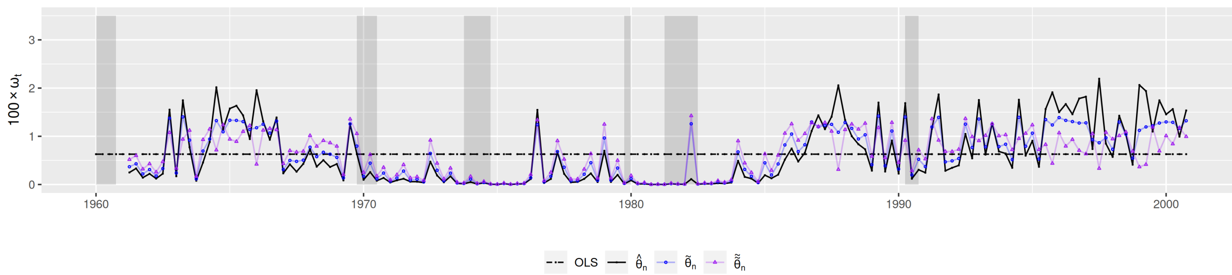
Note: a) OLS estimates, b) $\hat{\theta}_n$ robust estimates without bias correction, c) $\tilde{\theta}_n$ robust estimates with bias correction, d) $\tilde{\tilde{\theta}}_n$ robust estimates with repeated bias correction. Bands: estimates $\pm$ one standard error.

Figure G8: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates ($\nu = 10$)



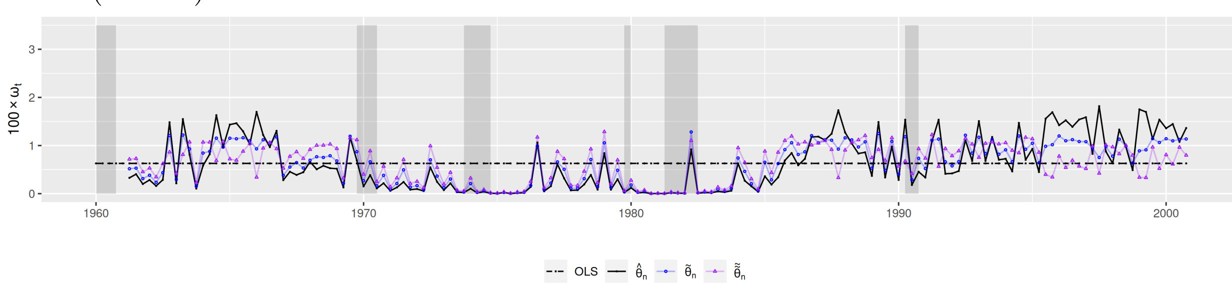
Note: Estimation weights ω_t implicitely used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

Figure G9: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates ($\nu = 15$)



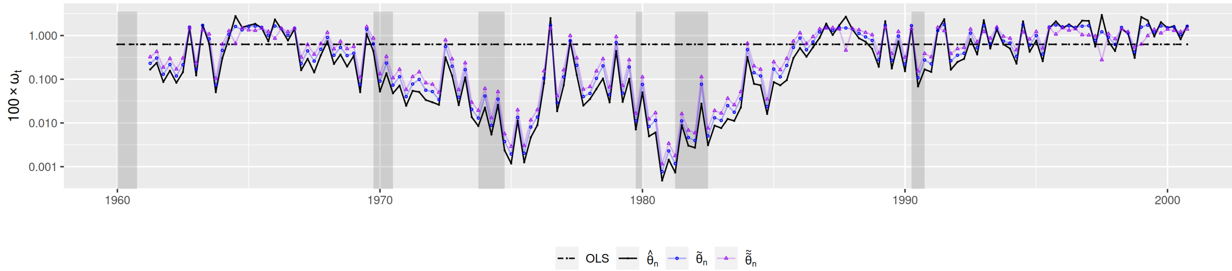
Note: Estimation weights ω_t implicitely used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

Figure G10: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates ($\nu = 20$)



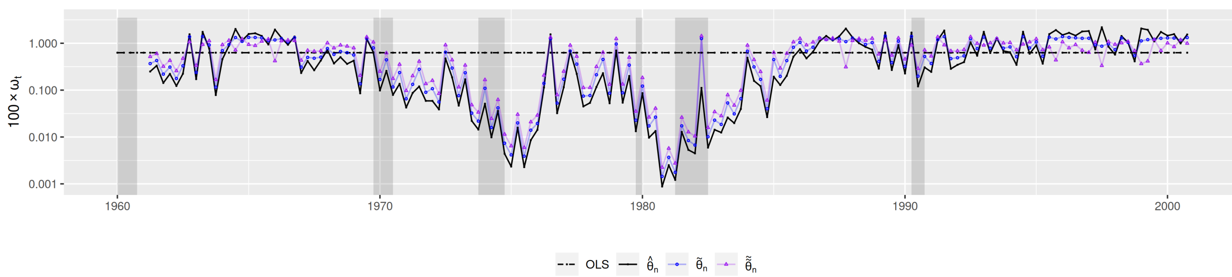
Note: Estimation weights ω_t implicitely used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

Figure G11: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates ($\nu = 10$, log scale)



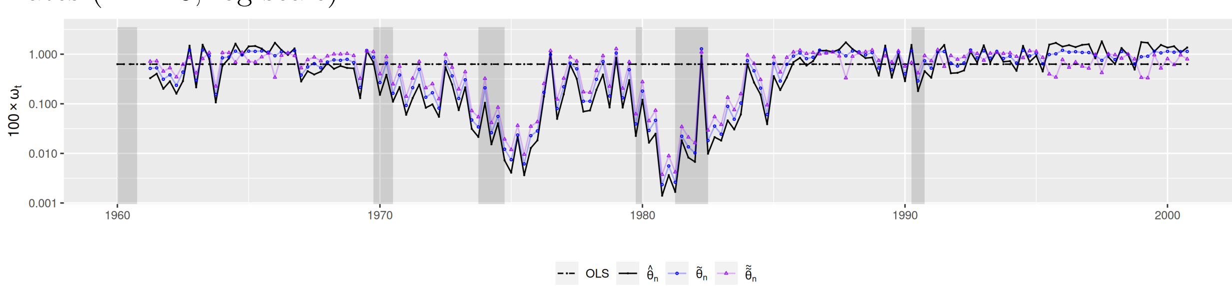
Note: Estimation weights ω_t implicitely used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

Figure G12: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates ($\nu = 15$, log scale)



Note: Estimation weights ω_t implicitely used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

Figure G13: Recursive VAR, Estimation Weights: OLS, Robust, and Bias-Corrected Estimates ($\nu = 20$, log scale)



Note: Estimation weights ω_t implicitely used to estimate θ. OLS (dashed/black): $\omega_t = 1/n$. Robust estimates $\hat{\theta}_n$ (solid/black). Bias-corrected robust estimates $\tilde{\theta}_n$ (solid/circle/blue). Repeated bias-corrected robust estimates $\tilde{\tilde{\theta}}_n$ (solid/triangle/purple). Shaded vertical bars = NBER recession dates.

G.2 Additional Results for Inflation and Openness

Table G10: Weights used in estimation ($y = \pi/100$) – 1/2

Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$
Algeria	0.88	1.04	0.95	0.78	Ethiopia	0.88	0.58	0.83	1.14
Argentina	0.88	0.02	0.04	0.09	Fiji	0.88	1.05	0.90	0.81
Australia	0.88	0.97	1.08	1.24	Finland	0.88	1.04	0.94	0.78
Austria	0.88	0.87	1.02	1.09	France	0.88	1.01	1.01	0.92
Bahrain	0.88	0.98	0.95	0.82	Gabon	0.88	1.06	0.89	0.86
Bangladesh	0.88	1.05	0.92	0.81	Gambia	0.88	0.90	1.02	0.91
Barbados	0.88	0.91	1.03	0.99	Germany	0.88	0.80	1.03	1.22
Belgium	0.88	0.98	0.98	0.85	Ghana	0.88	0.25	0.45	0.82
Benin	0.88	1.03	0.93	0.81	Greece	0.88	1.01	0.97	0.81
Bolivia	0.88	0.01	0.02	0.05	Guatemala	0.88	1.06	0.89	0.83
Botswana	0.88	1.06	0.89	0.85	Guyana	0.88	1.05	0.93	0.81
Brazil	0.88	0.04	0.07	0.16	Haiti	0.88	0.75	0.95	1.13
Burkina Faso	0.88	0.95	1.00	0.91	Honduras	0.88	0.96	0.97	0.86
Burma	0.88	0.71	0.95	1.20	Hong Kong	0.88	1.04	0.94	0.82
Burundi	0.88	0.70	0.91	1.14	Iceland	0.88	0.23	0.40	0.74
Cameroon	0.88	1.03	0.94	0.80	India	0.88	0.79	1.03	1.31
Canada	0.88	0.84	1.07	1.35	Indonesia	0.88	1.06	0.88	0.85
Central Afr. Rep.	0.88	0.98	1.01	0.98	Iran	0.88	1.04	0.94	0.84
Chile	0.88	0.16	0.30	0.59	Ireland	0.88	1.05	0.91	0.84
Colombia	0.88	0.93	1.12	0.91	Israel	0.88	0.03	0.05	0.09
Congo	0.88	1.06	0.88	0.87	Italy	0.88	1.05	0.90	0.86
Costa Rica	0.88	0.77	1.00	1.10	Ivory Coast	0.88	1.06	0.89	0.84
Cyprus	0.88	1.06	0.89	0.83	Jamaica	0.88	0.79	1.01	1.05
Denmark	0.88	0.99	0.99	0.94	Japan	0.88	0.78	1.02	1.27
Dominican Republic	0.88	1.06	0.89	0.81	Jordan	0.88	1.06	0.90	0.82
Ecuador	0.88	0.96	1.05	0.87	Kenya	0.88	1.03	0.94	0.81
Egypt	0.88	1.02	0.97	0.81	Korea	0.88	1.06	0.89	0.86
El Salvador	0.88	1.06	0.88	0.85	Kuwait	0.88	1.06	0.91	0.81

Note: $\hat{\theta}_n^{IV}$: IV estimates, $\hat{\theta}_n$: robust estimates, $\tilde{\theta}_n$: bias-corrected robust estimates, $\tilde{\tilde{\theta}}_n$: repeated bias-corrected robust estimates. $\hat{\nu}_n = 12.62$. Estimates for θ_2 reported using log(pcinc)/100 as a regressor. Sample size $n = 114$. All weights were multiplied by 100 for formatting.

Table G11: Weights used in estimation ($y = \pi/100$) – 2/2

Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$
Lesotho	0.88	0.90	1.02	1.14	Sierra Leone	0.88	0.83	1.02	0.91
Liberia	0.88	1.00	0.94	0.81	Singapore	0.88	0.86	0.93	1.02
Luxembourg	0.88	1.02	0.93	0.80	Somalia	0.88	0.56	0.90	1.25
Madagascar	0.88	1.06	0.89	0.82	South Africa	0.88	1.06	0.88	0.84
Malawi	0.88	1.03	0.94	0.82	Spain	0.88	1.06	0.88	0.85
Malaysia	0.88	1.00	0.97	0.77	Sri Lanka	0.88	1.06	0.88	0.86
Malta	0.88	0.85	1.03	0.85	Sudan	0.88	0.85	1.10	0.94
Mauritania	0.88	1.06	0.88	0.86	Suriman	0.88	1.06	0.88	0.87
Mauritius	0.88	0.92	1.01	0.87	Swaziland	0.88	0.89	1.02	1.03
Mexico	0.88	0.44	0.78	1.37	Sweden	0.88	1.03	0.96	0.80
Morocco	0.88	1.02	0.94	0.79	Switzerland	0.88	0.74	0.96	1.18
Nepal	0.88	0.90	1.00	1.03	Syria	0.88	1.06	0.89	0.87
Netherlands	0.88	0.88	1.02	1.04	Taiwan	0.88	0.97	0.97	0.88
New Zealand	0.88	1.06	0.88	0.83	Tanzania	0.88	1.04	0.90	0.86
Nicaragua	0.88	0.34	0.55	0.92	Thailand	0.88	0.98	1.01	0.86
Niger	0.88	1.06	0.88	0.85	Togo	0.88	1.00	0.95	0.81
Nigeria	0.88	1.06	0.88	0.86	Trinidad & Tobago	0.88	0.91	1.01	0.82
Norway	0.88	1.03	0.96	0.79	Tunisia	0.88	1.03	0.91	0.81
Oman	0.88	1.06	0.88	0.84	Turkey	0.88	0.66	1.04	1.47
Pakistan	0.88	1.02	0.97	0.84	Uganda	0.88	0.17	0.32	0.63
Panama	0.88	0.95	0.98	0.87	U.A. Emirates	0.88	1.06	0.89	0.79
Papua New Guinea	0.88	1.03	0.92	0.79	United Kingdom	0.88	1.06	0.90	0.78
Paraguay	0.88	1.05	0.91	0.85	United States	0.88	0.68	0.94	1.27
Peru	0.88	0.26	0.48	0.90	Uruguay	0.88	0.28	0.48	0.86
Philippines	0.88	1.06	0.88	0.87	Venezuela	0.88	1.06	0.90	0.86
Portugal	0.88	0.89	1.04	0.96	Yemen	0.88	1.06	0.90	0.85
Rwanda	0.88	0.85	0.99	1.13	Zaire	0.88	0.10	0.18	0.36
Saudi Arabia	0.88	1.06	0.91	0.76	Zambia	0.88	0.97	1.02	0.83
Senegal	0.88	1.06	0.89	0.84	Zimbabwe	0.88	1.04	0.92	0.80

Note: $\hat{\theta}_n^{IV}$: IV estimates, $\hat{\theta}_n$: robust estimates, $\tilde{\theta}_n$: bias-corrected robust estimates, $\tilde{\tilde{\theta}}_n$: repeated bias-corrected robust estimates. $\hat{\nu}_n = 12.62$. Estimates for θ_2 reported using log(pcinc)/100 as a regressor. Sample size $n = 114$. All weights were multiplied by 100 for formatting.

G.3 Additional Results for Segregation and Government Quality

Table G12: Weights used in Estimation (Ethnicity)

Country	$\hat{\theta}_n^{\mathrm{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	Country	$\hat{\theta}_n^{\mathrm{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$
Afghanistan	1.03	0.22	0.59	2.1	Kenya	1.03	0.95	1.51	2.2
Argentina	1.03	0.08	0.15	0.29	Korea	1.03	0.05	0.07	0.13
Armenia	1.03	2.57	1.99	0.39	Kyrgyzstan	1.03	1.87	2.22	1.45
Australia	1.03	0.93	0.97	1.29	Latvia	1.03	2.59	1.06	0.78
Austria	1.03	0.23	2.55	1.62	Lesotho	1.03	0.11	0.17	0.23
Bahrain	1.03	1.93	2.37	0.94	Lithuania	1.03	2.27	2.14	1.17
Bangladesh	1.03	0.02	0.03	0.08	Macedonia	1.03	2.59	1.82	1.33
Belarus	1.03	0.07	0.14	0.35	Malawi	1.03	2.59	1.69	0.63
Belgium	1.03	0.06	0.08	0.17	Mali	1.03	0.6	1.12	1.58
Belize	1.03	0.04	0.07	0.09	Mexico	1.03	2.59	1.04	0.58
Benin	1.03	0.09	0.15	0.28	Morocco	1.03	0.08	0.13	0.23
Bolivia	1.03	0.71	1.05	1.28	Nepal	1.03	2.1	1.89	0.44
Brazil	1.03	0.28	0.52	1.07	Netherlands	1.03	2.58	1.97	0.38
Bulgaria	1.03	0.15	0.22	0.35	New Zealand	1.03	0.29	0.47	1.33
Burkina_faso	1.03	0.04	0.19	0.3	Niger	1.03	1.98	2.21	1.86
Cambodia	1.03	2.28	1.24	1.97	Norway	1.03	2.42	2.73	3.22
Cameroon	1.03	0.17	0.3	0.57	Pakistan	1.03	2.48	1.51	1.79
Canada	1.03	1.92	1.93	0.43	Panama	1.03	1.27	1.88	1.69
Central African Republic	1.03	0.06	0.1	0.15	Paraguay	1.03	0.03	0.06	0.12
Chile	1.03	0.1	0.17	0.26	Peru	1.03	2	2.26	1.93
China	1.03	0.04	0.07	0.09	Philippines	1.03	2.56	1.17	0.95
Colombia	1.03	1.09	1.59	1.27	Portugal	1.03	2.55	1.84	1.08
Costa Rica	1.03	1.85	2.19	1.73	Qatar	1.03	0.05	0.09	0.29
Cote d'Ivoire	1.03	1.55	2.22	0.92	Romania	1.03	2.56	1.86	1.45
Croatia	1.03	2.59	1.15	0.96	Russia	1.03	0.59	1.16	1.47
Czech Republic	1.03	1.19	1.57	1.61	Rwanda	1.03	1.18	1.67	1.8
Denmark	1.03	0.94	2.37	2.85	Saudi Arabia	1.03	0.06	0.12	0.36
Ecuador	1.03	0.25	0.48	1.63	Senegal	1.03	1.3	2.15	1.85
Estonia	1.03	2.59	2.23	1.98	Slovakia	1.03	2.53	1.58	1.62
Ethiopia	1.03	0.05	0.08	0.14	Slovenia	1.03	0.92	1.32	1.9
Finland	1.03	0.11	0.05	0.26	South Africa	1.03	0.31	0.68	0.62
France	1.03	0.76	1.11	1.41	Spain	1.03	0.08	0.14	0.3
Gabon	1.03	0.05	0.1	0.22	Sri Lanka	1.03	0.37	0.86	1.69
Germany	1.03	0.19	0.13	0.18	Sweden	1.03	0.02	0.01	0.07
Ghana	1.03	2.16	2.19	1.38	Switzerland	1.03	0.04	0.24	1.44
Greece	1.03	0.1	0.27	0.5	Taiwan	1.03	0.07	0.09	0.12
Guatemala	1.03	0.34	0.65	1.75	Tajikistan	1.03	0.19	0.28	0.48
Guinea	1.03	2.41	1.75	0.86	Tanzania	1.03	0.06	0.13	0.23
Honduras	1.03	2.22	1.34	1.4	Togo	1.03	2.33	2.2	1.09
Hungary	1.03	0.36	0.63	0.79	Turkey	1.03	0.06	0.13	0.72
Iceland	1.03	0.01	0.01	0.06	Uganda	1.03	0.01	0.03	0.05
India	1.03	0.41	0.56	0.93	Ukraine	1.03	0.2	0.42	0.92
Indonesia	1.03	2.5	1.03	0.87	United Kingdom	1.03	1.82	1.62	1.72
Ireland	1.03	0.85	1.64	2.16	USA	1.03	0.79	1.65	1.91
Israel	1.03	0.34	1.69	1.84	Uzbekistan	1.03	0.16	0.28	0.8
Italy	1.03	0.52	1.01	1.6	Vietnam	1.03	0.11	0.17	0.19
Japan	1.03	2.59	1.29	2.54	Zambia	1.03	2.52	2.1	1.18
Jordan	1.03	2.45	1.29	1.92	Zimbabwe	1.03	2.45	2.25	2.01
Kazakhstan	1.03	0.18	0.32	0.79					

Table G13: Weights used in Estimation (Language)

Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$
Afghanistan	1.09	2.5	2.66	1.32	Lesotho	1.09	2.94	2.41	0.95
Armenia	1.09	0.26	0.4	0.52	Lithuania	1.09	0.03	0.07	0.22
Australia	1.09	2.28	2.58	2.14	Macedonia	1.09	3.1	1.38	1.45
Austria	1.09	0.01	0.02	0.62	Malawi	1.09	1.37	1.66	2.15
Bangladesh	1.09	0.01	0.01	0.04	Mali	1.09	0.12	0.23	0.45
Belarus	1.09	0.01	0.02	0.04	Mauritius	1.09	1.83	2.4	2.34
Belgium	1.09	2.98	1.61	0.43	Mexico	1.09	1.7	2.53	2.74
Belize	1.09	0.01	0.03	0.07	Morocco	1.09	0.01	0.02	0.06
Benin	1.09	0.05	0.11	0.26	Mozambique	1.09	0.12	0.24	0.55
Bolivia	1.09	0.24	0.57	1.65	Namibia	1.09	0.04	0.09	0.2
Brazil	1.09	0.12	0.26	0.91	Nepal	1.09	0.23	0.4	0.49
Bulgaria	1.09	0.79	1.65	2.51	New Zealand	1.09	0.05	0.11	0.29
Burkina Faso	1.09	0.07	0.1	4.06	Nicaragua	1.09	0.54	1.14	2.76
Cambodia	1.09	2.62	2.63	0.95	Niger	1.09	3.12	1.85	2.65
Cameroon	1.09	1.25	2.19	2.76	Nigeria	1.09	0.01	0.03	0.08
Canada	1.09	1.63	2.55	2.43	Norway	1.09	0.02	0.05	2.95
Central African Republic	1.09	0.04	0.09	0.2	Pakistan	1.09	0.02	0.04	0.1
Chile	1.09	0.01	0.03	0.06	Panama	1.09	0.47	0.77	1.2
China	1.09	0.03	0.08	0.22	Paraguay	1.09	0.06	0.12	0.35
Colombia	1.09	0.06	0.12	0.33	Peru	1.09	3.12	1.97	1.88
Costa Rica	1.09	3.12	1.73	2.34	Philippines	1.09	2.34	2.53	2.62
Cote d'Ivoire	1.09	3.1	1.19	0.42	Portugal	1.09	1.81	2.53	2.67
Croatia	1.09	0.61	1.13	1.39	Romania	1.09	2.9	2.58	2.17
Czech Republic	1.09	1.49	2.47	2.6	Russia	1.09	0.01	0.02	0.05
Denmark	1.09	0.07	0.16	3.39	Rwanda	1.09	0.03	0.05	0.12
Ecuador	1.09	0.25	0.56	1.78	Saudi Arabia	1.09	2.57	2.6	1.65
Estonia	1.09	2.77	2.54	0.54	Senegal	1.09	0.07	0.14	0.27
Ethiopia	1.09	0.02	0.04	0.1	Slovakia	1.09	1.83	2.54	1.45
Finland	1.09	0.25	0.51	0.11	Slovenia	1.09	3.11	1.71	1.56
Gabon	1.09	0.35	0.82	1.46	South Africa	1.09	0.01	0.02	0.03
Ghana	1.09	2.34	2.61	2.73	Spain	1.09	0.07	0.14	0.3
Guatemala	1.09	3.01	2.12	0.62	Sweden	1.09	0.01	0.02	0.13
Guinea	1.09	0.88	1.57	2.27	Switzerland	1.09	0	0.01	0.06
Haiti	1.09	0.02	0.04	0.1	Tajikistan	1.09	0.04	0.08	0.23
Honduras	1.09	2.01	2.7	0.27	Tanzania	1.09	0.01	0.02	0.05
Hungary	1.09	0.22	0.43	0.67	Thailand	1.09	0.02	0.03	0.07
Iceland	1.09	3.13	3.81	0.16	Togo	1.09	2.99	1.91	0.89
India	1.09	2.8	2.62	2.13	Turkey	1.09	3.07	1.12	0.68
Indonesia	1.09	3.08	1.65	1.76	Uganda	1.09	0	0.01	0.02
Italy	1.09	0.38	0.68	1.28	Ukraine	1.09	0.02	0.04	0.09
Japan	1.09	3	3.8	0.09	United Kingdom	1.09	2.44	2.48	2.49
Kazakhstan	1.09	0.06	0.1	0.24	USA	1.09	1.75	2.72	1.14
Kenya	1.09	0.26	0.61	1.52	Uzbekistan	1.09	0.02	0.04	0.12
Korea	1.09	0	0.01	0.01	Vietnam	1.09	0.06	0.14	0.35
Kyrgyzstan	1.09	3.01	2.6	2.74	Zambia	1.09	3.12	2.36	2.67
Latvia	1.09	1.56	2.42	1.96	Zimbabwe	1.09	0.01	0.01	0.02

Table G14: Weights used in Estimation (Religion)

Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$	Country	$\hat{\theta}_n^{\text{IV}}$	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\tilde{\tilde{\theta}}_n$
Armenia	1.09	2.5	2.66	1.32	Malawi	1.09	0.38	0.68	1.28
Australia	1.09	0.26	0.4	0.52	Mali	1.09	3	3.8	0.09
Austria	1.09	2.28	2.58	2.14	Mauritius	1.09	0.06	0.1	0.24
Bangladesh	1.09	0.01	0.02	0.62	Mexico	1.09	0.26	0.61	1.52
Belize	1.09	0.01	0.01	0.04	Mozambique	1.09	0	0.01	0.01
Benin	1.09	0.01	0.02	0.04	Namibia	1.09	3.01	2.6	2.74
Brazil	1.09	2.98	1.61	0.43	Nepal	1.09	1.56	2.42	1.96
Bulgaria	1.09	0.01	0.03	0.07	Netherlands	1.09	2.94	2.41	0.95
Burkina Faso	1.09	0.05	0.11	0.26	New Zealand	1.09	0.03	0.07	0.22
Cambodia	1.09	0.24	0.57	1.65	Nicaragua	1.09	3.1	1.38	1.45
Cameroon	1.09	0.12	0.26	0.91	Niger	1.09	1.37	1.66	2.15
Canada	1.09	0.79	1.65	2.51	Nigeria	1.09	0.12	0.23	0.45
Central African Republic	1.09	0.07	0.1	4.06	Pakistan	1.09	1.83	2.4	2.34
Chile	1.09	2.62	2.63	0.95	Paraguay	1.09	1.7	2.53	2.74
Cote d'Ivoire	1.09	1.25	2.19	2.76	Peru	1.09	0.01	0.02	0.06
Croatia	1.09	1.63	2.55	2.43	Philippines	1.09	0.12	0.24	0.55
Czech Republic	1.09	0.04	0.09	0.2	Portugal	1.09	0.04	0.09	0.2
Dominican Republic	1.09	0.01	0.03	0.06	qatar	1.09	0.23	0.4	0.49
Egypt	1.09	0.03	0.08	0.22	Romania	1.09	0.05	0.11	0.29
Estonia	1.09	0.06	0.12	0.33	Russia	1.09	0.54	1.14	2.76
Ethiopia	1.09	3.12	1.73	2.34	Rwanda	1.09	3.12	1.85	2.65
Gabon	1.09	3.1	1.19	0.42	Sao Tome	1.09	0.01	0.03	0.08
Ghana	1.09	0.61	1.13	1.39	Senegal	1.09	0.02	0.05	2.95
Guatemala	1.09	1.49	2.47	2.6	Slovakia	1.09	0.02	0.04	0.1
Guinea	1.09	0.07	0.16	3.39	Slovenia	1.09	0.47	0.77	1.2
Haiti	1.09	0.25	0.56	1.78	South Africa	1.09	0.06	0.12	0.35
Hungary	1.09	2.77	2.54	0.54	Sri Lanka	1.09	3.12	1.97	1.88
India	1.09	0.02	0.04	0.1	Switzerland	1.09	2.34	2.53	2.62
Indonesia	1.09	0.25	0.51	0.11	Tanzania	1.09	1.81	2.53	2.67
Iran	1.09	0.35	0.82	1.46	Thailand	1.09	2.9	2.58	2.17
Ireland	1.09	2.34	2.61	2.73	Togo	1.09	0.01	0.02	0.05
Israel	1.09	3.01	2.12	0.62	Turkey	1.09	0.03	0.05	0.12
Japan	1.09	0.88	1.57	2.27	Uganda	1.09	2.57	2.6	1.65
Kazakhstan	1.09	0.02	0.04	0.1	United Kingdom	1.09	0.07	0.14	0.27
Kenya	1.09	2.01	2.7	0.27	USA	1.09	1.83	2.54	1.45
Korea	1.09	0.22	0.43	0.67	Uzbekistan	1.09	3.11	1.71	1.56
Kyrgyzstan	1.09	3.13	3.81	0.16	Vietnam	1.09	0.01	0.02	0.03
Lithuania	1.09	2.8	2.62	2.13	Zambia	1.09	0.07	0.14	0.3
Madagascar	1.09	3.08	1.65	1.76	Zimbabwe	1.09	0.01	0.02	0.13

Appendix H Algorithms for computing $\hat{\psi}_n(\theta;\nu)$, $\hat{\theta}_n$, $\tilde{\theta}_n$

The following describes the algorithm used to compute $\hat{\psi}_n$ in the simulated and empirical examples. Algorithm 1 relies on explicit gradient calculations with respect to μ and Σ. The updates preserve symmetry and positive definiteness for Σ which makes the iterations more stable than a direct implementation of gradient-descent for instance. A line search is used to update $\psi_b \to \psi_{b+1}$, in practice searching over $\gamma \in \{0.1, 1\}$ provides good results more quickly. The initial $\mu_0 = 0$ is chosen specifically because $\hat{\mu}_n(\hat{\theta}_n;\nu) = 0$ is eventually the solution so that Algorithm 1 tends to speed up as θ gets closer to $\hat{\theta}_n$.

Algorithm 1 Computing $\hat{\psi}_n(\theta;\nu)$

---

1) **Inputs** (a) $\kappa_1, \kappa_2 > 0$, $\nu \geq 1$ (b) tol > 0, maxit ≥ 1, (c) $\mu_0 = 0$, $\Sigma_0 = I_d$.
2) **Iterations**
set $b = 0$, $\psi_0 = (\mu_0, \Sigma_0)$
repeat
    compute $\delta_t = \|g(z_t;\theta) - \mu_b\|^2_{\Sigma_b^{-1}}$, $w_t = (1 + p/\nu)(1 + \delta_t/\nu)$,
    normalize $w_t = \frac{w_t}{\kappa_1/\nu + \sum_t w_t}$, compute $\bar{\mu}_{b+1} = \sum_t w_t g(z_t;\theta)$
    compute $\bar{S} = (I_d + \kappa_2 \Sigma_b/\nu)^{-1}$, center $\bar{x}_t = g(z_t;\theta) - \mu$
    compute $\bar{\Sigma}_{b+1} = \bar{S}\left(\sum_t w_t \bar{x}_t \bar{x}_t' + \kappa_1 \mu\mu'/\nu\right)\bar{S}$
    minimize $Q_n(\gamma\psi_b + (1-\gamma)\bar{\psi}_{b+1};\nu)$ over $\gamma \in [0,1)$, $\bar{\psi}_{b+1} = (\bar{\mu}_{b+1}, \bar{\Sigma}_{b+1})$
    compute $\psi_{b+1} = \gamma^\star \psi_b + (1-\gamma^\star)\bar{\psi}_{b+1}$, $\gamma^\star$ is the arg-minimizer of Q_n above
    increment $b := b + 1$
until $|Q_n(\psi_b) - Q_n(\psi_{b+1})| <$ tol, or $b >$ maxit
3) **Output** estimates $\hat{\psi}_n(\theta;\nu) = \psi_{b+1}$, weights w_t

---

Algorithm 2 describes more specifically the steps used to minimize $\|\tilde{\mu}_n(\theta)\|^2_{W_n}$. It is a Gauss-Newton algorithm where the Jacobian is approximated using the weighted average representation rather than a more costly computation based on the implicit function Theorem. For OLS, $\tilde{G}_n(\theta) = -\sum_t \tilde{w}_t(\theta;\nu) x_t x_t'$, and IV $\tilde{G}_n(\theta) = -\sum_t \tilde{w}_t(\theta;\nu) z_t x_t'$. Although the Jacobian $\tilde{G}_n(\theta)$ is inexact, the Gauss-Newton algorithm performed well in the simulated and empirical applications. The Algorithm is essentially the same when computing $\hat{\theta}_n$ or $\tilde{\tilde{\theta}}_n$.

Algorithm 2 Computing $\tilde{\theta}_n$

---

1) **Inputs** (a) $\kappa_1, \kappa_2 > 0$, $\nu \geq 1$ (b) tol > 0, maxit ≥ 1, $\gamma \in (0,1)$ (c) inital guess θ_0.
2) **Iterations**
set $b = 0$,
repeat
    compute $\hat{\psi}_n(\theta_b;\nu), \hat{\psi}_n(\theta_b;\nu/2)$
    compute $\tilde{\mu}_n(\theta) = 2\hat{\mu}_n(\theta;\nu) - \hat{\mu}_n(\theta;\nu/2)$ and $\tilde{w}_t(\theta;\nu) = 2w_t(\theta;\nu) - w_t(\theta;\nu/2)$
    compute $\tilde{G}_n(\theta) = \sum_t \tilde{w}_t(\theta;\nu)\partial_\theta g(z_t;\theta)$
    update $\theta_{b+1} = \theta_b - \gamma\left(\tilde{G}_n(\theta)' W_n \tilde{G}_n(\theta)\right)^{-1} \tilde{G}_n(\theta)' W_n \tilde{\mu}_n(\theta_b)$
    increment $b := b + 1$
until $\|\tilde{\mu}_n(\theta_{b+1})\|_{W_n} <$ tol, or $b >$ maxit
3) **Output** estimates $\tilde{\theta}_n = \theta_{b+1}$, weights $\tilde{w}_t$
