# Indirect inference with(out) constraints

David T. Frazier
Department of Econometrics and Business Statistics, Monash University

Eric Renault
Department of Economics, University of Warwick

Indirect Inference (I-I) estimation of structural parameters $\theta$ requires matching observed and simulated statistics, which are most often generated using an auxiliary model that depends on instrumental parameters $\beta$. The estimators of the instrumental parameters will encapsulate the statistical information used for inference about the structural parameters. As such, artificially constraining these parameters may restrict the ability of the auxiliary model to accurately replicate features in the structural data, which may lead to a range of issues, such as a loss of identification. However, in certain situations the parameters $\beta$ naturally come with a set of $q$ restrictions. Examples include settings where $\beta$ must be estimated subject to $q$ possibly strict inequality constraints $g(\beta) > 0$, such as, when I-I is based on GARCH auxiliary models. In these settings, we propose a novel I-I approach that uses appropriately modified unconstrained auxiliary statistics, which are simple to compute and always exists. We state the relevant asymptotic theory for this I-I approach without constraints and show that it can be reinterpreted as a standard implementation of I-I through a properly modified binding function. Several examples that have featured in the literature illustrate our approach.

Keywords. Inequality restrictions, constrained estimation, parameters on the boundary, indirect inference, stochastic volatility.

JEL classification. C10, C13, C15.

## 1. Introduction

The indirect estimation procedures of Gourieroux, Monfort, and Renault (1993) (hereafter, GMR), Smith (1993), and Gallant and Tauchen (1996) (hereafter, GT) provide convenient estimation methods when efficient estimation of a fully parametric structural model is a daunting task due to the intractability of the likelihood function. GMR motivate Indirect Inference (I-I) by arguing that in such cases a natural procedure is to replace the likelihood function by another criterion based on some convenient auxiliary (or naive) model that is simpler but possibly misspecified. The overall aim of I-I is then to conduct correct inference "based on this incorrect criterion."

As described by Jiang and Turnbull (2004), the "essential ingredients" of I-I are as follows:

(i) A parametric model for data generation, with distribution $P_\theta$ that depends on an unknown vector $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ of parameters of interest. This model is the so-called structural model and $\theta$ is the vector of structural parameters.

(ii) An intermediate or auxiliary statistic, say $\hat{\beta}_T$, of dimension $d_\beta \geq d_\theta$, which is a functional of the observed sample $\{y_t\}_{t=1}^T$.

(iii) A bridge (or binding) relationship $\beta = b(\theta)$ defined between the true unknown value $\theta^0$ of the structural parameters and $\beta^0 = \text{plim}_{T\to\infty} \hat{\beta}_T$, where the unknown quantity $\beta^0 = b(\theta^0)$ is referred to as the pseudo-true value of the auxiliary parameters.

(iv) With the auxiliary statistic $\hat{\beta}_T$ replacing $\beta$, the bridge relationship is used to compute an I-I estimator of $\theta$ by "inverting" $b(\theta)$.

Jiang and Turnbull (2004) acknowledge that "the choice of an intermediate statistic $\hat{\beta}_T$ is not necessarily unique," however, the authors argue that "in any given situation there is often a natural one to use." Herein, we question this traditional interpretation of I-I when it pertains to examples where the choice of "intermediate statistic" is ambiguous due to the fact that the parameters of the auxiliary model used in I-I must be estimated subject to a vector of inequality constraints.

In this commonly encountered situation, the choice of appropriate intermediate statistics for I-I can be ambiguous for several reasons: first, as noted by Calzolari, Fiorentini, and Sentana (2004) (hereafter, CFS), the pseudo-likelihood function of the auxiliary model may not be well behaved when certain parameter restrictions are violated, and hence, without these additional restrictions $\hat{\beta}_T$ cannot be obtained; second, if the pseudo-true value $\beta^0$ is on (or near) the boundary of the parameter space defined by the inequality constraints, the intermediate statistic $\hat{\beta}_T$ may be insufficient to identify $\theta^0$; lastly, even if identification of $\theta^0$ is possible, if $\beta^0$ is on (or near) the boundary of the parameter space defined by the constraints, pseudo-maximum likelihood (hereafter, PML) estimation will lead to an intermediate statistic that is not well suited for I-I because it is not asymptotically normal (see Andrews (1999) and CFS for details).

The question then is how to choose the intermediate statistics so as to guarantee consistent and asymptotically normal I-I estimators of $\theta^0$ even though $\beta^0$ can lie on (or near) the boundary of the parameter space defined by these inequality constraints? One approach, which is proposed by CFS, is to consider as our intermediate statistic a "well-behaved" linear combination of the constrained PML estimates of $\beta$, say $\hat{\beta}_T^r$, and the Kuhn–Tucker (hereafter, KT) multipliers, say $\hat{\lambda}_T$, corresponding to the inequality constraints. CFS demonstrate that one can use these linear combinations as intermediate statistics to produce I-I estimators of $\theta^0$ with asymptotically Gaussian limits. In addition, CFS show that imposing additional inequality restrictions on the auxiliary model will never decrease the efficiency of the resulting I-I estimator, so long as the corresponding KT multipliers are included in the vector of intermediate statistics. While correct, until now the reason behind this phenomena has not been completely understood.

Our first contribution is to demonstrate that the asymptotic normality of the CFS I-I estimator is not due to the information brought by the constraints but is a direct consequence of the relationship between $\hat{\beta}_T^r$ and $\hat{\lambda}_T$. In particular, we demonstrate that the "well-behaved" linear combinations of $\hat{\beta}_T^r$ and $\hat{\lambda}_T$ put forward by CFS as new "auxiliary parameters" is asymptotically equivalent to a new feasible unconstrained PML estimator of the auxiliary model, which is always well-defined and asymptotically normal. Therefore, adding constraints to an auxiliary model does not increase the information about the structural parameters because this information was already contained in the (unconstrained) auxiliary model.[1]

Using this new feasible unconstrained estimator, we propose a computationally simple unconstrained I-I estimation strategy that does not enforce the inequality constraints on the auxiliary model. Even though our new I-I estimators do not enforce the inequality constraints on the auxiliary parameters, our I-I estimators are asymptotically equivalent to the constrained I-I estimators proposed by CFS. Moreover, we demonstrate that the standard asymptotic Gaussian distribution of our I-I estimators remains valid, even if the pseudo-true value of the auxiliary parameters, $\beta^0$, is on or near the boundary of the parameter space.

Our second contribution is to make rigorous the notion of pseudo-true values of the auxiliary parameters on or near the boundary of the auxiliary parameter space. While the approach of CFS treats the case of $\beta^0$ on the boundary, their approach is dichotomous: $\beta^0$ is either on the boundary, in which case $\hat{\beta}_T^r$ is not asymptotically normal, or it is not on the boundary, in which case $\hat{\beta}_T^r$ is asymptotically normal. This binary treatment of parameters near the boundary cannot capture cases where $\beta^0$ is close enough to the boundary of the parameter space to render standard asymptotic approximations unreliable. In such cases, modeling the pseudo-true value as a sequence that is close to but not on the boundary, for any finite sample size, can provide more reliable asymptotic approximations.

Following, among others, Moon and Schorfheide (2009), Andrews and Cheng (2012), and Ketz (2018), we consider a drifting sequence of true data generating processes (DGPs), which, in turn, admits a sequence of drifting pseudo-true values for the auxiliary parameters. This sequence of pseudo-true auxiliary parameters is then employed to make rigorous the notion of parameters on or near the boundary within I-I estimation. This treatment requires us to revisit the limit theory of Andrews (1999) to accommodate cases where the drifting pseudo-true values cause a "boundary bias" that results in the score of the auxiliary model losing its asymptotic mean-zero property; we refer the reader to Section 2 for precise details and discussion. However, even in the case where boundary bias occurs, our proposed I-I estimator still displays standard Gaussian asymptotics. The intuition behind this result is simple: I-I, by generating simulated data that mimics the observed data, carries out an implicit bias correction that alleviates the impact of the auxiliary parameters being close to the boundary.

---

[1]However, we must acknowledge that there are cases where equality constraints on the auxiliary parameters are indeed necessary to obtain a well-defined inverse of the binding function (see Gospodinov, Komunjer, and Ng (2017) for a recent example).

We apply this new I-I approach to a range of examples that have featured in the I-I literature: a stochastic volatility model with a GARCH(1, 1) auxiliary model (see, e.g., CFS); $\alpha$-stable models with a skewed Student-t auxiliary model (see, e.g., Garcia, Renault, and Veredas (2011)); and continuous-time jump-diffusion models for returns with a Student-t GARCH auxiliary model. In each example, we require that the auxiliary parameters satisfy a vector of inequality constraints, and, in each example, we demonstrate that empirically plausible values of the structural parameters lead to estimates of the auxiliary parameters that are near the boundary of the parameter space. We then demonstrate that our I-I approach can easily be applied to obtain estimators of the structural parameters that have good finite-sample performance.

The remainder of the paper is organized as follows. In Section 2, we discuss constrained auxiliary models and give three classes of empirically relevant examples from the I-I literature where the constraints imposed on the auxiliary model are known to bind, at least in some cases. In addition, we present the particular drifting DGP framework considered in this paper, which generalizes the approach of Andrews (1999) to consider drifting pseudo-true values that can capture boundary affects for any finite-sample size. Within this particular setup, we demonstrate that a well-defined unconstrained auxiliary parameter estimator, which contains the same amount of information as the linear combinations of constrained auxiliary estimates and KT multipliers used in CFS as auxiliary parameters, always exists and can be readily used for the purpose of I-I. Section 3 uses this unconstrained auxiliary estimator to propose novel I-I estimators and demonstrates that this unconstrained estimator is asymptotically equivalent to the constrained I-I approach proposed in CFS. In Section 4, we consider a series of Monte Carlo examples and an empirical application that demonstrates the performance of this approach and makes clear the empirical relevance of our approach. Section 5 concludes and all proofs are relegated to the Appendix.

## 2. Inequality constraints on the auxiliary model

We observe a sample $\{y_T : T \geq 1\}$ generated from a strictly stationary and ergodic probability model $P_\theta$ depending on an unknown parameter $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, with $\Theta$ compact. Conditional on observed data $\mathbf{Y}_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$, the model admits the conditional density $p(y_t|\mathbf{Y}_{t-1}; \theta)$. We are interested in conducting inference on $\theta$ in situations where maximum likelihood estimation based on $p(y_t|\mathbf{Y}_{t-1}; \theta)$ is infeasible or otherwise unattractive, but simulation from $p(y_t|\mathbf{Y}_{t-1}; \theta)$ is relatively simple.

I-I proposes to estimate $\theta$ by targeting consistent parameter estimates of a simpler auxiliary model $f(y_t|\mathbf{Y}_{t-1}; \beta)$, with auxiliary parameters $\beta \in \mathbf{B} \subset \mathbb{R}^{d_\beta}$, with $\mathbf{B}$ compact, and where $d_\beta \geq d_\theta$. Let $Q_T(\beta)$ denote the sample auxiliary objective function associated with $f(y_t|\mathbf{Y}_{t-1}; \beta)$ and the observed sample $\{y_t\}_{t=1}^T$.

We concern ourselves with situations where, to ensure estimates obtained from $Q_T(\beta)$ are well behaved, $\beta$ must be estimated subject to a vector of inequality restrictions:

$$g(\beta) \geq 0.$$

The constraint function $g : \check{\mathbf{B}} \to \mathbb{R}^q$, with $\check{\mathbf{B}} \subset \mathbb{R}^{d_\beta}$ an open set containing $\mathbf{B}$, is known and continuously differentiable on $\check{\mathbf{B}}$. Throughout the remainder the notation, $g(\beta) \geq 0$ is taken to mean $g_j(\beta) \geq 0$, $j = 1, \ldots, q$. From the inequality constraints, we define a restricted, or constrained, parameter space as

$$\mathbf{B}^r := \big\{ \beta \in \mathbf{B} : g(\beta) \geq 0 \big\}.$$

Throughout, we allow the number of constraints ($q$) to be greater than or less than the number of parameters ($d_\beta$).

We assume throughout that $\mathbf{B}^r$ has a nonempty interior, which precludes equality constraints. However, this assumption is immaterial since, up to an abuse of notation, if the problem originally featured a mix of equality constraints, say $\{g_l(\beta) = 0 : 1 \leq l \leq q_1\}$, and inequality constraints, say $\{g_k(\beta) \geq 0 : 1 \leq k \leq q_2\}$, with $q_1 + q_2 = q$, we could always redefine $g(\beta) \geq 0$ to be only the inequality constraints $\{g_k(\beta) \geq 0 : 1 \leq k \leq q_2\}$ that remain active after imposing the equality constraints and eliminating some corresponding components of $\beta$.

Subsequently, we can define the constrained estimator of the auxiliary parameters as

$$\hat{\beta}_T^r := \arg\max_{\beta \in \mathbf{B}} Q_T(\beta) \quad \text{s.t.} \quad g(\beta) \geq 0.$$

$$:= \arg\max_{\beta \in \mathbf{B}^r} Q_T(\beta).$$

Consistent I-I estimation requires that $\hat{\beta}_T^r$ be a consistent estimator of an appropriately defined pseudo-true parameter value, generically denoted by $\beta^0$ and satisfying $g(\beta^0) \geq 0$. As discussed by CFS, if these inequality constraints bind at $\beta^0$, in the sense that for at least one $j$,

$$g_j(\beta^0) = 0, \quad 1 \leq j \leq q,$$

I-I based on $\hat{\beta}_T^r$ may result in estimators with a non-Gaussian limit distribution.

In certain examples, such as those considered in the following section, in order for estimators based on $Q_T(\beta)$ to be well behaved, certain inequalities must be strict, in that $g_j(\beta) > 0$ for some $1 \leq j \leq q$. We will ensure that $\hat{\beta}_T^r$ satisfies this property, at least with probability one for $T$ large enough, by assuming that we have a drifting DGP where, for any finite $T$, the pseudo-true value of the optimization program belongs to $\text{Int}(\mathbf{B}^r)$. In so doing, when some constraints are actually binding, this precisely means that the pseudo-true value is then "near the boundary"; we refer to, for example, Andrews and Cheng (2012) and Ketz (2018) for a similar use of this terminology.

Before going further with the precise mathematical framework, we first give examples where constraints on the auxiliary models used within I-I estimation feature in the empirical literature.

EXAMPLE 1 (Stochastic Volatility). We begin with the classic example of a log-normal stochastic volatility (SV) model and a GARCH(1, 1) auxiliary model with Gaussian or

Student-t innovations.[2] The log-normal stochastic volatility model is defined as follows:

$$y_t = \sqrt{h_t} e_t, \quad t = 1, \ldots, T, \tag{1}$$

$$\ln(h_t) = \alpha + \delta \ln(h_{t-1}) + \sigma_v v_t,$$

where $|\delta| < 1$, $\sigma_v > 0$, $(e_t, v_t)' \sim_{\text{i.i.d.}} N(0, \text{Id}_2)$ and we denote the structural parameters as $\theta = (\alpha, \delta, \sigma_v)'$. We observe a series $\{y_t\}_{t=1}^T$ from the SV model in (1) and our goal is to conduct inference on $\theta$.

We follow CFS and consider as our auxiliary model the GARCH(1, 1) model:

$$y_t = \sqrt{h_t} \epsilon_t, \tag{2}$$

$$h_t = \psi + \varphi y_{t-1}^2 + \pi h_{t-1}.$$

Common specifications for the errors $\epsilon_t$ in (2) are $\epsilon_t \sim_{\text{i.i.d.}} N(0, 1)$ or, for $v(\eta) := [(1/\eta - 1/2)/1/\eta]$, $\epsilon_t \sim_{\text{i.i.d.}} v(\eta)^{1/2} t_{1/\eta}$, where $t_{1/\eta}$ denotes the Student-t distribution with $1/\eta$ degrees of freedom, so that $v(\eta)^{1/2} t_{1/\eta}$ denotes a Student-t with unit variance (which requires that $\eta < 1/2$). We denote the auxiliary parameters as $\beta$, with $\beta = (\psi, \varphi, \pi)'$ if $\epsilon_t \sim_{\text{i.i.d.}} N(0, 1)$ and $\beta = (\psi, \varphi, \pi, \eta)'$ otherwise. The GARCH(1, 1) model is very useful as an auxiliary model as it can capture many of the structural ideas associated with (1), such as thick tails and volatility clustering, while yielding closed form formulas for the score and Hessian based on the pseudo-log-likelihood $Q_T(\beta)$.

However, the GARCH(1, 1) auxiliary model must be estimated subject to inequality constraints to ensure that the pseudo-maximum likelihood estimator of $\beta$ is well behaved. The set of inequality constraints for the auxiliary model can be stated as

$$\psi \geq 0, \qquad \varphi \geq 0, \qquad \pi \geq 0, \qquad \varphi + \pi \leq 1, \tag{3}$$

with the added constraint $0 \leq \eta \leq 1/2$ when $\epsilon_t$ is distributed as Student-t with $1/\eta$ degrees of freedom. Moreover, we stress that the quasi-likelihood is not even well-defined when all the parameters, $(\varphi, \psi, \pi)'$ are simultaneously on the boundary, since the conditional variance must be strictly positive. More generally, Francq and Zakoian (2007) stress that to establish asymptotic normality of the QMLE for the GARCH parameters, a key regularity condition is that "the true parameter must lie in the interior of the parameter space." This statement can obviously be extended to the pseudo-true value of the GARCH parameters. To enforce the above strict inequalities on the auxiliary parameters, CFS require (see their footnote five on p. 960) that the GARCH parameters in their auxiliary model satisfy

$$\varphi \geq 0.025, \qquad \eta \leq 0.499. \tag{4}$$

However, such constraints are effectively arbitrary and do not necessarily reflect the true nature of the constraints in (3).

---

[2]Several authors consider I-I estimation for this model using GARCH($p, q$) models and we refer the reader to Engle and Lee (1996), Monfardini (1998), Pastorello, Renault, and Touzi (2000), and CFS for examples.

EXAMPLE 2 ($\alpha$-Stable Random Variables). The class of $\alpha$-stable distributions is often used to capture random variables that display heavy-tailed features, such as stock returns data. The distribution of a random variable, $y_t$, from the $\alpha$-stable class is characterized by four parameters: $\alpha$—the tail index, which captures the "heavyness" of the tail; $\gamma$—the skewness parameter; $\mu$—the location parameter; and $\sigma$—the scale parameter. Denoting $\theta = (\alpha, \gamma, \mu, \sigma)'$, we have that

$$\Theta := (0, 2) \times (-1, 1) \times \mathbb{R} \times [0, \infty).$$

We note here that values of $\alpha < 2$ ensure that the variance of the random variable is not finite, while if $\alpha \leq 1$ both the mean and the variance are not finite. We refer the reader to Samoradnitsky (2017) for a book length treatment on $\alpha$-stable random variables.

An interesting feature of the $\alpha$-stable class is its lack of a closed-form density function, which makes the application of maximum likelihood methods to estimate $\theta$ difficult (see, e.g., Garcia, Renault, and Veredas (2011) for a discussion). The difficulty of maximum likelihood estimation has led to the development of I-I estimators for $\theta$ that first postulate an auxiliary model with parameters $\beta$ that can roughly match the parameters of the $\alpha$-stable distribution and for which consistent estimators of these parameters can easily be obtained.

Following Garcia, Renault, and Veredas (2011), one such class of auxiliary models is the skewed Student-t (hereafter, skew-t) distribution developed by Fernandez and Steel (1998):

$$f(y; \beta) = \frac{\dfrac{\Gamma\left(\dfrac{\nu}{2} + \dfrac{1}{2}\right)}{\Gamma(\nu/2)}}{\sqrt{\nu}} \frac{1}{\ell\left(\eta + \dfrac{1}{\eta}\right)} \left\{1 + \frac{1}{\nu}\left(\frac{y - \omega}{\ell}\right)^2\left[\frac{1}{\eta^2}\mathbb{1}[y \geq \omega] + \eta^2\mathbb{1}[y < \omega]\right]\right\}^{-\frac{\nu+1}{2}}.$$

The degree of freedom parameter $\nu$ captures tail thickness, $\eta$ captures skewness, and $\omega$ and $\ell$ denote the location and scale parameters. Clearly, the parameters $\theta$ and $\beta$ are closely related. Moreover, the close match between the parameters of the $\alpha$-stable and skew-t distributions should lead to well behaved and nearly efficient I-I estimators.

In many empirical applications, estimates of $\alpha$ are often near 2; recall that a value of $\alpha < 2$ implies that the unconditional variance is not finite. This feature is potentially troubling for I-I estimation since Garcia, Renault, and Veredas (2011) demonstrate that when $\alpha$ is larger than about 1.9, the parameter $\nu$ in the skew-t auxiliary model, and by association the $\alpha$ parameter in the structural model, becomes poorly identified. In particular, Garcia, Renault, and Veredas (2011) argue that "for $\alpha$ close to 2, we may expect that observed data will give the spurious feeling that variance is finite, which would imply a normal distribution corresponding to $\nu = +\infty$ in a Student framework. This is why we will constrain the auxiliary parameter." Since the authors assume that $\alpha < 2$, the authors constrain the skew-t auxiliary model in a similar fashion and impose the inequality constraint $\nu \leq 2$ on the auxiliary parameter. Numerical results presented in Garcia, Renault, and Veredas (2011) demonstrate that a constrained version of I-I, which uses

the inequality constraint $\nu \leq 2$, produces estimators that are better behaved than those where the constraint on the auxiliary parameter $\nu$ is not maintained. A similar I-I estimation strategy has also been employed by Lombardi and Calzolari (2009) for estimation of $\alpha$-stable stochastic volatility models and by Calzolari and Halbleib (2018) for estimation of $\alpha$-stable factor models.

EXAMPLE 3 (Stochastic Volatility Jump-Diffusion (SVJD) Models). The stylized facts of time-varying and autocorrelated volatility, allied with non-Gaussian return distributions, are now extensively documented in the literature on financial returns. However, it is often the case that standard volatility models, such as the one treated in Example 1, cannot completely capture the variability of daily returns in periods of extreme volatility, such as during the 2008–2009 financial crisis. A common approach to address this issue is to consider the inclusion of so-called "jump" processes within existing volatility models. The inclusion of the jump process allows volatility models to exhibit periods of high volatility without significantly altering the interpretation of the model. The literature on modeling returns under the assumption of nonnegligible jumps is now extensive, and we refer the reader to Ait-Sahalia and Jacod (2014) for a textbook treatment.

An important class of widely used continuous-time stochastic volatility models in finance is the mean reverting stochastic volatility jump diffusion (SVJD) model. Let $P_t$ denote the asset price at time $t > 0$, and let $p_t := \ln(P_t)$. In the SVJD model, the evolution of $p_t$ follows a bivariate jump diffusion process, with a representative example of the SVJD model being

$$dp_t = \mu\, dt + \exp(V_t/2)\, dW_t^p + dJ_t,$$
$$dV_t = \kappa(\eta - V_t)\, dt + \sigma_v\, dW_t^v, \tag{5}$$

where $dW_t^v$, $dW_t^p$ are independent standard Brownian motion processes, $dJ_t$ is a jump component with $dJ_t := Z_t\, dN_t$, $Z_t$ denotes the jump size and $dN_t$ is a Poisson process with constant jump intensity. The SVJD model captures two important empirical features of asset prices: one, return volatility exhibits strong serial dependence; two, price jumps exist due to the arrival of unanticipated market news.

The SVJD model is a workhorse of empirical finance, where nonparametric approaches are commonly used to obtain high-frequency measures of variability, such as integrated volatility and quadratic variation (we refer the reader to Andersen, Bollerslev, and Diebold (2010) for a discussion of volatility measures in this model, and its many generalizations). On the other hand, inference on the unknown parameters is hindered by the latent nature of the volatilities, which ensures that estimation techniques based on the likelihood are computationally demanding. Luckily, simulation-based procedures, such as I-I, can bypass the calculation of the likelihood by simulating directly from the process in (5); for empirical applications of the SVJD model using simulation-based inference techniques; see, for example, Eraker (2001), Andersen, Benzoni, and Lund (2002), Creel and Kristensen (2015).

For the purpose of I-I estimation, a useful class of auxiliary models for capturing the behavior in (5) would be the GARCH model class in equation (2). As discussed in Example 1, the use of GARCH auxiliary models requires imposing several inequality restrictions. However, even for this simple version of the SVJD model in equation (5), GARCH

auxiliary models can often yield estimates of the auxiliary parameters for which the inequality constraints in (3) bind. For example, GARCH-based estimators of $\varphi + \pi$ are often very close to unity.[3] Therefore, as argued in CFS, additional information about the auxiliary parameters would be required to successfully identify the structural parameters.

### 2.1 *Parameters near the boundary*

Before presenting our new approach to I-I with constraints, we clarify what is meant by auxiliary parameters on the boundary of the parameter space. Recall the definition

$$\mathbf{B}^r := \big\{ \beta \in \mathbf{B} : g(\beta) \geq 0 \big\}.$$

We say that auxiliary parameters are on the boundary if $\beta^0$, the pseudo-true value of the auxiliary parameters, is on the boundary of $\mathbf{B}^r$.

   We rely on a drifting DGP to capture the behavior of extremum estimators when $\beta^0$ is on the boundary of $\mathbf{B}^r$. In particular, we consider that the DGP of the structural model is indexed by a sequence of drifting "true" values $\{\theta_T\} := \{\theta_T : T \geq 1\}$ that satisfy[4]

$$\{\theta_T\} \in \Theta[\mathbb{N}^+] := \Big\{ \{\theta_T \in \Theta : T \geq 1\} : \lim_{T \to \infty} \theta_T = \theta^0 \in \text{Int}(\Theta) \Big\}.$$

The population objective function for the auxiliary model, calculated under this DGP, is denoted by $\mathcal{Q}(\theta_T, \beta)$. Using $\mathcal{Q}(\theta_T, \beta)$, we define the sequence of pseudo-true auxiliary parameters

$$b(\theta_T) := \arg\max_{\beta \in \mathbf{B}^r} \mathcal{Q}(\theta_T, \beta), \quad \text{where } \{\theta_T\} \in \Theta[\mathbb{N}^+].$$

The notation $\theta_T \mapsto b(\theta_T)$ clarifies that this map depends on the sequence $\{\theta_T\}$. To maintain notational simplicity, when no confusion is likely to result, we denote these pseudo-true auxiliary parameters as $\beta_T^0 := b(\theta_T)$.

   The drifting sequences $\{\theta_T\}$ and $\{\beta_T^0\}$ allow us to capture auxiliary parameters near the boundary of $\mathbf{B}^r$ using the set:

$$\Gamma(\theta^0, \beta^0) := \Big\{ \{\theta_T\} \in \Theta[\mathbb{N}^+] : \beta_T^0 \in \text{Int}(\mathbf{B}^r), \lim_{T \to \infty} \beta_T^0 := \beta^0 \in \mathbf{B}^r \Big\},$$

and by restricting our analysis to DGPs satisfying

$$\{\theta_T\} \in \Gamma(\theta^0, \beta^0). \tag{6}$$

This construction enforces that $\beta_T^0$ belongs to the interior of $\mathbf{B}^r$ but allows $\beta^0$ to lie on the boundary of $\mathbf{B}^r$.[5]

---

[3]We refer the reader to the Monte Carlo section for numerical evidence of this statement as it pertains to the GARCH auxiliary model in equation (2).

[4]Formally, this assumption implies that the observed data, viewed as a triangular array $\{y_{t,T} : T \geq 1,$ $t = 1, \ldots, T\}$, comes from a stationary process depending on $T$.

[5]We note that, when $\beta^0$ is in the interior of the parameter set, the concept of a drifting DGP is hardly useful; one can then assume $b(\theta_T) = \beta^0$ for all $T$ sufficiently large. For instance, in the illustrative stochastic volatility example, a drifting true value for $\psi^0$ is not necessary; when $\varphi^0$ is on the boundary ($\varphi^0 = 0$), $\psi^0$ must be strictly positive and its constrained estimator (the sample mean of $y_t^2$) will automatically fulfill this inequality constraint.

Our need to consider a drifting DGP partly arises because we have in mind cases when $\beta$ must satisfy some strict inequalities in order for $Q_T(\beta)$ to be well behaved in finite-samples, but where the population analogue $\mathcal{Q}(\theta_T, \beta)$ remains well behaved for all $\beta \in \mathbf{B}^r$. The definition of $\Gamma(\theta^0, \beta^0)$ ensures this by requiring that $\beta_T^0 \in \mathrm{Int}(\mathbf{B}^r)$, so that $Q_T(\beta_T^0)$ remains well-defined, while $\lim_T \mathcal{Q}(\theta_T, \beta_T^0)$ is well-defined under standard continuity assumptions. As an illustration of why we require this drifting DGP, consider the stochastic volatility example, and recall that the auxiliary parameter $\varphi$ must be strictly positive to ensure that the pseudo-true value of $\beta$ is identified. To enforce this condition, CFS require the ad hoc condition (4), while we enforce this condition by imposing the high-level condition (6).

It is worth noting that we maintain the assumption that $b(\theta_T)$ always fulfills the constraints, and, by continuity, $\beta^0$ must also fulfill them, while it may violate the strict inequality constraints we implicitly need to maintain in finite samples. However, we may expect that all KT multipliers still converge to zero, in contrast to the setting considered in the asymptotic theory of CFS.

## 2.2 *Standard auxiliary estimators*

I-I with constrained auxiliary parameters relies on the constrained estimator $\hat{\beta}_T^r$ which satisfies

$$Q_T(\hat{\beta}_T^r) \geq \sup_{\beta \in \mathbf{B}^r} Q_T(\beta) + o_P(1/T),$$

which we can obtain through the Lagrangian function

$$\mathcal{L}_T(\beta, \lambda) := Q_T(\beta) + g(\beta)'\lambda$$

for $\lambda \in \mathbb{R}^q$ a vector of KT multipliers. Under differentiability conditions, $\hat{\beta}_T^r$ and the associated KT multipliers $\hat{\lambda}_T$ solve the first-order conditions

$$\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta} + \frac{\partial g'(\hat{\beta}_T^r)}{\partial \beta} \cdot \hat{\lambda}_T = 0, \tag{7}$$

with the slackness conditions

$$\hat{\lambda}_{j,T} \cdot g_j(\hat{\beta}_T^r) = 0 \quad \text{for all } j = 1, \dots, q, \tag{8}$$

$$g(\hat{\beta}_T^r) \geq 0, \qquad \hat{\lambda}_T \geq 0.$$

Building on the theory of constrained estimation discussed in Andrews (1999), under the following sufficient conditions, the estimators $\hat{\beta}_T^r$ and $\hat{\lambda}_T$ are $\sqrt{T}$-consistent estimators of $\beta_T^0$ and 0, respectively.[6]

ASSUMPTION A0. *Under* $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$: (i) $\sup_{\beta \in \mathbf{B}^r} |Q_T(\beta) - \mathcal{Q}(\theta_T, \beta)| = o_P(1)$; (ii) *for all* $\varepsilon > 0$, $\liminf_T \{\mathcal{Q}(\theta_T, \beta_T^0) - \sup_{\beta \in \mathbf{B}^r : \|\beta - \beta_T^0\| > \varepsilon} \mathcal{Q}(\theta_T, \beta)\} > 0$.

---

[6]These assumptions are similar to those given in CFS (see their Assumptions 1 and 3), but are adapted to accommodate our drifting DGP setting.

Assumption A1. *Under $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$:*

(i) $\beta \mapsto Q_T(\beta)$ *has continuous partial derivatives of order two on* $\mathrm{Int}(\mathbf{B}^r)$ *with probability one.*

(ii) *For $\mathcal{J}^0$ a nonstochastic $(d_\beta \times d_\beta)$ positive-definite matrix, and for any $\gamma > 0$:*

$$\sup_{\beta \in \mathbf{B}^r : \|\beta - \beta_T^0\| \le \frac{\gamma}{\sqrt{T}}} \left\| \frac{\partial^2 Q_T(\beta)}{\partial \beta \, \partial \beta'} + \mathcal{J}^0 \right\| = o_P(1).$$

(iii) *For some $\delta^0 \in \mathbb{R}^{d_\beta}$, and for $\mathcal{I}^0$ a nonstochastic $(d_\beta \times d_\beta)$ positive-definite matrix,*

$$\sqrt{T} \partial Q_T(\beta_T^0)/\partial \beta \to_d \aleph(\delta^0, \mathcal{I}^0).$$

(iv) (a) $g(\beta)$ *is continuously differentiable for $\beta \in \check{\mathbf{B}}$, for $\check{\mathbf{B}}$ open and $\mathbf{B} \subseteq \check{\mathbf{B}}$; (b) there exists a $0 \le \tilde{q} \le \min\{q, d_\beta\}$ and a function $\tilde{g} : \check{\mathbf{B}} \to \mathbb{R}^{\tilde{q}}$, a $\tilde{q}$-dimensional subvector of $g(\beta)$, such that, for all $T$ large enough, $\tilde{g}(\beta_T^0)$ contains all the zero entries of $g(\beta_T^0)$ and* $\mathrm{rank}(\partial \tilde{g}(\beta_T^0)'/\partial \beta) = \tilde{q}$.

Assumption A2. *Consider the quadratic expansion*:

$$Q_T(\beta) = Q_T(\beta_T^0) + \frac{\partial Q_T(\beta_T^0)}{\partial \beta'}(\beta - \beta_T^0) + \frac{1}{2}(\beta - \beta_T^0)' \frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta \, \partial \beta'}(\beta - \beta_T^0) + R_T(\beta).$$

*Under $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$: for any sequence $\gamma_T = o(1)$,*

$$\sup_{\beta \in \mathbf{B}^r : \|\beta - \beta_T^0\| \le \gamma_T} \left\{ \frac{|R_T(\beta)|}{\left[1 + \sqrt{T}\|\beta - \beta_T^0\|\right]^2} \right\} = o_P(1/T).$$

The above assumptions are similar to those employed by Andrews (1999) to deduce his Theorem 1. However, Assumption A1(iii) is novel and is maintained to accommodate our drifting DGP. In particular, Assumption A1(iii) allows for the drifting behavior of $\beta_T^0$ to contaminate the limiting distribution of the scaled pseudo-score $\sqrt{T}\partial Q_T(\beta_T^0)/\partial \beta$. In particular, Assumption A1(iii) can capture cases where the proximity of $\beta_T^0$ to the boundary causes a "boundary bias," whereby the pseudo-score looses its asymptotic mean-zero property, typically because $\sqrt{T}(\beta_T^0 - \beta^0)$ is $O(1)$ and not $o(1)$. This framework is often used when one wishes to accurately capture the behavior of estimators when the estimated parameter values are close to, but potentially not on, the boundary of the parameter space; while we have suggested several examples where this phenomena may be in evidence, we refer the interested reader to Ketz (2018) for further examples and discussion. Assumption A1(iv) deals with the behavior of the constraint function $g(\beta)$. Part (a) is standard, while part (b) is required since we allow the number of inequality constraints to be larger than the number of auxiliary parameters (i.e., $q > d_\beta$). Indeed, one particularly important example of this phenomena is the GARCH auxiliary model discussed in Section 2. Intuitively, $\tilde{g}(\beta)$ includes all the constraints that are active at $\beta_T^0$.

Under the above assumptions, the following result holds.

LEMMA 1. *For $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, under Assumptions* A0–A2, $\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) = O_P(1)$ *and* $\sqrt{T}\hat{\lambda}_T = O_P(1)$.

While $\hat{\beta}_T^r$ is $\sqrt{T}$-consistent, it has rightly been stressed in CFS that the constraints $g(\beta) \geq 0$ may produce some singularity (and nonnormality) in the asymptotic distribution of $\hat{\beta}_T^r$ and, therefore, $\hat{\beta}_T^r$ may not be appropriate for I-I estimation. This issue is exacerbated under our setup, since not only may the asymptotic distribution of $\sqrt{T}(\hat{\beta}_T^r - \beta_T^0)$ be nonnormal, but under Assumption A1(iii), the asymptotic distribution of $\sqrt{T}(\hat{\beta}_T^r - \beta_T^0)$ may have a nonzero asymptotic mean. For the former reason, that is, nonnormality, CFS search for a seemingly ad hoc linear combination of the constrained estimator $\hat{\beta}_T^r$ and the vector $\hat{\lambda}_T$ of KT multipliers that is asymptotically normal (see Proposition 2 in CFS, p. 950). To elucidate the implications of this statement, first consider the following (infeasible) quadratic objective function based on the quadratic expansion in Assumption A2:

$$M_T(\beta) = Q_T(\beta_T^0) + \frac{\partial Q_T(\beta_T^0)}{\partial \beta'}(\beta - \beta_T^0) + \frac{1}{2}(\beta - \beta_T^0)'\frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta \, \partial \beta'}(\beta - \beta_T^0).$$

Since $\beta_T^0$ is in the interior of the parameter space, $M_T(\beta)$ is well-defined for any $\beta \in \mathbb{R}^{d_\beta}$ and is uniquely maximized by the solution of the first-order conditions,

$$\ddot{\beta}_T = \beta_T^0 + J_T^{-1}\frac{\partial Q_T(\beta_T^0)}{\partial \beta}, \quad \text{where } J_T = -\frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta \, \partial \beta'}.$$

Note that, with an abuse of language, $\ddot{\beta}_T$ could be dubbed an "unconstrained estimator" since the constraints $g(\beta) \geq 0$ are never taken into account in its definition. The advantage of this "estimator" is that it always exists, since $\beta_T^0$ is an interior point the quadratic approximation always exists, while a general unconstrained estimator may not even exist. However, calling $\ddot{\beta}_T$ an estimator is an abuse of language since it is not feasible to compute $M_T(\beta)$ since the pseudo-true value $\beta_T^0$ is unknown.

Our first key result of this section is to demonstrate that, under our drifting DGP setup, the linear combination of auxiliary parameters put forward in CFS is tightly related to this (potentially) infeasible unconstrained estimator.

PROPOSITION 1. *For $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, under Assumptions* A0–A2:

$$J_T\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) - \frac{\partial g'(\beta_T^0)}{\partial \beta}\sqrt{T}\hat{\lambda}_T = J_T\sqrt{T}(\ddot{\beta}_T - \beta_T^0) + o_P(1). \tag{9}$$

*The remainder term $o_P(1)$ in* (9) *is identically zero when the criterion function $Q_T(\beta)$ is quadratic and the constraints $g(\beta)$ are linear.*

The LHS of equation (9) is identical to the so-called "linear combinations [of the constrained estimator and KT multipliers] that are asymptotically well behaved" in Proposition 2 of CFS (p. 950). By "well behaved" CFS essentially mean asymptotically normal,

whereas separately the constrained estimator and the KT multipliers may not be asymptotically normal when the parameters are close to or on the boundary of $\mathbf{B}^r$.[7] Proposition 1 demonstrates that the linear combinations studied in CFS are well behaved, precisely because they correspond (asymptotically) to the unconstrained extremum estimator; that is, equation (9) demonstrates that by combining the auxiliary parameters and KT multipliers to create the well-behaved linear combinations, we are just back to unconstrained estimation.

### 2.3 *Asymptotically normal feasible unconstrained estimation*

Proposition 1 demonstrates that the linear combinations of $\hat{\beta}_T^r$ and $\hat{\lambda}_T$ that lead to asymptotically normal auxiliary parameters are asymptotically equivalent to the infeasible unconstrained estimator $\ddot{\beta}_T$. Therefore, a feasible version of $\ddot{\beta}_T$ would provide an asymptotically equivalent alternative to the ad hoc combination of constrained estimators and KT multipliers used in CFS. To deduce such an estimator, we first recall that $\ddot{\beta}_T$ is actually the global maximizer of the quadratic objective function $M_T(\beta)$, which depends on the infeasible $\beta_T^0$. This suggests that a feasible unconstrained estimator can be obtained by replacing $\beta_T^0$ in $M_T(\beta)$ by a consistent estimator. By Lemma 1, the constrained estimator $\hat{\beta}_T^r$ is a consistent estimator of $\beta_T^0$, and we can replace $\beta_T^0$ in the quadratic objective function $M_T(\beta)$ by $\hat{\beta}_T^r$, and define the feasible unconstrained estimator

$$\widehat{\beta}_T = \arg \max_{\beta \in \mathbb{R}^{d_\beta}} \left[ Q_T(\hat{\beta}_T^r) + \frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta'}(\beta - \hat{\beta}_T^r) + \frac{1}{2}(\beta - \hat{\beta}_T^r)' \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'}(\beta - \hat{\beta}_T^r) \right].$$

Interestingly, $\widehat{\beta}_T$ is obtained simply by taking a Newton-step away from $\hat{\beta}_T^r$:

$$\widehat{\beta}_T = \hat{\beta}_T^r - \left[ \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta},$$

so that obtaining $\widehat{\beta}_T$ is extremely simple in practice. Throughout the remainder, we refer to $\widehat{\beta}_T$ as the feasible unconstrained (FUNC) estimator of $\beta_T^0$.

Before the FUNC estimator $\widehat{\beta}_T$ can be used for the purpose of I-I, we must understand its asymptotic properties. The asymptotic behavior of $\widehat{\beta}_T$ can be determined by analyzing the asymptotic behavior of the quadratic expansion in Assumption A2. We now give the main result of this section: the FUNC estimator $\widehat{\beta}_T$ is asymptotically equivalent to $\ddot{\beta}_T$, and thus to the well-behaved linear combinations employed in CFS as auxiliary parameters.

THEOREM 1. *For $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, under Assumptions* A0–A2, $\sqrt{T}(\widehat{\beta}_T - \ddot{\beta}_T) = o_P(1)$.

---

[7]When the constraints $g(\cdot)$ are nonlinear, CFS actually consider more complicated linear combinations involving the second derivatives of $g(\beta)$. However, these additional terms will cancel out when working, as we do in this section, under the assumption that the constraints are fulfilled in the population; in this case, the vector of KT multipliers actually converge to zero, and kill the additional terms in CFS.

Ketz (2018) has proven a similar result in the framework of a drifting true value similar to ours. For the sake of being self-contained, we provide our own proof of this result. Before concluding, we note that, by the result of Theorem 1, the FUNC estimator $\widehat{\beta}_T$ allows us to rewrite the decomposition (9) as follows:

$$J_T\sqrt{T}\left(\hat{\beta}_T^r - \beta_T^0\right) - \frac{\partial g'\left(\beta_T^0\right)}{\partial\beta}\sqrt{T}\hat{\lambda}_T = J_T\sqrt{T}\left(\widehat{\beta}_T - \beta_T^0\right) + o_P(1). \tag{10}$$

That is, by working with the computationally friendly FUNC estimator $\widehat{\beta}_T$ we convey exactly the same information as the complicated linear combination of constrained estimators and KT multipliers considered by CFS. The implications of this remark for the purpose of I-I are discussed in the subsequent sections.

## 3. Indirect inference with(out) constraints

The key input of I-I is a set of $H$ simulated paths $\{\tilde{y}_t^{(h)}(\theta)\}_{t=1}^T$, $h = 1, \ldots, H$. From this input, there are several ways to perform I-I. Our focus of interest in this section is to compare four strategies. The first two strategies are based on the score matching approach of GT. The approach of CFS and the approach proposed in this paper will produce two distinct, albeit asymptotically equivalent, variants of the score-matching approach. As already mentioned in the comments of Theorem 1, we differ from CFS in that we will not incorporate, explicitly, the KT multipliers as additional auxiliary parameters for I-I since the FUNC estimator $\widehat{\beta}_T$ carries the same information.

The last two strategies are based on the GMR approach of minimum distance between auxiliary parameters. These two strategies differ regarding the parameters to match: constrained estimators of $\beta$ augmented by KT multipliers, as in CFS, or the user-friendly FUNC estimator proposed in this paper.

By analogy with the trinity of tests, we will dub "Wald approach" the minimum distance approach while the score-matching approach will simply be called "Score approach." Note that CFS dub CMD (Classical Minimum Distance) the Wald approach and GMM (Generalized Method of Moments) the Score approach. GMR have shown that in classical circumstances (I-I without constraints) the two approaches are asymptotically equivalent. This equivalence will be revisited in the present context.

### 3.1 *Score-based indirect inference with(out) constraints*

Given $H$ simulated paths $\{\tilde{y}_t^{(h)}(\theta)\}_{t=1}^T$, $h = 1, \ldots, H$, a simulated version of the auxiliary criterion, denoted by $Q_{TH}(\theta, \beta)$, can then be constructed for use in I-I. To fix ideas, say we have in mind auxiliary parameters $\beta$ defined as M-estimators that maximize the criterion

$$Q_T(\beta) = \frac{1}{T}\sum_{t=1+l}^T q(y_t, y_{t-1}, \ldots, y_{t-l}; \beta).$$

The simulated auxiliary criterion $Q_{TH}(\theta, \beta)$ is then constructed by averaging over the $H$ paths[8]

$$Q_{TH}(\theta, \beta) = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{T} \sum_{t=l+1}^{T} q(\tilde{y}_t^{(h)}(\theta), \tilde{y}_{t-1}^{(h)}(\theta), \dots, y_{t-l}^{(h)}(\theta); \beta). \tag{11}$$

Given $Q_{TH}(\theta, \beta)$, under sufficient smoothness conditions, the gradient (w.r.t. $\beta$) for the simulated version of the quadratic criterion function $M_T(\beta)$ is given by

$$\frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\theta, \beta_T^0)}{\partial \beta\, \partial \beta'}(\beta - \beta_T^0).$$

Replacing the infeasible $\beta_T^0$ by $\hat{\beta}_T^r$, and evaluating this gradient at $\beta = \widehat{\beta}_T$, we can then use the resulting estimating equations

$$\begin{aligned}
\bar{m}_{TH}[\theta; \widehat{\beta}_T] &= \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta\, \partial \beta'}(\widehat{\beta}_T - \hat{\beta}_T^r) \\
&= \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} - \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta\, \partial \beta'}\left[\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta\, \partial \beta'}\right]^{-1} \frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta}
\end{aligned} \tag{12}$$

to carry out a score-based I-I approach. In the absence of constraints for $\theta$, this approach yields the following I-I estimator:

$$\widehat{\theta}_{T,H}^s(W) = \arg\min_{\theta \in \Theta} \bar{m}_{TH}[\theta; \widehat{\beta}_T]' \cdot W \cdot \bar{m}_{TH}[\theta; \widehat{\beta}_T], \tag{13}$$

where $W$ is a positive-definite $(d_\beta \times d_\beta)$ weighting matrix.

In contrast to the I-I estimator in (13), the key idea of the CFS I-I strategy is to incorporate the KT multipliers by considering the modified estimating equations[9]

$$m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T] = \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} + \frac{\partial g'(\hat{\beta}_T^r)}{\partial \beta} \cdot \hat{\lambda}_T.$$

However, by plugging in the vector $\hat{\lambda}_T$ of KT multipliers, as given by the first-order conditions, (7), we obtain

$$m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T] = \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} - \frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta}. \tag{14}$$

Then, for any positive-definite $(d_\beta \times d_\beta)$ weighting matrix $W$, CFS compute their so-called "restricted" score-based I-I estimator as

$$\widehat{\theta}_{T,H}^{\text{CFS}}(W) = \arg\min_{\theta \in \Theta} m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T]' \cdot W \cdot m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T]. \tag{15}$$

---

[8]Note that our use of $Q_{TH}(\cdot)$ is a slight abuse of notation since, in the case of a dynamic model, the probability distribution of $Q_{TH}(\cdot)$ depends separately on $T$ and $H$ and not only on the product $TH$. This abuse of notation is immaterial for first-order asymptotics.

[9]Note that CFS actually define this estimator only for "$H = \infty$."

CFS refer to their I-I estimator $\widehat{\theta}_{T,H}^{\text{CFS}}(W)$ as a restricted estimator, while we dub our I-I estimator $\widehat{\theta}_{T,H}^{s}(W)$ an "unrestricted" estimator since we employ the unrestricted estimator $\widehat{\beta}_T$.[10]

The first key result of this section is to demonstrate that the restricted terminology employed by CFS is potentially misleading: both the restricted equations used by CFS and our unrestricted equations are asymptotically equivalent to the estimating equations that would be used in an *unconstrained* (but infeasible) GT-type score-based I-I approach. That is, the estimating equations used in both the CFS approach and our approach are equivalent to the following unrestricted estimating equations that match simulated data at the unconstrained, but infeasible, estimator $\ddot{\beta}_T$:

$$\frac{\partial Q_{TH}(\theta, \ddot{\beta}_T)}{\partial \beta}, \quad \text{where } \ddot{\beta}_T = \beta_T^0 + J_T^{-1}\partial Q_T(\beta_T^0)/\partial\beta. \tag{16}$$

Demonstrating equivalence between the three sets of estimating equations, (12), (14), and (16), requires the following assumption.

ASSUMPTION A3.   *Under* $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, *the following are satisfied for any fixed* $H \geq 1$:

(i) *For all* $\theta \in \Theta$, $Q_{TH}(\theta, \beta)$ *has continuous partial derivatives (in* $\beta$) *of order two on* $\text{Int}(\mathbf{B}^r)$ *with probability one.*

(ii) *For* $\delta^0$ *and* $\mathcal{I}^0$ *defined in Assumption* A1:

$$\sqrt{T}\partial Q_{TH}(\theta_T, \beta_T^0)/\partial\beta \to_d \aleph(\delta^0, \mathcal{I}^0/H).$$

(iii) *There exists a continuous matrix function* $\theta \mapsto \mathcal{J}(\theta, \beta^0)$ *such that, for all* $\theta \in \Theta$, $\mathcal{J}(\theta, \beta^0)$ *is positive-definite and for any* $\gamma > 0$:

$$\sup_{\theta \in \Theta} \sup_{\beta \in \mathbf{B}^r : \|\beta - \beta_T^0\| \leq \frac{\gamma}{\sqrt{T}}} \left\| \frac{\partial^2 Q_{TH}(\theta, \beta)}{\partial\beta\,\partial\beta'} + \mathcal{J}(\theta, \beta^0) \right\| = o_P(1).$$

While the contents of Assumption A3 are relatively straightforward, we would like to point out that Assumption A3(ii) requires that, if we were to simulate under the true value of the structural parameters, then asymptotically the behavior of $\sqrt{T}\partial Q_{TH}(\theta_T, \beta_T^0)/\partial\beta$ and $\sqrt{T}\partial Q_T(\beta_T^0)/\partial\beta$ must agree. Implicitly, such an assumption requires that the models we are simulating data from be correctly specified, at least asymptotically.

Under Assumptions A0–A3, we have the following result.

PROPOSITION 2.   *For* $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, *under Assumptions* A0–A3, *for any given* $H \geq 1$,

$$\sup_{\theta \in \Theta} \left\| \bar{m}_{TH}[\theta; \widehat{\beta}_T] - \partial Q_{TH}(\theta, \ddot{\beta}_T)/\partial\beta \right\| = o_P(1/\sqrt{T})$$

$$= \sup_{\theta \in \Theta} \left\| m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T] - \partial Q_{TH}(\theta, \ddot{\beta}_T)/\partial\beta \right\|,$$

*and it follows that* $\sup_{\theta \in \Theta} \|m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T] - \bar{m}_{TH}[\theta; \widehat{\beta}_T]\| = o_P(1/\sqrt{T})$.

---

[10]We note here that both $\bar{m}_{TH}[\theta; \widehat{\beta}_T]$ and $m_{TH}^{\text{CFS}}[\theta; \hat{\lambda}_T]$ depend on the constrained estimator $\hat{\beta}_T^r$. While it is an abuse of notation to subsumed this dependence, we believe this avoids notational clutter and allows us to easily differentiate between the two estimating equations.

From Proposition 2, we conclude that both the restricted estimating equations considered by CFS, and the unrestricted estimating equations proposed herein, are equivalent (uniformly over $\Theta$, at first-order) to unrestricted GT-type estimating equations. Therefore, the central message of Proposition 2 is that constraints on the auxiliary model should have no effect on the choice of moments to match within score-based I-I estimation.

To understand the significance of this asymptotic equivalence, recall that if the auxiliary parameters are near the boundary, the traditional approach based on minimizing, over $\Theta$, the estimating equations $\partial Q_{TH}(\theta, \hat{\beta}_T^r)/\partial \beta$ may not deliver an asymptotically Gaussian estimator of $\theta^0$. In contrast, an I-I approach based on the unconstrained (but infeasible) GT-type estimating equations in (16) would deliver an asymptotically Gaussian estimator of $\theta^0$: from a Taylor series expansion, under Assumptions A0–A3, we can conclude that

$$\sqrt{T}\frac{\partial Q_{TH}(\theta_T, \ddot{\beta}_T)}{\partial \beta} = \sqrt{T}\frac{\partial Q_{TH}(\theta_T, \beta_T^0)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\theta_T, \beta_T^0)}{\partial \beta \, \partial \beta'}\sqrt{T}(\ddot{\beta}_T - \beta_T^0) + o_P(1)$$

$$= \sqrt{T}\frac{\partial Q_{TH}(\theta_T, \beta_T^0)}{\partial \beta}$$

$$+ \frac{\partial^2 Q_{TH}(\theta_T, \beta_T^0)}{\partial \beta \, \partial \beta'}\left\{ -\frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta \, \partial \beta'} \right\}^{-1}\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial \beta} + o_P(1)$$

$$= \sqrt{T}\frac{\partial Q_{TH}(\theta_T, \beta_T^0)}{\partial \beta} - \sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial \beta} + o_P(1). \tag{17}$$

By Assumptions A1(iii) and A3(ii), and the independence of the observed and simulated data, the right-hand side term in equation (17) is asymptotically Gaussian with zero mean. Therefore, a direct consequence of Proposition 2 and equation (17) is that our unrestricted estimating equations $\bar{m}_{TH}[\theta; \widehat{\beta}_T]$, when evaluated at $\theta_T$, are also asymptotically Gaussian even when the auxiliary parameters are near the boundary.

Given the asymptotic equivalence derived in Proposition 2, we would expect that our I-I estimator $\widehat{\theta}_{T,H}^s(W)$ will be asymptotically equivalent to the CFS I-I estimator $\widehat{\theta}_{T,H}^{\mathrm{CFS}}(W)$. To detail such an equivalence result, we must maintain the following standard assumption for consistency of extremum estimators.

ASSUMPTION A4. *Under* $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, *for any* $H \geq 1$:

(i) *For any* $\beta \in \mathrm{Int}(\mathbf{B}^r)$, *the function* $\theta \mapsto \partial Q_{TH}(\theta, \beta)/\partial \beta$ *is continuous on* $\Theta$.

(ii) *There exists a vector function* $L(\theta, \beta^0)$ *such that, for any* $\gamma > 0$,

$$\sup_{\theta \in \Theta} \sup_{\|\beta - \beta_T^0\| \leq \frac{\gamma}{\sqrt{T}}} \left\| \frac{\partial Q_{TH}(\theta, \beta)}{\partial \beta} - L(\theta, \beta^0) \right\| = O_P(1/\sqrt{T}).$$

(iii) $L(\theta, \beta^0) = 0 \Longleftrightarrow \theta = \theta^0$.

Assumption A4 is a standard identification assumption, and is implicitly maintained in CFS. However, our explicit treatment of parameters near the boundary forces us to be more cautious. To see that, let us discuss the content of the identification Assumption A4 in the context of the stochastic volatility model example in Section 2. For sake of expositional simplicity, let us consider an auxiliary model based on conditional normality, with $\beta = (\psi, \varphi, \pi)'$. CFS rightly recall that $\pi$ becomes asymptotically underidentified when $\varphi = 0$. CFS circumvent this issue by assuming $\varphi \geq 0.025$. In contrast, we propose in this paper an explicit treatment of parameters near the boundary, which may allow the asymptotic true value $\beta^0 = (\psi^0, \varphi^0, \pi^0)'$ to be such that $\varphi^0 = 0$. The reader can easily check that this specific value does not prevent $\partial Q_{TH}(\theta, \beta^0)/\partial \beta$ from having a well-defined probability limit $L(\theta, \beta^0)$. Consider a trial true value $\theta^0 = (\alpha^0, \delta^0, \sigma_v^0)'$ with $\delta^0 = 0$. Then, $y_t$ is homoskedastic and

$$L(\theta^0, \beta^0) = 0,$$

with $\beta^0 = (\psi^0, \varphi^0, \pi^0)'$ and with

$$\psi^0 = \mathrm{Var}(y_t) = \alpha^0 \quad \text{and} \quad \varphi^0 = \pi^0 = 0.$$

But, if $\theta = (\alpha, \delta, \sigma_v)'$ with $\delta \neq 0$, then, $y_t$ is conditionally heteroskedastic and obviously:

$$L(\theta, \beta^0) \neq 0.$$

From this toy example, we conclude that Assumptions A4 is sensible.[11]

Assumptions A0–A4 allow us to prove the following result.

PROPOSITION 3. *For $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, under Assumptions A0–A4, for any given $H \geq 1$ and any positive-definite matrix $W$, $\mathrm{plim}_{T \to \infty} \widehat{\theta}_{T,H}^s(W) = \mathrm{plim}_{T \to \infty} \widehat{\theta}_{T,H}^{\mathrm{CFS}}(W) = \theta^0$, and $\|\widehat{\theta}_{T,H}^s(W) - \widehat{\theta}_{T,H}^{\mathrm{CFS}}(W)\| = o_P(1/\sqrt{T})$.*

Since our unrestricted I-I estimator $\widehat{\theta}_{T,H}^s(W)$ is asymptotically equivalent to the restricted I-I estimator $\widehat{\theta}_{T,H}^{\mathrm{CFS}}(W)$, we will set the focus on the former. By doing so, we confirm the discussion given earlier and in Section 2: when it comes to the choice of the moments to match, we do not really care about constrained estimation of the auxiliary model.

To prove asymptotic normality of the resulting estimator, a local identification assumption is required to complete the global Assumption A4(iii).

ASSUMPTION A5. *The vector function $\theta \mapsto L(\theta, \beta^0)$ is continuously differentiable on $\mathrm{Int}(\Theta)$ and, for all $T \geq 1$, $\mathrm{rank}(\partial L(\theta_T, \beta^0)/\partial \theta') = d_\theta = \mathrm{rank}(\partial L(\theta^0, \beta^0)/\partial \theta')$.*

---

[11]The reader may wonder how to identify $\sigma_v$ in the homoskedastic case. This actually requires matching the kurtosis since in the general case the unconditional kurtosis is

$$\frac{\mathrm{Var}(h_t)}{[E(h_t)]^2} = \exp\left(\frac{\sigma_v^2}{1 - \delta^2}\right) - 1.$$

This kurtosis matching is implicitly performed when using a Student-t conditional distribution as an auxiliary model.

THEOREM 2. *For $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$, under Assumptions A0–A5, for any given $H \geq 1$ and any positive-definite matrix $W$*

$$\sqrt{T}\big(\widehat{\theta}^s_{T,H}(W) - \theta_T\big) \to_d \aleph\bigg(0, \bigg(1 + \frac{1}{H}\bigg)\Omega_W\bigg),$$

*where, recalling $\mathcal{I}^0 = \lim_{T \to \infty} \mathrm{Var}[\sqrt{T}\partial Q_T(\beta^0_T)/\partial\beta]$,*

$$\Omega_W = A_W^{-1} B_W A_W^{-1},$$

$$A_W = \frac{\partial L(\theta^0, \beta^0)'}{\partial\theta} W \frac{\partial L(\theta^0, \beta^0)}{\partial\theta'}, \qquad B_W = \frac{\partial L(\theta^0, \beta^0)'}{\partial\theta} W \mathcal{I}^0 W \frac{\partial L(\theta^0, \beta^0)}{\partial\theta'}.$$

*The optimal weighting matrix $W$ is given by*

$$W^* = \big[\mathcal{I}^0\big]^{-1},$$

*and leads to an optimal I-I estimator with asymptotic variance[12]*

$$\bigg(1 + \frac{1}{H}\bigg)\Omega^* = \bigg(1 + \frac{1}{H}\bigg)\bigg(\frac{\partial L(\theta^0, \beta^0)'}{\partial\theta}\big[\mathcal{I}^0\big]^{-1}\frac{\partial L(\theta^0, \beta^0)}{\partial\theta'}\bigg)^{-1}.$$

Theorem 2 demonstrates that even though, due to the boundary bias in Assumption A2, the term $\sqrt{T}(\theta_T - \theta^0)$ may not converge to zero, the term $\sqrt{T}(\widehat{\theta}^s_{T,H}(W) - \theta_T)$ still converges to a zero-mean Gaussian random variable. This result comes about from the structure of $\bar{m}_{TH}[\theta; \widehat{\beta}_T]$ and the results of Proposition 2, which imply that

$$\sqrt{T}\bar{m}_{TH}[\theta_T; \widehat{\beta}_T] = \sqrt{T}\frac{\partial Q_{TH}(\theta_T, \beta^0_T)}{\partial\beta} - \sqrt{T}\frac{\partial Q_T(\beta^0_T)}{\partial\beta} + o_P(1).$$

Then Assumptions A1(iii) and A3(ii), together with the independence between the observed and simulated data, ensure that $\{\sqrt{T}\partial Q_{TH}(\theta_T, \beta^0_T)/\partial\beta - \sqrt{T}\partial Q_T(\beta^0_T)/\partial\beta\}$ is asymptotically Gaussian with zero mean. That is, even though $\sqrt{T}\partial Q_T(\beta^0_T)/\partial\beta$ has a nonzero asymptotic mean, because I-I seeks to simulate data so that the observed and simulated scores agree, in the sense that their normed difference is small, this nonzero asymptotic mean is "knocked out" and does not contaminate the asymptotic distribution of $\sqrt{T}(\hat{\theta}^s_{T,H}(W) - \theta_T)$.

It is also important to note that the above formulas are identical to those given in GMR, confirming that we actually perform I-I *without* constraints. To see this, let $Q_T(\theta, \beta)$ denote the simulated auxiliary criterion function calculated using the single simulated path ($H = 1$) $\{\tilde{y}_t(\theta)\}^T_{t=1}$, that is, with reference to equation (11)

$$Q_T(\theta, \beta) := \frac{1}{T}\sum_{t=l+1}^{T} q\big(\tilde{y}^{(h)}_t(\theta), \tilde{y}^{(h)}_{t-1}(\theta), \ldots, y^{(h)}_{t-l}(\theta); \beta\big),$$

---

[12]The reader may notice that the formula for $\Omega^*$ given above differs from that given in Proposition 4 of CFS, denoted as $\mathcal{C}^r_0$ in their equation (8). However, it is simple to verify that the two coincide when the constraints $g(\beta) \geq 0$ are satisfied at $\beta^0_T$ since the KT multipliers will be zero in the limit.

and consider the constrained estimator

$$\tilde{\beta}_T^r(\theta) = \arg\max_{\beta \in \mathbf{B}^r} Q_T(\theta, \beta).$$

For sake of interpretation, let us consider the simplest case without boundary problems. Then the constrained estimator $\tilde{\beta}_T^r(\theta)$ converges towards a (nondrifting) pseudo-true value $b(\theta)$ that is in the interior of the parameter set. Then, while KT multipliers converge to zero, we have

$$\plim_{T \to \infty} \frac{\partial Q_T(\theta, \tilde{\beta}_T^r(\theta))}{\partial \beta} = L(\theta, b(\theta)) = 0, \quad \forall \theta \in \Theta.$$

In particular, by differentiating the above and assuming, following Assumption A1,

$$\frac{\partial L(\theta^0, \beta^0)}{\partial \beta'} = \plim_{T \to \infty} \frac{\partial^2 Q_T(\beta^0)}{\partial \beta \, \partial \beta'} = -\mathcal{J}^0$$

we obtain

$$\frac{\partial L(\theta^0, \beta^0)}{\partial \theta'} + \frac{\partial L(\theta^0, \beta^0)}{\partial \beta'} \frac{\partial b(\theta^0)}{\partial \theta'} = 0$$

and so

$$\frac{\partial L(\theta^0, \beta^0)}{\partial \theta'} = \mathcal{J}^0 \frac{\partial b(\theta^0)}{\partial \theta'}.$$

This relationship between $\partial L(\theta^0, \beta^0)/\partial \theta'$ and $\partial b(\theta)/\partial \theta'$ allows us to rewrite the asymptotic variance of $\sqrt{T}(\widehat{\theta}_{T,H}^s(W) - \theta_T)$ as

$$\Omega_W = A_W^{-1} B_W A_W^{-1},$$

$$A_W = \frac{\partial b(\theta^0)'}{\partial \theta} \mathcal{J}^{0'} W \mathcal{J}^0 \frac{\partial b(\theta^0)}{\partial \theta'}, \qquad B_W = \frac{\partial b(\theta^0)'}{\partial \theta} \mathcal{J}^{0'} W \mathcal{I} W \mathcal{J}^0 \frac{\partial b(\theta^0)}{\partial \theta'}.$$

Therefore,

$$\Omega^* = \left\{ \frac{\partial b'(\theta^0)}{\partial \theta} \mathcal{J}^{0'} [\mathcal{I}^0]^{-1} \mathcal{J}^0 \frac{\partial b(\theta^0)}{\partial \theta'} \right\}^{-1},$$

and we recognize the familiar formula given by GMR (see their Proposition 4) for the asymptotic variance of the optimal I-I estimator.

### 3.2 *Wald-based indirect inference with(out) constraints*

The aforementioned tight connection with the results of GMR suggest that it should be possible to perform I-I *without* constraints in an alternative, albeit asymptotically equivalent, manner using the Wald approach and our well-behaved unconstrained estimator $\widehat{\beta}_T$. The philosophy of the Wald approach to I-I would then amount to compute an unconstrained estimator $\tilde{\beta}_{TH}(\theta)$ on simulated data (for any given value $\theta$ of the structural parameters) and then to minimize, in some norm, $\widehat{\beta}_T - \tilde{\beta}_{TH}(\theta)$. We show in this section that this approach may work, but requires care in the definition of $\tilde{\beta}_{TH}(\theta)$.

3.2.1 *A first solution: The CFS strategy*    The Wald-based I-I strategy of CFS, which uses constrained auxiliary parameter estimates, can be reinterpreted as a minimum distance I-I approach based on a vector of unconstrained auxiliary parameter estimates. To see this, first define

$$\tilde{\beta}^r_{TH}(\theta) = \arg \max_{\beta \in \mathbf{B}^r} Q_{TH}(\theta, \beta),$$

and let $\tilde{\lambda}_{TH}(\theta)$ be the vector of KT multipliers delivered by this constrained optimization. The Wald-based estimator of CFS is then given by

$$\check{\theta}^{\mathrm{CFS}}_{T,H}(W) = \arg \min_{\theta \in \Theta} \begin{bmatrix} \hat{\beta}^r_T - \tilde{\beta}^r_{TH}(\theta) \\ \hat{\lambda}_T - \tilde{\lambda}_{TH}(\theta) \end{bmatrix}' K^{r\prime}_0 \cdot W^\boxplus \cdot K^r_0 \begin{bmatrix} \hat{\beta}^r_T - \tilde{\beta}^r_{TH}(\theta) \\ \hat{\lambda}_T - \tilde{\lambda}_{TH}(\theta) \end{bmatrix}, \tag{18}$$

where, since we are under the assumption that the constraints are fulfilled,

$$W^\boxplus = \begin{bmatrix} W & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}, \qquad K^r_0 = \begin{bmatrix} K^r_{0,1} \\ \mathbf{O} \end{bmatrix}, \qquad K^r_{0,1} = \begin{bmatrix} -\mathcal{J}^0 & \vdots & \dfrac{\partial g'(\beta^0_T)}{\partial \beta} \end{bmatrix}.$$

Recall that we have simplified the exposition by considering only auxiliary parameter estimates $\tilde{\beta}^r_{TH}(\theta)$ defined as above. Alternatively, we could consider $H$ auxiliary parameters based on a single simulated paths of length $T$: for $h = 1, \ldots, H$

$$\tilde{\beta}^{r(h)}_T(\theta) = \arg \max_{\beta \in \mathbf{B}^r} Q_T(\theta, \beta),$$

and then compute[13]

$$\bar{\beta}^r_{T,H}(\theta) = \frac{1}{H} \sum_{h=1}^{H} \tilde{\beta}^{(h)}_T(\theta).$$

Note that the estimator $\bar{\beta}^r_{T,H}(\theta)$ fulfills the constraints if $\mathbf{B}^r$ is a convex set. A sufficient condition for that is to assume that the set $\mathbf{B}$ is convex and the functions $g_j(\cdot)$, $j = 1, \ldots, q$ defining the constraints are concave. However, the satisfaction of this condition is immaterial for the validity of an I-I estimator based on $\bar{\beta}^r_{T,H}(\theta)$.

It must be acknowledged that a more complicated definition for $K^r_{0,1}$ is given in CFS. However, this complication is immaterial in our setting as we work under the assumption that the constraints are fulfilled, and thus the population (resp., estimated) vector of KT multipliers is zero (resp., $O_P(1/\sqrt{T})$). As a matter of fact, the above estimator becomes feasible only when $K^r_{0,1}$ is replaced by a consistent estimator like

$$\hat{K}^r_{0,1,T} = \begin{bmatrix} \dfrac{\partial^2 Q_T(\hat{\beta}^r_T)}{\partial \beta \, \partial \beta'} & \vdots & \dfrac{\partial g'(\hat{\beta}^r_T)}{\partial \beta} \end{bmatrix} = \begin{bmatrix} -\hat{J}_T & \vdots & \dfrac{\partial g'(\hat{\beta}^r_T)}{\partial \beta} \end{bmatrix}.$$

---

[13]Extending the results of GMR, we can conclude that $\tilde{\beta}^r_{TH}(\theta)$ and $\bar{\beta}^r_{T,H}(\theta)$ are asymptotically equivalent and would lead to asymptotically equivalent I-I estimators of $\theta$. However, the results of Gourieroux, Renault, and Touzi (1999) suggest that an I-I estimator based on $\bar{\beta}^r_{T,H}(\theta)$ will have better finite sample properties, at the cost of performing $H$ optimizations in the auxiliary model instead of only just one. This discussion is beyond the scope of this paper.

Hence, for the sake of feasibility, we should rather consider

$$\check{\theta}_{T,H}^{\mathrm{CFS}}(W) = \arg\min_{\theta \in \Theta} \begin{bmatrix} \hat{\beta}_T^r - \tilde{\beta}_{TH}^r(\theta) \\ \hat{\lambda}_T - \tilde{\lambda}_{TH}(\theta) \end{bmatrix}' \hat{K}_{0,1,T}^{r\prime} \cdot W \cdot \hat{K}_{0,1,T}^r \begin{bmatrix} \hat{\beta}_T^r - \tilde{\beta}_{TH}^r(\theta) \\ \hat{\lambda}_T - \tilde{\lambda}_{TH}(\theta) \end{bmatrix}. \tag{19}$$

Since the two estimators (18) and (19) are obviously asymptotically equivalent, we simplify the exposition by denoting them identically, even though only (19) is feasible.

Just as with the score-based approach to I-I, we can now interpret the Wald-based I-I estimator of CFS as I-I *without* constraints. To do so, note that

$$\hat{K}_{0,1,T}^r \begin{bmatrix} \hat{\beta}_T^r - \beta_T^0 \\ \hat{\lambda}_T \end{bmatrix} = \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} (\hat{\beta}_T^r - \beta_T^0) + \frac{\partial g'(\hat{\beta}_T^r)}{\partial \beta} \hat{\lambda}_T$$

$$= -\hat{J}_T(\widehat{\beta}_T - \beta_T^0) + o_P(1/\sqrt{T}),$$

where the second equality follows from equation (10). Therefore, an asymptotically equivalent version of the CFS Wald-based I-I estimator could be computed as

$$\bar{\theta}_{T,H}^{\mathrm{CFS}}(W) = \arg\min_{\theta \in \Theta} (\widehat{\beta}_T - \tilde{\beta}_{TH}^{\mathrm{CFS}}(\theta))' \hat{J}_T W \hat{J}_T (\widehat{\beta}_T - \tilde{\beta}_{TH}^{\mathrm{CFS}}(\theta)),$$

where we define $\tilde{\beta}_{TH}^{\mathrm{CFS}}(\theta)$ as

$$\tilde{\beta}_{TH}^{\mathrm{CFS}}(\theta) = \tilde{\beta}_{TH}^r(\theta) + \left[ \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \frac{\partial g'(\hat{\beta}_T^r)}{\partial \beta} \tilde{\lambda}_{TH}(\theta). \tag{20}$$

Note that the notation $\tilde{\beta}_{TH}^{\mathrm{CFS}}(\theta)$ is justified by analogy with the relationships

$$\widehat{\beta}_T = \hat{\beta}_T^r - \left[ \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta} \tag{21}$$

$$= \hat{\beta}_T^r + \left[ \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \frac{\partial g'(\hat{\beta}_T^r)}{\partial \beta} \hat{\lambda}_T. \tag{22}$$

This reinterpretation of the so-called "restricted" Wald approach to I-I, as dubbed by CFS, is an unconstrained I-I approach based (through equation (19)) on our FUNC estimator. Therefore, we have a similar message to the score-based approach. This is confirmed by Propositions 5 and 6 of CFS, which yield the following insights.

(i) For any choice of the positive definite weighting matrix $W$ (or more generally for any sequence of sample dependent positive-definite weighting matrices $W_T$ with a positive-definite limit), the score-based I-I estimator $\widehat{\theta}_{T,H}^{\mathrm{CFS}}(W)$ and the Wald-based I-I estimator $\check{\theta}_{T,H}^{\mathrm{CFS}}(W)$ are asymptotically equivalent.

(ii) For $T$ sufficiently large, the two estimators are numerically equal in the case of an auxiliary model that just identifies the structural parameters because $d_\beta = d_\theta$.

Point (i) above revisits the results of GMR (see their Section 2.5, p. S91), demonstrating that, for any choice of the weighting matrix $W$, the score-based approach

with weighting matrix $W$ is asymptotically equivalent to the Wald-based approach with weighting matrix $\hat{J}_T W \hat{J}_T$ as in the definition of $\widehat{\theta}_T^{\text{CFS}}(W)$. In the case of a just identified auxiliary model, the choice of the weighting matrix is immaterial and point (ii) calls to mind Proposition 4.1 in Gourieroux and Monfort (1996). Once more, this similarity to the results of GMR and Gourieroux and Monfort (1996) confirms that we are actually performing I-I *without* constraints. In addition, since our unrestricted score-based I-I estimator $\widehat{\theta}_{T,H}^s(W)$ is asymptotically equivalent to the restricted score-based estimator $\widehat{\theta}_{T,H}^{\text{CFS}}(W)$ (see Theorem 2), it is also (by point (i) above) asymptotically equivalent to the alternative aforementioned Wald-based estimators of CFS: $\check{\theta}_{T,H}^{\text{CFS}}(W)$ and $\bar{\theta}_{T,H}^{\text{CFS}}(W)$.

3.2.2 *A second solution: Back to the score*   The previous subsection revisited the Wald-based CFS estimator by resorting to a definition of $\tilde{\beta}_{TH}(\theta)$ that mimics, on simulated data, the alternative definition of the FUNC estimator given in equation (22). We can alternatively use a definition of $\tilde{\beta}_{TH}(\theta)$ that mimics equation (21). To see this, recall that our score-based approach was focused on minimizing, in some norm,

$$
\begin{aligned}
\bar{m}_{TH}[\theta; \widehat{\beta}_T] &= \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} (\widehat{\beta}_T - \hat{\beta}_T^r) \\
&= \left[ \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right] \left\{ \widehat{\beta}_T - \hat{\beta}_T^r + \left[ \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} \right\} \\
&= \left[ \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right] \{ \widehat{\beta}_T - \tilde{\beta}_{T,H}^c(\theta) \},
\end{aligned}
$$

where

$$
\tilde{\beta}_{TH}^c(\theta) = \hat{\beta}_T^r - \left[ \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta}. \tag{23}
$$

Let us acknowledge, however, an important difference of philosophy between the definitions of $\tilde{\beta}_{TH}^{\text{CFS}}(\theta)$ and $\tilde{\beta}_{TH}^c(\theta)$. In the former case, we make a Newton–Raphson improvement of $\tilde{\beta}_{TH}^r(\theta)$, while in the latter case we remain true to $\hat{\beta}_T^r$. In this respect, we obviously set the focus on score matching and, as a consequence, a comparison with our score-based approach is straightforward. More precisely, if we define another Wald-based I-I estimator, the solution of

$$
\widehat{\theta}_{T,H}^c(W) = \arg\min_{\theta \in \Theta} (\widehat{\beta}_T - \tilde{\beta}_{TH}^c(\theta))' \hat{J}_T W \hat{J}_T (\widehat{\beta}_T - \tilde{\beta}_{TH}^c(\theta)),
$$

we see that, from the formulas above, this minimization program can be equivalently written as

$$
\min_{\theta \in \Theta} \bar{m}_{TH}[\theta; \widehat{\beta}_T] \left[ \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \hat{J}_T W \hat{J}_T \left[ \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} \bar{m}_{TH}[\theta; \widehat{\beta}_T], \tag{24}
$$

which is nothing but minimizing a certain norm of $\bar{m}_{TH}[\theta; \widehat{\beta}_T]$ exactly as in equation (13).[14] As observed by a referee, for the above Wald-based I-I estimator, the auxiliary

---

[14]It might be argued that we are not exactly minimizing a norm w.r.t. $\theta$ since the weighting matrix itself depends on $\theta$. However, it must be realized that this is immaterial, both for consistency and asymptotic

parameters are never estimated on simulated data, only on the observed data, which can be convenient if the chosen auxiliary model is computationally challenging in some manner, and signifies that we are basically back to a "score-based" I-I approach.

The asymptotic distribution of $\widehat{\theta}^c_{T,H}(W)$ obviously depends on the limit of the weighting matrix sequence given by

$$\plim_{T\to\infty}\left[\frac{\partial^2 Q_{TH}(\theta_T,\hat{\beta}^r_T)}{\partial\beta\,\partial\beta'}\right]^{-1}\hat{J}_T W \hat{J}_T\left[\frac{\partial^2 Q_{TH}(\theta_T,\hat{\beta}^r_T)}{\partial\beta\,\partial\beta'}\right]^{-1}=W.$$

We can then conclude that this Wald-based I-I estimator $\widehat{\theta}^c_{T,H}(W)$ is asymptotically equivalent to the score-based I-I estimator $\widehat{\theta}^s_{T,H}(W)$ introduced in Section 3.1. In other words, all I-I estimators discussed so far (for the same weighting matrix $W$) are asymptotically equivalent, exactly as in GMR. Furthermore, as in Theorem 2 above, the optimal choice of $W$ is $W^* = [\mathcal{I}^0]^{-1}$.

Interestingly enough, our unconstrained view of I-I results in numerical equivalence between this Wald-based I-I estimator and our score-based I-I estimator when the dimension of the auxiliary and structural parameters are equal.

THEOREM 3. *For $T$ sufficiently large and in the case of a just identified auxiliary model ($d_\beta = d_\theta$), the estimators $\widehat{\theta}^s_{T,H}(W)$ and $\widehat{\theta}^c_{T,H}(W^*)$ are numerically identical irrespective of the choice of weighting matrix (i.e., $W \neq W^*$).*

To conclude this subsection, it is worth comparing, in more detail, the two definitions of $\tilde{\beta}_{TH}(\theta)$ that have delivered Wald-based I-I estimators (by calibration against the FUNC estimator) that are asymptotically equivalent to the score-based approach:

$$\tilde{\beta}^{\text{CFS}}_{TH}(\theta) = \tilde{\beta}^r_{TH}(\theta) + \left[\frac{\partial^2 Q_T(\hat{\beta}^r_T)}{\partial\beta\,\partial\beta'}\right]^{-1}\frac{\partial g'(\hat{\beta}^r_T)}{\partial\beta}\tilde{\lambda}_{TH}(\theta),$$

$$\tilde{\beta}^c_{TH}(\theta) = \hat{\beta}^r_T - \left[\frac{\partial^2 Q_{TH}(\theta,\hat{\beta}^r_T)}{\partial\beta\,\partial\beta'}\right]^{-1}\frac{\partial Q_{TH}(\theta,\hat{\beta}^r_T)}{\partial\beta}.$$

Since $\tilde{\beta}^{\text{CFS}}_{TH}(\theta)$ is based on constrained estimation on the simulated path, through the computation of $\tilde{\beta}^r_{TH}(\theta)$ and $\tilde{\lambda}_{TH}(\theta)$, one may wish to revisit $\tilde{\beta}^c_{TH}(\theta)$ by also using constrained estimators on the simulated path, that is by instead computing

$$\tilde{\beta}^{\text{func}}_{TH}(\theta) = \tilde{\beta}^r_{TH}(\theta) - \left[\frac{\partial^2 Q_{TH}(\theta,\tilde{\beta}^r_{TH}(\theta))}{\partial\beta\,\partial\beta'}\right]^{-1}\frac{\partial Q_{TH}(\theta,\tilde{\beta}^r_{TH}(\theta))}{\partial\beta}$$

$$= \tilde{\beta}^r_{TH}(\theta) + \left[\frac{\partial^2 Q_{TH}(\theta,\tilde{\beta}^r_{TH}(\theta))}{\partial\beta\,\partial\beta'}\right]^{-1}\frac{\partial g'(\tilde{\beta}^r_{TH}(\theta))}{\partial\beta}\tilde{\lambda}_{TH}(\theta).$$

$\tilde{\beta}^{\text{func}}_{TH}(\theta)$ is the FUNC estimator computed on the simulated path, and it seems sensible to match it against the FUNC estimator $\widehat{\beta}_T$ computed on the observed data.

---

distribution, to replace the occurrence of $\theta$ in the weighting matrix by a first-step consistent estimator. This argument is quite similar to the one of equivalence between continuously updated GMM (Hansen, Heaton, and Yaron (1996)) and efficient two-step GMM.

However, this approach will not deliver a consistent estimator of $\theta^0$ in general. To see this, note that $\tilde{\beta}_{TH}^{\text{CFS}}(\theta)$ and $\tilde{\beta}_{TH}^{\text{func}}(\theta)$ both set the focus on the same linear combination of $\tilde{\beta}_{TH}^r(\theta)$ and $\tilde{\lambda}_{TH}(\theta)$. However, while $\tilde{\beta}_{TH}^{\text{CFS}}(\theta)$ is guaranteed to end up with a consistent estimator for the coefficients of this linear combination, the coefficients in $\tilde{\beta}_{TH}^{\text{func}}(\theta)$ themselves depend on the unknown $\theta$. As a consequence, setting the focus on $[\widehat{\beta}_T - \tilde{\beta}_{TH}^{\text{func}}(\theta)]$ alone, or some norm thereof, can induce an additional perverse solution in the limit; that is, the limiting estimating equations, because of their nonlinear dependence on $\theta$, can admit an additional solution $\bar{\theta}$ with $\bar{\theta} \neq \theta^0$. As such, an I-I strategy based on $\tilde{\beta}_{TH}^{\text{func}}(\theta)$ above may not identify $\theta^0$.[15]

## 4. Illustrative examples

### 4.1 *Stochastic volatility*

In this section, we apply our score-based I-I approach to estimate the parameters of the stochastic volatility (SV) model:

$$y_t = \sqrt{h_t} e_t, \tag{25}$$

$$\ln(h_t) = \alpha + \delta \ln(h_{t-1}) + \sigma_v v_t, \tag{26}$$

where $0 < \delta < 1$, $\sigma_v > 0$, $(e_t, v_t)' \sim_{\text{i.i.d.}} N(0, \text{Id}_2)$ and $\theta = (\alpha, \delta, \sigma_v)'$. We observe a series $\{y_t\}_{t=1}^T$ from the SV model in (25)–(26) and our goal is to conduct inference on $\theta$.

Following the discussion in Section 2, we consider the GARCH(1, 1) auxiliary model

$$y_t = \sqrt{h_t} \epsilon_t, \tag{27}$$

$$h_t = \psi + \varphi y_{t-1}^2 + \pi h_{t-1}^2,$$

where the errors $\epsilon_t$ in (27) are $\epsilon_t \sim_{\text{i.i.d.}} N(0, 1)$. The auxiliary parameters are denoted by $\beta$, with $\beta = (\psi, \varphi, \pi)'$. As mentioned in Section 2, to ensure the GARCH(1, 1) auxiliary model is well-behaved CFS require the following inequality constraint:

$$\varphi \geq 0.025.$$

Unlike the approach of CFS, by considering drifting sequences of auxiliary parameters, we allow the constrained estimator to fully reach the boundary of constrained space, in the limit. That is, instead, we assume the true auxiliary parameters satisfy the inequality

$$\varphi_T^0 \geq o(1).$$

4.1.1 *Monte Carlo design* To assess the performance of our proposed I-I estimation strategy, we follow the Monte Carlo design of Jacquier, Polson, and Rossi (1994) (JPR, hereafter), also used in CFS. In particular, we consider two sets of structural parameters: $\theta^{0,1} = (-0.736, 0.90, 0.363)'$ and $\theta^{0,2} = (-0.147, 0.98, 0.0614)'$. These particular values for

---

[15]Frazier and Renault (2017) gave additional examples of settings where such perverse roots can arise in nonlinear econometric models.

$\theta^0$ are related to the unconditional coefficient of variation $\kappa$ for the unobserved level of volatility $h_t$, where

$$\kappa^2 = \frac{\text{Var}(h_t)}{(E[h_t])^2} = \exp\left(\frac{\sigma_v^2}{1 - \delta}\right) - 1.$$

In the first design, we have $\kappa^2 = 1$, which roughly represents lower-frequency returns (say, weekly or monthly returns); for the second design, we set $\kappa^2 = 0.1$, which roughly corresponds to higher-frequency returns (say, daily returns).

As noted in CFS, the choice of the Gaussian auxiliary model, in conjunction with the constraints, means that the GARCH(1, 1) model is not well equipped to handle the thick-tailed behavior exhibited by series generated from the log-normal SV model. Intuitively, this means that the constraints on the auxiliary parameters are likely to be binding since this auxiliary model is a crude approximation of the structural model. However, it is not certain if the inadequacy of the Gaussian GARCH(1, 1) auxiliary model in this case, which was originally noted in Kim, Shephard, and Chib (1998), is due to the model itself, the bindings constraints or a mixture of both issues. In this way, the FUNC based auxiliary estimator may be able to mitigate these issues since it captures, in some sense, the impact of the constraints.

The score based I-I objective function does not require a weighting matrix as we are in the just identified setting; that is, we choose $W = I$. For computational simplicity, we fix the number of data replications to be $H = 10$ across all Monte Carlo designs.[16] We illustrate the performance of our proposed I-I estimator across three different sample sizes, $T = 500, 1000, 2000$, and consider 1000 Monte Carlo replications for each sample size/parameter specification, leading to six separate specifications in total.

### 4.1.2 *Monte Carlo results*

*Simulation design one:* $\theta^{0,1} = (-0.736, 0.90, 0.363)'$   To understand the difference between the constrained and unconstrained auxiliary estimators, Table 1 contains the frequency of binding constraints for the GARCH(1, 1) auxiliary parameter estimates when we allow the boundary of the constrained space to drift, which replicates the behavior of a drifting DGP. Recalling that, in their assessment, CFS employ the constraint $\varphi \geq 0.025$. We choose a drifting boundary of $\varphi \geq \bar{\varphi}_T$ so that the drifting pseudo-true value $\beta_T^0$ lies in the interior of the constrained space, but $\beta^0$ can lie on the boundary. To ensure that our approach is comparable to the approach of CFS, we set $\bar{\varphi}_T := T^{-0.5}$, which ensure that, at least for $T = 2000$ the two constraints are comparable.

For each replication, we calculate the auxiliary estimator $\hat{\beta}_T^r$ subject to the constraints in (3), where we require the constraint $\varphi \geq \bar{\varphi}_T$, and calculate $\widehat{\beta}_T$ by taking a Newton-step from $\hat{\beta}_T^r$. While no constraints are used in the calculation of $\widehat{\beta}_T$, it is informative to ascertain the number of times this estimator would have caused the con-

---

[16]Optimization is carried out using an iterative Gauss–Seidel grid search approach. Starting values were obtained by first running a crude grid search over $\Theta$ and choosing the corresponding grid values that minimized the I-I objective function. Only one iteration of the minimization procedure was carried out and more efficient estimates could be obtained by considering multiple iterations.

TABLE 1. Binding constraints for auxiliary estimators $\hat{\beta}_T^r$ and $\widehat{\beta}_T$ in design one: $\theta^{0,1} = (-0.736, 0.90, 0.363)'$. All terms are in percentages. For $\widehat{\beta}_T$, the values represent the percentage where the FUNC estimator would have caused the constraint to bind or be violated.

| | $T = 500$ | | $T = 1000$ | | $T = 2000$ | |
|---|---|---|---|---|---|---|
| | $\hat{\beta}_T^r$ | $\widehat{\beta}_T$ | $\hat{\beta}_T^r$ | $\widehat{\beta}_T$ | $\hat{\beta}_T^r$ | $\widehat{\beta}_T$ |
| $\psi \geq 0$ | 0.00% | 1.20% | 0.00% | 0.20% | 0.00% | 1.50% |
| $\varphi \geq T^{-0.5}$ | 11.10% | 1.20% | 4.50% | 0.20% | 0.60% | 0.20% |
| $\pi \geq 0$ | 3.80% | 0.80% | 0.40% | 1.90% | 0.70% | 1.10% |
| $\varphi + \pi \leq 1$ | 1.10% | 1.90% | 0.30% | 0.50% | 0.00% | 0.20% |

straints to bind or be violated, as this will tell us, to some extent, what using the unconstrained $\widehat{\beta}_T$ buys us, at least in comparison with $\hat{\beta}_T^r$.

Table 1 demonstrates that under the first design, the constraints for the auxiliary model are binding in a nonnegligible portion on the replications. Interestingly, the FUNC estimator violates the constraint $\varphi \geq \bar{\varphi}_T$ less frequently than the constrained estimator. This is important given the results of Francq and Zakoian (2009), which demonstrate that the constraint $\varphi > 0$ is misleading, suggesting to conclude "that the GARCH$(1, 1)$ model is sufficient for financial data," while additional ARCH lags would be relevant. As clearly explained by Francq and Zakoian (2009), this misleading conclusion is due to the fact that "as a result of the positive constraints, it is possible that the fitted GARCH$(1, 1)$ models" deliver a zero constrained estimator for the second ARCH lag while the score at this value is strongly positive. By construction, the FUNC estimator makes a Newton correction to take this positive score into account. Note that, in contrast, the FUNC estimator violates the stationarity constraint $\varphi + \pi < 1$ more frequently than this constraint binds for the constrained estimator. It is worth realizing that this constraint may be irrelevant since, when the GARCH$(1, 1)$ model is misspecified, the fact that the pseudo-true value violates the stationarity constraint does not imply that the process is nonstationary. This finding actually confirms a point made by Kim, Shephard, and Chib (1998): when the data generating process is stochastic volatility (with true volatility persistence of $\delta^0 = 0.90$), estimating the pseudo-true value for a GARCH$(1, 1)$ model can often deliver estimates of volatility persistence with $\varphi + \pi > 0.90$.

Summary statistics for the resulting score-based I-I parameter estimates are collected in Table 2. The results show that this I-I approach behaves well in finite samples, regardless of the constraints for the auxiliary model. To further understand the finite-sample properties of these estimators, in panels A and B of Figure 1 we plot the sampling distributions of the $\delta$ and $\sigma_v$ estimators across the three different sample sizes.[17] The results in Figure 1 are similar to those reported on p. 963 in CFS.

*Simulation design two:* $\theta^{0,2} = (-0.141, 0.98, 0.0614)'$   Analyzing the frequency of binding constraints for the second Monte Carlo design, we find a very similar story to the

---

[17]To ensure that all plots adequately represent the various sampling distributions and neatly fit in the same figure, we have thrown out 1.5% of the lower tail observations for each series.

TABLE 2. Summary statistics for I-I estimates based on the proposed score approach in design one: $\theta^{0,1} = (-0.736, 0.90, 0.363)'$. STD—Monte Carlo standard deviation of the replications. RMSE—root mean squared error of the replications. M. Bias—mean bias of the replications.

| | $T = 500$ | | | $T = 1000$ | | | $T = 2000$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | STD | RMSE | M. Bias | STD | RMSE | M. Bias | STD | RMSE | M. Bias |
| $\alpha$ | 0.2849 | 0.2854 | 0.0178 | 0.1996 | 0.1996 | −0.0052 | 0.1439 | 0.1439 | −0.0011 |
| $\delta$ | 0.1299 | 0.1397 | −0.0503 | 0.0857 | 0.0897 | −0.0266 | 0.0381 | 0.0392 | −0.0092 |
| $\sigma_v$ | 0.0945 | 0.0961 | −0.0178 | 0.0447 | 0.0477 | −0.0003 | 0.0333 | 0.0336 | 0.0047 |

first Monte Carlo design. Under this design, there are an even larger number of replications where the constraint $\varphi \geq \bar{\varphi}_T$ binds for the constrained estimator and is violated for the FUNC estimator. As discussed by CFS, a small unconditional coefficient of variation for volatility creates a more challenging estimation problem, which seems to have had an impact on the frequency of binding constraints in this GARCH(1, 1) auxiliary model. The aforementioned work of Francq and Zakoian (2009) may suggest that a GARCH(1, $q$), $q > 1$, would have provided more informative parameters to match for I-I. Again, the fact that the FUNC estimator appears to be much less impacted by this
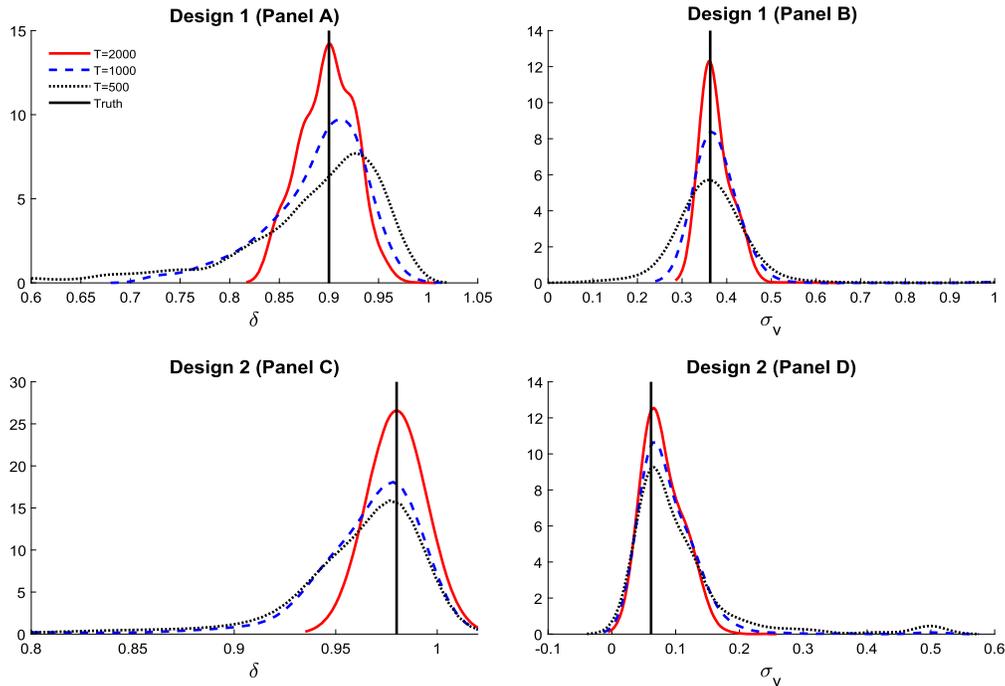


FIGURE 1. Sampling distribution for the I-I estimator of $\delta$ and $\sigma_v$ under Monte Carlo designs one, $\theta^{0,1} = (-0.736, 0.90, 0.363)'$, and two, $\theta^{0,2} = (-0.141, 0.98, 0.0614)'$. The thick line corresponds to a sample size of $T = 2000$, the thick dashed line corresponds to $T = 1000$, and the thin dashed line corresponds to $T = 500$. The vertical line represents the true value in the simulation.

TABLE 3. Binding constraints for auxiliary estimators $\hat{\beta}_T^r$ and $\widehat{\beta}_T$ in design two: $\theta^{0,2} = (-0.141, 0.98, 0.0614)'$. All terms are in percentages. For $\widehat{\beta}_T$, the values represent the percentage where the FUNC estimator would have caused the constraint to bind or be violated.

|  | T = 500 | | T = 1000 | | T = 2000 | |
|---|---|---|---|---|---|---|
|  | $\hat{\beta}_T^r$ | $\widehat{\beta}_T$ | $\hat{\beta}_T^r$ | $\widehat{\beta}_T$ | $\hat{\beta}_T^r$ | $\widehat{\beta}_T$ |
| $\psi \geq 0$ | 0.00% | 6.10% | 0.00% | 2.40% | 0.00% | 0.30% |
| $\varphi \geq T^{-0.5}$ | 29.50% | 6.10% | 20.70% | 2.40% | 10.90% | 0.30% |
| $\pi \geq 0$ | 0.00% | 6.40% | 0.00% | 2.50% | 0.00% | 0.50% |
| $\varphi + \pi \leq 1$ | 0.30% | 6.50% | 0.00% | 2.60% | 0.10% | 0.30% |

constraint (five fewer times for a sample size of 500, and ten less times for larger sample sizes) is good news regarding its ability to capture the relevant information in the data. In addition, we find that there are a relatively large number of replications where the FUNC estimator does not satisfy the constraint $\varphi + \pi \leq 1$: in about seven percent of the samples the FUNC estimator violates this constraint (at the sample size of $T = 500$). This is not surprising since, as explained above, as discussed in Kim, Shephard, and Chib (1998), we would expect the estimator of $\varphi + \pi$ to be larger than 0.98 (the true value of the volatility persistence, $\delta$, under this design).

Summary statistics for the resulting score-based I-I parameter estimates are collected in Table 4, with the results reflecting the same conclusions as those obtained in the first Monte Carlo design. The sampling distributions of the $\delta$ and $\sigma_v$ estimators in the second Monte Carlo design are contained in panels C and D of Figure 1. Again, the figures demonstrate that this approach works well.

### 4.2 $\alpha$-Stable model

In this section, we apply our I-I approach to data generated i.i.d. from the $\alpha$-stable distribution. We recall that the $\alpha$-stable distribution is characterized by four parameters: $\alpha$—the tail index; $\gamma$—the skewness parameter; $\mu$—the location; and $\sigma$—the scale. To

TABLE 4. Summary statistics for I-I estimates based on the proposed score approach in design two: $\theta^{0,2} = (-0.141, 0.98, 0.0614)'$. STD—Monte Carlo standard deviation of the replications. RMSE— root mean squared error of the replications. M. Bias—mean bias of the replications.

|  | T = 500 | | | T = 1000 | | | T = 2000 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | STD | RMSE | M. Bias | STD | RMSE | M. Bias | STD | RMSE | M. Bias |
| $\alpha$ | 0.5576 | 0.5576 | 0.0037 | 0.3857 | 0.3860 | −0.0171 | 0.2822 | 0.2822 | 0.0018 |
| $\delta$ | 0.0592 | 0.0657 | −0.0287 | 0.0303 | 0.0321 | −0.0107 | 0.0096 | 0.0102 | −0.0034 |
| $\sigma_v$ | 0.0804 | 0.0805 | 0.0054 | 0.0358 | 0.0359 | 0.0030 | 0.0097 | 0.0097 | 0.0008 |

simplify the analysis, we assume that the location parameter $\mu$ is known and fix its value at $\mu = 0$.[18]

The $\alpha$-stable distribution has no closed-form density representation, and thus maximum likelihood estimation of the unknown parameters is difficult, which has led several authors to apply I-I to estimate the unknown parameters. Following the discussion in Section 2, we take as our auxiliary model the skewed Student-t (hereafter, skew-t) distribution of Fernandez and Steel (1998):

$$f(y; \beta) = \frac{\Gamma\left(\dfrac{\nu}{2} + \dfrac{1}{2}\right)}{\dfrac{\Gamma(\nu/2)}{\sqrt{\nu}}} \frac{1}{\ell\left(\eta + \dfrac{1}{\eta}\right)} \left\{ 1 + \frac{1}{\nu}\left(\frac{y - \omega}{\ell}\right)^2 \left[\frac{1}{\eta^2} \mathbb{1}[y \geq \omega] + \eta^2 \mathbb{1}[y < \omega]\right] \right\}^{-\frac{\nu+1}{2}},$$

where $\nu$ captures tail thickness, $\eta$ captures skewness, and the location and scale parameters are $\omega$ and $\ell$, respectively.

4.2.1 *Monte Carlo design*   We consider a Monte Carlo design that is similar to Garcia, Renault, and Veredas (2011). We fix $(\mu, \sigma, \gamma)'$ at $(0, 0.5, 0)'$ and consider two different true values for $\alpha^0$: (1) $\alpha^0 = 1.90$ and (2) $\alpha^0 = 1.95$. We consider two sample sizes, $T = 500$ and $T = 1000$, and generate 1000 Monte Carlo replications for each sample size.

As noted in Section 2, when $\alpha$ is close to, or larger than, 1.9, the parameter $\nu$ in the skew-t auxiliary model can become poorly identified in small samples and can result in I-I estimators with poor behavior.[19] Simulation results in Garcia, Renault, and Veredas (2011) demonstrate that, for data generated from the $\alpha$-stable distribution with $\alpha < 2$ but close to 2, the unconstrained PMLE for $\nu$ is non-Gaussian and numerically unstable in small samples. To circumvent this issue, the authors impose $\nu \leq 2$ within the skew-t auxiliary model, and then use as auxiliary statistics for I-I inference the auxiliary PMLE for $\beta$ and the corresponding KT multiplier $\lambda$ associated with the inequality constraint on $\nu$. We refer the interested reader to Section 2 for further discussion on the need for this constraint.

Instead of imposing an arbitrary constraint on the auxiliary model, which may limit its identifying power, we propose to use the FUNC estimator. In this context, the FUNC estimator does not display the same numerical instability observed in the unconstrained estimator for $\nu$: the calculation of the FUNC estimator is based on the numerically stable constrained estimator. Therefore, we argue that the FUNC estimator represents a practically useful medium between the fully unconstrained, and numerically unstable, auxiliary estimator and the standard constrained auxiliary estimator.

4.2.2 *Monte Carlo results*   Under both Monte Carlo designs, the constraint on the auxiliary estimator ($\hat{\beta}_T^r$) for $\nu$ binds across all the replications, which implies that if one were to use the constrained I-I approach suggested by CFS, identification of $\alpha^0$ is completely determined by the KT multiplier associated with the constraint. Likewise, across all replications and both Monte Carlo designs, the FUNC estimator ($\widehat{\beta}_T$) would have

---

[18]This is not overly restrictive as reliable estimators of the location parameter exist that can easily be employed before the analysis.

[19]A value of $\alpha < 2$ implies that the unconditional variance of the random variable is not finite.

caused the constraints to bind or be violated for both Monte Carlo designs estimator. Therefore, we can conclude that the constrained version of this auxiliary model is unable to identify $\alpha^0$ by itself, and instead must rely on additional auxiliary statistics that are not associated with the auxiliary model; that is, the KT multiplier in this case.

In contrast, the FUNC estimator only uses the information contained in the (unconstrained) auxiliary model. Since the FUNC estimator is unconstrained, it is not surprising that this estimated auxiliary parameter always violates the constraint $\nu \leq 2$. Given the above results for the constrained and FUNC-based auxiliary estimators, it is clear that the pseudo-true value defined by this constrained optimization program is on the boundary of the parameter space. As such, an alternative strategy would be to consider a skew-t auxiliary model that imposes the equality constraint $\nu = 2$.

Before presenting the Monte Carlo results, we remark that if the pseudo-true value of the auxiliary parameters were to violate the constraint $\nu \leq 2$, the auxiliary score (evaluated at this pseudo-true value) would not be zero in the limit, and, as such, this case would be outside the scope of our theoretical analysis. While such scenarios could be accommodated by extending our theoretical framework, at the cost of additional notation and technical arguments, this extension is not germane to the main message of the paper. Therefore, for the sake of brevity, we do not consider such cases further.

Table 5 reports the standard deviation (STD), root-mean-squared error (RMSE) and mean bias (M. Bias) associated with the parameter estimates from our unconstrained I-I estimation approach; namely, $\alpha$, $\gamma$, $\sigma$. The results demonstrate that our approach produces estimates with reliable finite-sample properties across both sample sizes.

In addition, Figure 2 contains kernel density estimates of the standardized parameter estimates across the Monte Carlo replications for the case of $\alpha^0 = 1.95$, the corresponding results for the design where $\alpha^0 = 1.90$ are very similar and are not reported for the sake of brevity. The results demonstrate that the standardized estimators have a roughly Gaussian shape, even though they are computed from $\alpha$-stable random vari-

TABLE 5. Summary statistics for Monte Carlo estimators across both sample sizes for design one ($\alpha^0 = 1.90$) and two ($\alpha^0 = 1.95$), denoted as D1 and D2 in the table. STD—Monte Carlo standard deviation of the replications. RMSE—root mean squared error of the replications. M. Bias—mean bias of the replications. Across both designs $\gamma^0 = 0$ and $\sigma^0 = 0.5$. Under both designs, the constrain $\nu \leq 2$ was binding in all Monte Carlo replications for both $\hat{\beta}_T^r$ and $\widehat{\beta}_T$.

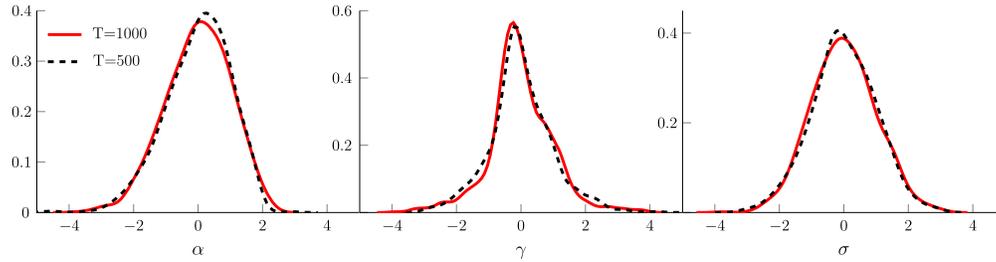| | $T = 500$ | | | $T = 1000$ | | |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\gamma$ | $\sigma$ | $\alpha$ | $\gamma$ | $\sigma$ |
| D1 | | | | | | |
| STD | 0.0817 | 0.0003 | 0.0295 | 0.0612 | 0.0002 | 0.0210 |
| RMSE | 0.0819 | 0.0003 | 0.0295 | 0.0611 | 0.0003 | 0.0210 |
| M. Bias | −0.0060 | 0.0002 | −0.0016 | −0.0016 | 0.0001 | 0.0007 |
| D2 | | | | | | |
| STD | 0.0637 | 0.0002 | 0.0272 | 0.0471 | 0.0002 | 0.0195 |
| RMSE | 0.0646 | 0.0003 | 0.0272 | 0.0474 | 0.0003 | 0.0195 |
| M. Bias | −0.0108 | 0.0002 | −0.0007 | −0.0058 | 0.0001 | 0.0005 |

FIGURE 2. Sampling distribution for the standardized I-I estimator of $\theta$ in the $\alpha$-stable Monte Carlo experiments: $\theta^0 = (\alpha^0, \gamma^0, \sigma^0)' = (1.95, 0.0, 0.5)'$. The results under the Monte Carlo experiment with $\theta^0 = (1.90, 0.0, 0.5)'$ are similar and, therefore, not presented for brevity.

ables. We refer the reader to Garcia, Renault, and Veredas (2011) for theoretical justification of this phenomena.

### 4.3 *Stochastic Volatility Jump Diffusion (SVJD) model*

Motivated by the now well-established empirical findings of time-varying volatility and the existence of jumps in returns data, we explore here a continuous-time specification for financial returns. We consider that returns evolve in continuous-time according to a mean reverting stochastic volatility model, which follows an Ornstein–Uhlenbeck process, and where returns themselves exhibit random jumps. We first demonstrate, via Monte Carlo results, that in this empirically relevant model, the class of GARCH auxiliary models will often deliver estimated auxiliary parameters that are near the boundary of the parameter space. We then demonstrate that our I-I approach delivers reliable estimators of the corresponding structural parameters even though the original auxiliary parameters are near the boundary of the parameter space. Lastly, we use our I-I approach to conduct inference on the parameters of the SVJD model to determine whether or not there exists significant evidence of jumps in daily S&P500 returns.

For $P_t$ denoting the asset price at time $t > 0$, and $p_t := \ln(P_t)$, assume that $p_t$ evolves according to the bivariate diffusion process

$$dp_t = \mu \, dt + \exp(V_t/2) \, dW_t^p + dJ_t^p,$$

$$dV_t = \kappa(\eta - V_t) \, dt + \sigma_v \, dW_t^v,$$

$$dJ_t^p = Z_t \, dN_t, \quad Z_t \sim N\big(\mu_j, \sigma_j^2\big), \Pr[dN_t = 1] = \bar{\lambda}_j \, dt + o(t), \tag{28}$$

where $dW_t^v$ and $dW_t^p$ are correlated Brownian motion processes, with correlation $\rho$, and $dN_t$ is a Poisson process with intensity $\bar{\lambda}_j$, and $Z_t \sim N(\mu_j, \sigma_j^2)$. We collect the unknown parameters into $\theta = (\mu, \kappa, \eta, \sigma_v^2, \bar{\lambda}_j, \mu_j, \sigma_j^2, \rho)'$ and consider inference on $\theta$ using I-I.

Following the analysis in Example 1, we take as our auxiliary model the GARCH model with Student-t errors: for $r_t := p_t - p_{t-1}$ denoting log-returns,

$$r_t = \mu_a + \sqrt{h_t}\epsilon_t,$$

$$h_t = \psi + \varphi(r_{t-1} - \mu_a)^2 + \pi h_{t-1},$$

where, for $v(\eta) := [(1/\eta - 1/2)/1/\eta]$, $\epsilon_t \sim_{\text{i.i.d.}} v(\eta)^{1/2}t_{1/\eta}$, and $v(\eta)^{1/2}t_{1/\eta}$ denotes a Student-t with unit variance. The auxiliary GARCH model is unable to identify the jump parameters in (28), and so we supplement the auxiliary GARCH model with additional summary statistics based on both bipower variation and realized jump variation: for $r_{t,i}$ denoting the $i$th, out of $M$, equally-spaced intraday returns observed on day $t$, bipower variation is defined as

$$\text{BV}_t := \frac{\pi}{2}\left(\frac{M}{M-1}\right)\sum_{i=2}^{M}|r_{t,i}||r_{t,(i-1)}|,$$

and jump-variation is defined as

$$\text{JV}_t := \max\{\text{RV}_t - \text{BV}_t, 0\}, \quad \text{RV}_t = \sum_{i=1}^{M}r_{t,i}^2,$$

where $\text{RV}_t$ denotes realized volatility. For I-I estimation, we then consider the additional summary statistics: for $\overline{\text{JV}} := \frac{1}{T}\sum_{t=1}^{T}\text{JV}_t$,

$$S_1 := \frac{1}{T}\sum_{t=1}^{T}\text{sign}(r_t)\sqrt{\text{JV}_t}, \qquad S_2 := \frac{1}{T}\sum_{t=1}^{T}(\text{JV}_t - \overline{\text{JV}}_t)^2,$$

$$S_3 := \frac{1}{T}\sum_{t=2}^{T}(\text{JV}_t - \overline{\text{JV}}_t)(\text{JV}_{t-1} - \overline{\text{JV}}_t)$$

which correspond to the mean, variance and covariance of the realized jump variation. We note here that Frazier, Maneesoonthorn, Martin, and McCabe (2019) have used these statistics to help identify the jump-process parameters in a discrete-time version of this SVJD model for daily S&P500 returns data.

Estimating GARCH models on daily data can often lead to estimated values of $\varphi$ and $\pi$ such that the constraint $\varphi + \pi \leq 1$ is very close to binding. To ensure numerical stability of the optimization procedure, in practice this constraint is often implemented as $\varphi + \pi \leq 1 - c$, for $c > 0$ and small.

With this point in mind, and similar to the $\alpha$-stable example, there is every reason to suspect that a lack of variation in the estimated constrained auxiliary parameters of $\varphi$ and $\pi$, due to the need to satisfy the constraint $\varphi + \pi \leq 1$, may cause identification issues for the I-I estimator of the structural parameters; that is, if there are many parameter combinations for $\theta$ that yield estimated constrained auxiliary parameters close to the boundary of the constraints, then those simulated data sets may be difficult for the I-I objective function to distinguish. Therefore, we argue that an appropriate approach to the use of GARCH models in such settings is to employ our FUNC-based II approach.

4.3.1 *Monte Carlo design*   The observed data is generated from the model in equation (28) using an Euler discretization scheme with step size $1/\delta$. Log-prices are then gener-

ated according to the recursive scheme

$$p_{t,(i+1)/\delta} = p_{t,i/\delta} + \mu\frac{1}{\delta} + \exp(V_{t,i/\delta}/2)\epsilon^p_{t,i}\frac{1}{\delta} + Z_{t,i}\Delta N_{t,i/\delta},$$

$$V_{t,(i+1)/\delta} = V_{t,i/\delta} + \kappa(\eta - V_{t,i/\delta})\frac{1}{\delta} + \frac{\sigma_v}{\delta}\left(\rho\epsilon^p_{t,i} + \sqrt{1-\rho^2}\epsilon^v_{t,i}\right),$$

where $(\epsilon^p_{t,i}, \epsilon^v_{t,i})'$ is bivariate standard normal, $Z_{t,i} \sim_{\text{i.i.d.}} N(\mu_j, \sigma_j^2)$ and $\Delta N_{t,i/\delta}$ is drawn from a Poisson distribution with intensity $\bar{\lambda}_j/\delta$.

   We simulate data from the above Euler approximation at the (approximate) 1 minute frequency, $\delta = 400$, and record daily returns, and 10 minute intraday returns. The remaining simulations are discarded. We retain $T = 500$ trading days for the Monte Carlo, which amounts to approximately 2 years of daily returns. For the intraday returns, we use 10-minute returns to calculate our measures of realized volatility $\text{RV}_t^{10} := \sum_{t=1}^M r_{t,i}^2$, $M = 40$, and our corresponding realized jump measure $\text{JV}_t$. For the data simulation, we consider an initialization period of 400 periods, or one trading day. Following the design of Creel and Kristensen (2015), we set the unknown parameters to $\theta^0 = (0.00, 0.02, 0.25, 0.20, 0.10, 0.00, 0.50, -0.10)'$.

   For I-I estimation, we also consider a Euler discretization scheme at the one-minute sampling frequency, although finer-sampling schemes may lead to estimators with better properties. Since this scheme already requires simulating a large number of data points, to ease the computational burden of the I-I estimator, we limit the analysis to consider only a single simulated path, that is, $H = 1$. This will induce some efficiency loss in the resulting estimators, however, that is the price to pay for computational convenience.

4.3.2 *Monte Carlo results*   First, we analyze the frequency at which the estimated auxiliary parameters cause the constraint $\varphi + \pi \leq 1$ to bind. From Table 6, we see that in 26% of the simulations the constraint was binding for the constrained auxiliary estimators, while the FUNC estimator either led to a value of the estimated parameters that was greater than or equal to one. As explained in Section 4.1, a true daily volatility persistence $\exp(-\kappa) \approx 0.98$ leads us to expect an estimator of $\varphi + \pi$ even larger than 0.98.

TABLE 6. Summary statistics for Monte Carlo estimators for the SVJD example. The sample size is $T = 500$. STD—Monte Carlo standard deviation of the replications. RMSE—root mean squared error of the replications. M. Bias—mean bias of the replications. For this Monte Carlo example, the constraint on $\varphi + \pi \leq 1$ was binding in 26% of the Monte Carlo replications, for both the constrained estimator $\hat{\beta}^r_T$ and the FUNC estimator $\widehat{\beta}_T$.

| $\theta$ | $\mu$ | $\kappa$ | $\eta$ | $\sigma_v$ | $\bar{\lambda}_j$ | $\mu_j$ | $\sigma_j^2$ | $\rho$ |
|---|---|---|---|---|---|---|---|---|
| $\theta^0$ | 0.0000 | 0.0200 | 0.2500 | 0.2000 | 0.1000 | 0.0000 | 0.5000 | $-0.1000$ |
| STD | 0.0010 | 0.0068 | 0.1005 | 0.1193 | 0.0195 | 0.0010 | 0.1099 | 0.0204 |
| RMSE | 0.0010 | 0.0068 | 0.1006 | 0.1193 | 0.0197 | 0.0010 | 0.1098 | 0.0204 |
| M.Bias | 0.0000 | $-0.0002$ | 0.0062 | $-0.0048$ | 0.0027 | 0.0000 | $-0.0005$ | $-0.0014$ |

Moreover, this behavior is amplified in the presence of jumps, and leads to an uncon-strained estimations of $\varphi + \pi$ that if frequently larger than unity.

While this behavior of the FUNC estimator may seem counterintuitive, recall that the condition $\varphi + \pi < 1$ is required for stationarity in the GARCH model, under the explicit assumption that the true DGP is GARCH. Therefore, in this example the constraint $\varphi + \pi < 1$ is meaningless since the true DGP is not GARCH. Indeed, the process in this simulation is stationary even though the estimated values can satisfy $\varphi + \pi \geq 1$ in any finite-sample. This example clearly demonstrates that if we simply used the constrained GARCH auxiliary estimators, we could artificially limit the identifying power of this auxiliary model.

Using our FUNC-based I-I approach, we estimate the structural parameters of the SVJD model and report the results in Table 6. The corresponding estimators display low mean bias and reasonable values for the RMSE, especially given that we have used only a single simulated path for I-I. Similar to the previous two examples, the results of this section demonstrate that our approach is able to achieve identification without the need to resort to an I-I approach that utilizes the KT multipliers associated with the constraints for the auxiliary model.

4.3.3 *Empirical illustration: S&P500 data*   To further illustrate our approach, we apply our method to the SVJD model with leverage effects based on demeaned S&P500 returns observed at the daily frequency, between 3 January 2017 and 3 January 2019, which consists of 501 daily observations. High-frequency intraday returns are used to build realized volatility and bi-power variation estimators at the 5-minute sampling frequency. We source the data from the Oxford-Man Institutes "realised library," which contains daily returns on several important financial indices, and nonparametric volatility measures (Gerd, Lunde, Shephard, and Sheppard (2009)).

The auxiliary model is again taken to be the GARCH model with standardized Student-t errors (and unit variance): for $r_t := p_t - p_{t-1}$ denoting log-returns on the S&P500 index, with $t = 1, \ldots, 501$ denoting the daily frequency,

$$r_t = \mu_a + \sqrt{h_t}\epsilon_t,$$

$$h_t = \psi + \varphi(r_{t-1} - \mu_a)^2 + \pi h_{t-1}.$$

Again, for $\overline{\text{JV}} := \frac{1}{T}\sum_{t=1}^{T} \text{JV}_t$, we augment this model with the summary statistics

$$S_1 := \frac{1}{T}\sum_{t=1}^{T}\text{sign}(r_t)\sqrt{\text{JV}_t}, \qquad S_2 := \frac{1}{T}\sum_{t=1}^{T}(\text{JV}_t - \overline{\text{JV}})^2,$$

$$S_3 := \frac{1}{T}\sum_{t=2}^{T}(\text{JV}_t - \overline{\text{JV}})(\text{JV}_{t-1} - \overline{\text{JV}})$$

that capture the jump components of the SVJD model. While this simple SVJD model has now been generalized in several directions, for example, with the inclusion of auto-correlated jumps (Fulop, Li, and Yu (2014), Ait-Sahalia, Cacho-Diaz, and Laeven (2015),

and Maneesoonthorn, Forbes, and Martin (2017)), this simpler SVJD model is still empirically relevant since, if the corresponding jumps components are not statistically significant, it is highly unlikely that these more complicated modeling approaches are necessary.

Maximum likelihood-based inference on the full set of static parameters in this SVJD model,

$$\theta = \left( \kappa, \eta, \sigma_v^2, \bar{\lambda}_j, \mu_j, \sigma_j^2, \rho \right)',$$

is challenging, due to the existence of the latent volatilities.[20] In contrast, I-I is straightforward due to our ability to cheaply simulate data from this model. Therefore, we consider estimation of the model in (28) using our I-I approach base on the GARCH auxiliary model, and where the statistics $(S_1, S_2, S_3)'$ yield auxiliary moments that enable us to identify the jump components. Similar to the Monte Carlo example, the I-I procedure uses a simulation frequency of 1 minute.

The estimated values of $\theta$ obtained using this setup and the daily S&P500 data are given in Table 7, along with the corresponding standard errors. The standard errors are calculated using a block bootstrap approach, with 999 bootstrap replications and with a block length of twenty-five observations. Given the relatively short length of the time series, we believe these bootstrap standard errors are likely more reliable than those obtained from the asymptotic formula presented in Section 3.

Before analyzing the results in the table, we first note that the estimated auxiliary parameters for the GARCH model are such that $\varphi + \pi \approx 1$ for both the constrained and FUNC-based auxiliary estimators. Recall that, even though in this example the FUNC and constrained auxiliary estimators are similar, the FUNC estimator is guaranteed to be asymptotically normal, whereas the constrained estimator will in general not be asymptotically normal. Therefore, we contend that, even though the two estimators are similar in this small scale example, the use of the FUNC estimator for I-I is a safer choice than the constrained estimator.

Analyzing the results for $\theta$ in Table 7, we see that the majority of the coefficients are statistically significant and have the correct signs and magnitudes, with the majority of the results being similar to those obtained elsewhere; see, for example, Creel and Kristensen (2015). In particular, the results suggest that the jump process has a significant jump frequency, but that the resulting jump sizes are small, negative, and have large

TABLE 7. Estimators (Est) and standard errors (STD) for the SVJD S&P500 exercise. For this data set, the constraint on $\varphi + \pi \leq 1$ was binding for both the constrained estimator $\hat{\beta}_T^r$ and the FUNC estimator $\widehat{\beta}_T$.

| $\theta$ | $\kappa$ | $\eta$ | $\sigma_v$ | $\bar{\lambda}_j$ | $\mu_j$ | $\sigma_j^2$ | $\rho$ |
|---|---|---|---|---|---|---|---|
| Est | 0.2876 | 0.5945 | 0.1166 | 0.1278 | $-0.0034$ | 1.2650 | $-0.6131$ |
| STD | 0.0836 | 0.0539 | 0.0366 | 0.0383 | 0.0203 | 0.2902 | 0.1717 |

---

[20]Since the data is demeaned before hand, we do not estimate $\mu$ in this example.

variability. In addition, these results also suggest that the resulting jump components $(\mu_j, \sigma_j^2)$ are difficult to accurately measure.

## 5. Conclusion

The overall message of this paper can be summarized as follows: application of the I-I methodology may require the imposition of certain constraints on the auxiliary parameters, however, one must bear in mind that the behavior of I-I estimators for the structural parameters can be adversely affected by the constraints placed on the auxiliary parameters. In place of these constrained auxiliary parameters, our proposed strategy is to use, for the purpose of I-I, a FUNC (Feasible UNConstrained) estimator of the auxiliary parameters, which in spite of being unconstrained, is always well-defined.

This FUNC estimator leads to simple score and Wald-based I-I approaches, which have been shown to be asymptotically equivalent, at least to first order, with the approach based on constrained auxiliary parameters proposed by CFS. Several Monte Carlo studies demonstrate the good finite-sample properties of this approach, and document that our I-I estimator can deliver robust estimators of the corresponding structural parameters, even in cases where the pseudo-true value of the auxiliary parameters is on the boundary of the parameter space.

## Appendix A: Proofs of main results

Proof of Lemma 1. We first prove that $\|\hat{\beta}_T^r - \beta_T^0\| = o_P(1)$. The argument follows the standard approach. Under $\{\theta_T\} \in \Gamma(\theta^0, \beta^0)$,

$$
\begin{aligned}
0 &\leq \mathcal{Q}(\theta_T, \beta_T^0) - \mathcal{Q}(\theta_T, \hat{\beta}_T^r) \\
&= \mathcal{Q}(\theta_T, \beta_T^0) - \mathcal{Q}(\theta_T, \hat{\beta}_T^r) + Q_T(\hat{\beta}_T^r) - Q_T(\hat{\beta}_T^r) + Q_T(\beta_T^0) - Q_T(\beta_T^0) \\
&\leq 2 \sup_{\beta \in \mathbf{B}} |Q_T(\beta) - \mathcal{Q}(\theta_T, \beta)| + o_P(1),
\end{aligned}
$$

where the $o_P(1)$ follows from the fact that $Q_T(\hat{\beta}_T^r) \geq Q_T(\beta_T^0) + o_P(1)$. From the uniform convergence in Assumption A0(i), we can conclude

$$
0 \leq \mathcal{Q}(\theta_T, \beta_T^0) - \mathcal{Q}(\theta_T, \hat{\beta}_T^r) \leq o_P(1).
$$

The result then follows from the identification condition in Assumption A0(ii).

We next show $\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) = O_P(1)$. The quadratic expansion

$$
Q_T(\beta) = Q_T(\beta_T^0) + \frac{\partial Q_T(\beta_T^0)}{\partial \beta'}(\beta - \beta_T^0) + \frac{1}{2}(\beta - \beta_T^0)'\frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta \, \partial \beta'}(\beta - \beta_T^0) + R_T(\beta),
$$

can be rewritten as

$$
Q_T(\hat{\beta}_T^r) = Q_T(\beta_T^0) + \frac{1}{T}\kappa_T' J_T^{-1/2} \sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial \beta} - \frac{1}{2T}\|\kappa_T\|^2 + R_T(\hat{\beta}_T^r), \tag{29}
$$

where

$$J_T = -\frac{\partial^2 Q_T(\beta_T^0)}{\partial\beta\,\partial\beta'}, \qquad \kappa_T = J_T^{1/2}\sqrt{T}(\hat\beta_T^r - \beta_T^0).$$

From the definition of $\hat\beta_T^r$, applying the quadratic expansion in (29)

$$o_P(1) \le T\cdot[Q_T(\hat\beta_T^r) - Q_T(\beta_T^0)]$$

$$= \kappa_T' J_T^{-1/2}\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta} - \frac{1}{2}\|\kappa_T\|^2 + T\cdot R_T(\hat\beta_T^r).$$

However, by Assumption A2, since $\|\hat\beta_T^r - \beta_T^0\| = o_P(1)$, we have that

$$|T\cdot R_T(\hat\beta_T^r)| \le (1 + \|\sqrt{T}(\hat\beta_T^r - \beta_T^0)\|)^2 o_P(1).$$

Applying the above, we have

$$\kappa_T' J_T^{-1/2}\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta} - \frac{1}{2}\|\kappa_T\|^2 + T\cdot R_T(\hat\beta_T^r)$$

$$= O_P(\|\kappa_T\|) - \frac{1}{2}\|\kappa_T\|^2 + (1 + \|J_T^{-1/2}\kappa_T\|)^2 o_P(1)$$

$$= O_P(\|\kappa_T\|) - \frac{1}{2}\|\kappa_T\|^2 + o_P(\|\kappa_T\|) + o_P(\|\kappa_T\|^2) + o_P(1),$$

so that we may rewrite the above as

$$\|\kappa_T\|^2 \le 2\|\kappa_T\|[o_P(1) + O_P(1)] + o_P(1).$$

Defining $\zeta_T := [o_P(1) + O_P(1)] \equiv O_P(1)$ and $x_T := \|\kappa_T\|$, we end up with the inequality

$$x_T^2 - 2x_T\zeta_T + o_P(1) \le 0,$$

which is satisfied for $x_T$ in the interval $\zeta_T \pm \sqrt{\zeta_T^2 + o_P(1)}$; that is, for $x_T$ in the interval $[0, 2\zeta_T + o_P(1)]$. Hence, $\zeta_T = O_P(1)$ implies that

$$x_T = \|\kappa_T\| = \|J_T^{1/2}\sqrt{T}(\hat\beta_T^r - \beta_T^0)\| = O_P(1)$$

which implies that

$$\sqrt{T}(\hat\beta_T^r - \beta_T^0) = O_P(1).$$

Now, we prove $\sqrt{T}\hat\lambda_T = O_P(1)$. First, consider the case where $q > d_\beta$. By assumption, at most there are $\tilde q$ dimensions of $g(\beta_T^0)$ that are precisely zero, which are all contained in the vector $\tilde g(\beta_T^0)$. As shown above, $\sqrt{T}(\hat\beta_T^r - \beta_T^0) = O_p(1)$, and we can then be sure that asymptotically, with probability one, all zero entries of $g(\hat\beta_T^r)$ are also included in $\tilde g(\hat\beta_T^r)$. Define $\tilde\lambda_T$ to be the $\tilde q$-dimensional subvector of $\hat\lambda_T$ that corresponds to the entries of $g$ that are in $\tilde g$. By the slackness conditions of the Kuhn–Tucker optimization problem, asymptotically, with probability one, since $\tilde g(\hat\beta_T^r)$ contains all the zero entries of $g(\hat\beta_T^r)$, $\tilde\lambda_T$ contains all the possible nonzero entries of $\hat\lambda_T$.

With these definitions, the Kuhn–Tucker first-order conditions can be stated as

$$\sqrt{T}\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta'} + \frac{\partial \tilde{g}'(\hat{\beta}_T^r)}{\partial \beta}\sqrt{T}\tilde{\lambda}_T = 0.$$

For some intermediate value $\bar{\beta}_T$, a first-order expansions gives

$$\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial \beta'} + \frac{\partial^2 Q_T(\bar{\beta}_T)}{\partial \beta\,\partial \beta'}\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) + \frac{\partial \tilde{g}'(\beta_T^0)}{\partial \beta}\sqrt{T}\tilde{\lambda}_T + o_P(\sqrt{T}\tilde{\lambda}_T) = 0,$$

where the $o_P(\sqrt{T}\tilde{\lambda}_T)$ term follows by the first part of Lemma 1 and the continuity of $\partial g(\beta)/\partial \beta'$ in Assumption A1(iv). Since $\partial \tilde{g}'(\beta_T^0)/\partial \beta$ is full column-rank, we have, for some intermediate value $\bar{\beta}_T$,

$$\sqrt{T}\tilde{\lambda}_T = -\left[\frac{\partial \tilde{g}(\beta_T^0)}{\partial \beta'}\frac{\partial \tilde{g}'(\beta_T^0)}{\partial \beta}\right]^{-1}\frac{\partial \tilde{g}(\beta_T^0)}{\partial \beta'}\sqrt{T}\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta} + o_P(\sqrt{T}\tilde{\lambda}_T)$$

$$= -\left[\frac{\partial \tilde{g}(\beta_T^0)}{\partial \beta'}\frac{\partial \tilde{g}'(\beta_T^0)}{\partial \beta}\right]^{-1}\frac{\partial \tilde{g}(\beta_T^0)}{\partial \beta'}\left[\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial \beta} + \frac{\partial^2 Q_T(\bar{\beta}_T)}{\partial \beta\,\partial \beta'}\sqrt{T}(\hat{\beta}_T^r - \beta_T^0)\right]$$

$$\quad + o_P(\sqrt{T}\tilde{\lambda}_T)$$

$$= O_P(1) + o_P(\sqrt{T}\tilde{\lambda}_T),$$

where the last line follows from Assumption A1(iii) and $\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) = O_P(1)$.

In the case where $q \leq d_\beta$, the above arguments remain valid if we take $\tilde{q} = q$, $\tilde{g}(\beta) = g(\beta)$, $\tilde{\lambda}_T = \hat{\lambda}_T$, and note that $\partial \tilde{g}'(\beta_T^0)/\partial \beta$ has full column-rank $q$. □

PROOF OF PROPOSITION 1. A first-order expansions of the first-order conditions (7) give

$$\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial \beta'} + \frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta\,\partial \beta'}\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) + \frac{\partial g'(\beta_T^0)}{\partial \beta}\sqrt{T}\hat{\lambda}_T = o_P(1). \tag{30}$$

Recalling the definition of the infeasible unconstrained estimator $\ddot{\beta}_T$, we can rewrite the LHS of the above equation as follows (while the RHS is exactly zero when the function $Q_T$ is quadratic and the constraints are linear):

$$J_T\sqrt{T}(\ddot{\beta}_T - \beta_T^0) - J_T\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) + \frac{\partial g_T'(\beta_T^0)}{\partial \beta}\sqrt{T}\hat{\lambda}_T = o_P(1).$$

By Lemma 1, and Assumption A1, all three terms of the LHS of the above equality are all $O_P(1)$. We deduce that

$$J_T\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) - \frac{\partial g'(\beta_T^0)}{\partial \beta}\sqrt{T}\hat{\lambda}_T = J_T\sqrt{T}(\ddot{\beta}_T - \beta_T^0) + o_P(1). \tag{31}$$

Moreover, as already noted above, the remainder term $o_P(1)$ in (31) is zero when the criterion function $Q_T$ is quadratic and the constraints are linear. □

Proof of Theorem 1.  By definition,

$$\sqrt{T}\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial\beta} + \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\sqrt{T}(\widehat{\beta}_T - \hat{\beta}_T^r) = 0.$$

Therefore, by a Taylor expansion of the first term around the true value $\beta_T^0$:

$$\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta} + \frac{\partial^2 Q_T(\beta_T^0)}{\partial\beta\,\partial\beta'}\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) + \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\sqrt{T}(\widehat{\beta}_T - \hat{\beta}_T^r) = o_P(1)$$

and then, since $\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) = O_P(1)$, we can obviously simplify the above decomposition to obtain

$$\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta} = -\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\sqrt{T}(\widehat{\beta}_T - \beta_T^0) + o_P(1).$$

Since by Assumption A1, we know that

$$\plim_{T\to\infty}\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'} = -\mathcal{J}^0$$

we can conclude that $\sqrt{T}(\widehat{\beta}_T - \beta_T^0) = O_P(1)$ and

$$\sqrt{T}(\widehat{\beta}_T - \beta_T^0) = [\mathcal{J}^0]^{-1}\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta} + o_P(1).$$

By comparison with the definition of $\ddot{\beta}_T$,

$$\sqrt{T}(\ddot{\beta}_T - \beta_T^0) = [J_T]^{-1}\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta}$$

we have the announced equivalence between estimators.                                    $\square$

Proof of Proposition 2.  For $\beta_T^*$ a component-by-component intermediate value between $\beta_T^0$ and $\hat{\beta}_T^r$, by Assumption A1(ii), we deduce that

$$\left\|\frac{\partial^2 Q_T(\beta_T^*)}{\partial\beta\,\partial\beta'}\left[\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\right]^{-1} - \mathrm{Id}_{d_\beta}\right\| = o_P(1),$$

where $\mathrm{Id}_{d_\beta}$ denotes the $d_\beta \times d_\beta$ identity matrix. Moreover, this bound does not depend on $\theta$. We apply the fact that this bound remains uniformly valid on $\Theta$ when quantities are multiplied by continuous functions of $\theta$ over compact $\Theta$.

For all $\theta \in \Theta$,

$$\sqrt{T}\bar{m}_{TH}[\theta; \widehat{\beta}_T] = \sqrt{T}\frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial\beta} - \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\left[\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\right]^{-1}\sqrt{T}\frac{\partial Q_T(\beta_T^0)}{\partial\beta}$$

$$+ \left\{\frac{\partial^2 Q_{TH}(\theta, \beta_T(\theta))}{\partial\beta\,\partial\beta'} - \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\right\}\sqrt{T}(\hat{\beta}_T^r - \beta_T^0) + o_P(1),$$

where $\beta_T(\theta)$ is a component-by-component intermediate value between $\beta_T^0$ and $\hat{\beta}_T^r$, and where the bound $o_P(1)$ does not depend on $\theta$. By Assumption A3(iii),

$$\sup_{\theta \in \Theta} \left\| \frac{\partial^2 Q_{TH}(\theta, \beta_T(\theta))}{\partial \beta \, \partial \beta'} - \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right\| = o_P(1),$$

so that, by Assumption A3(iii) and Assumption A1(ii), we have

$$\sup_{\theta \in \Theta} \left\| \frac{\partial^2 Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \left[ \frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial \beta \, \partial \beta'} \right]^{-1} - \mathrm{Id}_{d_\beta} \right\| = o_P(1).$$

Therefore, we can conclude that

$$\sup_{\theta \in \Theta} \left\| \sqrt{T} \bar{m}_{TH}[\theta; \widehat{\beta}_T] - \left\{ \sqrt{T} \frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} - \sqrt{T} \frac{\partial Q_T(\beta_T^0)}{\partial \beta} \right\} \right\| = o_P(1).$$

A similar argument would allow us to prove

$$\sup_{\theta \in \Theta} \left\| \sqrt{T} m_{TH}^{\mathrm{CFS}}[\theta; \hat{\lambda}_T] - \left\{ \sqrt{T} \frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} - \sqrt{T} \frac{\partial Q_T(\beta_T^0)}{\partial \beta} \right\} \right\| = o_P(1).$$

Now, revisiting the definition of $\partial Q_{TH}(\theta, \ddot{\beta}_T)/\partial \beta$, for all $\theta \in \Theta$,

$$\sqrt{T} \frac{\partial Q_{TH}(\theta, \ddot{\beta}_T)}{\partial \beta} = \sqrt{T} \frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\theta, \beta_T(\theta))}{\partial \beta \, \partial \beta'} \sqrt{T} (\ddot{\beta}_T - \beta_T^0)$$

$$= \sqrt{T} \frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\theta, \beta_T(\theta))}{\partial \beta \, \partial \beta'} \left\{ -\frac{\partial^2 Q_T(\beta_T^0)}{\partial \beta \, \partial \beta'} \right\}^{-1} \sqrt{T} \frac{\partial Q_T(\beta_T^0)}{\partial \beta},$$

where $\beta_T(\theta)$ is a component-by-component intermediate value between $\beta_T^0$ and $\ddot{\beta}_T$. Applying a similar argument to the one above, and using the fact that $\sqrt{T}(\ddot{\beta}_T - \beta_T^0) = O_P(1)$, we have

$$\sup_{\theta \in \Theta} \left\| \sqrt{T} \frac{\partial Q_{TH}(\theta, \ddot{\beta}_T)}{\partial \beta} - \left\{ \sqrt{T} \frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} - \sqrt{T} \frac{\partial Q_T(\beta_T^0)}{\partial \beta} \right\} \right\| = o_P(1).$$

Therefore, the three estimating equations $\sqrt{T} \bar{m}_{TH}[\theta; \widehat{\beta}_T]$, $\sqrt{T} m_{TH}^{\mathrm{CFS}}[\theta; \hat{\lambda}_T]$ and $\sqrt{T} \frac{\partial Q_{TH}(\theta, \ddot{\beta}_T)}{\partial \beta}$ are each asymptotically equivalent to $\{ \sqrt{T} \frac{\partial Q_{TH}(\theta, \beta_T^0)}{\partial \beta} - \sqrt{T} \frac{\partial Q_T(\beta_T^0)}{\partial \beta} \}$, and the result follows. □

PROOF OF PROPOSITION 3. (i) We first prove that $\widehat{\theta}_{T,H}^{\mathrm{CFS}}(W)$ is consistent. By Assumption A4(ii), $m_{TH}^{\mathrm{CFS}}[\theta; \hat{\lambda}_T]$ converges in probability, uniformly on $\theta \in \Theta$, toward

$$\operatorname*{plim}_{T \to \infty} \left\{ \frac{\partial Q_{TH}(\theta, \hat{\beta}_T^r)}{\partial \beta} - \frac{\partial Q_T(\hat{\beta}_T^r)}{\partial \beta} \right\} = L(\theta, \beta^0) - L(\theta^0, \beta^0) = L(\theta, \beta^0).$$

The identification Assumption A4(iii), jointly with compactness of $\Theta$ and the continuity assumption (Assumption A4(i)), then yields

$$\operatorname*{plim}_{T\to\infty}\{\widehat{\theta}_{T,H}^{\text{CFS}}(W)\}=\theta^0.$$

(ii) By comparing (12) and (14), we have

$$\bar{m}_{TH}[\theta;\widehat{\beta}_T]-m_{TH}^{\text{CFS}}[\theta;\hat{\lambda}_T]=\frac{\partial^2 Q_{TH}(\theta,\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}(\widehat{\beta}_T-\hat{\beta}_T^r)+\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial\beta}$$

$$=\left\{\text{Id}_{d_\beta}-\frac{\partial^2 Q_{TH}(\theta,\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\left[\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\right]^{-1}\right\}\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial\beta}.$$

By Assumptions A1(ii) and A3(iii), this difference converges, uniformly on $\theta\in\Theta$, toward

$$-\mathcal{J}(\theta,\beta^0)[\mathcal{J}^0]^{-1}L(\theta^0,\beta^0)+L(\theta^0,\beta^0)=0,$$

where $\mathcal{J}^0=\mathcal{J}(\theta^0,\beta^0)$. Then, by a standard argument (see, e.g., Pakes and Pollard (1989, p. 1038)), we deduce that

$$\operatorname*{plim}_{T\to\infty}\{\widehat{\theta}_{T,H}^{s}(W)\}=\operatorname*{plim}_{T\to\infty}\{\widehat{\theta}_{T,H}^{\text{CFS}}(W)\}=\theta^0.$$

(iii) By Assumptions A1(ii) and A3(iii),

$$\sup_{\|\theta-\theta^0\|\leq\gamma/\sqrt{T}}\left\|-\frac{\partial^2 Q_{TH}(\theta,\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\left[\frac{\partial^2 Q_T(\hat{\beta}_T^r)}{\partial\beta\,\partial\beta'}\right]^{-1}+\text{Id}_{d_\beta}\right\|=o_P(1).$$

Then deduce from the above decomposition that

$$\sup_{\|\theta-\theta^0\|\leq\gamma/\sqrt{T}}\|\bar{m}_{TH}[\theta;\widehat{\beta}_T]-m_{TH}^{\text{CFS}}[\theta;\hat{\lambda}_T]\|=o_P\left(\left\|\frac{\partial Q_T(\hat{\beta}_T^r)}{\partial\beta}\right\|\right)=o_P\left(\frac{1}{\sqrt{T}}\right),$$

which in turn implies that

$$\sup_{\|\theta-\theta^0\|\leq\gamma/\sqrt{T}}\left|S_T^{\text{unr}}(\theta)-S_T^{\text{res}}(\theta)\right|=o_P(1/T),$$

for $S_T^{\text{unr}}(\theta)$ and $S_T^{\text{res}}(\theta)$, respectively, the objective functions minimized in (13) and (15) to define the estimators $\widehat{\theta}_{T,H}^{s}(W)$ and $\widehat{\theta}_{T,H}^{\text{CFS}}(W)$, respectively.

It is then a standard argument (see, e.g., Pakes and Pollard (1989, p. 1040)) to deduce that, using the asymptotic normality in Assumption A1(iii), the corresponding extremum estimators are asymptotically equivalent: $\|\widehat{\theta}_{T,H}^{s}(W)-\widehat{\theta}_{T,H}^{\text{CFS}}(W)\|=o_P(1/\sqrt{T})$.   □

PROOF OF THEOREM 2. We first show that $\sqrt{T}(\widehat{\theta}_{T,H}^{s}(W)-\theta_T)=O_P(1)$. Let $\widehat{\theta}_T:=\widehat{\theta}_{T,H}^{s}(W)$ and define, for a vector $x$, $\|x\|_W=\sqrt{x'Wx}$. By the triangle inequality,

$$\|L(\widehat{\theta}_T,\beta_T^0)-L(\theta_T,\beta_T^0)\|_W\leq\|L(\widehat{\theta}_T,\beta_T^0)-L(\theta_T,\beta_T^0)-\bar{m}_{TH}[\widehat{\theta}_T;\widehat{\beta}_T]\|_W$$

$$+\|\bar{m}_{TH}[\widehat{\theta}_T;\widehat{\beta}_T]\|_W. \tag{32}$$

Consider the first term in (32). Recall the definition of $\bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T]$, and apply the definitions of $\widehat{\beta}_T$, $\hat{\beta}^r_T$, and the uniform convergence in Assumption A3(iii), to deduce

$$\bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T] = \frac{\partial Q_{TH}(\widehat{\theta}_T, \hat{\beta}^r_T)}{\partial \beta} + \frac{\partial^2 Q_{TH}(\widehat{\theta}_T, \hat{\beta}^r_T)}{\partial \beta \, \partial \beta'}(\widehat{\beta}_T - \hat{\beta}^r_T)$$

$$= \frac{\partial Q_{TH}(\widehat{\theta}_T, \hat{\beta}^r_T)}{\partial \beta} - \mathcal{J}(\theta_T, \beta^0_T)[\mathcal{J}^0]^{-1} \frac{\partial Q_T(\hat{\beta}^r_T)}{\partial \beta} + o_P(1/\sqrt{T}). \quad (33)$$

Applying equation (33) and the triangle inequality, we obtain

$$\left\| L(\widehat{\theta}_T, \beta^0_T) - L(\theta_T, \beta^0_T) - \bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T] \right\|_W \leq \sup_{\theta \in \Theta, \|\beta - \beta^0_T\| \leq \frac{\gamma}{\sqrt{T}}} \left\| \frac{\partial Q_{TH}(\theta, \beta)}{\partial \beta} - L(\theta, \beta^0_T) \right\|_W$$

$$+ \left\| \frac{\partial Q_T(\hat{\beta}^r_T)}{\partial \beta} - L(\theta_T, \beta^0_T) \right\|_W. \quad (34)$$

In (34), the first term is $O_P(1/\sqrt{T})$ by Assumption A4(ii), and the second term is $O_P(1/\sqrt{T})$ by Assumption A1(iii) and Lemma 1. Analyzing the second term in equation (32), note that, by definition

$$\left\| \bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T] \right\|_W \leq \left\| \bar{m}_{TH}[\theta_T; \widehat{\beta}_T] \right\|_W.$$

Apply the same decomposition in (33) to the term $\bar{m}_{TH}[\theta_T; \widehat{\beta}_T]$, and the triangle inequality to obtain

$$\left\| \bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T] \right\|_W \leq \left\| \bar{m}_{TH}[\theta_T; \widehat{\beta}_T] \right\|_W \leq \left\| \frac{\partial Q_{TH}(\theta^0_T, \beta^0_T)}{\partial \beta} \right\|_W + O_P(1/\sqrt{T}).$$

From Assumption A3(ii), we have $\|\partial Q_{TH}(\theta^0_T, \beta^0_T)/\partial \beta\| = O_P(1/\sqrt{T})$, which yields

$$\left\| \bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T] \right\|_W \leq O_P(1/\sqrt{T}). \quad (35)$$

Applying the results in (34) and (35) into (33), we arrive at $\|L(\widehat{\theta}_T, \beta^0_T) - L(\theta, \beta^0_T)\|_W = O_P(1/\sqrt{T})$. From the local identification Assumption A5, we then have that, for some constant $C > 0$,

$$C\|\widehat{\theta}_T - \theta_T\| \leq \left\| L(\widehat{\theta}_T, \beta^0_T) - L(\theta_T, \beta^0_T) \right\|_W = O_P(1/\sqrt{T}).$$

Having proven $\sqrt{T}(\widehat{\theta}_T - \theta_T) = O_P(1/\sqrt{T})$, the remainder of the proof proceeds through a standard first-order Taylor series of the first-order conditions. First, apply the result of Proposition 2 to obtain

$$0 = \left[ \frac{\partial^2 Q_{TH}(\widehat{\theta}_T, \hat{\beta}^r_T)}{\partial \beta \, \partial \theta'} \right]' W \sqrt{T} \bar{m}_{TH}[\widehat{\theta}_T; \widehat{\beta}_T] + o_P(1)$$

$$= \left[ \frac{\partial^2 Q_{TH}(\widehat{\theta}_T, \hat{\beta}^r_T)}{\partial \beta \, \partial \theta'} \right]' W \sqrt{T} \frac{\partial Q_{TH}(\widehat{\theta}_T, \ddot{\beta}_T)}{\partial \beta} + o_P(1).$$

Now, disregarding terms of smaller order than $O_P(1/\sqrt{T})$, a Taylor series expansion of $\sqrt{T}\partial Q_{TH}(\widehat{\theta}_T, \ddot{\beta}_T)/\partial\beta$ around $\theta_T$ yields

$$0 = \left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]' W\sqrt{T}\left\{\frac{\partial Q_{TH}(\theta_T, \ddot{\beta}_T)}{\partial\beta} + \frac{\partial Q_{TH}(\theta_T, \ddot{\beta}_T)}{\partial\beta\,\partial\theta'}(\widehat{\theta}_T - \theta_T)\right\} + o_P(1).$$

Expanding $\sqrt{T}\partial Q_{TH}(\theta_T, \ddot{\beta}_T)/\partial\beta$ as in equation (17), and making use of the uniform convergence in Assumption A3(iii), we obtain (up to an $o_P(1)$ term)

$$0 = \left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]' W\sqrt{T}\left\{\frac{\partial Q_{TH}(\theta_T, \beta_T^0)}{\partial\beta} - \frac{\partial Q_T(\beta_T^0)}{\partial\beta}\right\}$$
$$+ \left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]' W\left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]\sqrt{T}(\widehat{\theta}_T - \theta_T).$$

Rearranging terms and making use of Assumption A5,

$$\sqrt{T}(\widehat{\theta}_T - \theta_T) = -\left\{\left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]' W\left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]\right\}^{-1}$$
$$\times \left[\frac{\partial L(\theta_T, \beta_T^0)}{\partial\theta'}\right]' W\sqrt{T}\left\{\frac{\partial Q_{TH}(\theta_T, \beta_T^0)}{\partial\beta} - \frac{\partial Q_T(\beta_T^0)}{\partial\beta}\right\}.$$

By Assumptions A1(iii) and A3(ii), the term in brackets is an asymptotically Gaussian mean-zero random variable. The stated result then follows.   □

PROOF OF THEOREM 3.  The result follows from the following sequence of arguments: (i) $\widehat{\theta}_{T,H}^c$ solves $\widehat{\beta}_T = \tilde{\beta}_{TH}^c(\theta)$; (ii) $\widehat{\theta}_{T,H}^s$ solves $0 = \bar{m}_{TH}[\theta, \widehat{\beta}_T]$; (iii) from (ii) and the structure of $\bar{m}_{TH}[\theta, \widehat{\beta}_T]$ we have, re-arranging $0 = \bar{m}_{TH}[\widehat{\theta}_T^s, \widehat{\beta}_T]$ and solving for $\widehat{\beta}_T$,

$$\widehat{\beta}_T = \hat{\beta}_T^r - \left[\frac{\partial^2 Q_{TH}[\widehat{\theta}_T^s, \hat{\beta}_T^r]}{\partial\beta\,\partial\beta'}\right]^{-1}\frac{\partial Q_{TH}[\widehat{\theta}_T^s, \hat{\beta}_T^r]}{\partial\beta} = \tilde{\beta}_{TH}^c(\widehat{\theta}_T^s),$$

where the last equality follows from the definition of $\tilde{\beta}_{TH}^c(\theta)$. Therefore, from (i) we have $\widehat{\beta}_T = \tilde{\beta}_{TH}^c(\widehat{\theta}_T^c)$ and from (iii) we have $\widehat{\beta}_T = \tilde{\beta}_{TH}^c(\widehat{\theta}_T^c) = \tilde{\beta}_{TH}^c(\widehat{\theta}_T^s)$.   □

## REFERENCES

Ait-Sahalia, Y., J. Cacho-Diaz, and R. J. A. Laeven (2015), "Modeling financial contagion using mutually exciting jump processes." *Journal of Financial Economics*, 117 (3), 585–606. [147]

Ait-Sahalia, Y. and J. Jacod (2014), *High-Frequency Financial Econometrics*. Princeton University Press, Princeton, NJ. [120]

Andersen, T. G., L. Benzoni, and J. Lund (2002), "An empirical investigation of continuous-time equity return models." *Journal of Finance*, 57 (3), 1239–1284. [120]

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2010), "Parametric and nonparametric volatility measurement." In *Handbook of Financial Econometrics: Tools and Techniques*, 67–137, North-Holland, Amsterdam. [120]

Andrews, D. W. K. (1999), "Estimation when a parameter is on a boundary." *Econometrica*, 67 (6), 1341–1383. [114, 115, 116, 122, 123]

Andrews, D. W. K. and X. Cheng (2012), "Estimation and inference with weak, semi-strong, and strong identification." *Econometrica*, 80 (5), 2153–2211. [115, 117]

Calzolari, G., G. Fiorentini, and E. Sentana (2004), "Constrained indirect estimation." *The Review of Economic Studies*, 71 (4), 945–973. [114]

Calzolari, G. and R. Halbleib (2018), "Estimating stable latent factor models by indirect inference." *Journal of Econometrics*, 205 (1), 280–301. [120]

Creel, M. and D. Kristensen (2015), "ABC of SV: Limited information likelihood inference in stochastic volatility jump-diffusion models." *Journal of Empirical Finance*, 31, 85–108. [120, 146, 148]

Engle, R. F. and G. G. J. Lee (1996), "Estimating diffusion models of stochastic volatility." In *Modeling Stock Market Volatility: Bridging the Gap to Continuous Time* (P. E. Rossi, ed.), Academic Press, San Diego, CA. [118]

Eraker, B. (2001), "MCMC analysis of diffusion models with application to finance." *Journal of Business & Economic Statistics*, 19 (2), 177–191. [120]

Fernandez, C. and M. F. J. Steel (1998), "On Bayesian modeling of fat tails and skewness." *Journal of the American Statistical Association*, 93 (441), 359–371. [119, 142]

Francq, C. and J.-M. Zakoian (2007), "Quasi-maximum likelihood estimation in GARCH processes when some coefficients are equal to zero." *Stochastic Processes and their Applications*, 117 (9), 1265–1284. [118]

Francq, C. and J.-M. Zakoian (2009), "Testing the nullity of GARCH coefficients: Correction of the standard tests and relative efficiency comparisons." *Journal of the American Statistical Association*, 104, 313–324. [139, 140]

Frazier, D. T. and E. Renault (2020), "Supplement to 'Indirect inference with(out) constraints'." *Quantitative Economics Supplemental Material*, 11, https://doi.org/10.3982/QE986.

Frazier, D. T., W. Maneesoonthorn, G. M. Martin, and B. P. M. McCabe (2019), "Approximate Bayesian forecasting." *International Journal of Forecasting*, 35 (2), 521–539. [145]

Frazier, D. T. and E. Renault (2017), "Efficient two-step estimation via targeting." *Journal of Econometrics*, 201 (2), 212–227. [137]

Fulop, A., J. Li, and J. Yu (2014), "Self-exciting jumps, learning, and asset pricing implications." *The Review of Financial Studies*, 28 (3), 876–912. [147]

Gallant, A. R. and G. Tauchen (1996), "Which moments to match." *Econometric Theory*, 12, 657–681. [113]

Garcia, R., E. Renault, and D. Veredas (2011), "Estimation of stable distributions by indirect inference." *Journal of Econometrics*, 161 (2), 325–337. [116, 119, 142, 144]

Gerd, H., A. Lunde, N. Shephard, and K. Sheppard (2009), "Oxford-Man Institute's realized library." [147]

Gospodinov, N., I. Komunjer, and S. Ng (2017), "Simulated minimum distance estimation of dynamic models with errors-in-variables." *Journal of Econometrics*, 2, 181–193. [115]

Gourieroux, C., A. Monfort, and E. Renault (1993), "Indirect inference." *Journal of Applied Econometrics*, 85, S85–S118. [113]

Gourieroux, C. and A. Monfort (1996), *Simulation-Based Econometric Methods*. Oxford University Press, Oxford. [135]

Gourieroux, C., E. Renault, and N. Touzi (1999), "Calibration by simulation for small sample bias correction." In *Simulation-Based Inference in Econometrics, Methods and Applications* (R. Mariano, T. Schuermann, and M. Weeks, eds.), 328–358, Cambridge University Press, Cambridge. [133]

Hansen, L. P., J. Heaton, and A. Yaron (1996), "Finite-sample properties of some alternative GMM estimators." *Journal of Business & Economic Statistics*, 14 (3), 262–280. [136]

Jacquier, E., N. G. Polson, and P. E. Rossi (1994), "Bayesian analysis of stochastic volatility models (with discussion)." *Journal of Business & Economic Statistics*, 12, 371–417. [137]

Jiang, W. and B. Turnbull (2004), "The indirect method: Inference based on intermediate statistics—A synthesis and examples." *Statistical Science*, 19 (2), 239–263. [114]

Ketz, P. (2018), "Subvector inference when the true parameter vector may be near or at the boundary." *Journal of Econometrics*, 27 (2), 285–306. [115, 117, 123, 126]

Kim, S., N. Shephard, and S. Chib (1998), "Stochastic volatility: Likelihood inference and comparison with ARCH models." *The Review of Economic Studies*, 65 (3), 361–393. [138, 139, 141]

Lombardi, M. J. and G. Calzolari (2009), "Indirect estimation of alpha-stable stochastic volatility models." *Computational Statistics & Data Analysis*, 53 (6), 2298–2308. [120]

Maneesoonthorn, W., C. S. Forbes, and G. M. Martin (2017), "Inference on self-exciting jumps in prices and volatility using high-frequency measures." *Journal of Applied Econometrics*, 32 (3), 504–532. [148]

Monfardini, C. (1998), "Estimating stochastic volatility models through indirect inference." *The Econometrics Journal*, 1 (1), 113–128. [118]

Moon, H. R. and F. Schorfheide (2009), "Estimation with overidentifying inequality moment conditions." *Journal of Econometrics*, 153, 136–154. [115]

Pakes, A. and D. Pollard (1989), "Simulation and the asymptotics of optimization estimators." *Econometrica*, 57 (5), 1027–1057. [154]

Pastorello, S., E. Renault, and N. Touzi (2000), "Statistical inference for random-variance option pricing." *Journal of Business & Economic Statistics*, 18 (3), 358–367. [118]

Samoradnitsky, G. (2017), *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*. Routledge, New York, NY. [119]

Smith, A. A. (1993), "Estimating nonlinear time-series models using simulated vector autoregressions." *Journal of Applied Econometrics*, 8 (S1), S63–S84. [113]