No. 2009–25

GENERALIZED METHOD OF TRIMMED MOMENTS

By Pavel Čížek

April 2009

ISSN 0924-7815

GENERALIZED METHOD OF TRIMMED MOMENTS

PAVEL ČÍŽEK[1]

Department of Econometrics and Operation Research

Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands[1]

Abstract. High breakdown-point regression estimators protect against large errors and data contamination. We adapt and generalize the concept of trimming used by many of these robust estimators so that it can be employed in the context of the generalized method of moments. The proposed generalized method of trimmed moments (GMTM) offers a globally robust estimation approach (contrary to existing only locally robust estimators) applicable in econometric models identified and estimated using moment conditions. We derive the consistency and asymptotic distribution of GMTM in a general setting, propose a robust test of overidentifying conditions, and demonstrate the application of GMTM in the instrumental variable regression. We also compare the finite-sample performance of GMTM and existing estimators by means of Monte Carlo simulation.

Keywords: asymptotic normality, generalized method of moments, instrumental variables regression, robust estimation, trimming

JEL codes: C13, C20, C30, C12

1. Introduction

The generalized method of moments (GMM; Hansen, 1982) and related procedures are important econometric tools for estimation and inference in models based on moment conditions. During last two decades, the estimation by GMM has been enhanced in many areas, which include primarily its behavior in small and moderate samples (e.g., Altonji and Segal, 1996; Imbens et al., 1998; Newey and Smith, 2004) and its robustness against small deviations from the assumed model (e.g., Ronchetti and Trojani, 2001; Honore and Hu, 2004; Lo and Ronchetti, 2006). In this paper, we concentrate on the second area and propose the generalized method of trimmed moments that is, contrary to most existing methods, robust to large deviations from the model and that can achieve practically the same variance of estimates as the original GMM in many situations. By being robust to small or large deviations from the model, we mean how large is the smallest fraction of a sample that, if modified in

Date:

[1]*Corresponding author.* Tel.: +31 13 466 8723. Fax: +31 13 466 3280. Email: P.Cizek@uvt.nl.

some way (e.g., by data contamination or heterogeneity not presumed by the model), can arbitrarily change the estimates under consideration. This measure is called breakdown point (see Rousseeuw and Leroy, 1987, for the standard definition and Genton and Lucas, 2003, for a discussion of the breakdown point under dependence) and it is asymptotically equal to zero for the majority of typically used GMM estimators (see Ronchetti and Trojani, 2001, for a discussion of the robust properties of GMM).

The need for robust estimation methods have been demonstrated in various contexts both theoretically by Krasker and Welsch (1985), Hampel et al. (1986), Peracchi (1990), Hubert and Rousseeuw (1997), Krishnakumar and Ronchetti (1997), Ferretti et al. (1999), Cantoni and Ronchetti (2001), Genton and Ronchetti (2003), Bramati and Croux (2007), and Čížek (2008b), for instance, and in real (GMM) applications by Knez and Ready (1997), Temple (1998), Sakata and White (1998), Dell'Aquila et al. (2003), and Czellar et al. (2007), for instance. In the case of GMM and its particular applications such as the linear instrumental variable (IV) regression, existing research concentrates on the quantile-based GMM estimation (e.g., see Amemiya, 1982, Honore and Hu, 2004, and Chernozhukov and Hansen, 2008, in IV regression) and on the M-estimation (e.g., see Krasker, 1986, Peracchi, 1991, and Krishnakumar and Ronchetti, 1997, in simultaneous equation models; Müller and Kim, 2005, and Wagenvoort and Waldmann, 2002, in linear panel data; and Ronchetti and Trojani, 2001, and Ortelli and Trojani, 2005, for general GMM estimation). All mentioned robust methods applicable in models estimated by IV or GMM are however only locally robust and usually cannot withstand large deviations from the model (see He et al., 1990, Čížek, 2008c, and Section 4 for the methods based on quantile regression and Maronna et al., 1979, and Ronchetti and Trojani, 2001, for the GMM based on M-estimation). Even though the M-estimators can be made more robust by means of one-step estimation (Simpson et al., 1992) as in Wagenvoort and Waldmann (2002), such a procedure nevertheless requires an initial highly robust estimator, which is not available for general method-of-moments estimation so far.

Hence, we aim to propose a high breakdown-point estimator for models based on general nonlinear moment conditions. Motivated by the least trimmed squares (Rousseeuw, 1985), maximum trimmed likelihood (Hadi and Luceno, 1997), and general trimmed extremum (Čížek, 2008a) regression estimators, which eliminate the influence of deviating observations on estimates by trimming the observations from estimators' objective functions,

we propose the generalized method of trimmed moments (GMTM). For a given model, the GMTM method relies on the moment conditions characterizing the model that are extended in order to include trimming of observations inconsistent with the original moment conditions. Because GMTM represents a very general concept, we demonstrate several ways to create trimmed moment conditions in the case of linear IV regression and discuss a data-dependent choice of trimming designed to minimize the number of trimmed observations. Furthermore, since the proposed trimming of observations in the moment conditions depends implicitly on the underlying parameter values and is thus endogenous, GMTM requires new asymptotic theory. We therefore study the consistency and asymptotic distribution of GMTM, discuss its implications for the estimation, and propose a GMTM analog of the test of overidentifying conditions (Hansen, 1982). On the other hand, the breakdown properties of GMTM will not be derived in general because they are model- and data-dependent in nonlinear models or under dependence (Genton and Lucas, 2003); we discuss the robust properties of GMTM only in the linear IV regression. Finally, we also do not address here questions concerning weak identification in the context of (robust) GMM estimation, although the extension of the current results along the lines of Stock and Wright (2000) is relatively straightforward.

In the rest of the paper, we first propose the GMTM estimator in Section 2, where we also provide various examples of GMTM in linear IV regression and discuss how the number of trimmed observations can be chosen in a data-dependent way. Assumptions needed for studying the asymptotic properties of GMTM as well as the main asymptotic results are summarized in Section 3. Later, the proposed and some existing estimators are studied by means of Monte Carlo simulations in Section 4. The proofs are provided in Appendix.

2. Generalized method of trimmed moments

Let us now introduce the generalized method of trimmed moments (Section 2.1) and demonstrate its use in the context of linear IV regression (Section 2.2). Later, a data-dependent choice of the trimming amount is discussed (Section 2.3).

2.1. **Generalized method of trimmed moments estimator.**

To introduce the idea of trimming, let us consider data $\{d_i\}_{i=1}^n = \{(y_i, x_i)\}_{i=1}^n$ and a linear regression model with intercept

$$y_i = x_i^\top \beta + \varepsilon_i, \tag{2.1}$$

where $\beta \in \mathbb{R}^p$ denotes the vector of unknown parameters. Assuming $\mathsf{E}(\varepsilon_i|x_i) = 0$ and $\mathsf{E}(x_i x_i^\top) > 0$, the standard least squares (LS) estimator $\hat{\beta}_n^{(LS)}$ is consistent, but very non-robust: being a linear function of y_i, a single outlying observation can arbitrarily change the value of $\hat{\beta}_n^{(LS)}$ and its breakdown point is thus at most $1/n$ and equals asymptotically zero (He et al., 1990).

To achieve a high breakdown point, many robust methods exclude (or downweight) observations unlikely under a model from their objective functions (e.g., Hadi and Luceno, 1997; Stromberg et al., 2000; and Čížek, 2008a). A well-known alternative to LS is, for example, the least trimmed squares (LTS) estimator (Rousseeuw, 1985), which minimizes the trimmed sum of the h_n smallest squared residuals:

$$\hat{\beta}_n^{(LTS)} = \arg\min_{\beta \in B} \sum_{j=1}^{h_n} e_{(j)}^2(\beta), \tag{2.2}$$

where $e_{(j)}^2(\beta)$ represents the jth smallest order statistics of squared residuals $e_i^2(\beta) = (y_i - x_i^\top \beta)^2$, $i = 1, \ldots, n$, and $n/2 < h_n \leq n$ is the trimming amount. By (endogenously) excluding $n - h_n$ observations from the objective function, LTS becomes insensitive to the presence of data inconsistent with the linear model. In general, $n/2 < h_n$ because we cannot distinguish which part of the data should be fit by the model and which one should be rejected if $h_n \leq n/2$. Thus for the maximum amount of trimming, $(n - h_n)/n \to 1/2$ as $n \to \infty$ and the breakdown point of LTS then converges asymptotically to 1/2, the maximum possible value for affine-equivariant estimators (Rousseeuw and Leroy, 1987).

The LTS estimator can be alternatively expressed also by means of moment conditions. If $e_i(\beta)$ is continuously distributed, Čížek (2006) showed that (2.2) can be expressed as

$$\hat{\beta}_n^{(LTS)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} e_i^2(\beta) \cdot I\{e_i^2(\beta) \leq e_{(h_n)}^2(\beta)\}$$

and that the corresponding first-order conditions for the LTS estimator are

$$2 \sum_{i=1}^{n} e_i(\beta) x_i \cdot I\{e_i^2(\beta) \leq e_{(h_n)}^2(\beta)\} = 0, \tag{2.3}$$

where $I(\cdot)$ is the indicator function. Note that the normal equations (2.3) consist of two parts: one corresponding to the LS moment conditions, $\sum_{i=1}^n e_i(\beta) x_i = 0$, and another one performing trimming of observations with large squared residuals, $I\{e_i^2(\beta) \leq e_{(h_n)}^2(\beta)\} = 0$, where $e_{(h_n)}^2(\beta)$ approximates a quantile of the distribution of squared residuals $e_i^2(\beta)$.

To generalize, let us now consider a stationary data sequence $\{d_i\}_{i=1}^n$, $d_i \in \mathbb{R}^k$, and a function $s : \mathbb{R}^k \times B \to \mathbb{R}^M$ that imposes a set of unconditional moment conditions

$$\mathsf{E}\, s(d_i; \beta^0) = 0 \tag{2.4}$$

on the underlying model. We also assume that $\beta^0 \in B \subset \mathbb{R}^p$ is the unique solution of (2.4) and that the number M of conditions is equal to or larger than the number p of parameters. The GMM estimator proposed by Hansen (1982) is then defined by

$$\hat{\beta}_n^{(GMM)} = \arg\min_{\beta \in B} Q_n^W(\beta) = \arg\min_{\beta \in B} \left[\frac{1}{n}\sum_{i=1}^n s(d_i;\beta)\right]^\top W \left[\frac{1}{n}\sum_{i=1}^n s(d_i;\beta)\right], \tag{2.5}$$

where W is a positive definite $M \times M$ matrix and $\sum_{i=1}^n s(d_i;\beta)/n$ represents the sample equivalent of (2.4).

Typically relying on an unbounded moment function s, the GMM estimator is not robust as a single data point can have an arbitrarily large influence of the GMM estimates (Ronchetti and Trojani, 2001). To improve robust properties of GMM, trimming of observations similar to (2.3) could be employed. Therefore, we now propose to base the estimation on the trimmed moment conditions

$$\mathsf{E}\left[s(d_i;\beta^0) \cdot I\{r(d_i;\beta^0) \le G_{\beta^0}^{-1}(\lambda)\}\right] = 0 \tag{2.6}$$

instead of conditions (2.4), where function $s(d_i;\beta)$ represents the original moment condition, $r(d_i;\beta) : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}$ is a general trimming function that ranks observations and determines their inclusion in or trimming from the objective function, and $G_\beta^{-1}(\lambda)$ denotes the λ-quantile of the distribution of $r(d_i;\beta)$, $1/2 < \lambda \le 1$; λ is referred here as the trimming constant. For example in the case of linear regression (2.1), the LTS estimator (2.2) corresponds to setting $s(d_i;\beta) = (y_i - x_i^\top\beta)x_i = e_i(\beta)x_i$ and $r(d_i;\beta) = (y_i - x_i^\top\beta)^2 = e_i^2(\beta)$, see equation (2.3). In general, the trimming function $r(d_i;\beta)$ should be designed so that its small values indicate likely observations ("good fit", small squared residuals, high likelihood) and its large values indicate unlikely observations ("bad fit", large squared residuals, low likelihood) in a given model (Čížek, 2008a). Apart from weak regularity assumptions, the only other requirement on $r(d_i;\beta)$ is that the trimmed moment equation (2.6) holds.

To construct a sample equivalent of (2.6), $G_\beta^{-1}(\lambda)$ is replaced by the $[\lambda n]$th smallest order statistics of $r(d_i;\beta)$, where $[t]$ represents t rounded to the closest integer value. Consequently,

the generalized method of trimmed moments can be defined by

$$\hat{\beta}_n^{(GMTM,\lambda)} = \arg\min_{\beta \in B} Q_n^{W,\lambda}(\beta), \tag{2.7}$$

where

(2.8)

$$Q_n^{W,\lambda}(\beta) = \left[\frac{1}{n}\sum_{i=1}^{n} s(d_i;\beta) I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\}\right]^\top W \left[\frac{1}{n}\sum_{i=1}^{n} s(d_i;\beta) I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\}\right].$$

Although this definition is analogous to the standard GMM, the use of trimmed moments (2.6), which trim observations depending on the values of all variables d_i and model parameters β, requires a new asymptotic theory and results that establish the behavior of the proposed GMTM method (see Section 3). Further note that, for the asymptotic analysis of GMTM, we can assume $\lambda \in (0, 1\rangle$, whereas the robustness and equivariance properties of GMTM impose $\lambda \in \langle 1/2, 1\rangle$, $\lambda = 1/2$ being the most robust choice in many continuous-response models (e.g., see Müller and Neykov, 2003, for the case of generalized linear models). Thus, λ close to $1/2$ can produce very robust consistent estimates, but on the other hand, it will probably lead to much larger variances of estimates than $\lambda = 1$, that is, the original GMM (2.5) without any trimming. A data-dependent choice of λ, which combines high robustness and small variances of estimates, is discussed later in Section 2.3.

2.2. **Linear IV regression.** To demonstrate possible implementations and uses of trimming, let us consider the linear IV regression model with $y_i = x_i^\top \beta + \varepsilon_i$ as in (2.1), $\mathsf{E}(\varepsilon_i | x_i) \neq 0$, and $\mathsf{E}(\varepsilon_i | z_i) = 0$, where z_i represents a vector of instrumental variables; the data vector d_i equals then to $d_i = (y_i, x_i^\top, z_i^\top)^\top$. The standard IV and GMM estimators are based on the identification condition $\mathsf{E}(\varepsilon_i | z_i) = 0$ (together with other assumptions such as $\dim(z_i) \ge \dim(x_i)$ and x_i and z_i being correlated), which implies the unconditional moment conditions $\mathsf{E}(\varepsilon_i z_i) = 0$ and

$$\mathsf{E}\, s^{IV}(d_i;\beta) = \mathsf{E}[(y_i - x_i^\top \beta) z_i] = 0. \tag{2.9}$$

In the case of exact identification, $\dim(z_i) = \dim(x_i)$, $\beta = \{\mathsf{E}(z_i x_i^\top)\}^{-1} \mathsf{E}(z_i y_i)$, and the sample analog is

$$\hat{\beta}_n^{(IV)} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right).$$

Being a linear function of responses y_i like LS, the IV estimator is obviously very sensitive to outliers as even a single large observation can arbitrarily change the estimate $\hat{\beta}_n^{(IV)}$ as noted already by Krasker and Welsch (1985).

A robust alternative can be provided by the proposed GMTM estimator (2.7), which solves the trimmed moment equations (2.6):

$$(2.10) \quad \mathsf{E}\Big[s^{IV}(d_i;\beta)\cdot I\{r(d_i;\beta)\le G_\beta^{-1}(\lambda)\}\Big] = \mathsf{E}\Big[(y_i - x_i^\top\beta)z_i\cdot I\{r(d_i;\beta)\le G_\beta^{-1}(\lambda)\}\Big] = 0.$$

This trimmed instrumental variable (TIV) estimator however requires a choice of the trimming function $r(d_i;\beta)$. Analogously to LTS in (2.3), a (seemingly) straightforward protection against outliers in the dependent variable y_i could be implemented by setting $r(d_i;\beta) \equiv r^e(d_i;\beta) = (y_i - x_i^\top\beta)^2$ in (2.10). The corresponding GMTM estimator using $s \equiv s^{IV}$ and $r \equiv r^e$ will be denoted TIV-TE and corresponds to the linear IV method by Víšek (2006).

Before analyzing the robust properties of TIV-TE, let us discuss the parameter identification. Similarly to the linear regression case and LTS, the standard (2.9) and trimmed (2.10) moment conditions identify the same set of parameters if the distribution function of $\varepsilon_i = y_i - x_i^\top\beta^0$ is symmetric because the trimming by $r^e(d_i;\beta^0) = \varepsilon_i^2$ is symmetric around zero (Čížek, 2006). If the underlying distribution of ε_i is not symmetric, the slope estimates are still identified and consistently estimated, see Marazzi and Yohai (2004). On the other hand, the trimmed equation for intercept β_0 identifies instead of the usual mean value $\beta_0 = \mathsf{E}\,y_i - (\beta_1,\dots\beta_{p-1})\,\mathsf{E}(x_{1i},\dots,x_{p-1i})^\top$ a different value $\tilde{\beta}_0 = \beta_0 + \mathsf{E}\{\varepsilon_i I(\varepsilon_i \le G_{\beta^0}^{-1}(\lambda)\} \ne \beta_0$, where $\beta = (\beta_0,\dots,\beta_{p-1})^\top$ and $x_i = (1, x_{1i},\dots,x_{p-1i})^\top$. The lack of "mean identification" is a common feature of practically all positive breakdown-point regression estimators applicable under asymmetric errors: for example, the median regression (Bassett and Koenker, 1978) estimates medians rather than means and generalized S-estimators (Croux et al., 1994; Stromberg et al., 2000) do not identify intercept at all. If the intercept estimate is needed, one can use $\tilde{\beta}_0$, use some other intercept estimate such as the median, or compute β_0 by evaluating $\mathsf{E}\{\varepsilon_i I(\varepsilon_i \le G_{\beta^0}^{-1}(\lambda)\}$ for an assumed parametric family of ε_i distributions as in Marazzi and Yohai (2004).

Returning to the robust properties of the TIV-TE estimator, it protects against the extreme influence of observations with large residuals $r^e(d_i,\beta)$ on the estimates by trimming them from the moment equation (2.10). This mechanism is similar in spirit to the IV estimators based

on the median conditions (Med-IV) such as

$$\mathsf{E}\{\mathrm{sgn}(y_i - x_i^\top \beta) z_i\} = 0 \tag{2.11}$$

(Honore and Hu, 2004) in the sense that $\mathrm{sgn}(y_i - x_i^\top \beta)$ is not influenced by large values of residuals (only by their signs). In both cases, the protection against large values of residuals however does not guarantee that estimates cannot be arbitrarily changed, for example, if additionally atypical or erroneous values of instruments z_i occur in data: a large value of a particular instrument value z_i gives a disproportionally large weight to the residual $y_i - x_i^\top \beta$ in (2.9), (2.10), or (2.11), which can lead to an estimation bias and possible breakdown of an estimator even in the presence of a single contaminated observation (cf. He et al., 1990, and Wagenvoort and Waldmann, 2002).

On the other hand, the results of He et al. (1990) for the quantile-regression and M-estimators indicate that the estimators can reach a positive (although design-dependent) breakdown point if the values of the instruments z_i in (2.10) for $r \equiv r^e$ or in (2.11) are bounded. Since transforming the instruments z_i does not invalidate the consistency of GMM as long as the employed moment conditions stay valid, one way to add protection against atypical values in z_i is their standardization. Specifically, we propose replacing z_i by $z_i/\|z_i\|$ and using $s^{SIV}(d_i;\beta) = (y_i - x_i^\top \beta) z_i/\|z_i\|$ to obtain trimmed moment conditions

$$\mathsf{E}\Big[s^{SIV}(d_i;\beta) \cdot I\{r^e(d_i;\beta) \le G_\beta^{-1}(\lambda)\}\Big] = \mathsf{E}\Big[(y_i - x_i^\top \beta) z_i/\|z_i\| \cdot I\{r^e(d_i;\beta) \le G_\beta^{-1}(\lambda)\}\Big] = 0. \tag{2.12}$$

The corresponding GMTM estimator using $s \equiv s^{SIV}$ and $r \equiv r^e$ will be denoted TIV-TESZ.

While normalizing instruments can make the TIV-TE and Med-IV estimators globally robust (although the size of the breakdown point generally depends on the design of z_i), the generality of GMTM also allows for another protection against observations "incompatible" with the moment conditions (2.9). For example, we can trim observations with large contributions to the moment conditions (because a single large value can arbitrarily change the sample average). Defining $r(d_i;\beta) \equiv r^{ez}(d_i;\beta) = \|(y_i - x_i^\top \beta) z_i\|^2$ as the Euclidean norm of the moment contribution $(y_i - x_i^\top \beta) z_i$, we can use the trimmed moment conditions based on the original moments $s^{IV}(d_i;\beta) = (y_i - x_i^\top \beta) z_i$ with the unmodified instruments:

$$\mathsf{E}\Big[s^{IV}(d_i;\beta) I\{r^{ez}(d_i;\beta) \le G_\beta^{-1}(\lambda)\}\Big] = \mathsf{E}\Big[(y_i - x_i^\top \beta) z_i \cdot I\{r^{ez}(d_i;\beta) \le G_\beta^{-1}(\lambda)\}\Big] = 0. \tag{2.13}$$

The corresponding GMTM estimator using $s \equiv s^{IV}$ and $r \equiv r^{ez}$ will be denoted TIV-TETZ. The main advantage of this approach is its generality compared to TIV-TESZ or Med-IV. Whereas the robustness of TIV-TESZ achieved by standardizing the instruments relies on the linearity of the moment conditions $s^{IV}(d_i;\beta)$, the trimming by the norm of the moment contribution $\|s^{IV}(d_i;\beta)\|$ is applicable in general nonlinear models. On the other hand, note that the previous discussion of the intercept and slopes identification also applies to (2.13) because trimming $r^{ez}(d_i;\beta^0)$ is symmetric with respect to $\varepsilon_i = y_i - x_i^\top\beta^0$ conditional on z_i. More detailed comparison of the proposed robust IV estimators is in Section 4.

2.3. **Adaptive choice of trimming.** While trimming 30% or 50% observations can well protect estimates against the influence of outliers, erroneous, and atypical observations, eliminating many observations from an estimator's objective function will intuitively lead to a worse performance of the estimator: less observations imply a higher variance. On the other hand, the moment conditions (2.4) usually depend on the (unobservable) error term expressed as a function of observables, $\varepsilon_i = e(d_i;\beta^0)$, and trimming will typically protect against observations unlikely in a given model, that is, observations with large values of regression residuals $e(d_i;\beta)$. For example in the linear IV regression, the moment conditions (2.9) equal $\mathsf{E}\{e(d_i;\beta)z_i\} = \mathsf{E}\{(y_i - x_i^\top\beta)z_i\} = 0$ and trimming in the TIV estimators depends on $e^2(d_i;\beta) = (y_i - x_i^\top\beta)^2$. Therefore, the choice of the trimming constant λ in (2.6)–(2.8) can be made data-dependent by looking at the tail behavior of $e(d_i;\beta)$ as proposed by Gervini and Yohai (2002).

Specifically, even though GMM estimators do not typically require the error term ε_i to be from a specific parametric family of distributions, GMM for a given model often performs optimally under some specific parametric distribution $\varepsilon_i \sim F_\theta, \theta \in \Theta$. For example, LS in the standard linear regression (2.1) require only $\mathsf{E}(\varepsilon_i|x_i) = 0$, but LS perform optimally if the error term is normally distributed, $\varepsilon_i \sim N(0,\theta), \theta \in \mathbb{R}_+$. Consequently, we can determine the fraction $\hat\lambda_n$ of sample observations having residuals consistent with the assumption $\varepsilon_i \sim F_\theta$ (in its tail) and trim only remaining $n - [\hat\lambda_n n]$ observations in GMTM.

Such an adaptive choice of trimming was proposed by Gervini and Yohai (2002) in linear regression. Let us assume that we obtain initial robust estimates $\hat\beta_n^0$ and $\hat\theta_n^0$ of the regression parameters β and distribution parameters θ, for example, by using GMTM with $\lambda = 1/2$ and $\hat\theta_n^0 = 1.4826 \cdot \mathrm{MAD}_{i=1,\dots,n}\, e_i(d_i;\hat\beta_n^0)$ if $F_\theta \equiv N(0,\theta)$, where MAD denotes the median

absolute deviation. The choice of trimming is then done by comparing the empirical distribution function $\hat{F}_n^0$ of the absolute residuals $|e(d_i;\hat{\beta}_n^0)|$ and the estimated optimal distribution function $\hat{F}_{|\cdot|}(z) = F_{\hat{\theta}_n^0}(z) - F_{\hat{\theta}_n^0}(-z)$ of $|\varepsilon_i|$ under the assumption $\varepsilon_i \sim F_\theta$, where F_θ is symmetric (equivalently, squared residuals can be used). The two distributions are compared by measuring the largest difference between $\hat{F}_n^0$ and $\hat{F}_{|\cdot|}$ in the tail of the distributions,

$$d_n = \sup_{t \geq c} \max\{0, \hat{F}_{|\cdot|}(t) - \hat{F}_n^0(t)\}, \tag{2.14}$$

where the cut-off point c equals 99% or 99.5% quantile of $\hat{F}_{|\cdot|}$. Using this measure, the data-dependent choice of trimming is determined by $\hat{\lambda}_n = 1 - d_n$. In the linear regression (2.1), GMTM with this data-dependent choice of trimming corresponds to LTS with the same choice of trimming, is asymptotically equivalent to LS under normality, and at the same time, it preserves the breakdown point of the initial estimator $\hat{\beta}_n^0$ (Gervini and Yohai, 2002). It also performs very well under various light- and heavy-tailed distributions and under heteroscedasticity despite "assuming" the same distribution for all data in (2.14) (Čížek, 2007a).

Finally, let us note that the comparison of the empirical and optimal distributions in (2.14) was done for the absolute values of residuals, $|e(d_i;\beta)|$, as proposed by Gervini and Yohai (2002) because the trimming by the TIV estimators in Section 2.2 depends on $e^2(d_i;\beta)$, which is symmetric around 0 and is equivalent to trimming using $|e(d_i;\beta)|$. In a general case with a possibly asymmetric distribution F_θ and trimming, we can construct $\hat{\lambda}_n$ by comparing the empirical distribution function $\hat{F}_n^0$ of the residuals $e(d_i;\hat{\beta}_n^0)$ and the distribution function $\hat{F}(z) = F_{\hat{\theta}_n^0}(z)$ of ε_i under the assumption $\varepsilon_i \sim F_\theta$ in both tails, for example:

$$d_n = \sup_{t \leq \underline{c}} \max\{0, \hat{F}_n^0(t) - \hat{F}(t)\} + \sup_{t \geq \overline{c}} \max\{0, \hat{F}(t) - \hat{F}_n^0(t)\}, \tag{2.15}$$

where $\underline{c}$ and $\overline{c}$ represent the 0.5% and 99.5% quantiles of $\hat{F}$, respectively.

3. Asymptotic properties of GMTM

In this section, we introduce the assumptions for the asymptotic analysis of GMTM (Section 3.1), derive the main asymptotic properties of GMTM (Section 3.2), and propose a test of overidentifying conditions (Section 3.3).

3.1. **Assumptions.** Let us now complement the GMTM definition first by some notation and definitions and later by assumptions on the random variables and moment and trimming functions needed for further analysis.

First, we refer to the distribution function of $r(d_i;\beta)$ in (2.6) as $G_\beta(z)$ and to the corresponding probability density function as $g_\beta(z)$ if it exists. We also use a simpler notation $G \equiv G_{\beta^0}$ and $g \equiv g_{\beta^0}$ at the true parameter value β^0. Whenever we need to refer to the quantile function corresponding to G_β, notation G_β^{-1} is used. Next, because the derivatives of functions $s(d;\beta)$ and $r(d;\beta)$ are taken only with respect to β here, we denote them simply by $s'(d;\beta)$, $r'(d;\beta)$, ... meaning $\partial s(d;\beta)/\partial\beta^\top$, $\partial r(d;\beta)/\partial\beta$, ... We also need a notation for an open δ-neighborhood of a point x in a Euclidean space $\mathbb{R}^l$: $U(x,\delta) = \left\{ z \in \mathbb{R}^l \middle| \, \|z - x\| < \delta \right\}$.

Second, let us introduce the concept of β-mixing, which is central to the distributional assumptions made in this paper. A sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$ is said to be absolutely regular (or β-mixing) if $\beta_m = \sup_{t\in\mathbb{N}} \mathsf{E} \sup_{B\in\sigma_{t+m}^f} |P(B|\sigma_t^p) - P(B)| \to 0$ as $m \to \infty$, where the σ-algebras $\sigma_t^p = \sigma(X_t, X_{t-1}, \ldots)$ and $\sigma_t^f = \sigma(X_t, X_{t+1}, \ldots)$; see Davidson (1994) or Arcones and Yu (1994) for details. Numbers $\beta_m, m \in \mathbb{N}$, are called mixing coefficients.

Now, I specify all the assumptions necessary to derive the consistency and asymptotic normality of GMTM (a smaller subset of assumptions sufficient for the consistency of GMTM is discussed at the end of the section). They form three groups: distributional Assumptions D for random variables d_i, Assumptions F concerning properties of the moment function $s(d;\beta)$ and auxiliary trimming function $r(d;\beta)$, and finally, identification Assumptions I.

Assumptions D.

D1: Random variables $\{d_i\}_{i\in\mathbb{N}}$ form a strongly stationary absolutely regular sequence of random vectors with mixing coefficients satisfying $m^{r_\beta/(r_\beta-2)} (\log m)^{2(r_\beta-1)/(r_\beta-2)} \beta_m \to 0$ as $m \to +\infty$ for some $r_\beta > 2$.

D2: The distribution function G_β of $r(d_i;\beta)$ is absolutely continuous for any $\beta \in B$.

D3: Assume that for $m_G = \inf_{\beta\in B} G_\beta^{-1}(\lambda)$ and $M_G = \sup_{\beta\in B} G_\beta^{-1}(\lambda)$, it holds that

$$M_{gg} = \sup_{\beta\in B} \sup_{z\in(m_G-\delta_g, M_G+\delta_g)} g_\beta(z) < \infty$$

and

$$m_{gg} = \inf_{\beta\in B} \inf_{z\in(-\delta_g,\delta_g)} g_\beta\left(G_\beta^{-1}(\lambda) + z\right) > 0$$

for some $\delta_g > 0$.

Having a general moment function $s(d;\beta)$, Assumption D1 is one of relatively weak conditions for the uniform central limit theorem used by Andrews (1993) and Arcones and Yu (1994), for instance. Assumption D2 indicates that at least one random variable has to be continuously distributed so that trimming by $r(d_i;\beta)$ is well defined (note though that the absolute continuity of G_β is really necessary only in a neighborhood of its λ-quantile $G_\beta^{-1}(\lambda)$ as used in Assumption D3; see its discussion below for more details). Moreover, Assumption D2 is purposely formulated in a simple way, which however seem to exclude distributional variation such as heteroscedasticity across observations. That is not the case: for example, if d_i includes both observable and unobservable random variables u_i driving heteroscedasticity in data, then the distribution of $r(d_i;\beta)$ conditional on u_i changes across different realizations (observations) of u_i even though $r(d_i;\beta)$ as a function of observables does not explicitly refer to unobservables u_i contained in d_i. Nevertheless, we do not need the conditional distributions of $r(d_i;\beta)$ at each $i \in \mathbb{N}$ to study the behavior of trimmed moments, but rather the unconditional univariate distribution function of $r(d_i;\beta)$, which "averages out" all differences in distribution across observations (there is one common trimming point for all $r(d_i;\beta)$). Alternatively, if d_i contains only observed quantities and the distribution of $r(d_i;\beta)$ varies with $i \in \mathbb{N}$, we could define $G_\beta = \lim_{n\to\infty} G_\beta^n$, where G_β^n denotes the distribution function of $r(d_{U_n};\beta)$ and U_n is the random variable attaining all values $1,\dots,n$ with probability $1/n$.

Further, Assumption D3 formalizes two things. First, the density function g_β has to be bounded uniformly in $\beta \in B$, which prevents distribution G_β to become or to be arbitrarily close to a discrete or singular one for some $\beta \in B$. Second, the density function has to be positive in a neighborhood of the λ-quantile of G_β, that is, around the chosen "trimming" point of the $r(d_i;\beta)$ distribution. In a less general setting when the structure of a model is known and $r(d_i;\beta)$ is differentiable, Assumption D3 is usually implied by $G \equiv G_{\beta^0}$ being absolutely continuous with a density function $g \equiv g_{\beta^0}$ positive, bounded, and differentiable around $G^{-1}(\lambda)$; see Čížek (2006) for nonlinear regression. Let us recall here that differentiability of the density function g is a standard condition needed for the asymptotic analysis of rank statistics (e.g., see Hössjer, 1994, and Zinde-Walsh, 2002).

Next, several conditions on the moment function $s(d;\beta)$ and auxiliary trimming function $r(d;\beta)$ have to be specified. The GMTM concept aims to add robust qualities to moment estimators that lack robustness, but preferably possess other desirable properties such as asymptotic normality and some kind of optimality. Since an estimator's objective function

typically has to be smooth to guarantee such properties, we will assume that both functions $s(d;\beta)$ and $r(d;\beta)$ are differentiable, at least in a neighborhood $U(\beta^0,\delta)$ of β^0. Similarly to the GMM estimator, the asymptotic variance of GMTM will then depend on the expectations of the moment function and its derivatives (cf. Manski, 1988). Specifically, it will depend on the variance of the trimmed moment equations (cf. Hansen, 1982, p. 1042),

(3.1)
$$V_s(\lambda) = \mathsf{E}\left[\sum_{k=-\infty}^{\infty} s(d_i;\beta^0)s(d_{i-k};\beta^0)^\top \cdot I\{r(d_i;\beta^0) \leq G^{-1}(\lambda)\}I\{r(d_{i-k};\beta^0) \leq G^{-1}(\lambda)\}\right],$$

and on the expected value of the derivative of the moment equations with respect to parameters β, which, by the product rule, consists of the trimmed derivative of the moment function,

$$J_s(\lambda) = \mathsf{E}\left[s'(d_i;\beta^0) \cdot I\{r(d_i;\beta^0) \leq G^{-1}(\lambda)\}\right], \tag{3.2}$$

and the derivative of the expectation of the trimming indicator in the moment equations,

$$J_I(\lambda) = \frac{\partial}{\partial\beta^\top}\mathsf{E}\left[s(d_i;\beta^0) \cdot I\{r(d_i;\beta) \leq G_\beta^{-1}(\lambda)\}\right]\Big|_{\beta=\beta^0}. \tag{3.3}$$

Assumptions F. Let us assume that there are a positive constant $\delta > 0$, a neighborhood $U(\beta^0,\delta)$, and an integer $n_0 \in \mathbb{N}$ such that the following assumptions hold.

F1: Let $s(d_i;\beta)$ and $r(d_i;\beta)$ be continuous (uniformly over any compact subset of the support of (x,y)) in $\beta \in B$, $r(d_i;\beta)$ be differentiable in β on $U(\beta^0,\delta)$ almost surely, and $s(d_i;\beta)$ be twice differentiable in β on $U(\beta^0,\delta)$ almost surely.

F2: Let $\{s(d_i;\beta)|\beta \in U(\beta^0,\delta)\}$, $\{s'(d_i;\beta)|\beta \in U(\beta^0,\delta)\}$, and $\{r(d_i;\beta)|\beta \in U(\beta^0,\delta)\}$ form VC classes of functions. Moreover, let us assume that the trimmed envelopes $E_s^k(x) = \sup_{\beta\in U(\beta^0,\delta)} \sup_{n\geq n_0} \|s^{(k)}(d_i;\beta) \cdot I\{r(d_i;\beta) \leq r_{([\lambda n])}(\beta)\})\|$ have finite r_β-th moments for $k \in \{0,1\}$.

F3: Expectations $\mathsf{E}\sup_{\beta\in B}|r_{([\lambda n])}(\beta)|$, $\mathsf{E}\sup_{\beta\in B}\sup_{n\geq n_0}\|s(d_i;\beta)\cdot I\{r(d_i;\beta) \leq r_{([\lambda n])}(\beta)\}\|$, $\mathsf{E}\sup_{\beta\in U(\beta^0,\delta)}\sup_{n\geq n_0}\|\partial s(d_i;\beta)/\partial\beta_k \cdot I\{r(d_i;\beta) \leq r_{([\lambda n])}(\beta)\}\|$, and $\mathsf{E}\sup_{n\geq n_0}\|\partial^2 s(d_i;\beta^0)/\partial\beta_k\partial\beta_l \cdot I\{r(d_i;\beta^0) \leq r_{([\lambda n])}(\beta^0)\}\|$ exist and are finite for $k,l = 1,\ldots,p$. Moreover, assume that $J_s(\lambda)$ and $J_s(\lambda) + J_I(\lambda)$ are full-rank matrices and $V_s(\lambda)$ is a non-singular positive definite matrix.

F4: Conditional expectation

$$\mathsf{E}\left\{ \sup_{\beta\in U(\beta^0,\delta)} \left\| s(d_i;\beta^0)\cdot I\{r(d_i;\beta)\in\mathcal{I}(\beta)\}\right\| \,\middle|\, \exists\beta\in U(\beta^0,\delta): r(d_i;\beta)\in\mathcal{I}(\beta)\right\}, \tag{3.4}$$

where $\mathcal{I}(\beta)=\{z:|z-G^{-1}(\lambda)|\leq|z-r_{([\lambda n])}(\beta)|\}$, is uniformly bounded for $n\geq n_0$.

As already discussed, the differentiability of the moment and trimming functions are standard assumptions. On the other hand, Assumption F2, which facilitates deriving the convergence rate of the order statistics in this general framework, limits the class of functions $s(d;\beta)$, $s'(d;\beta)$, and $r(d;\beta)$ to VC classes (see Powell, 1984, and Van der Vaart and Wellner, 1996, for a definition). Although limited, they cover many common functions including polynomial, logarithmic, and exponential functions, their sums, products, maxima and minima, monotonic transformations, and so on. For example, trimming functions having a single-index form $\tau^k(x_i^\top\beta)$ with a monotonic link function τ and $k\in\mathbb{N}$ are covered by Assumption F2.

Further, let us discuss Assumptions F2 and F3 concerning the existence of various expectations. First, the expectations $V_s(\lambda)$, $J_s(\lambda)$, and $J_I(\lambda)$ are trimmed forms of the standard expectations (variances) that appear in the asymptotic variances of extremum estimators (e.g., see Pakes and Pollard, 1989, and Čížek, 2008a). Next, we assume that the trimmed derivatives of the moment function $s(d;\beta)$ have an integrable majorant in some small neighborhood $U(\beta^0,\delta)$. This is not very restrictive given that those expectation have to exist at β^0, that is for $\delta=0$, and the derivatives are continuous. Additionally, we have to assume the existence of integrable majorants of the trimming function and trimmed moment function on the whole parametric space B. The identification assumptions presented below however require that the parametric space B is compact and thus bounded, which makes Assumption F3 much less strict (alternatively, one can assume that $\sup_{\beta\in B}\mathsf{E}\,|r(d_i;\beta)|^{1+\varepsilon}$ is finite for some $\varepsilon>0$). The assumptions of the bounded parametric space and the existence of the integrable majorants of $r(d;\beta)$ and trimmed $s(d;\beta)$ can be relaxed only if the moment conditions are linear in the parameters, at least conditionally (cf. Manski, 1988).

Additionally, the proof of $\sqrt{n}$ consistency requires an unusual regularity assumption Assumption F4, which is one of the (weak) links between the moment function $s(d;\beta)$ and auxiliary trimming function $r(d;\beta)$. This assumption is however not very restrictive and would usually follow from the fact that the moment conditions have finite expectations, see Čížek (2008a) for a discussion.

Finally, we introduce the identification conditions.

Assumptions I.

I1: B is a compact parametric space.

I2: W is a positive definite matrix.

I3: For any $n \in \mathbb{N}$, it holds that $\mathsf{E}\left[s(d_i;\beta)\cdot I\{r(d_i;\beta)\le r_{([\lambda n])}(\beta)\}\right]=0$ if and only if $\beta=\beta^0$, and for any $\delta>0$, that

$$\inf_{\beta\in B\setminus U(\beta^0,\delta)}\left\|\mathsf{E}\left[s(d_i;\beta)\cdot I\{r(d_i;\beta)\le r_{([\lambda n])}(\beta)\}\right]\right\|>0$$

While Assumptions I2 and I3 guarantee that the GMTM objective function (2.8) has a global minimum at β^0, Assumption I3 primarily states that the employed trimming does not invalidate the moment equations under consideration, see (2.4) and (2.6). Note that the identification Assumption I3 can be relaxed by allowing for more solutions of equation (2.6); the GMTM estimate $\hat{\beta}_n$ will then converge to one of the solutions rather than to a unique one.

To close this section, let us note that Assumptions D, F, and I are sufficient to prove the asymptotic normality of GMTM. If only consistency is required, one can omit all assumptions concerning the derivatives of the functions $s(d_i;\beta)$ and $r(d_i;\beta)$ (Assumptions F), Assumption F2 on VC classes, Assumption F4, and also weaken Assumption D1, since centered $s(d_i;\beta)$ can form an $L^{1+\delta}$-mixingale in the most general case (Andrews, 1988).

3.2. **Consistency and asymptotic normality.** Let us now present the main asymptotic results concerning GMTM: its consistency and asymptotic distribution. In all cases, we split the sample trimmed moment conditions to two parts:

$$\begin{aligned} S_n^{\lambda}(\beta) &= \frac{1}{n}\sum_{i=1}^{n} s(d_i;\beta)\cdot I\{r(d_i;\beta)\le r_{([\lambda n])}(\beta)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} s(d_i;\beta)\cdot\Big[I\{r(d_i;\beta)\le r_{([\lambda n])}(\beta)\}-I\{r(d_i;\beta)\le G_\beta^{-1}(\lambda)\}\Big] \qquad (3.5)\\ &+ \frac{1}{n}\sum_{i=1}^{n} s(d_i;\beta)\cdot I\{r(d_i;\beta)\le G_\beta^{-1}(\lambda)\}. \qquad (3.6)\end{aligned}$$

Whereas the first term (3.5) on the right-hand side will be shown to be small because of the convergence of order statistics to quantiles, $r_{([\lambda n])}(\beta)\to G_\beta^{-1}(\lambda)$, the second term (3.6) on the right-hand side will be dealt with by standard asymptotic tools and shown to converge to

$$S^{\lambda}(\beta)=\mathsf{E}\left[s(d_i;\beta)\cdot I\{r(d_i;\beta)\le G_\beta^{-1}(\lambda)\}\right].$$

First, using the uniform law of large numbers for trimmed sums, we prove the consistency of the GMTM estimator $\hat{\beta}_n^{(GMTM,\lambda)}$ minimizing (2.8) on the parametric space B.

Theorem 1. *Let $s(d_i;\beta)$ and $r(d_i;\beta)$ be continuous functions with integrable majorants as specified in Assumptions F1 and F3 and let Assumptions D and I hold. Then the GMTM estimator $\hat{\beta}_n^{(GMTM,\lambda)}$ is weakly consistent, that is, $\hat{\beta}_n^{(GMTM,\lambda)} \to \beta^0$ in probability as $n \to +\infty$.*

Proof: See the Appendix. □

Next, the asymptotic distribution of GMTM will be studied. To derive it, one has to study the behavior of the moment equations $S_n^\lambda(\beta)$ in a neighborhood of β^0 and to prove their asymptotic linearity, that is, the linearity of $S_n^\lambda(\beta^0 - n^{-\frac{1}{2}}t) - S_n^\lambda(\beta^0)$ as a function of t for $n \to \infty$. Once the $\sqrt{n}$ consistency of GMTM is established (Lemma 5 in the Appendix), the asymptotic linearity of GMTM and the decomposition (3.5)–(3.6) allow us to apply the central limit theorem, which results in the asymptotic normality of GMTM.

Theorem 2. *Let Assumptions D, F, and I hold. Then the GMTM estimator $\hat{\beta}_n^{(GMTM,\lambda)}$ is asymptotically normal, that is, $\sqrt{n}\left(\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0\right) \xrightarrow{F} N(0, V(\lambda))$ as $n \to +\infty$, where*

$$V(\lambda) = \Big[J_s(\lambda)^\top W\{J_s(\lambda) + J_I(\lambda)\}\Big]^{-1} \cdot J_s(\lambda)^\top W V_s(\lambda) W J_s(\lambda) \cdot \Big[J_s(\lambda)^\top W\{J_s(\lambda) + J_I(\lambda)\}\Big]^{-1\top}.$$

Proof: See the Appendix. □

Comparing the asymptotic variances of GMTM and GMM, we see that the variance matrix $V(\lambda)$ of GMTM depends on $J_I(\lambda)$ in an asymmetric way. Consequently, it is not possible to find a generally optimal choice of the weighting matrix W as in the case of GMM (Hansen, 1982). Moreover, while the other matrices $V_s(\lambda)$ and $J_s(\lambda)$ needed to evaluate $V(\lambda)$ for an estimate $\hat{\beta}_n$ can be estimated by the corresponding sample means, for example by

$$V_s(\lambda) = \frac{1}{n}\sum_{i=1}^{n} s(d_i;\hat{\beta}_n) s(d_i;\hat{\beta}_n)^\top \cdot I\{r(d_i;\hat{\beta}_n) \le r_{([\lambda n])}(\hat{\beta}_n)\} \tag{3.7}$$

and

$$J_s(\lambda) = \frac{1}{n}\sum_{i=1}^{n} s'(d_i;\hat{\beta}_n) \cdot I\{r(d_i;\hat{\beta}_n) \le r_{([\lambda n])}(\hat{\beta}_n)\} \tag{3.8}$$

for independent observations (see Hansen, 1982, and Newey and West, 1987, for a general discussion of the $V_s(\lambda)$ estimation), the matrix $J_I(\lambda)$ defined in (3.3) is difficult to estimate, which limits the use of the formula for $V(\lambda)$ derived in Theorem 2. To facilitate the variance estimation in typical situations such as the IV estimation discussed in Section 2.2, we impose

additional restrictions on the random variables entering the trimming function, for example, that the dependent variable conditionally on the explanatory variables is continuously distributed, and derive a practically relevant expression for $J_I(\lambda)$. Without loss of generality, we will also assume that the trimming function $r(d_i;\beta)$ is a square of some function $h(d_i;\beta)$ because $r(d_i;\beta)$, measuring a norm of random variables, is typically non-negative and any monotonic transformation of $r(d_i;\beta)$ does not affect ordering of $r(d_1;\beta),\dots,r(d_n;\beta)$.

Lemma 3. *Consider the assumptions of Theorem 2 and let us assume that* $r(d_i,\beta) = h^2(d_i;\beta) = \{h_1(d_i) + h_2(v_i;\beta)\}^2$, *where* v_i *denotes a subset of variables* d_i *such that* $h_1(d_i)|v_i$ *is absolutely continuously distributed with density* f_{v_i} *and independent of parameters* β. *The density function* f_{v_i} *is assumed to be uniformly bounded and differentiable on* $U(\sqrt{G^{-1}(\lambda)},\delta_f)$ *for some* $\delta_f > 0$. *Additionally, we normalize* $h_2(v_i;\beta^0) = 0$ *and assume that* $h_2(v_i;\beta)$ *is twice differentiable in* β *on* $U(\beta^0,\delta)$, $h_2''(v_i;\beta^0) = 0$, *and possesses derivatives with uniformly bounded expectations,* $\sup_{\beta\in U(\beta^0,\delta)} \mathsf{E}\,|h_2^{(k)}(v_i;\beta)|^\eta < K_h \in \mathbb{R}$ *for* $k = 1,2$ *and* $\eta > 1$. *The final assumption is*

$$\mathsf{E}\left\{ s(d_i;\beta^0)\big|\operatorname{sgn} h_1(d_i), |h_1(d_i)| = \sqrt{G^{-1}(\lambda)}, v_i\right\} = \operatorname{sgn} h_1(d_i)\cdot\tilde{s}(v_i) \tag{3.9}$$

and $\|s(d_i;\beta^0) - \operatorname{sgn} h_1(d_i)\cdot\tilde{s}(v_i)\| \le d\{|h_1(d_i)| - \sqrt{G^{-1}(\lambda)}\}\bar{s}(v_i)$, *where* d *is a locally Lipschitz norm on* $\mathbb{R}$ *and* $\bar{s}(v_i)$ *has the finite first moment. Then it holds that*

$$J_I(\lambda) = -\mathop{\mathsf{E}}_{v}\left\{\tilde{s}(v_i)h_2'(v_i;\beta^0)^\top\cdot\left[f_{v_i}\left(-\sqrt{G^{-1}(\lambda)}\right) + f_{v_i}\left(\sqrt{G^{-1}(\lambda)}\right)\right]\right\}. \tag{3.10}$$

Proof: See the Appendix. □

Lemma 3 covers, for example, the linear IV regression and TIV estimators introduced in Section 2.2: TIV-TE(SZ) corresponds to

$$r(d_i;\beta) = (y_i - x_i^\top\beta)^2 = \{[y_i - x_i^\top\beta^0] + [x_i^\top(\beta^0-\beta)]\}^2 = \{\varepsilon_i + x_i^\top(\beta^0-\beta)\}^2 \tag{3.11}$$

and TIV-TETZ corresponds to

$$r(d_i;\beta) = \|(y_i - x_i^\top\beta)z_i\|^2 = \{[y_i - x_i^\top\beta^0] + [x_i^\top(\beta^0-\beta)]\}^2\|z_i\|^2 = \{\varepsilon_i\|z_i\| + x_i^\top(\beta^0-\beta)\|z_i\|\}^2.$$

To discuss the assumptions of Lemma 3, let us consider the TIV-TE estimator, see (3.11): $v_i = (x_i^\top, z_i^\top)^\top$, $s(d_i;\beta^0) = (y_i - x_i^\top\beta^0)z_i = \varepsilon_i z_i$, $h_1(d_i) = y_i - x_i^\top\beta^0 = \varepsilon_i$, $h_2(v_i;\beta) = x_i^\top(\beta^0-\beta)$, $h_2''(v_i;\beta^0) = 0$, and the density function f_{v_i} describes the conditional distribution $\varepsilon_i|v_i$. Assumption (3.9) just means that $s(d_i;\beta^0) = \varepsilon_i z_i$ depends on $h_1(d_i) = \varepsilon_i$ only by

means of $\operatorname{sgn}\varepsilon_i$ once we fix the value of trimming function at β^0: $r(d_i;\beta^0) = h_1^2(d_i) = \varepsilon_i^2 = G^{-1}(\lambda)$. This is however trivially satisfied in this case and $\tilde{s}(v_i) = \sqrt{G^{-1}(\lambda)}z_i$. Consequently, $|s(d_i;\beta^0) - \operatorname{sgn} h_1(d_i)\cdot\tilde{s}(v_i)| \le ||\varepsilon_i| - \sqrt{G^{-1}(\lambda)}|\|z_i\|$ and the existence of assumed moments follows from Assumptions D1 and F3. Under these assumptions and for ε_i being identically distributed with a density function f for simplicity, Lemma 3 implies that

$$J_I(\lambda) = \sqrt{G^{-1}(\lambda)}\left\{f\left[-\sqrt{G^{-1}(\lambda)}\right] + f\left[\sqrt{G^{-1}(\lambda)}\right]\right\}\cdot \mathsf{E}(z_i x_i^\top),$$

where G denotes the distribution of ε_i^2. The matrix $J_I(\lambda)$ can be estimated in this case using any consistent nonparametric density estimator for the density f at points $\pm\sqrt{G^{-1}(\lambda)}$, which are in turn consistently estimated by $\pm\sqrt{r_{([\lambda n])}(\hat{\beta}_n^{(GMTM,\lambda)})}$ (Čížek, 2008a, Lemma A.2).

In a general case, the estimation of the GMTM asymptotic variance $V(\lambda)$ has to be done by bootstrap. Theoretically, bootstrap can be used for GMTM in the same situations as for the original GMM estimator. However to preserve the robust properties of GMTM also in the case of variance estimation, a weighted bootstrap has to be used to prevent bootstrap samples containing an improportionally large share of contaminated observations (Salibian-Barrera and Zamar, 2002) unless a parametric bootstrap can be employed.

3.3. **Test of overidentifying conditions.** Similarly to the seminal paper by Hansen (1982), we also design a test for the validity of overidentifying trimmed conditions if the number of moment restrictions M is greater than the number p of the estimated parameters β. Specifically, we consider the statistics of the form $T_n = nS_n^\lambda(\hat{\beta}_n)^\top\Theta_n^{-1}S_n^\lambda(\hat{\beta}_n)$ and find a matrix Θ_n such that T_n asymptotically follows the χ^2_{M-p} distribution with $M-p$ degrees of freedom. Contrary to the standard GMM case, the matrix Θ_n will require the computation of all elements of the GMTM variance matrix because there is no optimal choice of the weighting matrix W resulting in a simple form of $V(\lambda)$ and Θ_n. On the other hand, let us note that the proposed test will be robust to outliers and atypical values of instruments because Θ_n will depend only $V_s(\lambda)$, $J_s(\lambda)$, $J_I(\lambda)$, and W. Hence, the test statistics T_n is related to data only by means of the trimmed moment conditions S_n^λ and matrices $V_s(\lambda)$, $J_s(\lambda)$, $J_I(\lambda)$, that is, only via quantities containing the trimming indicators $I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\}$ and $I\{r(d_i;\beta) \le G_\beta^{-1}(\lambda)\}$.

Theorem 4. *Under the assumptions of Theorem 2 and $M > p \ge 1$, let*

$$\Pi(\lambda) = \{J_s(\lambda) + J_I(\lambda)\}\left[J_s(\lambda)^\top W\{J_s(\lambda) + J_I(\lambda)\}\right]^{-1} J_s(\lambda)^\top W$$

and let $\hat{\Pi}_n(\lambda)$ *and* $\hat{V}_{sn}(\lambda)$ *be consistent estimates of* $\Pi(\lambda)$ *and* $V_s(\lambda)$*, respectively. Then the test statistics* T_n*,*

$$T_n = nS_n^\lambda(\hat{\beta}_n)^\top \left\{ [I - \hat{\Pi}_n(\lambda)]\hat{V}_{sn}(\lambda)[I - \hat{\Pi}_n(\lambda)]^\top \right\}^- S_n^\lambda(\hat{\beta}_n),$$

converges in distribution to the χ^2 *distribution with* $M - p$ *degrees of freedom,* $T_n \sim \chi^2_{M-p}$*, where the notation* A^- *means the Moore-Penrose generalized inverse of matrix* A*.*

Proof: See the Appendix. □

Theorem 4 is straightforward to apply if all matrices $J_s(\lambda)$, $J_I(\lambda)$, and $V_s(\lambda)$ can be directly estimated, for example, using Lemma 3. Otherwise, the variance matrix $V(\lambda)$ of GMTM is estimated by some resampling method and $J_s(\lambda) + J_I(\lambda)$ has to be "reconstructed" from the knowledge of estimates $\hat{V}_n(\lambda)$, $\hat{J}_s(\lambda)$, and $\hat{V}_s(\lambda)$. In particular, if $A^{1/2}$ denotes the square root of a positive semidefinite matrix A, one can employ the variance formula derived in Theorem 2 and show that

$$J_s(\lambda)^\top W\{J_s(\lambda) + J_I(\lambda)\} = [J_s(\lambda)^\top W V_s(\lambda) W J_s(\lambda)]^{1/2} V^{-1/2}(\lambda). \tag{3.12}$$

(The square roots of matrices can be obtained by the Choleski decomposition, for instance.) If $J_s(\lambda) + J_I(\lambda)$ itself is needed, one can solve the joint system of linear equations (3.12) obtained for at least $\lceil M/p \rceil$ different values of W (such that a sufficient number of equations for $J_s(\lambda) + J_I(\lambda)$ is generated).

4. Monte Carlo Simulations

In this section, we study and compare performance of some existing GMM and proposed GMTM estimators by means of simulations. We will first discuss the models and estimators used in the comparison (Section 4.1). Later, we compare all methods using data with and without aberrant observations (Sections 4.2 and 4.3).

4.1. **Simulated models and estimation methods.** Various existing and proposed estimators will be compared in the context of the linear IV model. Let us first discuss the estimation methods compared in simulations. We compare the standard estimators including LS and GMM with the methods proposed in Section 2.2: TIV-TE, TIV-TESZ, and TIV-TETZ both with the fixed trimming $\lambda = 0.55$ and the data-dependent amount of trimming $\hat{\lambda}_n$ using $N(0, \sigma^2)$ as the reference distribution, see Section 2.3; the choice of trimming is

indicated in brackets, for example, TIV-TE(0.55). Because we study here the robust properties of GMM estimators, we also include two median IV estimators: the Med-IV estimator by Honore and Hu (2004) and the instrumental variable quantile estimator (IV-Quant) introduced by Chernozhukov and Hansen (2008) at $\tau = 0.5$ quantile. Finally, the robust properties of Med-IV could benefit from the standardization of instruments introduced for TIV-TESZ in Section 2.2, and therefore, we also propose and use Med-IV using instruments normalized to have a unit Euclidean norm (IV-Quant cannot benefit from such a transformation); this method is referred to as Med-IV-SZ. Please note that all presented estimates are one-step GMM estimates using the identity weighting matrix $W = I$ because: (i) the two-stage least squares weighting matrix converges to the identity matrix in our setup; (ii) this choice improves robustness of all methods (even standard ones) in simulations as weights cannot be influenced by atypical values of instruments; and (iii) the two-step GMM estimates with estimated optimal weighting matrix $\hat{W}$ do not improve estimation results except for two models containing heteroscedasticity, where this improvement is rather limited (at most 7% decrease in the median squared error) and does not influence the qualitative results of the study.

All methods are compared using the linear regression model with an endogenous variable. As the results do not qualitatively depend on the number of included variables, we use here the following simple model:

$$y_i = 1 + x_{1i} - x_{2i} + \varepsilon_i, \tag{4.1}$$

$$x_{2i} = (1 + z_{1i} + z_{2i})/\sqrt{2} + \nu_i, \tag{4.2}$$

where y_i is the dependent variable and x_{2i} represents the endogenous variable because error terms $\varepsilon_i \sim F$ and $\nu_i \sim N(0,1)$ are correlated, $\mathsf{cor}(\varepsilon_i, \nu_i) = \rho = 0.5$ (the results are insensitive to the value of ρ). The distribution function F of ε_i can be normal $N(0, \sigma_\varepsilon^2)$ with a constant variance or variance depending of other variables (heteroscedasticity), Student $Std(d)$ with d degrees of freedom, or double exponential $DExp(\lambda)$ with a rate λ. The remaining variables $x_{1i} \sim N(0,1)$ and $z_{1i}, z_{i2} \sim N(0,1)$ are exogenous, independent of each other, and represent the exogenous and instrumental variables, respectively. Furthermore, data are contaminated by erroneous observations in some cases. Then α denotes the fraction of sample being contaminated. For the corresponding $[\alpha n]$ observations, an additional error term ϵ_i following the uniform distribution on $(-30, 30)$, $\epsilon_i \sim U(-30, 30)$, is added to y_i: $y_i = 1 + x_{1i} - x_{2i} + \varepsilon_i + \epsilon_i$. (Note that this definition does not invalidate the moment conditions used by standard IV

TABLE 1. The MSE of estimates for the linear IV regression model with normally distributed errors, $\varepsilon_i \sim N(0,1)$, and sample sizes $n = 50, 100, 200,$ and 400.

MSE	Sample size			
Estimator	$n = 50$	$n = 100$	$n = 200$	$n = 400$
GMM	0.056	0.025	0.013	0.007
IV-Quant	0.120	0.053	0.028	0.014
Med-IV	0.094	0.040	0.023	0.012
Med-IV-SZ	0.106	0.044	0.025	0.011
TIV-TE(0.55)	0.343	0.174	0.098	0.055
TIV-TE($\hat{\lambda}_n$)	0.068	0.028	0.016	0.007
TIV-TESZ(0.55)	0.325	0.183	0.105	0.054
TIV-TESZ($\hat{\lambda}_n$)	0.073	0.030	0.016	0.008
TIV-TETZ(0.55)	0.361	0.177	0.118	0.071
TIV-TETZ($\hat{\lambda}_n$)	0.073	0.029	0.015	0.008

estimators yet.) In some setups, the values of explanatory or instrumental variables x_{1i}, x_{2i}, z_{1i}, or z_{2i} are additionally shifted by $\Delta = 10$ for contaminated observations so that the model (4.1)–(4.2) does not hold anymore for these observations. This is referred to as contamination with leverage points in x_{1i}, x_{2i}, z_{1i}, or z_{2i}, respectively.

The results presented in the following sections were obtained for samples sizes $n = 50, \ldots, 400$ and are based on 1000 simulated samples. To summarize the estimation results, we use the median of squared errors (MSE).

4.2. **Clean data.** The first discussed experiment concerns the model (4.1)–(4.2) using normally distributed errors $\varepsilon_i \sim N(0,1)$ and no contamination. Results for sample sizes $n = 50, 100, 200,$ and 400 are summarized in Table 1. First of all, all estimators are consistent in this setting. Comparing various quantile IV estimators, they all perform similarly with Med-IV being the best one and they exhibit approximately two times higher MSEs than the standard GMM estimator. Looking at the trimmed estimators TIV with the fixed amount of trimming $\lambda = 0.55$, they perform poorly in terms of MSEs since they neglect almost half of all observations. On the other hand, all trimmed estimators with the adaptive choice of trimming $\hat{\lambda}_n$ outperform the quantile IV estimators and can match the standard GMM at the large sample size $n = 400$. Finally, one can observe that the qualitative results, that is, the ordering of methods by their MSEs, do not significantly change for different samples. For the sake of brevity, we therefore restrict to $n = 200$ in what follows.

Next, let us consider the IV model with other error distributions such as normal, Student, and double exponential, and additionally, with heteroscedastic normally distributed errors.

TABLE 2. The MSE of estimates for the linear IV regression model with errors following the Gaussian, Student, and double exponential distributions and sample size $n = 200$. The random variable u follows the uniform distribution, $u \sim U(0.25, 4)$, and $w_z = z_1 + z_2$.

MSE	Distribution of ε_i				
Estimator	$N(0,1)$	$N(0,e^u)$	$N(0,e^{w_z})$	$Std(5)$	$DExp(1)$
GMM	0.013	0.166	0.034	0.021	0.027
IV-Quant	0.028	0.178	0.019	0.030	0.022
Med-IV	0.023	0.130	0.015	0.022	0.017
Med-IV-SZ	0.025	0.151	0.016	0.025	0.018
TIV-TE(0.55)	0.098	0.313	0.053	0.068	0.041
TIV-TE($\hat{\lambda}_n$)	0.016	0.165	0.025	0.019	0.025
TIV-TESZ(0.55)	0.105	0.356	0.060	0.076	0.045
TIV-TESZ($\hat{\lambda}_n$)	0.016	0.168	0.024	0.020	0.027
TIV-TETZ(0.55)	0.118	0.416	0.064	0.084	0.054
TIV-TETZ($\hat{\lambda}_n$)	0.015	0.166	0.018	0.020	0.025

The estimation results for $n = 200$ are presented in Table 2. The first three columns compare the performance of all estimators for normally distributed homoscedastic and heteroscedastic errors. The presence of heteroscedasticity leads to a worse results for GMM: the Med-IV(-SZ) method now exhibits the smallest MSE. Although the TIV estimates are usually worse than Med-IV in this scenario, the TIV estimators using adaptively chosen trimming match (the second column) or outperform (the third column) the standard GMM estimator. Furthermore, comparing all methods under the Student distribution (the fourth column), all TIV variants with adaptively chosen trimming are slightly better than the GMM and quantile IV estimators. The role reverses for the errors following the double exponential distribution (the fifth column), where the quantile IV estimators outperform GMM and TIV estimators in terms of MSE. Additionally, notice that both the absolute and relative differences between MSEs of the TIV estimates with the fixed and adaptive trimming are quite smaller in the last two cases than in the case of normal errors.

4.3. **Contaminated data.** We will now consider contaminated data with contamination levels $\alpha = 0.10, 0.25$, and 0.40, where there are either no leverage points (Table 3) or leverage points in the direction of the endogenous explanatory variable x_2 (Table 3), exogenous explanatory variable x_1 (Table 4), or instrumental variables z_1 and z_2 (Table 4).

Let us first discuss the simulation results summarized in Table 3. If there are no leverage points, the GMM moment conditions are correctly specified for all observations from the model (4.1)–(4.2) and only some observations exhibit a very large variance. All estimates are

TABLE 3. The MSE of estimates for contaminated data originating from the linear IV regression model with Gaussian errors and sample size $n = 200$. Contamination levels are $\alpha = 0.10, 0.25$, and 0.40 with no leverage points or leverage in the endogenous variable x_2.

MSE	Contamination, no leverage			Contamination, leverage in x_2		
Estimator	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.40$	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.40$
GMM	0.356	0.969	1.617	0.991	8.347	27.68
IV-Quant	0.034	0.047	0.076	0.037	0.071	0.170
Med-IV	0.024	0.040	0.060	0.032	0.060	0.247
Med-IV-SZ	0.027	0.041	0.066	0.031	0.067	0.234
TIV-TE(0.55)	0.072	0.059	0.042	0.087	0.058	0.044
TIV-TE($\hat{\lambda}_n$)	0.016	0.021	0.047	0.018	0.023	0.058
TIV-TESZ(0.55)	0.080	0.070	0.042	0.093	0.060	0.042
TIV-TESZ($\hat{\lambda}_n$)	0.018	0.023	0.051	0.018	0.025	0.058
TIV-TETZ(0.55)	0.095	0.072	0.044	0.103	0.075	0.046
TIV-TETZ($\hat{\lambda}_n$)	0.017	0.021	0.046	0.018	0.025	0.051

thus consistent, but high variability of observations with errors $\varepsilon_i + U(-30, 30)$ leads to large MSEs of GMM estimates. The MSEs of GMM obviously increase with α, but unreported results confirm that they decrease as the sample size n grows. All other estimates exhibit small MSEs, where TIV with fixed trimming is worst unless α is very high, TIV with the data-dependent trimming is best unless $\alpha = 0.40$, and the quantile IV estimates have about 1.5 larger MSEs than the best TIV estimates.

If we simulate data from model (4.1)–(4.2) and the values of the endogenous variable are shifted for contaminated data points, the model no longer holds for the contaminated data. The GMM estimates then exhibit a large bias and MSE, which increase with the level of contamination α, but do not decrease with a sample size (as unreported results show). The quantile IV estimators are influenced by contamination only to a small extent since the leverage does not occurs in any variable used as an instrument. An exception is the case with the $\alpha = 0.40$ level of contamination as the levels α above 0.30 are generally beyond the breakdown capabilities of the L_1 estimators (cf. He et al. 1990). The smallest MSEs can be attributed to TIV estimators, all of which outperform quantile IV estimators for $\alpha \geq 0.25$ (for $\alpha = 0.10$, only TIV with the data-dependent triming are better than the quantile IV estimators). Similarly to the previous simulation, the TIV estimators with fixed trimming $\lambda = 0.55$ are worse than those with the adaptive trimming $\hat{\lambda}_n$ unless $\alpha = 0.40$. The reason is that there is practically no benefit of the adaptive choice of trimming for $\alpha = 0.40$ because the initial estimator with fixed trimming excludes $100(1 - \lambda) = 45$ percent of observations from the GMTM objective function, which is almost the optimal amount of trimming.

TABLE 4. The MSE of estimates for contaminated data originating from the linear IV regression model with Gaussian errors and sample size $n = 200$. Contamination levels are $\alpha = 0.10, 0.25$, and 0.40 with leverage in the exogenous variable x_1 and in instrumental variables z_1 and z_2.

MSE	Contamination, leverage in x_1			Contamination, leverage in z_1, z_2		
Estimator	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.40$	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.40$
GMM	1.014	1.471	1.577	3.483	4.178	3.046
IV-Quant	0.181	0.510	0.794	7.301	5.717	4.645
Med-IV	0.164	0.470	0.738	2.453	52.12	56.31
Med-IV-SZ	0.036	0.090	0.216	0.030	0.047	0.092
TIV-TE(0.55)	0.094	0.116	0.134	0.100	0.156	0.605
TIV-TE($\hat{\lambda}_n$)	0.029	0.122	0.455	0.053	0.958	4.715
TIV-TESZ(0.55)	0.093	0.072	0.047	0.079	0.070	0.058
TIV-TESZ($\hat{\lambda}_n$)	0.018	0.026	0.039	0.016	0.028	0.073
TIV-TETZ(0.55)	0.095	0.072	0.033	0.091	0.079	0.035
TIV-TETZ($\hat{\lambda}_n$)	0.024	0.081	0.287	0.031	0.252	0.526

Next, we will study contaminated data with leverage in the space of exogenous variable x_1, see Table 4. In this case, the variables with the values shifted by Δ enter both the regression residuals and the set of instruments. The MSEs of GMM are large even for 10% contamination and increase with an increasing level α of contamination. Additionally, the IV-Quant and Med-IV are substantially affected by any level of contamination as well (though less than GMM) given that many robust estimators have MSEs below 0.1 in all experiments with contaminated normal data. The only exception to this is the proposed Med-IV-SZ estimator, which normalizes all instruments and is thus insensitive to leverage points at least for $\alpha \leq 0.25$. Considering TIV-TE, which is not protected anyhow against atypical values of instruments similarly to Med-IV, we see its MSE increase significantly with α, especially for the adaptive choice of trimming and $\alpha = 0.40$, where it is no longer reliable. On the other hand, the TIV variants that protect against atypical values of instruments, TIV-TESZ and TIV-TETZ, exhibit the same behavior as in previous experiments: the most stable and smallest MSEs of all methods irrespective of the level of contamination. The only exception to the rule is the TIV-TETZ method with the adaptive trimming for $\alpha = 0.40$ probably because the data-dependent choice of trimming proposed in Section 2.3 is designed only with the residual-trimming in mind, not with trimming by the moment values.

Finally, contaminated data with leverage in the space of instrumental variables z_1 and z_2 are considered (which technically satisfy the moment conditions for model (4.1)–(4.2)). The results summarized in Table 4 are structurally similar to those using leverage in x_1, but are more pronounced since more instrumental variables are affected while the residuals

are not affected in the case of consistent (robust) estimators. Hence, GMM, IV-Quant, and Med-IV show very large MSEs and only the proposed Med-IV-SZ is not influenced by contamination. Similarly, TIV-TE protecting only against large residuals is substantially influenced by contamination, whereas TIV-TESZ with both trimmings and TIV-TETZ with fixed trimming provide stable estimates with small MSEs. TIV-TETZ with the adaptive trimming is biased by contamination because of the adaptive choice based on residuals only.

Summarizing all results for the IV regression, there is only one method which always has the smallest or close to the smallest MSE and which is not influenced by contaminated data in any considered setup: TIV-TESZ with the adaptive choice of trimming. It matches or outperforms GMM for non-contaminated data, it is not significantly influenced by any combination of large residuals and large values of regression variables, and it always outperforms Med-IV-SZ in contaminated samples. Another candidate and successful method is TIV-TETZ, which represents a more generally applicable method than TIV-TESZ or Med-IV-SZ (see Section 2.2) because it does not rely on the normalization of instruments, which can be effectively applied only in linear models. TIV-TETZ would however require a different procedure to achieve good results both in clean and contaminated data. This could be a different adaptive-trimming procedure or a one-step M-estimator combining TIV-TETZ and the robust GMM of Ronchetti and Trojani (2001).

5. Conclusion

Complementing locally robust GMM methods by Ronchetti and Trojani (2001) and others, we proposed a globally robust generalized method of trimmed moments, which extends the applicability of high breakdown-point methods to a wide range of econometric models, including time series, panel data, and limited dependent variable models. We derived the asymptotic distribution of GMTM as well as an analogy of the Sargan test of overidentifying restrictions. Moreover, we also show in simulations that the data-dependent choice of trimming can make GMTM performing as well as the standard GMM estimator in a variety of situations, while being preferable for its robust properties. An alternative approach to efficient robust GMM estimation could combine GMTM as a starting estimator with one iteration step of the robust M-estimation-based GMM by Ronchetti and Trojani (2001).

On the other hand, we discussed only the most basic form of trimmed estimation, where observations are either included in or excluded from the GMTM objective function. Nevertheless, various weighted trimmed estimators as in Víšek (2006) and Čížek (2007a) are

straightforward to apply. Furthermore, we argued that the breakdown properties will be analogous to existing results concerning existing trimmed estimators such as LTS, for instance, which are typically studied and applied in the context of location or linear regression models. Although this applies in simple linear regression models, possible applications of GMTM can involve rather complex (non)linear models under dependency. Hence, the robust properties of GMTM in such models have to be further studied.

Finally, we did not address and left for further research recent developments of GMM and related methods that address, for example, improving finite-sample performance (e.g., the generalized empirical likelihood methods, see Newey and Smith, 2004) or inference in the presence of weak identification (e.g., Stock and Wright, 2000; Chao and Swanson, 2005).

Appendix

Here we present the proofs of lemmas and theorems presented in the paper. Additional notation is used: the moment and trimming functions are written as $s_i(\beta) = s(d_i;\beta)$ and $r_i(\beta) = r(d_i;\beta)$, respectively; the sample trimmed moment conditions are denoted $S_n^\lambda(\beta) = n^{-1}\sum_{i=1}^n s(d_i;\beta)\cdot I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\}$ and their asymptotic counterpart is $S^\lambda(\beta) = \mathsf{E}[s(d_i;\beta)\cdot I\{r(d_i;\beta) \le G_\beta^{-1}(\lambda)\}]$. Similarly, the limit of the GMTM objective function $Q_n^{W,\lambda}(\beta) = S_n^\lambda(\beta)^\top W S_n^\lambda(\beta)$ is denoted $Q^{W,\lambda}(\beta) = S^\lambda(\beta)^\top W S^\lambda(\beta)$. Finally, since we extensively study and use the indicators $I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\}$, $I\{r(d_i;\beta) \le G_\beta^{-1}(\lambda)\}$, and their differences, we define $\iota_{in}^\lambda(\beta) = I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\}$, $\nu_i^\lambda(\beta) = I\{r(d_i;\beta) \le G_\beta^{-1}(\lambda)\}$, and

$$\delta_{in}^\lambda(\beta) = \iota_{in}^\lambda(\beta) - \iota_{in}^\lambda(\beta^0) = I\{r(d_i;\beta) \le r_{([\lambda n])}(\beta)\} - I\{r(d_i;\beta^0) \le r_{([\lambda n])}(\beta^0)\}.$$

Then $S_n^\lambda(\beta) = n^{-1}\sum_{i=1}^n s(d_i;\beta)\cdot\iota_{in}^\lambda(\beta)$ and $S^\lambda(\beta) = \mathsf{E}\{s(d_i;\beta)\cdot\nu_i^\lambda(\beta)\}$. Note that S_n^λ and S^λ correspond to the symbols $S_{nn}' = S_n'/n$ and S' in the notation of Čížek (2008a) whose results for trimmed sums are used in the proofs.

We first present the proof of the consistency of GMTM.

Proof of Theorem 1: This is a standard proof of consistency based on the uniform law of large numbers and the convergence of the order statistics $r_{([\lambda n])}(\beta)$ to the corresponding

quantile $G_\beta^{-1}(\lambda)$. By definition, $P\Big(Q_n^{W,\lambda}\left(\hat{\beta}_n^{(GMTM,\lambda)}\right) < Q_n^{W,\lambda}\left(\beta^0\right)\Big) = 1$. For any $\delta > 0$,

$$\begin{aligned}
1 &= P\Big(Q_n^{W,\lambda}\left(\hat{\beta}_n^{(GMTM,\lambda)}\right) < Q_n^{W,\lambda}\left(\beta^0\right)\Big) \\
&= P\Big(Q_n^{W,\lambda}\left(\hat{\beta}_n^{(GMTM,\lambda)}\right) < Q_n^{W,\lambda}\left(\beta^0\right) \quad \text{and} \quad \hat{\beta}_n^{(GMTM,\lambda)} \in U(\beta^0,\delta)\Big) \\
&+ P\Big(Q_n^{W,\lambda}\left(\hat{\beta}_n^{(GMTM,\lambda)}\right) < Q_n^{W,\lambda}\left(\beta^0\right) \quad \text{and} \quad \hat{\beta}_n^{(GMTM,\lambda)} \in B\backslash U(\beta^0,\delta)\Big) \\
&\leq P\Big(\hat{\beta}_n^{(GMTM,\lambda)} \in U(\beta^0,\delta)\Big) + P\Big(\inf_{\beta\in B\backslash U(\beta^0,\delta)} Q_n^{W,\lambda}\left(\beta\right) < Q_n^{W,\lambda}\left(\beta^0\right)\Big).
\end{aligned}$$

Hence, $P\Big(\inf_{\beta\in B\backslash U(\beta^0,\delta)} Q_n^{W,\lambda}\left(\beta\right) < Q_n^{W,\lambda}\left(\beta^0\right)\Big) \to 0$ as $n \to +\infty$ implies $P(\hat{\beta}_n^{(GMTM,\lambda)} \in U(\beta^0,\delta)) \to 1$ as $n \to +\infty$, that is, the consistency of $\hat{\beta}_n^{(GMTM,\lambda)}$, because δ is an arbitrary positive number. To verify $P\Big(\inf_{\beta\in B\backslash U(\beta^0,\delta)} Q_n^{W,\lambda}\left(\beta\right) < Q_n^{W,\lambda}\left(\beta^0\right)\Big) \to 0$ note that

$$\begin{aligned}
& P\left(\inf_{\beta\in B\backslash U(\beta^0,\delta)}\left[Q_n^{W,\lambda}\left(\beta\right) - Q^{W,\lambda}(\beta) + Q^{W,\lambda}(\beta)\right] < Q_n^{W,\lambda}\left(\beta^0\right)\right) \\
\leq\ & P\left(\inf_{\beta\in B\backslash U(\beta^0,\delta)}\left[Q_n^{W,\lambda}\left(\beta\right) - Q^{W,\lambda}(\beta)\right] < Q_n^{W,\lambda}(\beta^0) - \inf_{\beta\in B\backslash U(\beta^0,\delta)} Q^{W,\lambda}(\beta)\right) \\
\leq\ & P\left(\sup_{\beta\in B}\left|Q_n^{W,\lambda}\left(\beta\right) - Q^{W,\lambda}(\beta)\right| > \inf_{\beta\in B\backslash U(\beta^0,\delta)} Q^{W,\lambda}(\beta) - Q_n^{W,\lambda}(\beta^0)\right) \\
\leq\ & P\left(2\sup_{\beta\in B}\left|Q_n^{W,\lambda}\left(\beta\right) - Q^{W,\lambda}(\beta)\right| > \inf_{\beta\in B\backslash U(\beta^0,\delta)} Q^{W,\lambda}(\beta) - Q^{W,\lambda}(\beta^0)\right).
\end{aligned}$$

Since Assumptions I2 and I3 imply for any $\delta > 0$ that there is $\alpha > 0$ such that $\inf_{\beta\in B\backslash U(\beta^0,\delta)} Q^{W,\lambda}(\beta) - Q^{W,\lambda}(\beta^0) > \alpha$, it is enough to show that $P\Big(\sup_{\beta\in B}\left|Q_n^{W,\lambda}\left(\beta\right) - Q^{W,\lambda}(\beta)\right| > \alpha\Big) \to 0$ as $n \to +\infty$ for all $\alpha > 0$.

To prove this, let us first note that it holds for the trimmed moment conditions

$$\begin{aligned}
(.1) \qquad S_n^\lambda\left(\beta\right) - S^\lambda(\beta) &= \frac{1}{n}\sum_{i=1}^n s_i(\beta)\left\{\iota_{in}^\lambda(\beta) - \nu_i^\lambda(\beta)\right\} \\
(.2) \qquad &+ \frac{1}{n}\sum_{i=1}^n\left\{s_i(\beta)\nu_i^\lambda(\beta) - \mathsf{E}\left[s_i(\beta)\nu_i^\lambda(\beta)\right]\right\}.
\end{aligned}$$

Using Assumptions D and F, we can apply Čížek (2008a, Corollary A.6) to the term (.1) and Čížek (2008a, Lemma A.1) to the term (.2) to show that both terms are asymptotically negligible in probability, that is, for $n \to +\infty$ and any $\alpha > 0$

$$(.3) \qquad P\left(\sup_{\beta\in B}\left|S_n^\lambda\left(\beta\right) - S^\lambda(\beta)\right| > \alpha\right) \to 0.$$

Next, the objective function of GMTM is a quadratic form $t^\top W t$ in moment conditions $t = S_n^\lambda(\beta)$ (cf. $Q_n^{W,\lambda}(\beta) = S_n^\lambda(\beta)^\top W S_n^\lambda(\beta)$). Moreover, the function $t^\top W t$ is locally Lipschitz: for $\|t_1\| \leq K$, $\|t_2\| \leq K$, and $K > 0$, it holds that ($t_1, t_2 \in \mathbb{R}^M$ and W is symmetric)

$$|t_1^\top W t_1 - t_2^\top W t_2| = |(t_1 - t_2)^\top W (t_1 + t_2)| \leq 2KW|t_1 - t_2|.$$

Because $S_n^\lambda(\beta) \to S^\lambda(\beta)$ in probability uniformly on B by (.3), $S^\lambda(\beta)$ is continuous on compact B by Assumptions F1 and I1, and thus $S^\lambda(\beta)$ is bounded, we can find for any $\varepsilon > 0$ some $n_0 \in \mathbb{N}$ and $K > \sup_{\beta \in B} S^\lambda(\beta)$ such that $P(|S_n^\lambda(\beta)| > K) < \varepsilon/2$ and $P(2KW \sup_{\beta \in B} \left|S_n^\lambda(\beta) - S^\lambda(\beta)\right| > \alpha) < \varepsilon/2$ for all $n \geq n_0$. Thus,

$$\begin{aligned} P\left(\sup_{\beta \in B} \left|Q_n^{W,\lambda}(\beta) - Q^{W,\lambda}(\beta)\right| > \alpha\right) &= P\left(\sup_{\beta \in B} \left|S_n^\lambda(\beta)^\top W S_n^\lambda(\beta) - S^\lambda(\beta)^\top W S^\lambda(\beta)\right| > \alpha\right) \\ &\leq P\left(2KW \sup_{\beta \in B} \left|S_n^\lambda(\beta) - S^\lambda(\beta)\right| > \alpha\right) + P(|S_n^\lambda(\beta)| > K) \\ &\leq \varepsilon/2 + \varepsilon/2 = \varepsilon, \end{aligned}$$

which concludes the proof as $\varepsilon > 0$ can be arbitrarily small. □

After proving the consistency of GMTM, we aim to derive its asymptotic distribution using the asymptotic linearity of moment conditions. To do so, we have to show first that the GMTM estimates converge at rate $n^{-\frac{1}{2}}$.

Lemma 5. *Let Assumptions D, F, and I hold. Then $\hat{\beta}_n^{(GMTM,\lambda)}$ is $\sqrt{n}$-consistent, that is, $\sqrt{n}\left(\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0\right) = \mathcal{O}_p(1)$ as $n \to +\infty$.*

Proof: We already know that $\hat{\beta}_n^{(GMTM,\lambda)}$ is consistent and $P\Big(\|\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0\| > \rho\Big) \to 0$ as $n \to +\infty$ for any $\rho > 0$ (Theorem 1). Moreover, we showed in the proof of Theorem 1 that $S_n^\lambda(\beta) \to S^\lambda(\beta)$ uniformly in probability as $n \to \infty$. The same argument can be now used also for the derivatives of the moment conditions. Since $s_i(\beta)$ is twice differentiable in $\beta \in U(\beta^0, \delta)$ and

$$\frac{\partial^k S_n^\lambda(\beta)}{\partial \beta^k} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^k s_i(\beta)}{\partial \beta^k} \iota_{in}^\lambda(\beta) \tag{.4}$$

almost surely for $k \in \{0,1,2\}$, see Čížek (2008a, Lemma 2.1), we can apply the decomposition (.1)–(.2) used for $S_n^\lambda(\beta)$ in the proof of Theorem 1 to derivatives $\partial^k S_n^\lambda(\beta)/\partial\beta^k$:

$$(.5) \quad \frac{\partial^k S_n^\lambda(\beta)}{\partial\beta^k} - \mathsf{E}\left\{\frac{\partial^k s_i(\beta)}{\partial\beta^k}\nu_i^\lambda(\beta)\right\} = \frac{1}{n}\sum_{i=1}^n \frac{\partial^k s_i(\beta)}{\partial\beta^k}\left\{\iota_{in}^\lambda(\beta) - \nu_i^\lambda(\beta)\right\}$$

$$(.6) \quad + \frac{1}{n}\sum_{i=1}^n \left\{\frac{\partial^k s_i(\beta)}{\partial\beta^k}\nu_i^\lambda(\beta) - \mathsf{E}\left[\frac{\partial^k s_i(\beta)}{\partial\beta^k}\nu_i^\lambda(\beta)\right]\right\}.$$

Subsequently, we again apply Čížek (2008a, Corollary A.6) and Čížek (2008a, Lemma A.1) to these terms to show that, as $n \to +\infty$, it holds uniformly on $U(\beta^0,\delta)$ in probability

$$(.7) \quad \frac{\partial^k S_n^\lambda(\beta)}{\partial\beta^k} \to \mathsf{E}\left\{\frac{\partial^k s_i(\beta)}{\partial\beta^k}\nu_i^\lambda(\beta)\right\} = S^{\lambda,k}(\beta).$$

Because the GMTM objective function $Q_n^{W,\lambda}(\beta)$ is a quadratic form in $S_n^\lambda(\beta)$ and $U(\beta^0,\delta)$ is bounded, the uniform convergence of $Q_n^{W,\lambda}(\beta)$ to $Q^{W,\lambda}(\beta)$ follows. The same applies to the derivatives of $Q_n^{W,\lambda}(\beta)$, which are quadratic forms in $\partial^k S_n^\lambda(\beta)/\partial\beta^k$, $k \in \{0,1,2\}$, and where we use notation $Q^{W,\lambda,k}(\beta) = \operatorname{p\,lim}_{n\to\infty} \partial^k Q_n^{W,\lambda}(\beta)/\partial\beta^k$ for $k \in \{1,2\}$. Specifically, denoting $H_{jn}^\lambda(\beta) = \partial^2 S_{jn}^\lambda(\beta)/\partial\beta\partial\beta^\top$ and $\Pi_{jn}^\lambda(\beta) = W S_{jn}^\lambda(\beta)$, $j = 1,\ldots,M$,

$$\frac{\partial^2 Q_n^{W,\lambda}(\beta^0)}{\partial\beta\partial\beta^\top} = 2\frac{\partial S_n^\lambda(\beta^0)}{\partial\beta^\top}^\top W \frac{\partial S_n^\lambda(\beta^0)}{\partial\beta^\top} + \sum_{j=1}^M \Pi_{jn}^\lambda(\beta^0) H_{jn}^\lambda(\beta^0)$$

see Abadir and Magnus (2005, p. 382), and we can conclude that

$$\frac{\partial^2 Q_n^{W,\lambda}(\beta^0)}{\partial\beta\partial\beta^\top} \to 2\,\mathsf{E}\left\{s_i'(\beta^0)\nu_i^\lambda(\beta^0)\right\}^\top W \left\{s_i'(\beta^0)\nu_i^\lambda(\beta^0)\right\} = J_s(\lambda)^\top W J_s(\lambda)$$

because $H_{jn}^\lambda(\beta^0) \to \mathsf{E}[\partial^2 S_{jn}^\lambda(\beta^0)/\partial\beta\partial\beta^\top]$ is bounded by Assumption F3 and $\Pi_{jn}^\lambda(\beta^0) \to W S_j^\lambda(\beta^0) = 0$ by Assumption I3 ($S_j^\lambda(\beta)$ denotes the jth component of vector $S^\lambda(\beta)$).

Consequently, $\partial^2 Q_n^{W,\lambda}(\beta^0)/\partial\beta\partial\beta^\top$ converges to a positive definite matrix $Q^{W,\lambda,2}(\beta^0) = J_s(\lambda)^\top W J_s(\lambda) > 0$ by Assumption F3 and I2, and therefore, there exists a constant $\rho, \delta > \rho > 0$, such that $\left\|Q^{W,\lambda,1}(\beta)\right\| \geq C\left\|\beta - \beta^0\right\|$ for all $\beta \in U(\beta^0,\rho)$ and some $C > 0$. Due to the consistency of $\hat\beta_n^{(GMTM,\lambda)}$, this implies that for any $\varepsilon > 0$ there is some $n_0 \in \mathbb{N}$ such that $\hat\beta_n^{(GMTM,\lambda)} \in U(\beta^0,\rho)$ and subsequently $\|Q^{W,\lambda,1}(\hat\beta_n^{(GMTM,\lambda)})\| \geq C\|\hat\beta_n^{(GMTM,\lambda)} - \beta^0\|$ for all $n > n_0$ with probability at least $1-\varepsilon$. Therefore, it is sufficient to show that $\sqrt{n}\|Q^{W,\lambda,1}(\hat\beta_n^{(GMTM,\lambda)})\| = \mathcal{O}_p(1)$ to prove the lemma.

Since $Q^{W,\lambda,1}(\beta) = 2S^{\lambda,1}(\beta)^\top W S^\lambda(\beta)$, we can analyze

$$\sqrt{n}S^{\lambda,1}(\hat{\beta}_n^{(GMTM,\lambda)})^\top W S^\lambda(\hat{\beta}_n^{(GMTM,\lambda)}) \tag{.8}$$

$$= \sqrt{n}S^{\lambda,1}(\hat{\beta}_n^{(GMTM,\lambda)})^\top W \left\{ S^\lambda(\hat{\beta}_n^{(GMTM,\lambda)}) - S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)}) \right\} \tag{.9}$$

$$+ \sqrt{n} \left\{ S^{\lambda,1}(\hat{\beta}_n^{(GMTM,\lambda)}) - \frac{\partial S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})}{\partial \beta^\top} \right\}^\top W S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)}) \tag{.10}$$

(recall that the first-order conditions imply $\partial S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})/\partial\beta^\top W S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)}) = 0$). We now verify that both terms on the right hand side of (.8)–(.10) are bounded in probability. Since the verification follows exactly the same steps for both terms, we will do it just for (.10). First, $\hat{\beta}_n^{(GMTM,\lambda)} \in U(\beta^0, \rho)$ with probability higher than $1 - \varepsilon$ for $n \geq n_0$. As we have shown that $S_n^\lambda(\beta) \to S^\lambda(\beta)$ in probability uniformly on $U(\beta^0, \delta)$, $|S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})| \leq \sup_{\beta \in U(\beta^0,\rho)} |S_n^\lambda(\beta)|$ is bounded in probability by Assumption F3. The other part of the expression can be bounded as follows. It holds with an arbirtarily high probability that

$$\sqrt{n} \left| S^{\lambda,1}(\hat{\beta}_n^{(GMTM,\lambda)}) - \frac{\partial S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})}{\partial \beta^\top} \right| \leq \sqrt{n} \sup_{\beta \in U(\beta^0,\rho)} \left| S^{\lambda,1}(\beta) - \frac{\partial S_n^\lambda(\beta)}{\partial \beta^\top} \right| \leq$$

$$\leq \frac{1}{\sqrt{n}} \sup_{\beta \in U(\beta^0,\rho)} \left| \sum_{i=1}^n \left\{ \mathsf{E}\left[s_i'(\beta)\nu_i^\lambda(\beta) \right] - s_i'(\beta)\nu_i^\lambda(\beta) \right\} \right| \tag{.11}$$

$$+ \frac{1}{\sqrt{n}} \sup_{\beta \in U(\beta^0,\rho)} \left| \sum_{i=1}^n s_i'(\beta) \left[\nu_i^\lambda(\beta) - \iota_{in}^\lambda(\beta) \right] \right|. \tag{.12}$$

The second term (.12) is bounded in probability due to Čížek (2008a, Corrolary A.6) under Assumptions D and F. The other part (.11) on the right-hand side can be bounded in probability by the following argument. Assumption F2 together with Van der Vaart and Wellner (1996, Lemma 2.6.18) imply that $\mathcal{F}_{n,\delta} = \left\{ s_i'(\beta)\nu_i^\lambda(\beta) : \beta \in U(\beta^0, \delta) \right\}$ forms a VC class of functions. Therefore, Assumptions D1 and F2 permit the use of the uniform central limit theorem of Arcones and Yu (1994), which implies that $\mathcal{F}_{n,\delta}$ converges in distribution to a Gaussian process with uniformly bounded and continuous paths and confirms that (.11) is bounded in probability, which concludes the proof. □

The proof of the asymptotic normality of GMTM follows.

Proof of Theorem 2: The asymptotic normality of GMTM is a direct consequence of its $\sqrt{n}$ consistency (Lemma 5) and the asymptotic linearity of the trimmed moment equations

following from Čížek (2008a, Lemma A.7) who proved under Assumptions D and F that

$$(.13) \qquad n^{-\frac{1}{2}} \sup_{t \in T_{\mathcal{M}}} |nS_n^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t) - nS_n^{\lambda}(\beta^0) + n\{J_s(\lambda) + J_I(\lambda)\}n^{-\frac{1}{2}}t| = o_p(1),$$

where $T_{\mathcal{M}} = \{t \in \mathbb{R}^p | \|t\| \leq \mathcal{M}\}$ and $\mathcal{M} > 0$.

Since $t_n = \sqrt{n}(\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0) = \mathcal{O}_p(1)$ as $n \to +\infty$ by Lemma 5, we can write

$$(.14) \qquad S_n^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t_n) - S_n^{\lambda}(\beta^0) + n^{-\frac{1}{2}}\{J_s(\lambda) + J_I(\lambda)\}t_n = o_p\left(n^{-\frac{1}{2}}\right)$$

with a probability arbitrarily close to one uniformly in $t_n \in T_{\mathcal{M}}$. Moreover, $\partial S_n^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t_n)/\partial\beta^{\top} \to J_s(\lambda)$ in probability as $n \to \infty$, see (.7) in the proof of Lemma 5 ($n^{-\frac{1}{2}}t_n = o_p(1)$). Hence, the first order conditions of GMTM (see also (.4)),

$$\frac{\partial Q_n^{W,\lambda}(\hat{\beta}_n^{(GMTM,\lambda)})}{\partial\beta} = \left[\frac{\partial S_n^{\lambda}(\hat{\beta}_n^{GMTM,\lambda})}{\partial\beta^{\top}}\right]^{\top} W S_n^{\lambda}(\hat{\beta}_n^{(GMTM,\lambda)}) = 0,$$

imply after substituting for $S_n^{\lambda}(\hat{\beta}_n^{(GMTM,\lambda)}) = S_n^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t_n)$ from equation (.14) and substituting $\partial S_n^{\lambda}(\hat{\beta}_n^{(GMTM,\lambda)})/\partial\beta^{\top} = \partial S_n^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t_n)/\partial\beta^{\top} = J_s(\lambda) + o_p(1)$ that

$$[J_s(\lambda) + o_p(1)]^{\top} W\left[S_n^{\lambda}(\beta^0) - n^{-\frac{1}{2}}\{J_s(\lambda) + J_I(\lambda)\}t_n + o_p\left(n^{-\frac{1}{2}}\right)\right] = 0.$$

Expressing now t_n from this equation results in

$$(.15) \quad t_n = \sqrt{n}(\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0) = \sqrt{n}\left[J_s(\lambda)^{\top} W\{J_s(\lambda) + J_I(\lambda)\}\right]^{-1} J_s(\lambda)^{\top} W S_n^{\lambda}(\beta^0) + o_p(1)$$

(note that $J_s(\lambda)$, $J_s(\lambda)+J_I(\lambda)$, and W are non-singular matrices by Assumptions F3 and I2).

To find the asymptotic distribution of the GMTM estimate, we thus have to analyze the asymptotic behavior of $\sqrt{n}S_n^{\lambda}(\beta^0)$ as all other terms on the right hand side of (.28) are constants (except for $o_p(1)$, of course). By definition, it follows that

$$\begin{aligned} \sqrt{n}S_n^{\lambda}(\beta^0) &= n^{-\frac{1}{2}}\sum_{i=1}^{n} s_i(\beta^0)\iota_{in}^{\lambda}(\beta^0) \\ (.16) \qquad &= n^{-\frac{1}{2}}\sum_{i=1}^{n} s_i(\beta^0)\left\{\iota_{in}^{\lambda}(\beta^0) - \nu_i^{\lambda}(\beta^0)\right\} \\ (.17) \qquad &+ n^{-\frac{1}{2}}\sum_{i=1}^{n} s_i(\beta^0)\nu_i^{\lambda}(\beta^0). \end{aligned}$$

First, we show that (.16) is asymptotically negligible in probability. Expectations

$$\mathsf{E}\left\|n^{\frac{1}{4}}s_i(\beta^0)\left\{\iota_{in}^{\lambda}(\beta^0) - \nu_i^{\lambda}(\beta^0)\right\}\right\|^l = \mathsf{E}\left\{n^{\frac{l}{4}}\left\|s_i(\beta^0)\right\|^l\left|\iota_{in}^{\lambda}(\beta^0) - \nu_i^{\lambda}(\beta^0)\right|\right\} = \mathcal{O}(1)$$

are bounded for $l = 1, 2$ due to Assumption F2 and Čížek (2008a, Corollary A.5). Assumptions D1, F2, and I3 further indicate that the summands in (.16) multiplied by $n^{\frac{1}{4}}$ form a stationary sequence of random variables with zero means and finite variances ($\nu_i^\lambda(\beta^0)$ is the probability limit of $\iota_{in}^\lambda(\beta^0)$, see the proof of Theorem 1 or Čížek, 2008a, Lemma A.1). Thus, the law of large numbers for mixingales (Davidson, 1994, Corollary 20.16) leads to

$$n^{-\frac{3}{4}} \sum_{i=1}^{n} n^{\frac{1}{4}} s_i(\beta^0) \left\{ \iota_{in}^\lambda(\beta^0) - \nu_i^\lambda(\beta^0) \right\} \to 0,$$

which implies that (.16) is negligible in probability as $n \to \infty$.

Second, the summands in (.17), $s_i(\beta^0)\nu_i^\lambda(\beta^0)$, form a stationary sequence of absolutely regular random variables with zero mean and finite second moments (Assumptions D1, F2, and I3). We can thus employ the central limit theorem for (.17) (e.g., Arcones and Yu, 1994, by Assumptions D1 and F2). This results directly in the asymptotic normality of $\sqrt{n} S_n^\lambda(\beta^0) \sim N(0, V_s(\lambda))$ (cf. Davidson, 1994, Theorem 25.3).

Using equation (.15), the asymptotic normality of $\hat{\beta}_n^{(GMTM,\lambda)}$ follows with the asymptotic variance given by (Davidson, 1994, Theorem 22.8)

$$V(\lambda) = \Big[J_s(\lambda)^\top W \{ J_s(\lambda) + J_I(\lambda) \} \Big]^{-1} J_s(\lambda)^\top W \cdot V_s(\lambda) \cdot W J_s(\lambda) \Big[J_s(\lambda)^\top W \{ J_s(\lambda) + J_I(\lambda) \} \Big]^{-1\top}.$$

□

Next, we attempt to derive an analytic form of $J_I(\lambda)$ in order to be able compute the asymptotic variance matrix $V(\lambda)$. To achieve this, we have to study probability that the trimming indicator $\iota_{in}^\lambda(\beta)$ changes if we use $\beta = \hat{\beta}_n^{(GMTM,\lambda)}$ instead of $\beta = \beta^0$.

Lemma 6. *Under the assumptions of Lemma 3, it holds for any $\beta \in U(\beta^0, n^{-\frac{1}{2}}\mathcal{M})$ and $\mathcal{M} > 0$ that*

$$\begin{aligned} &P\big(I\big(r_i(\beta) \le r_{([\lambda n])}(\beta)\big) \ne I\big(r_i(\beta^0) \le r_{([\lambda n])}(\beta^0)\big) \big| \, v_i \big) \\ &= \Big| h_2'(v_i; \beta^0)^\top (\beta - \beta^0) \Big| \cdot \Big\{ f_{v_i}\Big(-\sqrt{G^{-1}(\lambda)} \Big) + f_{v_i}\Big(\sqrt{G^{-1}(\lambda)} \Big) \Big\} + o_p\Big(n^{-\frac{1}{2}} \Big) \end{aligned}$$

in probability as $n \to \infty$ and

$$\begin{aligned} &\mathsf{E}\left\{ \operatorname{sgn} h(d_i; \beta^0) \cdot \big[I\big(r_i(\beta) \le r_{([\lambda n])}(\beta)\big) - I\big(r_i(\beta^0) \le r_{([\lambda n])}(\beta^0)\big) \big] \big| \, v_i \right\} \\ &= -h_2'(v_i; \beta^0)^\top (\beta - \beta^0) \cdot \Big\{ f_{v_i}\Big(-\sqrt{G^{-1}(\lambda)} \Big) + f_{v_i}\Big(\sqrt{G^{-1}(\lambda)} \Big) \Big\} + o_p\Big(n^{-\frac{1}{2}} \Big). \end{aligned}$$

Proof: To simplify notation, let us first denote $q_\lambda = \sqrt{G^{-1}(\lambda)}$ and recall that $\delta_{in}^\lambda(\beta) = \iota_{in}^\lambda(\beta) - \iota_{in}^\lambda(\beta^0)$. Our aim is then to compute $P(|\delta_{in}^\lambda(\beta)| = 1|v_i)$.

Consider first $P(\delta_{in}^\lambda(\beta) = -1|v_i)$. Apparently, $\delta_{in}^\lambda(\beta) = -1$ if and only if

$$r_i(\beta) = h^2(d_i;\beta) > r_{([\lambda n])}(\beta) \quad \text{and} \quad r_i(\beta^0) = h_i^2(d_i;\beta^0) \leq r_{([\lambda n])}(\beta^0),$$

which implies

(.18)

$$h(d_i;\beta) \in \left(-\infty, r_{([\lambda n])}^{1/2}(\beta)\right) \cup \left(r_{([\lambda n])}^{1/2}(\beta), \infty\right) \text{ and } h(d_i;\beta^0) \in \left\langle -r_{([\lambda n])}^{1/2}(\beta^0), r_{([\lambda n])}^{1/2}(\beta^0)\right\rangle.$$

By means of the Taylor expansion we can write ($h_2(v_i;\beta^0) = 0$)

$$h(d_i;\beta) = h_1(d_i) + h_2(v_i;\beta) = h_1(d_i) + h_2^{'}(v_i;\xi_1)^\top(\beta - \beta^0),$$

where $h(d_i;\beta^0) = h_1(d_i)$, $\xi_1 \in \left[\beta^0, \beta\right]_{\curlywedge}$, and $\left[\beta^0, \beta\right]_{\curlywedge}$ denotes a convex span of β and β^0. Combining this result with (.18) and denoting $\Delta_h(v_i;\beta) = h_2^{'}(v_i;\xi_1)^\top(\beta - \beta^0)$, we see that

$$(.19)\quad h_1(d_i) \in \left\langle -r_{([\lambda n])}^{1/2}(\beta^0), -r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta)\right) \cup \left(r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta), r_{([\lambda n])}^{1/2}(\beta^0)\right\rangle,$$

where the convention $(a,b) = \emptyset$ if $b < a$ is used. For $\delta_{in}^\lambda(\beta) = 1$, it is possible to analogously derive that

$$(.20)\quad h_1(d_i) \in \left\langle -r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta), -r_{([\lambda n])}^{1/2}(\beta^0)\right) \cup \left(r_{([\lambda n])}^{1/2}(\beta^0), r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta)\right\rangle.$$

Next, combining (.19) and (.20) allows us to express $P(|\delta_{in}^\lambda(\beta)| = 1|v_i)$ as

$$P\Big(h_1(d_i) \in \left[-r_{([\lambda n])}^{1/2}(\beta^0), -r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta)\right]_{\curlywedge} \cup \left[r_{([\lambda n])}^{1/2}(\beta^0), r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta)\right]_{\curlywedge}\Big| v_i\Big).$$

This can be further simplified using Čížek (2004, Lemma A.4) to

$$P\Big(h_1(d_i) \in \left[-r_{([\lambda n])}^{1/2}(\beta^0), -r_{([\lambda n])}^{1/2}(\beta^0) - \Delta_h(v_i;\beta)\right]_{\curlywedge} \cup \left[r_{([\lambda n])}^{1/2}(\beta^0), r_{([\lambda n])}^{1/2}(\beta^0) - \Delta_h(v_i;\beta)\right]_{\curlywedge}\Big| v_i\Big)$$
$$+ o_p\Big(n^{-\frac{1}{2}}\Big)$$

as f_{v_i} is uniformly bounded. At this point, let us note that, conditionally on v_i, $\delta_{in}^\lambda(\beta) \neq 0$ implies $\delta_{in}^\lambda(\beta) \cdot \operatorname{sgn} h(d_i;\beta^0) = -\operatorname{sgn}\Delta_h(v_i;\beta)$ with probability approaching 1 as $1 - \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ with $n \to \infty$. We prove it as follows. On the one hand, $\Delta_h(v_i;\beta)$ is (conditionally on v_i) bounded and converges to zero as $n \to \infty$ because $\beta \in U(\beta^0, n^{-\frac{1}{2}}\mathcal{M})$. We can thus choose $n_0 \in \mathbb{N}$ such that $|\Delta_h(v_i;\beta)| < q_\lambda/\sqrt{2}$ for all $n \geq n_0$. On the other hand, $P(r_{([\lambda n])}^{1/2}(\beta^0) <$

$q_\lambda/\sqrt{2}) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ by Čížek (2008a, Lemma A.3). Hence, it follows that we can write for $\Delta_h(v_i;\beta) < 0$ and $n \geq n_0$ with probability $1 - \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$, see (.19) and (.20):

$$\begin{aligned}
\delta_{in}^{\lambda}(\beta) = 1 \quad &\Longrightarrow \quad h(d_i,\beta^0) = h_1(d_i) \in \Big(r_{([\lambda n])}^{1/2}(\beta^0), r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta)\Big) \subset (0,+\infty), \\
\delta_{in}^{\lambda}(\beta) = -1 \quad &\Longrightarrow \quad h(d_i,\beta^0) = h_1(d_i) \in \Big(-r_{([\lambda n])}^{1/2}(\beta^0), -r_{([\lambda n])}^{1/2}(\beta) - \Delta_h(v_i;\beta)\Big) \subset (-\infty,0).
\end{aligned}$$

A similar discussion can be made for the case of $\Delta_h(v_i;\beta) > 0$.

Now, let us return to the analysis of (.21). Because the density function f_{v_i} of $h_1(d_i)|v_i$ is bounded and differentiable in a neighborhood of q_λ, we can rewrite probability (.21) as

$$(.22) \qquad P(\, h_1(d_i) \in [-q_\lambda - \xi_2, -q_\lambda - \xi_2 - \Delta_h(v_i;\beta)]_{\lessgtr} \cup [q_\lambda + \xi_2, q_\lambda + \xi_2 - \Delta_h(v_i;\beta)]_{\lessgtr} |\, v_i),$$

where $\xi_2 = r_{([\lambda n])}^{1/2}(\beta^0) - q_\lambda = \mathcal{O}_p\Big(n^{-\frac{1}{2}}\Big)$ as $n \to \infty$ by Čížek (2008a, Lemma A.2). The mean value theorem and Taylor expansion for the distribution function f_{v_i} further lead to

$$\begin{aligned}
&P(\, h_1(d_i) \in [-q_\lambda - \xi_2, -q_\lambda - \xi_2 - \Delta_h(v_i;\beta)]_{\lessgtr} \cup [q_\lambda + \xi_2, q_\lambda + \xi_2 - \Delta_h(v_i;\beta)]_{\lessgtr} |\, v_i) \\
&= |\Delta_h(v_i;\beta)| \cdot \{f_{v_i}(-q_\lambda) + f_{v_i}(q_\lambda) + f_{v_i}^{'}(\xi_3)[\xi_2 + \Delta_h(v_i;\beta)] + f_{v_i}^{'}(\xi_4)[\xi_1 + \Delta_h(v_i;\beta)] \\
&= \Big|h_2^{'}(v_i;\xi_1)^\top(\beta - \beta^0)\Big| \cdot \{f_{v_i}(-q_\lambda) + f_{v_i}(q_\lambda)\} + o_p\Big(n^{-\frac{1}{2}}\Big)
\end{aligned}$$

because of $\beta \in U(\beta^0, n^{-\frac{1}{2}}\mathcal{M})$ and the assumptions of the lemma. The first conclusion of the lemma now follows from

$$h_2^{'}(v_i;\xi_1) = h_2^{'}(v_i;\beta^0) + h_2^{''}(v_i;\zeta)(\xi_1 - \beta^0) = h_2^{'}(v_i;\beta^0) + o_p\Big(n^{-\frac{1}{2}}\Big)$$

since $\max\{\|\xi_1 - \beta^0\|, \|\zeta - \beta^0\|\} \leq \|\beta - \beta^0\| = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$ as $n \to \infty$ and $h_2^{''}(v_i;\zeta) \to h_2^{''}(v_i;\beta^0) = 0$ in probability. The second conclusion is a direct consequence of the note explaining $\delta_{in}^{\lambda}(\beta) \cdot \operatorname{sgn} h(d_i;\beta^0) = -\operatorname{sgn} \Delta_h(v_i;\beta)$. □

The derivation of an analytic form of $J_I(\lambda)$ follows.

Proof of Lemma 3: Using the definition of partial derivatives and Čížek (2008a, Lemma A.3), we can write (see Čížek, 2007b, Lemma A.7)

(.23)

$$J_I(\lambda) = \frac{\partial}{\partial \beta_j} \mathsf{E}\left[s_i(\beta^0)\nu_i^{\lambda}(\beta^0)\right]_{\beta=\beta^0} = \lim_{n\to\infty} \frac{1}{n^{-\frac{1}{2}}T} \mathsf{E}\left[s_i(\beta^0)\left\{\iota_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t_j) - \iota_{in}^{\lambda}(\beta^0)\right\}\right],$$

where $e_j = (0,\ldots,0,1,0,\ldots,0)^\top$ represent the jth basis vector of $\mathbb{R}^p$, $t_j = Te_j$ and $T \in \mathbb{R}$, and $j = 1,\ldots,p$. Thus, we can employ the results of Lemma 6 to derive $J_I(\lambda)$. To do so, let

us express for any $t \in \mathbb{R}^p$

$$\mathsf{E}\left[s_i(\beta^0)\left\{\iota_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t) - \iota_{in}^{\lambda}(\beta^0)\right\}\right] = \underset{v}{\mathsf{E}}\,\mathsf{E}\left[s_i(\beta^0)\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\Big|\, v_i\right].$$

If, conditional on v_i, $\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t) \neq 0$ for some value of t, the proof of Lemma 6, equation (.22), implies ($\|\xi_1 - \beta^0\| \leq \|t\|$)

$$\left|h_1(d_i) - \sqrt{G^{-1}(\lambda)}\right| \leq \left|h_2'(v_i;\xi_1)^\top n^{-\frac{1}{2}}t\right| + \mathcal{O}_p\left(n^{-1/2}\right). \tag{.24}$$

This motivates the following decomposition:

$$\begin{aligned}
&\underset{v}{\mathsf{E}}\,\mathsf{E}\Big[\, s_i(\beta^0)\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\Big|\, v_i\Big] \\
&= \underset{v}{\mathsf{E}}\,\mathsf{E}\Big[\left(s_i(\beta^0) - \mathsf{E}\Big\{\, s(d_i,\beta^0)\big|\,\operatorname{sgn} h_1(d_i), |h_1(d_i)| = \sqrt{G^{-1}(\lambda)}, v_i\Big\}\right)\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\Big|\, v_i\Big] \qquad (.25)\\
&= \underset{v}{\mathsf{E}}\,\mathsf{E}\Big[\mathsf{E}\Big\{\, s(d_i,\beta^0)\big|\,\operatorname{sgn} h_1(d_i), |h_1(d_i)| = \sqrt{G^{-1}(\lambda)}, v_i\Big\}\,\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\Big|\, v_i\Big]. \qquad (.26)
\end{aligned}$$

The first term (.25) will be shown to behave like to $o(n^{-\frac{1}{2}})$. We can namely bound the absolute value of (.25) using (.24) and Lemma 6 by

$$\begin{aligned}
&\underset{v}{\mathsf{E}}\,\mathsf{E}\Big[\,\Big|s_i(\beta^0) - \mathsf{E}\Big\{\, s(d_i,\beta^0)\big|\,\operatorname{sgn} h_1(d_i), |h_1(d_i)| = \sqrt{G^{-1}(\lambda)}, v_i\Big\}\Big|\,\Big|\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\Big|\,\Big|\, v_i\Big] \\
&\leq \underset{v}{\mathsf{E}}\Big[d\left\{\left|h_2'(v_i;\xi_1)^\top n^{-\frac{1}{2}}t\right| + \mathcal{O}_p\left(n^{-1/2}\right)\right\}\bar{s}(v_i)\,\mathsf{E}\left\{\left|\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\right|\Big|\, v_i\right\}\Big] \\
&\leq \mathcal{O}\left(n^{-\frac{1}{2}}\right)\underset{v}{\mathsf{E}}\Big[d\left(\left|h_2'(v_i;\xi_1)^\top n^{-\frac{1}{2}}t\right| + \mathcal{O}_p\left(n^{-\frac{1}{2}}\right)\right)\bar{s}(v_i)\left\{\left|h_2'(v_i;\beta^0)^\top t\right| + \mathcal{O}_p(1)\right\}\Big].
\end{aligned}$$

The last expectation is asymptotically negligible since $\sup_{\beta\in U(\beta^0,\delta)}\mathsf{E}\,|h_2'(v_i;\beta)|^{1+\delta} < K_h \in \mathbb{R}$, $\left|h_2'(v_i;\xi_1)^\top n^{-\frac{1}{2}}t\right| + \mathcal{O}_p\big(n^{-1/2}\big)$ is uniformly integrable (Davidson, 1994, Theorem 12.10), and thus asymptotically negligible both in probability and expectation (d is a locally Lipschitz norm).

Hence, we now have to deal only with term (.26), which by the assumptions of the lemma can be written as $\mathsf{E}_v\,\mathsf{E}[\operatorname{sgn} h_1(d_i)\tilde{s}(v_i)\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)|v_i]$. Using Lemma 6 and the uniform integrability of the moment and trimming functions and their derivatives (Assumption F3 and Davidson, 1994, Theorem 12.10), it follows that

$$\begin{aligned}
&\underset{v}{\mathsf{E}}\Big\{\tilde{s}(v_i)\cdot\mathsf{E}\Big[\operatorname{sgn} h_1(d_i)\delta_{in}^{\lambda}(\beta^0 - n^{-\frac{1}{2}}t)\Big|\, v_i\Big]\Big\} \\
= \;& -n^{-\frac{1}{2}}t\,\underset{v}{\mathsf{E}}\Big\{\tilde{s}(v_i)h_2'(v_i;\beta^0)^\top\cdot\left[f_{v_i}\left(-\sqrt{G^{-1}(\lambda)}\right) + f_{v_i}\left(\sqrt{G^{-1}(\lambda)}\right)\right]\Big\}.
\end{aligned}$$

Substituting back to (.23) results in the claim of the lemma:

$$J_I(\lambda) = -\mathop{\mathsf{E}}_{v}\Big\{\tilde{s}(v_i)h_2'(v_i;\beta^0)^\top \cdot \Big[f_{v_i}\left(-\sqrt{G^{-1}(\lambda)}\right) + f_{v_i}\left(\sqrt{G^{-1}(\lambda)}\right)\Big]\Big\}. \qquad \square$$

Finally, the test of overidentifying restrictions is derived.

Proof of Theorem 4: We showed in the proof of Theorem 2 for $t_n = \mathcal{O}_p(1)$ that

$$S_n^\lambda(\beta^0 - n^{-\frac{1}{2}}t_n) - S_n^\lambda(\beta^0) + n^{-\frac{1}{2}}\left\{J_s(\lambda) + J_I(\lambda)\right\}t_n = o_p\Big(n^{-\frac{1}{2}}\Big), \tag{.27}$$

see equation (.14), and that

$$t_n = \sqrt{n}(\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0) = \sqrt{n}\left[J_s(\lambda)^\top W\{J_s(\lambda) + J_I(\lambda)\}\right]^{-1} J_s(\lambda)^\top W S_n^\lambda(\beta^0) + o_p(1), \tag{.28}$$

see equation (.15). Substituting t_n from (.28) to (.27), multiplying the whole equation by $\sqrt{n}$, and using $t_n = \sqrt{n}(\hat{\beta}_n^{(GMTM,\lambda)} - \beta^0)$, we obtain that

$$\begin{aligned}\sqrt{n}S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)}) &= \sqrt{n}\left[I - \{J_s(\lambda) + J_I(\lambda)\}\left[J_s(\lambda)^\top W\{J_s(\lambda) + J_I(\lambda)\}\right]^{-1} J_s(\lambda)^\top W\right] S_n^\lambda(\beta^0) \\ &\quad + o_p(1) \\ &= [I - \Pi(\lambda)]\sqrt{n}S_n^\lambda(\beta^0) + o_p(1).\end{aligned}$$

At the same time, we showed in the proof of Theorem 2 that $\sqrt{n}S_n^\lambda(\beta^0)$ converges in distribution to a normally distributed random variable with variance $V_s(\lambda)$, see the discussion of (.16)–(.17). Consequently, we see that $\sqrt{n}S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})$ is asymptotically normally distributed with its asymptotic variance matrix equal to

$$\Sigma(\lambda) \;=\; [I - \Pi(\lambda)]\, V_s(\lambda)\, [I - \Pi(\lambda)]^\top,$$

which can be consistently estimated by $\hat{\Sigma}_n(\lambda) = \left[I - \hat{\Pi}_n(\lambda)\right]\hat{V}_{sn}(\lambda)\left[I - \hat{\Pi}_n(\lambda)\right]^\top$ by the assumptions of the theorem. Hence, the test statistics

$$T_n = \sqrt{n}S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})^\top \hat{\Sigma}_n^-(\lambda)\sqrt{n}S_n^\lambda(\hat{\beta}_n^{(GMTM,\lambda)})$$

has asymptotically the same distribution as $Z^\top\Sigma^-(\lambda)Z$, where $Z \sim N(0, \Sigma(\lambda))$ and the generalized inverses $\hat{\Sigma}_n^-(\lambda)$ and $\Sigma^-(\lambda)$ are defined in the same way in order to $\hat{\Sigma}_n^-(\lambda) \to \Sigma^-(\lambda)$ in probability. The distribution of the quadratic form $Z^\top\Sigma^-(\lambda)Z$ is the χ^2 distribution with the degrees of freedom equal to the rank of $\Sigma(\lambda)$. Due to Assumption F3, the rank of $\Sigma(\lambda)$ equals to the rank of $I - \Pi(\lambda)$ and thus to $M - \text{rank}\{\Pi(\lambda)\}$. Since $\Pi(\lambda)$ is idempotent, the

rank of an idempotent matrix equals its trace, and

$$\mathrm{tr}\{\Pi(\lambda)\} = \mathrm{tr}\left[\left[J_s(\lambda)^\top W\{J_s(\lambda)+J_I(\lambda)\}\right]^{-1} J_s(\lambda)^\top W\left\{J_s(\lambda)+J_I(\lambda)\right\}\right] = I_{p\times p},$$

it follows that $\mathrm{rank}\{\Sigma(\lambda)\} = M - p$ and thus asymptotically $T_n \sim \chi^2_{M-p}$. □

References

Abadir, K. and J. Magnus, 2005, Matrix Algebra. Cambridge University Press, New York.

Altonji, J. G. and L. M. Segal, 1996, Small-sample bias in GMM estimation of covariance structures. Journal of Business and Economic Statistics 14, 353–366.

Amemiya, T., 1982, Two stage least absolute deviations estimators. Econometrica 50, 689–711.

Andrews, D. W. K., 1993, An introduction to econometric applications of empirical process theory for dependent random variables. Econometric Reviews 12, 183–216.

Arcones, M. A. and B. Yu, 1994, Central limit theorems for empirical and U-processes of stationary mixing sequences. Journal of Theoretical Probability 7, 47–71.

Bassett, G. and R. Koenker, 1978, Asymptotic theory of least absolute error regression. Journal of the American Statistical Association 73, 618–622.

Bramati, M. C. and C. Croux, 2007, Robust estimators for the fixed effects panel data model. Econometrics Journal 10, 521–540.

Cantoni, E. and E. Ronchetti, 2001, Robust inference for generalized linear models. Journal of the American Statistical Association 96, 1022–1030.

Chao, J. C. and N. R. Swanson, 2005, Consistent estimation with a large number of weak instruments. Econometrica 73, 1673–1692.

Chernozhukov, V. and C. Hansen, 2008, Instrumental variable quantile regression: A robust inference approach. Journal of Econometrics 142, 379–398.

Čížek, P., 2004, Asymptotics of least trimmed squares regression. CentER Discussion paper 72/2004, Tilburg University, The Netherlands.

Čížek, P., 2006, Least trimmed squares in nonlinear regression under dependence. Journal of Statistical Planning and Inference 136, 3967–3988.

Čížek, P., 2007a, Efficient robust estimation of regression models. CentER Discussion paper 87/2007, Tilburg University, The Netherlands.

Čížek, P., 2007b, General trimmed estimation: robust approach to nonlinear and limited dependent variable models. CentER Discussion paper 1/2007, Tilburg University, The Netherlands.

Čížek, P., 2008a, General trimmed estimation: robust approach to nonlinear and limited dependent variable models. Econometric Theory 24, 1500–1529.

Čížek, P., 2008b, Robust and efficient adaptive estimation of binary-choice regression models. Journal of the American Statistical Association 103, 687–696.

Čížek, P., 2008c, Robust instrumental-variable estimators based on quantile conditions. Center discussion paper, Tilburg University, The Netherlands.

Croux, C., P. J. Rousseeuw, and O. Hössjer, 1994, Generalized S-estimators. Journal of the American Statistical Association 89, 1271–1281.

Czellar, V., G. A. Karolyi, and E. Ronchetti, 2007, Indirect robust estimation of the short-term interest rate process. Journal of Empirical Finance 14, 546–563.

Davidson, J., 1994, Stochastic Limit Theory. Oxford University Press, New York.

Dell'Aquila, R., E. Ronchetti, and F. Trojani, 2003, Robust GMM analysis of models for the short rate process. Journal of Empirical Finance 10, 373–397.

Ferretti, N., D. Kelmansky, V. J. Yohai, and R. H. Zamar, 1999, A class of locally and globally robust regression estimates. Journal of the American Statistical Association 94, 174–188.

Genton, M. G. and A. Lucas, 2003, Comprehensive definitions of breakdown points for independent and dependent observations. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 65, 81–94.

Genton, M. G. and E. Ronchetti, 2003, Robust indirect inference. Journal of the American Statistical Association 98, 67–76.

Gervini, D. and V. J. Yohai, 2002, A class of robust and fully efficient regression estimators. The Annals of Statistics 30, 583–616.

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

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, 1986, Robust statistics, the approach based on influence function. Wiley, New York.

Hansen, L. P., 1982, Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

He, X., J. Jurečková, R. Koenker, and S. Portnoy, 1990, Tail behavior of regression estimators and their breakdown points. Econometrica 58, 1195–1214.

Honore, B. E. and L. Hu, 2004, On the performance of some robust instrumental variables estimators. Journal of Business & Economic Statistics 22, 30–39.

Hössjer, O., 1994, Rank-based estimates in the linear model with high breakdown point. Journal of the American Statistical Association 89, 149–158.

Hubert, M. and P. J. Rousseeuw, 1997, Robust regression with both continuous and binary regressors. Journal of Statistical Planning and Inference 57, 153–163.

Imbens, G. W., R. H. Spady, and P. Johnson, 1998, Information theoretic approaches to inference in moment condition models. Econometrica 66, 333–357.

Knez, P. J. and M. J. Ready, 1997, On the robustness of size and book-to-market in cross-sectional regressions. The Journal of Finance 52, 1355–1382.

Krasker, W. S., 1986, Two-stage bounded-influence estimators for simultaneous-equations models. Journal of Business & Economic Statistics 4, 437–44.

Krasker, W. S. and R. E. Welsch, 1985, Resistant estimation for simultaneous-equations models using weighted instrumental variables. Econometrica 53, 1475–1488.

Krishnakumar, J. and E. Ronchetti, 1997, Robust estimators for simultaneous equations models. Journal of Econometrics 78, 295–314.

Lo, S. N. and E. Ronchetti, 2006, Robust small sample accurate inference in moment condition models. Cahiers du Département d'Econométrie 2006.04, Département d'Econométrie, Université de Genève.

Manski, C. F., 1988, Analog estimation methods in econometrics. Chapman and Hall, New York.

Marazzi, A. and V. J. Yohai, 2004, Adaptively truncated maximum likelihood regression with asymmetric errors. Journal of Statistical Planning and Inference 122, 271–291.

Maronna, R., O. Bustos, and V. J. Yohai, 1979, Bias and efficiency robustness of general M-estimators for regression with random carriers. In: Gasser, T. and M. Rossenblatt (Eds.) Smoothing techniques for curve estimation. Springer, pp. 91–111.

Müller, C. and T.-H. Kim, 2005, Two-stage Huber estimation. Working Papers. Serie AD 2005-17, Instituto Valenciano de Investigaciones Económicas, S.A. (Ivie).

Müller, C. H. and N. Neykov, 2003, Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. Journal of Statistical Planning and

Inference 116, 503–519.

Newey, W. K. and R. J. Smith, 2004, Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72, 219–255.

Newey, W. K. and K. D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Ortelli, C. and F. Trojani, 2005, Robust efficient method of moments. Journal of Econometrics 128, 69–97.

Pakes, A. and D. Pollard, 1989, Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057.

Peracchi, F., 1990, Bounded-influence estimators for the Tobit model. Journal of Econometrics 44, 107–126.

Peracchi, F., 1991, Bounded-influence estimators for the SURE model. Journal of Econometrics 48, 119–134.

Powell, J. L., 1984, Least absolute deviations estimation for the censored regression model. Journal of Econometrics 25, 303–325.

Ronchetti, E. and F. Trojani, 2001, Robust inference with GMM estimators. Journal of Econometrics 101, 37–69.

Rousseeuw, P. J., 1985, Multivariate estimation with high breakdown point. In: Grossman, W., G. Pflug, I. Vincze, and W. Wertz (Eds.) Mathematical statistics and applications, volume B. Reidel, Dordrecht, Netherlands, pp. 283–297.

Rousseeuw, P. J. and A. M. Leroy, 1987, Robust regression and outlier detection. Wiley, New York.

Sakata, S. and H. White, 1998, High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. Econometrica 66, 529–567.

Salibian-Barrera, M. and R. H. Zamar, 2002, Bootstrapping robust estimates of regression. The Annals of Statistics 30, 556–582.

Simpson, D. G., D. Ruppert, and R. J. Carroll, 1992, On one-step GM estimates and stability of inferences in linear regression. Journal of the American Statistical Association 87, 439–450.

Stock, J. H. and J. H. Wright, 2000, GMM with weak identification. Econometrica 68, 1055–1096.

Stromberg, A. J., O. Hössjer, and D. M. Hawkins, 2000, The least trimmed differences regression estimator and alternatives. Journal of the American Statistical Association 95, 853–864.

Temple, J. R. W., 1998, Robustness tests of the augmented Solow model. Journal of Applied Econometrics 13, 361–375.

Van der Vaart, A. W. and J. A. Wellner, 1996, Weak convergence and empirical processes: with applications to statistics. Springer, New York.

Víšek, J. Á., 2006, Instrumental weighted variables. Austrian Journal of Statistics 35, 379–387.

Wagenvoort, R. and R. Waldmann, 2002, On B-robust instrumental variable estimation of the linear model with panel data. Journal of Econometrics 106, 297–324.

Zinde-Walsh, V., 2002, Asymptotic theory for some high breakdown point estimators. Econometric Theory 18, 1172–1196.