Bootstrapping two-stage quasi-maximum likelihood estimators of time series models

Sílvia Gonçalves*, Ulrich Hounyo†, Andrew J. Patton‡, and Kevin Sheppard§

October 30, 2019

Abstract

This paper's main contribution is to theoretically justify the application of bootstrap methods in multistage quasi-maximum likelihood estimation involving time series data. Two consistency results are provided: consistency of the bootstrap distribution and consistency of bootstrap variance estimators. These results justify constructing bootstrap percentile intervals and computing bootstrap standard errors using multi-step quasi-maximum likelihood estimation, avoiding the need to analytically quantify the estimation uncertainty caused by the multistage estimation process. Our results should be useful for inference in many models in finance and economics such as multivariate copula models or large multivariate GARCH models, which are often estimated in stages.

1 Introduction

Many models in economics and finance are estimated by maximum likelihood in multiple stages. Examples include estimation of multivariate copula models, where we first estimate parameters related to the marginal distributions and then estimate the copula parameters (see Joe (1995) and Patton (2006), for example); estimation of large multivariate GARCH models such as the Dynamic Conditional Correlation (DCC) GARCH model of Engle and Sheppard (2001), where we first estimate univariate GARCH models for each asset and then, using transformed residuals resulting from the first stage, we estimate a conditional correlation estimator; regression models with generated regressors, one example being the popular two-pass regressions used to estimate risk premium parameters, etc. In all these cases, inference on parameters estimated at a later stage should account for estimation parameter uncertainty in earlier stages. The bootstrap has often been used in this context as it avoids the need to compute standard errors obtained from cumbersome analytical formulas. See e.g. Patton (2012) for bootstrap applications in copula models and Cochrane (2001, Chapter 15.2) for bootstrap inference in two-pass regressions.

Although the existing literature has studied the validity of the bootstrap for one-step quasi-maximum likelihood estimators (QMLE) under very general conditions on the time series dependence and heterogeneity (see, in particular, Gonçalves and White (2004), henceforth GW(2004)), no results seem to be available for multistep QMLE under this level of generality. For instance, Chen et al. (2003)

*Department of Economics, McGill University, Leacock Building, Room 506, 855 Sherbrooke Street West, Montreal, Quebec, H3A 2T7. Email: silvia.goncalves@mcgill.ca.

†Department of Economics, University at Albany, SUNY, Buiding 25, room 231, 1400 Washington Ave, Albany, New York, 12222, United States. Email: khounyo@albany.edu.

‡Department of Economics, Duke University, Durham, NC 27708, United States. Email: andrew.patton@duke.edu.

§Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom. Email: kevin.sheppard@economics.ox.ac.uk.

and Armstrong et al. (2014) consider the bootstrap for two-step nonlinear parametric and semiparametric models, but assume i.i.d. data. More recently, and also for i.i.d. observations, Cattaneo et al. (2019) consider bootstrap inference for two-step GMM estimators, where the first step depends on a large number of covariates. Even though these papers allow for very general estimators, e.g. in the form of nonparametric first step estimators, their results do not cover time series applications. Our main goal in this paper is to fill this gap. In particular, we focus on the simpler parametric multistep QML estimator (which is popular in finance) but show bootstrap validity under general time series dependence and heterogeneity.

We consider two approaches to this problem. The first consists of jointly resampling the contributions to the quasi-loglikelihood functions in the two (or multiple) stages. Since model misspecification at any stage can induce time series dependence in the scores of each model, we rely on the moving blocks bootstrap (MBB) of Künsch (1989) and Liu and Singh (1992). In particular, given a set of bootstrap indices generated with the MBB, we first obtain $\hat{\alpha}_n^*$, the bootstrap analogue of the first step estimator $\hat{\alpha}_n$, by maximizing the resampled version of the quasi-loglikelihood function of the first step problem. We then use these same bootstrap indices to resample the contributions to the criterion function of the second stage problem, evaluated at $\hat{\alpha}_n^*$. Maximizing this function with respect to β yields $\hat{\beta}_n^*$, the bootstrap analogue of the two-step QMLE $\hat{\beta}_n$.

While valid, this bootstrap method may be computationally very intensive because it requires two (or more) sets of maximization on each bootstrap sample. For this reason, we also propose a fast resampling method that avoids any optimization problem in the bootstrap world. In particular, our proposal is to resample the score function underlying the asymptotic linear representation of the two-step QMLE, evaluated at $\hat{\alpha}_n$ and $\hat{\beta}_n$. In contrast, the fast resampling method of Armstrong et al. (2014) resamples only the score vector of the second step model evaluated at $\hat{\alpha}_n^*$ and $\hat{\beta}_n$. Whereas their approach avoids the explicit characterization of the score vector of the first step estimation problem (which can be difficult in nonparametric models), it is more computationally intensive than our fast resampling method as it requires computation of $\hat{\alpha}_n^*$.

We prove two sets of results for both bootstrap methods. First, we show the consistency of the bootstrap distribution of $\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)$ as an estimator of the distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ under a set of regularity conditions that allow for time series dependence and heterogeneity of unknown forms. Our conditions are an extension to the two-step QMLE of the conditions used by Gonçalves and White (2004) to show the asymptotic validity of the MBB for inference on one-step QMLE for nonlinear dynamic models. These results justify the construction of bootstrap percentile intervals, but do not by themselves justify estimating the standard errors of the two-step QMLE by the bootstrap. Hence, we also show the consistency of bootstrap variance estimators. This entails verifying a certain uniform integrability condition and requires stronger model assumptions. In particular, we follow the approach of Kato (2011) and Cheng (2015) and rely on empirical process theory to prove our results. Their results apply to one-step M estimators with i.i.d. data and do not cover time series applications. They also do not cover two-step QMLE, even under the i.i.d. assumption. Similarly, although the results of Gonçalves and White (2005) allow for time series dependence, they are specific to the one-step least squares estimator. Thus, no results appear to be available regarding the consistency of bootstrap variance estimators of one or multi-stage QMLE with general time series dependence.

We provide a set of Monte Carlo simulations that illustrate the usefulness of our results. In particular, we consider estimation of a bivariate copula model, where estimation is done by stages and the parameter of interest is the copula parameter. In addition to the standard asymptotic theory-based interval that relies on analytical standard errors, we consider two types of bootstrap-based intervals: intervals that rely on bootstrap standard errors, but use the normal critical value, and bootstrap percentile intervals. We can summarize our results as follows. First, all methods tend to provide similarly good coverage probabilities, even for the smaller sample sizes. This is in agreement with

the theory of the bootstrap since none of the methods promises asymptotic refinements. Moreover, the model design does not allow for dynamic misspecification, which explains why we do not find larger finite sample distortions even when n is small. Second, the main difference among the different methods is their confidence interval lengths. In particular, the intervals based on the fully optimized bootstrap method tend to be narrower than the intervals based on either the fast resampling method or the asymptotic approach using analytical standard errors. This can be explained by the fact that the fully optimized bootstrap standard error estimator has a smaller mean squared error than the remaining methods, as our simulations show. Thus, although more computationally intensive than the fast resampling method, the fully optimized bootstrap intervals have better finite sample properties.

The rest of the paper is organized as follows. In Section 2, we present the framework and provide an example of a two-step QMLE based on the bivariate copula model. In Section 3, we describe our two bootstrap methods and prove their consistency in Section 4. Section 5 contains the simulation results and Section 6 concludes. Proofs are relegated to an online supplementary appendix. This Appendix also contains a general bootstrap consistency theorem which provides a set of bootstrap high level conditions (in the form of bootstrap uniform laws of large numbers and bootstrap central limit theorems) under which the bootstrap distribution of any two-step bootstrap M-estimator is asymptotically normal with the same asymptotic covariance matrix of the corresponding two-step M estimator. This result may be of independent interest as it applies to any two-step M estimator and does not require the first step estimators $\hat{\alpha}_n$ and $\hat{\alpha}_n^*$ to be QMLE, only assuming that they are asymptotically linear. In addition, its high level conditions can be verified for any bootstrap scheme as they are not specific to the moving blocks bootstrap.

2 Framework

2.1 Two-stage QMLE

Suppose $\left\{X_t : \Omega \to \mathbb{R}^l,\ t \in \mathbb{N}\right\}$ denotes a sequence of $\mathbb{R}^l$-valued random vectors defined on some probability space $(\Omega, \mathcal{F}, P)$. Let $\Theta = \mathcal{A} \times \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are compact subsets of finite dimensional Euclidean spaces. Given an observed sample $\{X_t : t = 1, \ldots, n\}$, our goal is to estimate a parameter vector $\beta_0 \in \mathcal{B} \subset \mathbb{R}^p$ by a two-stage quasi-maximum likelihood (2QMLE) estimator. For simplicity, we focus on the two-stage QMLE, but our results generalize easily to multi-stage QMLE's.

In the first step, we estimate $\alpha_0 \in \mathcal{A} \subset \mathbb{R}^k$ with

$$\hat{\alpha}_n = \arg\max_{\alpha \in \mathcal{A}} Q_{1n}(\alpha),$$

where

$$Q_{1n}(\alpha) \equiv n^{-1} \sum_{t=1}^{n} \log f_{1t}\left(X^t, \alpha\right),$$

with $X^t \equiv (X_1, \ldots, X_{t-1}, X_t)$, for some quasi-likelihood function $f_{1t}\left(X^t, \alpha\right) : \mathbb{R}^{lt} \times \mathcal{A} \to \mathbb{R}^+$. To simplify the notation, we sometimes write $f_{1t}(\alpha) \equiv f_{1t}\left(X^t, \alpha\right)$. A similar notation is used for any other function of X^t throughout.

In the second step, we estimate β with

$$\hat{\beta}_n = \arg\max_{\beta \in \mathcal{B}} Q_{2n}(\hat{\alpha}_n, \beta)$$

where

$$Q_{2n}(\alpha, \beta) \equiv n^{-1} \sum_{t=1}^{n} \log f_{2t}\left(X^t, \alpha, \beta\right),$$

for a conditional quasi-likelihood function $f_{2t}(\alpha,\beta) \equiv f_{2t}\left(X^t,\alpha,\beta\right) : \mathbb{R}^{lt} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}^+$. We allow for time heterogeneity in $f_{1t}(\alpha)$ and $f_{2t}(\alpha,\beta)$ (i.e. the functional forms may depend on t) and we also allow for the possibility that these functions depend on the past information up to time t (i.e. X^t is a vector of possibly growing dimension).

2.2 An example: copula models

An example of time series models that are often estimated in multiple stages are copula-based multivariate models. These models combine separately estimated marginal distributions via a copula function to form a joint distribution. When the parameters that characterize the marginal distributions are different from those that characterize the copula density function, estimation and inference can be done in stages. Our results can be useful in this context.

To illustrate, let $X_t \equiv (y_{1t}, y_{2t})'$ denote a random vector whose joint conditional density we would like to model. By the usual decomposition, we can write

$$\log g(X_1,\ldots,X_n,\theta) = \sum_{t=1}^{n} \log g_t\left(X_t|\mathcal{F}^{t-1},\theta\right),$$

where $g_t\left(X_t|\mathcal{F}^{t-1},\theta\right)$ is the conditional density function of X_t given $\mathcal{F}^{t-1}$. Suppose $y_{it}|\mathcal{F}^{t-1} \sim G_{it}(\alpha_i)$, some distribution function parametrized by a set of parameters α_i with density function $g_{it}(\alpha_i)$. The joint (conditional) pdf of X_t is then given by

$$g_t\left(X_t|\mathcal{F}^{t-1},\theta\right) = g_{1t}(y_{1t},\alpha_1)\, g_{2t}(y_{2t},\alpha_2)\, c_t\left(G_{1t}(y_{1t},\alpha_1), G_{2t}(y_{2t},\alpha_2),\beta\right),$$

where $c_t(\cdot,\cdot,\beta)$ is a copula density function parametrized by β, and $\theta = (\alpha_1,\alpha_2,\beta)'$ denotes the full set of parameters. It follows that the joint log likelihood function can be written as

$$\begin{aligned}\log g(X_1,\ldots,X_n,\theta) &= \sum_{t=1}^{n} \log g_{1t}\left(y_{1t}|\mathcal{F}^{t-1},\alpha_1\right) + \sum_{t=1}^{n} \log g_{2t}\left(y_{2t}|\mathcal{F}^{t-1},\alpha_2\right) \\ &\quad + \sum_{t=1}^{n} \log c_t\left(G_{1t}(y_{1t},\alpha_1), G_{2t}(y_{2t},\alpha_2)|\mathcal{F}^{t-1},\beta\right).\end{aligned}$$

When the parameters characterizing the marginals and the copula function are separable (i.e. the parameters that enter one marginal do not enter another marginal nor the copula function and there are no cross equation restrictions), we can estimate these parameters by stages. In particular, we first estimate α_i by QMLE:

$$\hat{\alpha}_{in} = \arg\max_{\alpha_i} \sum_{t=1}^{n} \log g_{it}\left(y_{it}|\mathcal{F}^{t-1},\alpha_i\right), \text{ for } i = 1,2,$$

and then estimate the copula parameters β in a second stage by

$$\hat{\beta}_n = \arg\max_{\beta} \sum_{t=1}^{n} \log c_t\left(G_{1t}(y_{1t},\hat{\alpha}_{1n}), G_{2t}(y_{2t},\hat{\alpha}_{2n})|\mathcal{F}^{t-1},\beta\right).$$

Thus, in our previous notation,

$$\begin{aligned}Q_{2n}(\hat{\alpha}_n,\beta) &= \sum_{t=1}^{n} \log f_{2t}\left(X^t,\hat{\alpha}_n,\beta\right), \text{ where } \hat{\alpha}_n = (\hat{\alpha}_{1n},\hat{\alpha}_{2n})', \text{ and} \\ f_{2t}\left(X^t,\hat{\alpha}_n,\beta\right) &\equiv c_t\left(G_{1t}(y_{1t},\hat{\alpha}_{1n}), G_{2t}(y_{2t},\hat{\alpha}_{2n})|\mathcal{F}^{t-1},\beta\right).\end{aligned}$$

The contributions to this quasi-log likelihood function depend on the sample on $X_t = (y_{1t}, y_{2t})'$ up to time t through the integral probability transforms $G_{it}(y_{it},\hat{\alpha}_{in})$.

2.3 Asymptotic properties of two-stage QMLE: a review

In this section, we review the asymptotic properties of the two-stage QMLE. These results are useful to understand the properties that the bootstrap needs to have in order to be asymptotically valid.

Let α_0 be the unique maximizer of $\bar{Q}_1(\alpha) \equiv \lim_{n\to\infty} E(Q_{1n}(\alpha))$ on $\mathcal{A}$ and let β_0 be the unique maximizer of $\bar{Q}_2(\alpha_0,\beta) \equiv \lim_{n\to\infty} E(Q_{2n}(\alpha_0,\beta))$ on $\mathcal{B}$. Then, under Assumption A in the online Appendix A.1, we can show that $\hat{\beta}_n \xrightarrow{P} \beta_0$ and

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} N\left(0, H_0^{-1} J_0 H_0^{-1}\right),$$

where

$$H_0 \equiv \lim_{n\to\infty} E\left(n^{-1}\sum_{t=1}^{n} \frac{\partial}{\partial\beta'} s_{2t}(\alpha_0,\beta_0)\right), \quad \text{with} \quad s_{2t}(\alpha_0,\beta_0) \equiv \frac{\partial}{\partial\beta}\log f_{2t}(\alpha_0,\beta_0),$$

and

$$J_0 \equiv \lim_{n\to\infty} Var\left(n^{-\frac{1}{2}}\sum_{t=1}^{n}\left(s_{2t}(\alpha_0,\beta_0) - F_0 A_0^{-1} s_{1t}(\alpha_0)\right)\right),$$

where

$$\begin{aligned} s_{1t}(\alpha_0) &\equiv \frac{\partial}{\partial\alpha}\log f_{1t}(\alpha_0), \quad A_0 \equiv \lim_{n\to\infty} E\left(n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\alpha'} s_{1t}(\alpha_0)\right), \text{ and} \\ F_0 &\equiv \lim_{n\to\infty} E\left(n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\alpha'} s_{2t}(\alpha_0,\beta_0)\right). \end{aligned}$$

As this result shows, the impact of the first stage estimation of α_0 is not negligible asymptotically except when $F_0 = 0$. This implies that we need to adjust the standard errors of $\hat{\beta}_n$ for the added estimation uncertainty of $\hat{\alpha}_n$. Although a consistent estimator of J_0 can be obtained by applying a HAC (heteroskedasticity and autocorrelation covariance) estimator to $\left\{s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right) - \hat{F}_n\hat{A}_n^{-1}s_{1t}(\hat{\alpha}_n)\right\}$ (where $\hat{F}_n$ and $\hat{A}_n$ are consistent estimators of F_0 and A_0) in practice the bootstrap is often used. Our goal is to provide a set of conditions that justify this practice in time series applications.

3 Bootstrap methods

The asymptotic validity of the bootstrap depends on its ability to mimic the asymptotic variance-covariance matrix of $\hat{\beta}_n$. The form of J_0 suggests that the bootstrap should replicate the time series dependence and the heterogeneity properties of the score vector $\left\{s_t(\alpha_0,\beta_0) \equiv s_{2t}(\alpha_0,\beta_0) - F_0 A_0^{-1} s_{1t}(\alpha_0)\right\}$. Model misspecification at any stage can induce time series dependence and our approach is to use a block bootstrap. In particular, we rely on the moving blocks bootstrap (MBB) of Künsch (1989) and Liu and Singh (1992). See also Gonçalves and White (2002, 2004, 2005) for the validity of the MBB under general time series dependence and heterogeneity.

We consider two different methods. One is based on resampling the contributions to the log likelihood functions $\{f_{1t}(\alpha)\}$ (which yields a bootstrap QMLE $\hat{\alpha}_n^*$) and $\{f_{2t}(\hat{\alpha}_n^*,\beta)\}$ (which is optimized over β to yield $\hat{\beta}_n^*$). The same bootstrap indices obtained with the MBB are used across the two stages, ensuring that this method mimics the time series dependence of the extended score. Because it requires two (or multi) sets of maximization, this method may be computationally intensive. For this reason, we also propose another bootstrap method which resamples directly the estimated score

$s_t\left(\hat{\alpha}_n, \hat{\beta}_n\right) \equiv s_{2t}\left(\hat{\alpha}_n, \hat{\beta}_n\right) - \hat{F}_n \hat{A}_n^{-1} s_{1t}\left(\hat{\alpha}_n\right)$. Our simulations show that this method is less efficient than the fully optimized bootstrap method. In particular, the fast resampling standard errors have larger mean squared errors compared to the fully optimized standard errors, especially for the smaller sample sizes. This translates into wider confidence intervals for the parameter of interest.

Both methods involve resampling certain functions of the data using the MBB to obtain new indices, which can be described as follows. For a generic time series $\{Z_t : t = 1, \ldots, n\}$, let $\ell = \ell_n \in \mathbb{N}$ $(1 \leq \ell < n)$ be a block length. Define $B_{t,\ell} = \{Z_t, Z_{t+1}, \ldots, Z_{t+\ell-1}\}$ as the block of ℓ consecutive observations starting at Z_t ($\ell = 1$ corresponds to the standard i.i.d. bootstrap). For simplicity take $n = k\ell$. The MBB draws $k = n/\ell$ blocks randomly with replacement from the set of overlapping blocks $\{B_{1,\ell}, \ldots, B_{n-\ell+1,\ell}\}$. Letting $I_1, \ldots, I_k$ be i.i.d. random variables distributed uniformly on $\{0, \ldots, n-\ell\}$, we have $\{Z_t^* = Z_{\tau_t}, t = 1, \ldots, n\}$, where τ_t is defined as $\{\tau_t\} \equiv \{I_1 + 1, \ldots, I_1 + \ell, \ldots, I_k + 1, \ldots, I_k + \ell\}$.

3.1 The fully optimized bootstrap method

The first method we consider requires resampling the contributions to the two (or more) likelihood functions f_{1t} and f_{2t} and then computing $\hat{\alpha}_n^*$ and $\hat{\beta}_n^*$ using these resampled log-likelihood functions. More specifically, the bootstrap analogue of $\hat{\alpha}_n$ is given by

$$\hat{\alpha}_n^* = \arg\max_{\alpha \in \mathcal{A}} Q_{1n}^*\left(\alpha\right),$$

where

$$Q_{1n}^*\left(\alpha\right) \equiv n^{-1} \sum_{t=1}^{n} \log f_{1t}^*\left(\alpha\right),$$

and $f_{1t}^*\left(\alpha\right) = f_{1,\tau_t}\left(\alpha\right) \equiv f_{1,\tau_t}\left(X^{\tau_t}, \alpha\right)$ is a resampled version of $f_{1t}\left(\alpha\right) \equiv f_{1t}\left(X^t, \alpha\right)$, where the indices τ_t are chosen by the bootstrap. Thus, we resample the functions $f_{1t}\left(\alpha\right)$ rather than the data directly. However, when $f_{1t}\left(\alpha\right) = f_1\left(Z_t, \alpha\right)$ where the function f_{1t} does not depend on t and is a function of $Z_t \equiv \left(X_t, X_{t-1}, \ldots, X_{t-k}\right)'$ for some finite $k \geq 0$, resampling $f_{1t}\left(\alpha\right)$ is equivalent to resampling the vector Z_t, i.e. $f_{1t}^*\left(\alpha\right) \equiv f_{1t}^*\left(X^{*t}, \alpha\right) = f_1\left(Z_t^*, \alpha\right) = f_1\left(Z_{\tau_t}, \alpha\right)$. To better appreciate the difference between resampling the functions $f_{1t}\left(\alpha\right)$ and the data, take for instance an $ARCH\left(1\right)$ model, where

$$\begin{aligned} X_t &= \sqrt{h_t}\varepsilon_t,\ \varepsilon_t | \mathcal{F}^{t-1} \sim N\left(0,1\right), \\ h_t\left(\alpha\right) &= \omega + aX_{t-1}^2,\ \text{where } \alpha = \left(\omega, a\right)'. \end{aligned}$$

The conditional log likelihood function of X_t given $\mathcal{F}^{t-1}$ is equal to $\log f_1\left(X^t, \alpha\right) = -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log h_t\left(\alpha\right) - \frac{1}{2}\frac{X_t^2}{h_t(\alpha)}$, where $X^t = \left(X_{t-1}, X_t\right)$. In this case, resampling the functions $f_{1t}\left(\alpha\right) \equiv f_1\left(X^t, \alpha\right)$ is equivalent to resampling the pairs $\left(X_{t-1}, X_t\right)$. But if instead we consider a GARCH(1,1) model with $h_t\left(\alpha\right) = \omega + aX_{t-1}^2 + bh_{t-1}\left(\alpha\right)$, where now $\alpha = \left(\omega, a, b\right)'$, then $h_t\left(\alpha\right)$ is potentially a function of the infinite history of X up to time t. In practice, we choose an initial guess for h_t, say h_0, and let $X^t = \left(X_{t-1}, X_{t-2}, \ldots, X_1, h_0\right)$. Because X^t now does not have a fixed dimension, we cannot resample "pairs" of data. Our approach is to evaluate the log likelihood function at the new set of indices $\{\tau_t\}$ generated with the bootstrap. Thus, e.g. for the first observation in the bootstrap world, we let $\log f_{1,\tau_1}\left(\alpha\right) = -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log h_{\tau_1}\left(\alpha\right) - \frac{1}{2}\frac{X_{\tau_1}^2}{h_{\tau_1}(\alpha)}$, where $h_{\tau_1}\left(\alpha\right) = \omega + aX_{\tau_1-1}^2 + bh_{\tau_1-1}\left(\alpha\right)$, a function of $X^{\tau_1} = \left(X_{\tau_1-1}, X_{\tau_1-2}, \ldots, X_1, h_0\right)$.

The second step bootstrap estimator $\hat{\beta}_n^*$ is obtained as

$$\hat{\beta}_n^* = \arg\max_{\beta \in \mathcal{B}} Q_{2n}^*\left(\hat{\alpha}_n^*, \beta\right),$$

where

$$Q_{2n}^{*}\left(\hat{\alpha}_{n}^{*},\beta\right)\equiv n^{-1}\sum_{t=1}^{n}\log f_{2t}^{*}\left(\hat{\alpha}_{n}^{*},\beta\right),$$

with $f_{2t}^{*}\left(\hat{\alpha}_{n}^{*},\beta\right)=f_{2,\tau_t}\left(\hat{\alpha}_{n}^{*},\beta\right)\equiv f_{2,\tau_t}\left(X^{\tau_t},\hat{\alpha}_{n}^{*},\beta\right)$. Thus, we resample the functions $f_{2t}\left(\alpha,\beta\right)\equiv f_{2t}\left(X^{t},\alpha,\beta\right)$ evaluated at $\alpha=\hat{\alpha}_{n}^{*}$ using the *same* indices τ_t used in computing $\hat{\alpha}_{n}^{*}$. Resampling both functions f_{1t} and f_{2t} with the same set of indices is important because this will preserve the form of dependence between the two functions. In particular, this will guarantee that the bootstrap is able to mimic the dependence in the score vector $s_t\left(\hat{\alpha}_n,\hat{\beta}_n\right)$. If instead we used two different sets of indices, say τ_{1t} and τ_{2t}, generated independently of each other, this would induce an independence between f_{1t}^{*} and f_{2t}^{*} which would not necessarily exist for the original functions.

3.2 A fast resampling method

Bootstrapping multi-stage extremum estimators can be computationally intensive as this may require solving multiple optimization problems for each resample. For this reason, we also consider a fast resampling method for bootstrapping two-step QMLE which has a closed form expression and avoids any numerical optimization. To describe this estimator, let

$$\hat{H}_n=n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\beta'}s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right),\ \hat{A}_n=n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\alpha'}s_{1t}\left(\hat{\alpha}_n\right),\ \text{and}\ \hat{F}_n=n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\alpha'}s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right).$$

The fast resample two-step QMLE is given by

$$\hat{\beta}_{1,n}^{*}=\hat{\beta}_n-\hat{H}_n^{-1}n^{-1}\sum_{t=1}^{n}s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right),$$

where $s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)$ is a resampled version of

$$s_t\left(\hat{\alpha}_n,\hat{\beta}_n\right)\equiv s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right)-\hat{F}_n\hat{A}_n^{-1}s_{1t}\left(\hat{\alpha}_n\right),$$

i.e. $s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)=s_{\tau_t}\left(\hat{\alpha}_n,\hat{\beta}_n\right)\equiv s_{2\tau_t}\left(\hat{\alpha}_n,\hat{\beta}_n\right)-\hat{F}_n\hat{A}_n^{-1}s_{1\tau_t}\left(\hat{\alpha}_n\right)$. Thus, $\hat{\beta}_{1,n}^{*}$ is a one-step bootstrap QMLE which updates $\hat{\beta}_n$ using the estimated Hessian $\hat{H}_n$ and the bootstrap scores $s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)$, evaluated at $\hat{\alpha}_n$ and $\hat{\beta}_n$.

A special case of $\hat{\beta}_{1,n}^{*}$ is a version of the one-step bootstrap QMLE considered in GW (2004) in the context of one-stage QMLE. In that paper, $\hat{\beta}_n$ does not depend on a first stage estimator $\hat{\alpha}_n$, implying that $s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right)=s_{2t}\left(\hat{\beta}_n\right)$ and $s_t\left(\hat{\alpha}_n,\hat{\beta}_n\right)=s_{2t}\left(\hat{\beta}_n\right)$. The only difference with respect to GW(2004) in this case is that our proposal only resamples the estimated scores and does not involve resampling the contributions to the Hessian $\hat{H}_n$ (instead, their one-step bootstrap QMLE involves resampling both; see also Davidson and MacKinnon (2004) and Andrews (2002) who proposed k-step bootstrap methods that resample the contributions to the Hessian and the score vector at each iteration, starting from the original estimators).

$\hat{\beta}_{1,n}^{*}$ is also related to a fast resampling approach proposed by Armstrong et al. (2014) in the context of a two-step GMM estimator with i.i.d. data, where the first step is a potentially nonparametric estimator (see also Chen et al. (2003) and Chen and Liao (2015)). In our context, it amounts to

$$\tilde{\beta}_{1,n}^{*}=\hat{\beta}_n-\hat{H}_n^{-1}n^{-1}\sum_{t=1}^{n}s_{2t}^{*}\left(\hat{\alpha}_n^{*},\hat{\beta}_n\right).$$

There are two main differences between $\hat{\beta}^*_{1,n}$ and $\tilde{\beta}^*_{1,n}$. First, $\tilde{\beta}^*_{1,n}$ requires computing $\hat{\alpha}^*_n$ whereas $\hat{\beta}^*_{1,n}$ does not. Hence, $\tilde{\beta}^*_{1,n}$ only avoids the computational burden of the second step and not of the first step. Instead, our method avoids computing $\hat{\alpha}^*_n$ for each resample and therefore is computationally more attractive. Second, $\tilde{\beta}^*_{1,n}$ resamples the scores of the second-stage model (evaluated at $\left(\hat{\alpha}^*_n, \hat{\beta}_n\right)$), whereas our method involves resampling $s_t\left(\hat{\alpha}_n, \hat{\beta}_n\right) = s_{2t}\left(\hat{\alpha}_n, \hat{\beta}_n\right) - \hat{F}_n \hat{A}_n^{-1} s_{1t}\left(\hat{\alpha}_n\right)$. We can think of this vector as an "extended" version of the scores for the second-stage, extended by the term $-\hat{F}_n \hat{A}_n^{-1} s_{1t}\left(\hat{\alpha}_n\right)$. This term corrects for the added uncertainty due to the first step. We note that it would not be valid to resample $s_{2t}\left(\hat{\alpha}_n, \hat{\beta}_n\right)$ unless $\hat{F}_n = 0$.

4 Bootstrap theory

We discuss two uses of the bootstrap for inference on β using $\hat{\beta}_n$. First, in Section 4.1 we consider using the bootstrap distribution of $\sqrt{n}\left(\hat{\beta}^*_n - \hat{\beta}_n\right)$ (or $\sqrt{n}\left(\hat{\beta}^*_{1,n} - \hat{\beta}_n\right)$) to approximate the quantiles of the distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$. This approach underlies the construction of percentile bootstrap intervals for β. Even though it does not promise asymptotic refinements, it is empirically attractive as it does not require computing any standard errors for $\hat{\beta}$. An alternative is to use the bootstrap to estimate standard errors, which we consider in Section 4.2.

4.1 Bootstrap distribution consistency

The first order asymptotic validity of the MBB based on the fully optimized bootstrap two-step QMLE $\hat{\beta}^*_n$ follows by showing that the bootstrap distribution of $\sqrt{n}\left(\hat{\beta}^*_n - \hat{\beta}_n\right)$ is consistent for the distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$.

This result requires we strengthen Assumption A as follows.

Assumption B For some $r > 2$ and some $\delta > 0$,

B.1: (i) $\{s_{1t}(\alpha_0)\}$ is $r+\delta$-dominated on $\mathcal{A}$ uniformly in t.
(ii) $\{s_{2t}(\alpha_0, \beta_0)\}$ is $r+\delta$-dominated on $\mathcal{A} \times \mathcal{B}$ uniformly in t.

B.2: $\{V_t\}$ is an α-mixing sequence of size $-\frac{(2+\delta)(r+\delta)}{r-2}$.

B.3: (i) The elements of $\{s_{1t}(\alpha)\}$ are $L_{2+\delta}$-NED on $\{V_t\}$ of size -1, uniformly on $(\mathcal{A}, \rho)$.
(ii) The elements of $\{s_{2t}(\alpha, \beta)\}$ are $L_{2+\delta}$-NED on $\{V_t\}$ of size -1, uniformly on $(\mathcal{A} \times \mathcal{B}, \rho)$.

B.4: (i) $n^{-1}\sum_{t=1}^n E\left(s_{1t}(\alpha_0)\right) E\left(s_{1t}(\alpha_0)\right)' = o\left(\ell_n^{-1}\right)$, where $\ell_n = o(n)$ and $\ell_n \to \infty$.
(ii) $n^{-1}\sum_{t=1}^n E\left(s_{2t}(\alpha_0, \beta_0)\right) E\left(s_{2t}(\alpha_0, \beta_0)\right)' = o\left(\ell_n^{-1}\right)$, where $\ell_n = o(n)$ and $\ell_n \to \infty$.

These assumptions are weaker than those used by GW (2004) (see their Assumption 2.1) and are sufficient to show that a bootstrap CLT applies to $\left\{s^*_{2t}(\alpha_0, \beta_0) - F_0 A_0^{-1} s^*_{1t}(\alpha_0)\right\}$, as shown by Gonçalves and de Jong (2003). Assumption B.1 requires a slight strengthening of the moment conditions on the scores $\{s_{1t}(\alpha)\}$ and $\{s_{2t}(\alpha, \beta)\}$ by comparison with Assumption A.5 (we now require slightly more than r moments versus r moments in A.5, where $r > 2$). Similarly, B.2 and B.3 strengthen the mixing and near epoch dependence assumptions stated in Assumption A.6 and A.7, respectively. In particular, we require the mixing coefficients on $\{V_t\}$ to be of size $-\frac{(2+\delta)(r+\delta)}{r-2}$ instead of $-\frac{2r}{r-2}$, and we require the scores $\{s_{1t}\}$ and $\{s_{2t}\}$ to be $L_{2+\delta}$-NED rather than L_2-NED. Assumption

B.4 is a restatement of Assumption 2.2 of Gonçalves and White (2002). As discussed by Gallant and White (1988), this assumption is satisfied when the models are correctly specified or when the scores $\{s_{1t}(X^t, \alpha_0)\}$ and $\{s_{2t}(X^t, \alpha_0, \beta_0)\}$ are stationary (this follows if $\{X_t\}$ is a strictly stationary process, the log-likelihood functions $\{f_{1t}(\alpha)\}$ and $\{f_{2t}(\alpha, \beta)\}$ depend only on a finite number of lags of X_t and there is no time heterogeneity on $\{f_{1t}\}$ and $\{f_{2t}\}$). Under this assumption, the bootstrap covariance matrix of the scaled average of $\{s^*_{2t}(\alpha_0, \beta_0) - F_0 A_0^{-1} s^*_{1t}(\alpha_0)\}$ converges to J_0, the correct asymptotic covariance matrix of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$.

In the following theorem, and throughout, we let E^*, Var^* and P^* denote the bootstrap expectation, variance and probability measure induced by the resampling, conditional on the original sample.

Theorem 4.1. *Let Assumption A as strengthened by Assumption B hold. If $\ell_n \to \infty$ and $\ell_n = o\left(n^{1/2}\right)$, then*

$$\sup_{x \in \mathbb{R}^p} \left| P^*\left(\sqrt{n}\left(\hat{\beta}^*_n - \hat{\beta}_n\right) \leq x\right) - P\left(\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \leq x\right) \right| = o_P(1). \tag{1}$$

To prove Theorem 4.1, we verify the conditions of Theorem A.4 in the online appendix. This result shows the consistency of the bootstrap distribution of a general two-step M estimator $\hat{\beta}^*_n$ (based on an asymptotically linear first step estimator $\hat{\alpha}^*_n$) under a set of bootstrap high level conditions (Assumption $\mathcal{B}^*$). We show that Assumption A strengthened by Assumption B verifies Assumption $\mathcal{B}^*$.

The first-order asymptotic validity of the fast resampling method is given in the next result. Its proof is a by-product of the proof of Theorem 4.1 and is omitted.

Theorem 4.2. *Under the same assumptions as in Theorem 4.1,*

$$\sup_{x \in \mathbb{R}^p} \left| P^*\left(\sqrt{n}\left(\hat{\beta}^*_{1,n} - \hat{\beta}_n\right) \leq x\right) - P\left(\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \leq x\right) \right| = o_P(1).$$

4.2 Bootstrap variance consistency

Bootstrap standard errors are often used in applied work as they are easy to compute, avoiding the need to look up complicated formulas. This is especially true in multistage estimation, where these formulas become quickly involved due to the need to keep track of the added uncertainty caused by each estimation stage. Instead, bootstrap standard errors are easily computed by Monte Carlo simulation. For instance, we can approximate the bootstrap variance estimator of the parameter $\hat{\beta}^*_{n,j}$, $Var^*\left(\hat{\beta}^*_{n,j}\right)$, with the sample variance obtained across B replications of $\hat{\beta}^*_{n,j}$,

$$\frac{1}{B}\sum_{k=1}^{B}\left(\hat{\beta}^{*(k)}_{n,j} - \overline{\hat{\beta}^{*(k)}_{n,j}}\right)^2, \text{ where } \overline{\hat{\beta}^{*(k)}_{n,j}} = \frac{1}{B}\sum_{k=1}^{B}\hat{\beta}^{*(k)}_{n,j}.$$

The corresponding bootstrap standard error is the square root of this expression.

The previous results (Theorems 4.1 and 4.2) do not justify by themselves the consistency of bootstrap standard errors based on $\hat{\beta}^*_n$ or $\hat{\beta}^*_{1,n}$. The reason is that convergence in distribution of a random sequence does not imply convergence of moments. For instance, Ghosh et al. (1984) and Shao (1992) give examples of the inconsistency of bootstrap variance estimators for the sample median and smooth functions of sample means, respectively, in the i.i.d. context.

The main goal of this section is to provide a theoretical justification for computing bootstrap standard errors in the context of two-step QMLE with time series data. The current bootstrap literature does not cover this case as it has either assumed i.i.d. data (as in Kato (2011) and Cheng

(2015), who prove the consistency of bootstrap variance estimators for one-step M-estimators) or has considered time series least squares estimators, as in Gonçalves and White (2005). No results appear to be available for multistage QMLE, even for i.i.d. data.

Given our previous bootstrap distribution consistency results, a sufficient condition for showing the consistency of the corresponding bootstrap standard errors is to show that a uniform integrability condition holds. In particular, to show that $Var^*\left(\sqrt{n}\hat{\beta}_n^*\right)$ is consistent, it suffices to show that $E^*\left|\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)\right|^{2+\delta} = O_P(1)$ for some small $\delta > 0$. Because $\hat{\beta}_{1,n}^*$ has a closed form expression, it is substantially easier to verify this condition for the fast resampling method than for the fully optimized bootstrap two-stage QMLE. For this reason, we focus on this estimator first.

We impose the following assumption.

Assumption B.4′ $E(s_{1t}(\alpha_0)) = 0$ and $E(s_{2t}(\alpha_0, \beta_0)) = 0$ for all $t = 1, \ldots, n$.

This is a mild strengthening of Assumption B.4, which is satisfied whenever the score functions are not heterogeneous and/or the data $\{X_t\}$ are stationary. We also impose a smoothness condition on the vector of scores $\{s_{1t}\}$ and $\{s_{2t}\}$ which we did not need for bootstrap distribution consistency.

Assumption B.5 (i) $\{s_{1t}(\alpha)\}$ is Lipschitz continuous on $\mathcal{A}$, a.s.-P with Lipschitz functions $\{L_{1t}\}$ that satisfy the condition $n^{-1}\sum_{t=1}^n E(L_{1t})^{2+\delta} = O(1)$.

(ii) $\{s_{2t}(\alpha, \beta)\}$ is Lipschitz continuous on $\mathcal{A} \times \mathcal{B}$, a.s.-$P$ with Lipschitz functions $\{L_{2t}\}$ satisfying the condition $n^{-1}\sum_{t=1}^n E(L_{2t})^{2+\delta} = O(1)$.

Theorem 4.3. *Under Assumptions A and B strengthened by B4′ and B5,* $Var^*\left(\sqrt{n}\hat{\beta}_{1,n}^*\right) \xrightarrow{P} H_0^{-1} J_0 H_0^{-1}$.

Next, we consider the fully optimized bootstrap estimator $\hat{\beta}_n^*$. Similarly to Theorem 4.3 , we prove the consistency of $Var^*\left(\sqrt{n}\hat{\beta}_n^*\right)$ by relying on Theorem 4.1 and showing that $E^*\left|\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)\right|^{2+\delta} = O_P(1)$ for some small $\delta > 0$. Because $\hat{\beta}_n^*$ does not have a closed form expression, this condition is much harder to verify than for $\hat{\beta}_{1,n}^*$ and requires a different method of proof and a different set of assumptions.

Our proof and regularity conditions are inspired by Kato (2011) and Cheng (2015). Kato (2011) shows the consistency of bootstrap moment estimators for M-estimators of parametric models, whereas Cheng (2015) allows for semiparametric models, where the parameter of interest is a finite dimensional parameter, but the model also contains a nuisance parameter that is potentially infinite dimensional. Both papers focus on one-step M-estimators and give sufficient conditions for bootstrap variance consistency that only cover i.i.d. data. Our contribution is to extend those results to multistage M-estimation with time series data.

To present our regularity conditions, we need to introduce more notation. First, because our proof is based on showing that the unconditional moment of $\left|\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)\right|^{2+\delta}$ is finite, we need to introduce the joint probability measure $\mathbb{P} = P \times P^*$ that accounts for the two sources of randomness in $\hat{\beta}_n^*$: the randomness that comes from the original data (and which is described by P) and the randomness that comes from the resampling, conditional on the original sample (described by P^*). In the following, we write $\mathbb{E}$ to denote expected value with respect to $\mathbb{P}$. Second, to prove that $\mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)\right|^{2+\delta} < \infty$, we assume the uniform square integrability of the original two-step QMLE estimator (i.e. we assume that $E\left|\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)\right|^{2+\delta} < \infty$) and provide regularity conditions that

allows us to show that $\mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right|^{2+\delta} < \infty$. We follow Kato (2011) and Cheng (2015) and use an argument that entails bounding the tail probability $\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right| > u\right)$ for large u. This requires empirical process theory and maximal inequalities. In particular, we impose bounds on the L_p-moments (with $p > 2+\delta$) of the supremum of certain empirical processes which we describe next.

For any class of functions $\mathcal{F} = \{f_t\}$, define the empirical process $\mathbb{G}_n f = n^{-1/2}\sum_{t=1}^n (f_t - Ef_t)$ and let its norm be given by $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$.

Our assumptions are as follows.

Assumption B.6

(i) For any $(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$, the log likelihood function $f_2(\cdot, \alpha, \beta)$ and its expectation $\bar{Q}_2(\alpha, \beta) \equiv E\left(\log f_2\left(X^t, \alpha, \beta\right)\right)$ are time invariant.

(ii) There exists a positive constant K independent of β for which for all $\beta \in \mathcal{B}$, $\bar{Q}_2(\alpha_0, \beta) - \bar{Q}_2(\alpha_0, \beta_0) \leq -K\|\beta - \beta_0\|^2$.

(iii) Given $\eta > 0$, define the class of functions

$$\mathcal{N}_\eta = \{\log f_2(\alpha, \beta) - \log f_2(\alpha, \beta_0) : \|\beta - \beta_0\| \leq \eta, (\alpha, \beta) \in \mathcal{A} \times \mathcal{B}\}.$$

Then, for some $p > 2+\delta$, and every $\eta > 0$, there exists a positive constant K such that

$$\left[E\left(\|\mathbb{G}_n\|_{\mathcal{N}_\eta}^p\right)\right]^{1/p} \leq K\eta, \tag{2}$$

and

$$\left(\ell^{-1}E_{Z,R}\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N Z_{R(i)}\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} \leq K\eta,$$

where $Z_i = \sum_{t=1}^{\ell}\left(f_{t+i-1} - E\left(f_{t+i-1}\right)\right)$, $f_{t+i-1} \in \mathcal{N}_\eta$, R denotes a random permutation uniformly distributed on Π_N, the set of permutations of $1, 2, \ldots, N = n - \ell + 1$, and $E_{Z,R}(\cdot)$ denotes the expectation with respect to $Z_1, \ldots, Z_N$ and R jointly.

(iv) The functions $\{\log f_2(\alpha, \beta)\}$, $\left\{\frac{\partial}{\partial\alpha'}\log f_2(\alpha, \beta)\right\}$ and $\left\{\frac{\partial}{\partial\alpha\partial\alpha'}\log f_2(\alpha, \beta)\right\}$ satisfy a Lipschitz continuity condition on $\mathcal{A} \times \mathcal{B}$, a.s.-$P$ with Lipschitz functions $\{L_t\}$, $\{L_{1t}\}$ and $\{L_{2t}\}$ such that $E|L_t|^p < \infty$, $E\left(|L_{1t}|^{\frac{\varepsilon}{\varepsilon-1}p}\right) < \infty$ and $E\left(|L_{2t}|^{\frac{\varepsilon}{\varepsilon-1}p}\right) < \infty$, respectively, for $p > 2+\delta$ as in (iii) and for some $\varepsilon > 1$.

(v) The first step estimator $\hat{\alpha}_n$ and its bootstrap analog $\hat{\alpha}_n^*$ are such that

$$E\left|\sqrt{n}\left(\hat{\alpha}_n - \alpha_0\right)\right|^{3\varepsilon p} = O(1) \text{ and } \mathbb{E}\left|\sqrt{n}\left(\hat{\alpha}_n^* - \hat{\alpha}_n\right)\right|^{3\varepsilon p} = O(1), \tag{3}$$

where $\varepsilon > 1$ and $p > 2+\delta$ are as defined in (iv).

(vi) $\sup_n E\left|\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)\right|^{2+\delta} < \infty$.

Assumption B.6 (i) assumes that the log-likelihood functions f_{2t} and the population criterion function $\bar{Q}_2(\alpha, \beta) \equiv E\left(\log f_2\left(X^t, \alpha, \beta\right)\right)$ are time invariant (the latter will follow from the first under stationarity of $\{X_t\}$). To understand Assumption B.6(ii), suppose that β is a scalar and assume that the function $\bar{Q}_2(\alpha_0, \beta)$ is twice differentiable with respect to β, e.g. $\bar{Q}_2(\alpha_0, \beta) = -E(y_t - \hat{x}_t\beta)^2$,

where $\hat{x}_t = x_t\alpha_0$, as in the generator regressor problem. Then, by a second-order Taylor expansion of $\bar{Q}_2(\alpha_0, \beta)$ around β_0, we get

$$\bar{Q}_2(\alpha_0, \beta) = \bar{Q}_2(\alpha_0, \beta_0) + \frac{\partial}{\partial\beta}\bar{Q}_2(\alpha_0, \beta_0)(\beta - \beta_0) + \frac{1}{2}\frac{\partial^2}{\partial\beta^2}\bar{Q}_2\left(\alpha_0, \ddot{\beta}\right)(\beta - \beta_0)^2,$$

where $\ddot{\beta}$ lies between β and β_0. Since (α_0, β_0) maximizes $\bar{Q}_2(\alpha, \beta)$, $\frac{\partial}{\partial\beta}\bar{Q}_2(\alpha_0, \beta_0) = 0$, implying that

$$\bar{Q}_2(\alpha_0, \beta) = \bar{Q}_2(\alpha_0, \beta_0) + \frac{1}{2}\frac{\partial^2}{\partial\beta^2}\bar{Q}_2\left(\alpha_0, \ddot{\beta}\right)(\beta - \beta_0)^2.$$

So, the condition will be satisfied if we can bound $\frac{\partial^2}{\partial\beta^2}\bar{Q}_2(\alpha_0, \beta)$ by a negative constant $-K$, for any value of β. For instance, this is true if $\bar{Q}_2(\alpha_0, \beta)$ is a quadratic function of β, as in the generator regressor problem. This is a strong condition since it imposes a global restriction on $\bar{Q}_2(\alpha_0, \beta)$, but it is crucial for controlling the tail probability $\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right| > u\right)$, as Kato (2011) and Cheng (2015) note. A similar condition is also used by Nishiyama (2010) to prove the moment convergence of the original M-estimator.

Assumption B.6(iii) (cf. equation (2)) is a high level condition on the empirical process $\mathbb{G}_n$. Cheng (2015) relies on a similar assumption to show the consistency of bootstrap one-step moment estimators of any order $p \geq 1$ for i.i.d. data. This so-called L_p-maximal inequality condition can be verified under more primitive conditions involving in particular the structure of the function class $\mathcal{N}_\eta$, e.g. Cheng (2015) shows that it is implied by a finite uniform entropy integral condition, which is verified when the functions in $\mathcal{N}_\eta$ are Lipschitz continuous. Our Assumption B.6(iii) adapts Cheng (2015)'s high level condition to the two-step QMLE context. It would be interesting to provide more primitive conditions that apply in our time series context.

Cheng (2015) also imposes a similar high level condition on $\mathbb{G}_n^*$, the bootstrap version of the empirical process $\mathbb{G}_n$, namely

$$\left[\mathbb{E}\left(\|\mathbb{G}_n^*\|_{\mathcal{N}_\eta}^p\right)\right]^{1/p} \leq K\eta, \tag{4}$$

a condition that he then verifies for several bootstrap methods, including the nonparametric i.i.d. bootstrap. Using arguments similar to those of Cheng (2015), we show in the online Appendix that the MBB satisfies this bootstrap maximal inequality condition under our assumptions. See Lemma A.1 in the online Appendix for this result.

Assumptions B.6(iv) and (v) are new to the two-step estimators we treat here. Part (iv) imposes a Lipschitz continuity condition on the score and the Hessian of $\log f_{2t}(\alpha, \beta)$ with respect to α. Part (v) imposes uniform integrability conditions on the first step estimator $\hat{\alpha}_n$ and its bootstrap analog $\hat{\alpha}_n^*$. Similarly, part (vi) assumes the uniform integrability condition on $\hat{\beta}_n$. These high level conditions could be derived from more primitive conditions such as the ones used by Cheng (2015) or Kato (2011), but we prefer to state them as high level conditions since our focus in on the second step bootstrap estimator $\hat{\beta}_n^*$. It is nevertheless interesting to note that stronger than usual uniform square integrability conditions on the first step estimators are imposed in order to verify the uniform square integrability condition on the second stage bootstrap estimator $\hat{\beta}_n^*$. In particular, we require the existence of a bit more than six moments for $\sqrt{n}(\hat{\alpha}_n - \alpha_0)$ and its bootstrap analogue. This is three times more than the number of moments for the second step estimators $\hat{\beta}_n$ and $\hat{\beta}_n^*$. When the log likelihood function f_{2t} is quadratic in α and β, Assumption B.6 (v) can be weaken as follows

$$E\left|\sqrt{n}(\hat{\alpha}_n - \alpha_0)\right|^{2\varepsilon p} = O(1) \text{ and } \mathbb{E}\left|\sqrt{n}(\hat{\alpha}_n^* - \hat{\alpha}_n)\right|^{2\varepsilon p} = O(1).$$

Under these assumptions, we can prove the following theorem.

Theorem 4.4. *Suppose Assumptions A and B strengthened by Assumption B6 holds. Then, for some* $\delta > 0$, $\sup_n \mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)\right|^{2+\delta} < \infty$, *implying that* $Var^*\left(\sqrt{n}\hat{\beta}_n^*\right) \xrightarrow{P} H_0^{-1} J_0 H_0^{-1}$.

5 Monte Carlo simulations

We here assess the properties of the bootstrap approximation proposed in Sections 3 and 4. We do so via detailed and realistic Monte Carlo simulations and we start by describing the design of the study. We consider a copula-based model. We focus on a bivariate distribution. Each variables marginal distribution is an AR(1)-GARCH(1,1) with standardized Student's t errors:

$$\begin{aligned} y_{it} &= \phi_{0,i} + \phi_{1,i} y_{i,t-1} + \varepsilon_{it} \\ \varepsilon_{it} &= \sigma_{it}\eta_{it} \\ \sigma_{it}^2 &= \widetilde{\omega}_i + \widetilde{\alpha}_i \varepsilon_{i,t-1}^2 + \widetilde{\beta}_i \sigma_{i,t-1}^2 \\ \eta_{it} &\sim iid\ t\left(0,1,\nu_i\right). \end{aligned}$$

We examine the case where the amount of dependence between the two variables y_1 and y_2 is related to the Clayton copula, with parameter $\beta = 1$, which roughly implies linear correlation of 0.5. See e.g., Nelsen (1999) and Patton (2012) for more on this copula. We use parameters similar to those found in applied work (Oh and Patton, 2013). Specifically, the parameters are set as follows:

$$\begin{aligned} \left[\phi_{0,i}, \phi_{1,i}\right] &= \left[0, 0.1\right], \text{ for } i = 1,2 \\ \left[\widetilde{\omega}_i, \widetilde{\alpha}_i, \widetilde{\beta}_i\right] &= \left[0.05, 0.05, 0.9\right], \text{ for } i = 1,2 \\ \nu_1 &= \nu_2 = \nu, \text{ such that } \nu \in \{6, 10, 30\}. \end{aligned}$$

Thus, we have three DGPs, which differ only in the value of the Students t parameter ν, which control the thickness of the tail of the distributions. Note that when $\nu \to \infty$, this implies that $\eta \sim N(0,1)$. We generate repeated trials of length $n \in \{200, 500, 2500\}$ from these processes and conduct bootstrap inference based on the fitted AR(1)-GARCH(1) model for each trial.

In the following, we define the marginal parameters, the copula parameter and the vector of all parameters as follows:

$$\begin{aligned} \text{Mean } i \text{ params } \phi_i &= \left[\phi_{0,i}, \phi_{1,i}\right]', \text{ for } i = 1,2 \\ \text{Vol } i \text{ params } \zeta_i &= \left[\widetilde{\omega}_i, \widetilde{\alpha}_i, \widetilde{\beta}_i\right]' \\ \text{All margin } i \text{ params } \alpha_i &\equiv \left[\phi_i', \zeta_i', \nu_i\right]' \\ \text{All params } \theta &\equiv \left[\alpha_1', \alpha_2', \beta\right]. \end{aligned}$$

It is easy to see that our bivariate density models constructed using copulas can be partitioned into elements relating only to a marginal distribution and elements that relate only to the copula. As pointed out by Joe (2005) and Patton (2006), when such a partition is not possible, the familiar one-stage maximum likelihood estimator is the natural estimator to employ. However, when this partitioning is possible as in our simulation setting, great computational savings may be achieved by employing a multi-stage estimator. Therefore, in the following we consider the multi-stage maximum likelihood estimator (MSMLE).

Our estimation steps are:

1. Estimate the conditional mean parameter ϕ_i using OLS (equivalent to using a QML with normal log likelihood and constant variance), conditioning on the realizations for $t = 1$. Obtain the estimated residuals $\hat{\varepsilon}_{it}$.

2. Estimate the conditional variance parameter ζ_i using QML (with normal log-likelihood) and the residuals from step 1, conditioning on the realizations for $t = 1$. Obtain the estimated standardized residuals $\hat{\eta}_{it}$.

3. Estimate ν using ML and the standardized residuals from step 2. Obtain the estimated probability integral transforms (PITs) $\hat{G}_{it}$.

4. Estimate β using ML and the estimated PITs from step 3.

More generally in a multivariate d-dimensional application (with $d \geq 2$) there are a total of $3d+1$ estimation steps: three steps for each marginal distribution, and 1 step for the copula.

To generate the bootstrap data, we use the moving blocks bootstrap. The number of Monte Carlo trials is 1,000 with $B = 999$ bootstrap replications each. We implement two resampling methods: the fully optimized bootstrap procedure BOOT1 and the fast resampling approach BOOT2. To select the block size, we rely on the asymptotic equivalence between the MBB and the Bartlett kernel variance estimators, and choose ℓ equal to the bandwidth chosen by Andrews's automatic procedure for the Bartlett kernel.

We consider two types of confidence intervals for the copula parameter β: asymptotic normal theory-based confidence intervals, computed by using the quantile of the standard normal distribution, and bootstrap percentile confidence intervals, which use the bootstrap methods (BOOT1 and BOOT2) to compute critical values for the non studentized statistics based on $\hat{\beta}_n$. The asymptotic normal theory-based confidence interval for β is given by

$$\hat{\beta}_n \pm 1.96 \cdot \widehat{SE}\left(\hat{\beta}_n\right), \tag{5}$$

where $\widehat{SE}\left(\hat{\beta}_n\right)$ is a consistent estimator of $SE\left(\hat{\beta}_n\right) = \sqrt{Var\left(\hat{\beta}_n\right)}$. Three choices are used to compute $\widehat{SE}\left(\hat{\beta}_n\right)$. Our first choice is infeasible in practice (but can be used for comparison): we set $\widehat{SE}\left(\hat{\beta}_n\right) = SE\left(\hat{\beta}_n\right)$, i.e., the true standard error of $\hat{\beta}_n$ (obtained via 10, 000 Monte Carlo replications).

For our second choice of $\widehat{SE}\left(\hat{\beta}_n\right)$, we use the multi-stage maximum likelihood (MSML) standard errors estimator as described in detail in Section 3.1.1 in Patton (2012) (cf. equation (41)). In particular, the asymptotic covariance matrix estimator of the MSMLE $\hat{\theta}_n$ is

$$\hat{V}^{\text{MSML}} = \widetilde{A}_n^{-1} \widetilde{B}_n \left(\widetilde{A}_n^{-1}\right)', \text{ where} \tag{6}$$

$$\widetilde{A}_n = n^{-1} \sum_{t=1}^{n} \hat{H}_t, \text{ with}$$

$$\hat{H}_t = \begin{bmatrix} \nabla^2_{11,t} & 0 & \cdots & 0 & 0 \\ 0 & \nabla^2_{22,t} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \nabla^2_{dd,t} & 0 \\ \nabla^2_{1c,t} & \nabla^2_{2c,t} & \cdots & \nabla^2_{dc,t} & \nabla^2_{cc,t} \end{bmatrix},$$

such that for $i = 1, \ldots, d$, and $t = 1, \ldots, n$,

$$\begin{aligned}
\underset{(6\times6)}{\nabla^2_{ii,t}} &\equiv \frac{\partial^2}{\partial\alpha_i\partial\alpha_i'}\log g_{it}\left(y_{it}, \hat{\alpha}_{i,n}\right) \\
\underset{(1\times6)}{\nabla^2_{ic,t}} &\equiv \frac{\partial^2}{\partial\beta\partial\alpha_i'}\log c_t\left(G_{1t}\left(y_{1t}, \hat{\alpha}_{1,n}\right), \ldots, G_{dt}\left(y_{dt}, \hat{\alpha}_{d,n}\right), \hat{\beta}_n\right) \\
\underset{(1\times1)}{\nabla^2_{cc,t}} &\equiv \frac{\partial^2}{\partial\beta\partial\beta'}\log c_t\left(G_{1t}\left(y_{1t}, \hat{\alpha}_{1,n}\right), \ldots, G_{dt}\left(y_{dt}, \hat{\alpha}_{d,n}\right), \hat{\beta}_n\right)
\end{aligned}$$

where $\log g_{it}$ and G_{it} are the complete log-likelihood and the CDF for y_{it}, respectively, whereas $\log c_t$ is the Clayton copula log-likelihood for $y_{1t}, \ldots, y_{dt}$. Notice that with the three-step estimation, the matrix $\nabla^2_{ii,t}$ is block diagonal, as the estimation of $\hat{\alpha}_i$ is done in stages

$$\begin{aligned}
\nabla^2_{ii,t} &= \begin{bmatrix} \nabla^2_{ii,m,t} & 0 & 0 \\ 0 & \nabla^2_{ii,v,t} & 0 \\ 0 & 0 & \nabla^2_{ii,p,t} \end{bmatrix} \\
\underset{(2\times2)}{\nabla^2_{ii,m,t}} &= \frac{\partial^2}{\partial\phi_i\partial\phi_i'}\log g^{(m)}_{it}\left(y_{it}, \hat{\phi}_{i,n}\right) \\
\underset{(3\times3)}{\nabla^2_{ii,v,t}} &= \frac{\partial^2}{\partial\zeta_i\partial\zeta_i'}\log g^{(v)}_{it}\left(\hat{\varepsilon}_{it}, \hat{\zeta}_{i,n}\right) \\
\underset{(1\times1)}{\nabla^2_{ii,p,t}} &= \frac{\partial^2}{\partial\nu_i^2}\log g^{(p)}_{it}\left(\hat{\eta}_{it}, \hat{\nu}_{i,n}\right)
\end{aligned}$$

where: $\log g^{(m)}_{it}$ is the normal log-likelihood with constant variance, $\log g^{(v)}_{it}$ is the normal log-likelihood used for QML estimation of the GARCH model and $\log g^{(p)}_{it}$ is the log-likelihood of the standardized Student's t distribution.

The $\widetilde{B}_n$ matrix is given as follows:

$$\begin{aligned}
\widetilde{B}_n &= n^{-1}\sum_{t=1}^{n}\hat{s}_t\hat{s}_t' + n^{-1}\sum_{h=1}^{l}\left(1 - \frac{h}{l+1}\right)\sum_{t=h+1}^{n}\left(\hat{s}_t\hat{s}_{t-h}' + \hat{s}_{t-h}\hat{s}_t'\right), \text{ where} \qquad (7) \\
\hat{s}_t &= \left[\hat{s}_{1t}, \ldots, \hat{s}_{dt}, \hat{s}_{ct}\right]' \\
\hat{s}_{it} &= \frac{\partial}{\partial\alpha_i}\log g_{it}\left(y_{it}, \hat{\alpha}_{i,n}\right) \\
\hat{s}_{ct} &= \frac{\partial}{\partial\beta}\log c_t\left(G_{1t}\left(y_{1t}, \hat{\alpha}_{1,n}\right), \ldots, G_{dt}\left(y_{dt}, \hat{\alpha}_{d,n}\right), \hat{\beta}_n\right).
\end{aligned}$$

Specifically, in our simulations to compute $\widetilde{B}_n$, we use a Bartlett kernel with bandwidth selected by the data-based rule (i.e., automatic procedure) from Andrews (1991).

We construct a MSMLE variance, asymptotic normal theory-based confidence interval for β as in (5) by using $\widehat{SE}\left(\hat{\beta}_n\right) = \widehat{SE}^{\text{MSML}}\left(\hat{\beta}_n\right)$, where $\widehat{SE}^{\text{MSML}}\left(\hat{\beta}_n\right)$ is the estimated standard error of $\hat{\beta}_n$, obtained via equation (6) (which has a sandwich form). Specifically, in our setting $\widehat{SE}^{\text{MSML}}\left(\hat{\beta}_n\right) = n^{-1/2}\sqrt{\hat{V}^{\text{MSML}}_{13,13}}$, where $\hat{V}^{\text{MSML}}_{13,13}$ is the element $(13,13)$ of $\hat{V}^{\text{MSML}}$.

In our third and fourth choices of $\widehat{SE}\left(\hat{\beta}_n\right)$, we use the proposed bootstrap approaches BOOT1 and BOOT2. In particular, a fully optimized bootstrap procedure variance, asymptotic normal

theory-based confidence interval for β can be obtained as in (5) with $\widehat{SE}\left(\hat{\beta}_n\right) = \widehat{SE}^{\text{BOOT1}}\left(\hat{\beta}_n\right)$, where $\widehat{SE}^{\text{BOOT1}}\left(\hat{\beta}_n\right)$ is the estimated standard error of $\hat{\beta}_n$ based on BOOT1. Similarly, a fast re-sampling procedure variance, asymptotic normal theory-based confidence interval for β is obtained by using $\widehat{SE}\left(\hat{\beta}_n\right) = \widehat{SE}^{\text{BOOT2}}\left(\hat{\beta}_n\right)$,where $\widehat{SE}^{\text{BOOT2}}\left(\hat{\beta}_n\right)$ is the estimated standard error of $\hat{\beta}_n$ based on BOOT2 (the fast resampling approach). Note that, the standard errors $\widehat{SE}^{\text{BOOT1}}\left(\hat{\beta}_n\right)$ and $\widehat{SE}^{\text{BOOT2}}\left(\hat{\beta}_n\right)$ are obtained by computing the statistics $\sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\beta}_n^{*(i)} - \overline{\hat{\beta}_n^{*}}\right)^2}$ with $\overline{\hat{\beta}_n^{*}} = \frac{1}{B}\sum_{j=1}^{B}\hat{\beta}_n^{*(j)}$, where B is the number of bootstrap replications.

The second set of intervals we consider are bootstrap percentile confidence intervals, which are very simple to compute since they avoid the need to explicitly compute standard errors. For each resampling approach (BOOT1 and BOOT2), a bootstrap percentile confidence intervals for β is given by $\hat{\beta}_n \pm q_{95}^{*}$, where q_{95}^{*} is the 95% quantile of the bootstrap distribution of $|\hat{\beta}_n^{*} - \hat{\beta}_n|$.

Table 1 gives the actual rates of 95% confidence intervals of the copula parameter β for the three DGPs, respectively. Results in Table 1 are not too sensitive to the value of the Students t parameter ν. All intervals have approximately the desired coverage rate, and we see only small differences among them. In particular, the fast resampling procedure BOOT2 and the MSML approach perform well even for the small sample size $n = 200$ (with coverage rate almost equal to the nominal). The fact that we do not have dynamic misspecification in our models explains why we do not get larger distortions for the smaller sample sizes using the asymptotic-based intervals. Indeed, the evidence of no serial correlation in the scores is confirmed by the average value of the block sizes chosen by Andrews (1991) method, which is equal to 1.80 in our simulations. However, as Table 2 suggests, there are notable differences among the different methods when considering their confidence interval lengths. This table clearly shows that the intervals based on BOOT1 (either using the CLT-based or the bootstrap percentile approach) tend to mimic the lengths of the CLT-Inf intervals for all DGP's and sample sizes, and both tend to display shorter intervals for the smaller sample sizes compared to CLT-MSML and the BOOT2 intervals.

Note that all three asymptotic normal theory-based confidence intervals differ only by the way that the estimated standard errors of $\hat{\beta}_n$ have been computed. In order to gain further insight into the "relatively" good performance of these asymptotic normal theory-based confidence intervals in finite samples, we compute the ratio of the estimated standard error over the true value and the mean-square error (MSE) of the estimated standard errors. The results are presented in Table 3. For small sample sizes, on average MSML and BOOT2 overestimate the standard errors, with the ratio of estimated standard error over the true value above 1. For instance, when $n = 200$ and $\nu = 10$, the ratio of estimated standard error over the true value based on MSML and BOOT2 are 1.21 and 1.22 for MSML and BOOT2, respectively, whereas this ratio is 1.01 for BOOT1. Consequently, the length of confidence intervals based on estimated standard error from MSML and BOOT2 are larger than the one based on BOOT1.

Table 1: Coverage Rates of nominal 95% intervals for β

	CLT-Inf	CLT-MSML	CLT-BOOT1	CLT-BOOT2	BOOT1	BOOT2
			$\nu = 6$			
$n = 200$	95.10	93.40	93.20	93.60	93.50	95.00
$n = 500$	94.90	94.30	94.20	94.30	94.30	94.70
$n = 2500$	94.20	93.60	93.80	93.50	94.00	93.90
			$\nu = 10$			
$n = 200$	95.00	93.40	94.60	93.90	94.70	94.50
$n = 500$	94.80	93.90	93.40	93.90	93.70	94.30
$n = 2500$	93.60	93.90	93.50	93.00	93.60	94.20
			$\nu = 30$			
$n = 200$	94.50	95.00	95.00	95.20	95.00	96.10
$n = 500$	95.50	93.80	93.50	93.90	93.40	95.20
$n = 2500$	94.70	93.70	94.20	93.70	94.40	94.90

Notes: CLT-Inf, CLT-MSML, CLT-BOOT1, and CLT-BOOT2 -intervals based on the normal using estimated standard error based on the true standard error, the MSML, the BOOT1 and the BOOT2, respectively; BOOT1 bootstrap percentile intervals based on the fully optimized procedure, BOOT2 bootstrap percentile intervals based on the fast resampling procedure. 1,000 Monte Carlo trials with 999 bootstrap replications each.

Table 2: Length confidence intervals of nominal 95% intervals for β

	CLT-Inf	CLT-MSML	CLT-BOOT1	CLT-BOOT2	BOOT1	BOOT2
			$\nu = 6$			
$n = 200$	0.63	0.74	0.64	0.75	0.64	0.77
$n = 500$	0.40	0.40	0.40	0.41	0.41	0.42
$n = 2500$	0.18	0.18	0.18	0.18	0.18	0.18
			$\nu = 10$			
$n = 200$	0.64	0.77	0.64	0.78	0.65	0.82
$n = 500$	0.42	0.42	0.41	0.42	0.41	0.43
$n = 2500$	0.18	0.18	0.18	0.18	0.18	0.19
			$\nu = 30$			
$n = 200$	0.65	0.75	0.65	0.76	0.65	0.80
$n = 500$	0.41	0.41	0.41	0.41	0.41	0.43
$n = 2500$	0.19	0.18	0.18	0.18	0.18	0.19

Notes: CLT-Inf, CLT-MSML, CLT-BOOT1, and CLT-BOOT2 -intervals based on the normal using estimated standard error based on the true standard error, the MSML, the BOOT1 and the BOOT2, respectively; BOOT1 bootstrap percentile intervals based on the fully optimized procedure, BOOT2 bootstrap percentile intervals based on the fast resampling procedure. 1,000 Monte Carlo trials with 999 bootstrap replications each.

Table 3: Comparison of standard errors estimation of $\hat{\beta}_n$

	$n = 200$			$n = 500$			$n = 2500$		
	MSML	BOOT1	BOOT2	MSML	BOOT1	BOOT2	MSML	BOOT1	BOOT2
				$\nu = 6$					
(MSE of estimated SE)$\cdot 10^3$	29.70	0.43	30.21	1.76	0.07	1.72	0.007	0.004	0.007
Ratio of SE over the true value	1.17	1.01	1.18	1.00	1.00	1.00	0.97	0.99	0.97
				$\nu = 10$					
(MSE of estimated SE)$\cdot 10^3$	121.18	0.38	125.15	3.14	0.08	3.09	0.005	0.004	0.006
Ratio of SE over the true value	1.21	1.01	1.22	1.00	0.97	1.00	0.97	1.00	0.98
				$\nu = 30$					
(MSE of estimated SE)$\cdot 10^3$	35.71	0.38	35.97	0.73	0.07	0.70	0.012	0.004	0.013
Ratio of SE over the true value	1.16	1.01	1.17	1.00	0.99	1.00	0.97	0.99	0.97

Notes: This table provides the MSE of estimated standard errors and the ratio of estimated standard errors over the true value (simulation based) of standard errors. We compute the ratio of standard error of an estimator $\hat{\theta}$ over the true value as: $S^{-1}\sum_{i=1}^{S}\frac{\hat{SE}_i^j(\hat{\theta})}{SE(\hat{\theta})}$, where S is the number of Monte Carlo replications, $i = 1, \ldots, S$, j =MSML, BOOT1, BOOT2 thus $\hat{SE}_i^j\left(\hat{\theta}\right)$ is the estimated value of the standard error of an estimator $\hat{\theta}$ on the ith Monte Carlo replication obtained by using the method j. $SE\left(\hat{\theta}\right)$ is defined as $SE\left(\hat{\theta}\right) = \sqrt{S^{-1}\sum_{i=1}^{S}\left(\hat{\theta}_i - S^{-1}\sum_{s=1}^{S}\hat{\theta}_s\right)^2}$, with $\hat{\theta}_i$ the estimated value of the parameter θ on the ith Monte Carlo replication. Similarly, we compute the MSE as $S^{-1}\sum_{i=1}^{S}\left(\hat{SE}_i^j\left(\hat{\theta}\right) - SE\left(\hat{\theta}\right)\right)^2$. Simulations where done with 1,000 Monte Carlo trials with 999 bootstrap replications each.

The gains associated with the bootstrap methods can be quite substantial when the main goal of the researcher and/or practitioner is to estimate the standard errors. Results in Table 3 are in favor of the bootstrap particularly for small sample sizes. More specifically, the full resampling method BOOT1 is better than using MSML and/or BOOT2 standard errors. For small samples, the bootstrap method BOOT1 estimates the standard error of the copulas parameter estimator $\hat{\beta}_n$ more precisely than the MSML and BOOT2 approaches. For large sample size, we have approximately the same performance for all three methods. For instance, when $n = 200$ and $\nu = 6$, the MSE of the estimated standard errors of $\hat{\beta}_n$ based on MSML, BOOT1 and BOOT2 are 0.03, $0.43 \cdot 10^{-3}$ and 0.03, respectively. Whereas, for $n = 2,500$ and $\nu = 6$, the MSE become $7.2 \cdot 10^{-6}$, $4.4 \cdot 10^{-6}$ and $7.5 \cdot 10^{-6}$, respectively. Thus we see that although all three methods MSML, BOOT1 and BOOT2 are asymptotically equivalent, and the full resampling method BOOT1 may be computationally much more demanding, in small samples, the improved estimates of the standard errors based on BOOT1 may outweigh the computational cost of the later. Overall, the performance of BOOT2 (the fast resampling method) is comparable to that of MSML, whereas BOOT1 outperforms BOOT2 and MSML and provides more accurate estimators of the standard errors, specifically, when the sample size is small.

6 Conclusions

This paper proposes and theoretically justifies bootstrap methods for inference on nonlinear dynamic models that are estimated by two (or more) steps of quasi-maximum likelihood. In particular, we show the consistency of the bootstrap distribution of the two-step QMLE using dependence and heterogeneity conditions similar to those used by Gonçalves and White (2004) for the one-step QMLE. In addition, we also prove the consistency of bootstrap standard errors for the two-step QMLE, a result that does not seem to be available even for i.i.d. data. This justifies the standard practice of computing bootstrap standard errors instead of computing analytical standard errors, which quickly become cumbersome in the multistage QML context. Our simulation results show that intervals based on bootstrap standard errors or bootstrap percentile intervals obtained with the fully optimized method that resamples the log likelihood functions jointly are shorter on average than intervals based only on asymptotic theory or on the fast resampling method we propose. Thus, although more computationally demanding, the fully optimized bootstrap method has better finite sample properties than the other methods we consider.

This supplementary appendix is organized as follows. First, we provide a set of primitive assumptions under which the asymptotic theory of the two-step QMLE (consistency and asymptotic distribution) follows. A set of definitions useful to understand our assumptions is also provided. Next, we provide asymptotic theory and bootstrap theory for general two-stage M estimators under a set of high level conditions (which include uniform laws of large numbers, central limit theorems and an asymptotic linear representation for $\hat{\alpha}_n$ and $\hat{\alpha}_n^*$). These results are instrumental in proving the results of Section 3. Then, we provide the proofs of our results. Finally, we provide two auxiliary lemmas used in the proof of Theorem 4.4, followed by their proofs.

A.1 Assumptions

We start by providing a set of definitions which are useful to understand our assumptions.

Definition 1. *We define $\{X_t\}$ to be L_q-NED on a mixing process $\{V_t\}$ if $E(X_t^q) < \infty$ and $v_k \equiv \sup_t \left\| X_t - E_{t-k}^{t+k}(X_t) \right\|_q \to 0$ as $k \to \infty$. Here, $\|X_t\|_p \equiv (E|X_t|^p)^{1/p}$ is the L_p norm and $E_{t-k}^{t+k}(\cdot) \equiv E\left(\cdot|\mathcal{F}_{t-k}^{t+k}\right)$, where $\mathcal{F}_{t-k}^{t+k} \equiv \sigma(V_{t-k}, \ldots, V_{t+k})$ is the σ-field generated by $V_{t-k}, \ldots, V_{t+k}$. If $v_k = O\left(k^{-a-\delta}\right)$ for some $\delta > 0$, we say $\{X_t\}$ is L_q-NED of size $-a$.*

Definition 2. $\{V_t\}$ *is strong mixing if*

$$\lambda_k \equiv \sup_m \sup_{\{A\in\mathcal{F}_{-\infty}^m, B\in\mathcal{F}_{m+k}^\infty\}} |P(A\cap B) - P(A)P(B)| \to 0$$

as $k\to\infty$ *suitably fast.*

Definition 3. *A random function* $f:\mathcal{X}\times\Theta\to\mathbb{R}$ *is Lipschitz continuous on* Θ *a.s.-*P *if for all* θ_1 *and* $\theta_2\in\Theta$, $|f_t(x,\theta_1) - f_t(x,\theta_2)| \le L_t(x)|\theta_1-\theta_2|$ *for all* x *in a set with probability one, for some function* $L_t(x)$ *such that* $\sup_n\left\{n^{-1}\sum_{t=1}^n E(L_t(x))\right\} = O(1)$.

Definition 4. *A sequence of random functions* $\{f_t:\mathcal{X}\times\Theta\to\mathbb{R}\}$ *is* $r-$*dominated on* Θ *uniformly in* t *if there exists* $D_t:\mathcal{X}\to\mathbb{R}$ *such that* $|f_t(x,\theta)|\le D_t(x)$ *for all* $\theta\in\Theta$ *and* D_t *is measurable such that* $\|D_t\|_r\le\Delta<\infty$ *for all* t.

Definition 5. *A sequence of random functions* $\{f_t:\mathcal{X}\times\Theta\to\mathbb{R}\}$ *is* L_q*-NED on* $\{V_t\}$ *of size* $-a$ *on* (Θ,ρ) *if for each* $\theta_0\in\Theta$ *there exists* $\delta_0>0$ *such that the random sequences* $\left\{\bar{f}_t(\delta) = \sup_{\eta^0(\delta)} f_t(x,\theta)\right\}$ *and* $\left\{\underline{f}_t(\delta) = \inf_{\eta^0(\delta)} f_t(x,\theta)\right\}$ *are* L_q*-NED on* $\{V_t\}$ *of size* $-a$ *for all* $0<\delta\le\delta_0$, *where* $\eta^0(\delta) = \{\theta\in\Theta:\rho(\theta,\theta_0)<\delta\}$.

In the following and throughout the appendix, K denotes a constant, which may change from line to line and from (in)equality to (in)equality.

The following set of assumptions extends the assumptions of GW (2004) to the two-step QMLE context and are used to prove our bootstrap results.

Assumption A

A.1: Let $(\Omega,\mathcal{F},P)$ be a complete probability space. The observed data are a realization of a stochastic process $\left\{X_t:\Omega\to\mathbb{R}^l,\ t\in\mathbb{N}\right\}$, with

$$X_t(\omega) = W_t(\ldots,V_{t-1}(\omega),V_t(\omega),V_{t+1}(\omega),\ldots),$$

$V_t:\Omega\to\mathbb{R}^v$, and $W_t:\times_{\tau=-\infty}^\infty\mathbb{R}^v\to\mathbb{R}^l$ is such that X_t is measurable for t.

A.2: (i) The functions $\left\{f_{1t}\left(X^t,\alpha\right)\right\}$ are such that $f_{1t}(\cdot,\alpha)$ is measurable for each $\alpha\in\mathcal{A}$, where $\mathcal{A}$ is a compact subset of $\mathbb{R}^k$, $f_{1t}\left(X^t,\cdot\right)$ is continuous on $\mathcal{A}$, a.s.-P, and $f_{1t}\left(X^t,\cdot\right)$ is twice continuously differentiable on $int(\mathcal{A})$, a.s-P.

(ii) The functions $\left\{f_{2t}\left(X^t,\alpha,\beta\right)\right\}$ are such that $f_{2t}(\cdot,\alpha,\beta)$ is measurable for each $(\alpha,\beta)\in\mathcal{A}\times\mathcal{B}$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^p$, $f_{2t}\left(X^t,\cdot,\cdot\right)$ is continuous on $\Theta=\mathcal{A}\times\mathcal{B}$, a.s.-$P$, and $f_{2t}\left(X^t,\cdot,\cdot\right)$ is twice continuously differentiable on $int(\Theta)$, a.s-P.

A.3: (i) α_0 is the unique maximizer of $\bar{Q}_1(\alpha)\equiv\lim_{n\to\infty}E(Q_{1n}(\alpha))$ on $\mathcal{A}$.

(ii) β_0 is the unique maximizer of $\bar{Q}_2(\alpha_0,\beta)\equiv\lim_{n\to\infty}E(Q_{2n}(\alpha_0,\beta))$ on $\mathcal{B}$.

(iii) $\theta_0=(\alpha_0,\beta_0)$ is interior to $\Theta=\mathcal{A}\times\mathcal{B}$.

A.4: (i) The functions $\left\{\log f_{1t}\left(X^t,\alpha\right)\right\}$ and $\left\{\frac{\partial}{\partial\alpha'}s_{1t}\left(X^t,\alpha\right)\right\}$ are Lipschitz continuous on $\mathcal{A}$, a.s.-P, where $s_{1t}\left(X^t,\alpha\right)\equiv\frac{\partial}{\partial\alpha}\log f_{1t}\left(X^t,\alpha\right)$.

(ii) The functions $\left\{\log f_{2t}\left(X^t,\alpha,\beta\right)\right\}$, $\left\{\frac{\partial}{\partial\beta'}s_{2t}\left(X^t,\alpha,\beta\right)\right\}$ and $\left\{\frac{\partial}{\partial\alpha'}s_{2t}\left(X^t,\alpha,\beta\right)\right\}$ are Lipschitz continuous on $\mathcal{A}\times\mathcal{B}$, a.s.-$P$, where $s_{2t}\left(X^t,\alpha,\beta\right)\equiv\frac{\partial}{\partial\beta}\log f_{2t}\left(X^t,\alpha,\beta\right)$.

A.5: For some $r > 2$,

(i) The functions $\{\log f_{1t}(X^t, \alpha)\}$, $\{s_{1t}(X^t, \alpha)\}$ and $\{\frac{\partial}{\partial \alpha'} s_{1t}(X^t, \alpha)\}$ are r-dominated on $\mathcal{A}$ uniformly in t.

(ii) The functions $\{\log f_{2t}(X^t, \alpha, \beta)\}$, $\{s_{2t}(X^t, \alpha, \beta)\}$, $\left\{\frac{\partial}{\partial \beta'} s_{2t}(X^t, \alpha, \beta)\right\}$ and $\{\frac{\partial}{\partial \alpha'} s_{2t}(X^t, \alpha, \beta)\}$ are r-dominated on $\Theta = \mathcal{A} \times \mathcal{B}$ uniformly in t.

A.6: $\{V_t\}$ is an α-mixing sequence of size $-\frac{2r}{r-2}$, with $r > 2$.

A.7: The elements of (i) $\{\log f_{1t}(X^t, \alpha)\}$ and $\{\frac{\partial}{\partial \alpha'} s_{1t}(X^t, \alpha)\}$ are L_2-NED on $\{V_t\}$ of size $-\frac{1}{2}$, and those of $\{s_{1t}(X^t, \alpha)\}$ are L_2-NED on $\{V_t\}$ of size -1, uniformly on $(\mathcal{A}, \rho)$, where ρ is a metric on $\mathbb{R}^k$;

(ii) $\{\log f_{2t}(X^t, \alpha, \beta)\}$, $\left\{\frac{\partial}{\partial \beta'} s_{2t}(X^t, \alpha, \beta)\right\}$ and $\{\frac{\partial}{\partial \alpha'} s_{2t}(X^t, \alpha, \beta)\}$ are L_2-NED on $\{V_t\}$ of size $-\frac{1}{2}$, and those of $\{s_{2t}(X^t, \alpha, \beta)\}$ are L_2-NED on $\{V_t\}$ of size -1, uniformly on $(\mathcal{A} \times \mathcal{B}, \rho)$, where ρ is a metric on $\mathbb{R}^k \times \mathbb{R}^p$.

A.8: (i) $A_0 \equiv \lim_{n\to\infty} E\left(n^{-1} \sum_{t=1}^n \frac{\partial}{\partial \alpha'} s_{1t}(X^t, \alpha_0)\right)$ is nonsingular and
$B_0 \equiv \lim_{n\to\infty} Var\left(n^{-\frac{1}{2}} \sum_{t=1}^n s_{1t}(X^t, \alpha_0)\right)$ is positive definite.

(ii) $H_0 \equiv \lim_{n\to\infty} E\left(n^{-1} \sum_{t=1}^n \frac{\partial}{\partial \beta'} s_{2t}(X^t, \alpha_0, \beta_0)\right)$ is nonsingular,

$$J_0 \equiv \lim_{n\to\infty} Var\left(n^{-\frac{1}{2}} \sum_{t=1}^n \left(s_{2t}(X^t, \alpha_0, \beta_0) - F_0 A_0^{-1} s_{1t}(X^t, \alpha_0)\right)\right)$$

is positive definite, and $F_0 \equiv \lim_{n\to\infty} E\left(n^{-1} \sum_{t=1}^n \frac{\partial}{\partial \alpha'} s_{2t}(X^t, \alpha_0, \beta_0)\right) < \infty$.

A.2 General results for two-step M-estimators

In this section, we provide results for a general two-step M estimator $\hat{\beta}_n$ based on a first step estimator $\hat{\alpha}_n$ which has an asymptotic linear representation. Specifically, in the first step, we estimate $\alpha_0 \in \mathcal{A} \subset \mathbb{R}^k$ with some asymptotically linear estimator $\hat{\alpha}_n$ (which does not need to be an M estimator; e.g. it could be a GMM estimator). In the second step, we estimate β_0 with

$$\hat{\beta}_n = \arg\min_{\beta \in \mathcal{B}} Q_{2n}(\hat{\alpha}_n, \beta),$$

where

$$Q_{2n}(\hat{\alpha}_n, \beta) \equiv n^{-1} \sum_{t=1}^n q_{2t}(X^t, \hat{\alpha}_n, \beta),$$

and $q_{2t} : \mathbb{R}^{lt} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is an objective function that depends on β and α and $X^t \equiv (X_1, \ldots, X_{t-1}, X_t)$. The two-step QMLE of Section 3 is a special case of $\hat{\beta}_n$ when $q_{2t}(X^t, \hat{\alpha}_n, \beta) = -\log f_{2t}(X^t, \hat{\alpha}_n, \beta)$, where f_{2t} denotes the conditional likelihood function of X_t given X^{t-1}, and $\hat{\alpha}_n$ is also a QMLE.

We follow White (1994) and Wooldridge (1994) and provide a set of high level conditions that allow us to derive general results.

Assumption $\mathcal{A}$.

$\mathcal{A}.1$ Let $(\Omega, \mathcal{F}, P)$ be a complete probability space. The observed data are a realization of a stochastic process $\{X_t : \Omega \to \mathbb{R}^l,\ t \in \mathbb{N}\}$.

$\mathcal{A}.2$ The functions $\left\{q_{2t}\left(X^t,\alpha,\beta\right)\right\}$ are such that $q_{2t}\left(\cdot,\alpha,\beta\right)$ is measurable for each $(\alpha,\beta)\in\mathcal{A}\times\mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are compact subsets of $\mathbb{R}^k$ and $\mathbb{R}^p$, respectively, and $q_{2t}\left(x^t,\cdot,\cdot\right)$ is continuous on $\mathcal{A}\times\mathcal{B}$ for all x^t in some set F_t with $P\left(F_t\right)=1$.

$\mathcal{A}.3$ (i) $\hat{\alpha}_n\to^P\alpha_0\in int\left(\mathcal{A}\right)$.

(ii) $\sqrt{n}\left(\hat{\alpha}_n-\alpha_0\right)=n^{-1/2}\sum_{t=1}^n\psi_t\left(X^t,\alpha_0\right)+o_P\left(1\right)$, for some function $\left\{\psi_t\left(X^t,\alpha_0\right)\right\}$ such that $\sqrt{n}\left(\hat{\alpha}_n-\alpha_0\right)=O_P\left(1\right)$.

$\mathcal{A}.4$ (i) $\bar{Q}_2\left(\alpha,\beta\right)\equiv\lim_{n\to\infty}E\left(Q_{2n}\left(\alpha,\beta\right)\right)$ exists and is continuous on $\mathcal{A}\times\mathcal{B}$.

(ii) β_0 is the unique minimizer of $\bar{Q}_2\left(\alpha_0,\beta\right)\equiv\lim_{n\to\infty}E\left(Q_{2n}\left(\alpha_0,\beta\right)\right)$ on $\mathcal{B}$.

(iii) $\beta_0\in int\left(\mathcal{B}\right)$.

$\mathcal{A}.5$ $\left\{q_{2t}\left(X^t,\alpha,\beta\right)\right\}$ satisfies a weak ULLN on $\mathcal{A}\times\mathcal{B}$ (i.e. $\sup_{\alpha,\beta}\left|Q_{2n}\left(\alpha,\beta\right)-\bar{Q}_2\left(\alpha,\beta\right)\right|=o_P\left(1\right)$).

$\mathcal{A}.6$ (i) $\left\{q_{2t}\left(X^t,\alpha,\beta\right)\right\}$ is twice continuously differentiable on $int\left(\mathcal{A}\right)\times int\left(\mathcal{B}\right)$.

(ii) The functions $\left\{\frac{\partial}{\partial\alpha'}\varphi_{2t}\left(X^t,\alpha,\beta\right)\right\}$ and $\left\{\frac{\partial}{\partial\beta'}\varphi_{2t}\left(X^t,\alpha,\beta\right)\right\}$ satisfy a weak ULLN on $\mathcal{A}\times\mathcal{B}$, where $\varphi_{2t}\left(X^t,\alpha,\beta\right)\equiv\frac{\partial}{\partial\beta}q_{2t}\left(X^t,\alpha,\beta\right)$.

$\mathcal{A}.7$ (i) $H_0\equiv\lim_{n\to\infty}E\left(n^{-1}\sum_{t=1}^n\frac{\partial}{\partial\beta'}\varphi_{2t}\left(X^t,\alpha_0,\beta_0\right)\right)>0$.

(ii) $F_0\equiv\lim_{n\to\infty}E\left(n^{-1}\sum_{t=1}^n\frac{\partial}{\partial\alpha'}\varphi_{2t}\left(X^t,\alpha_0,\beta_0\right)\right)<\infty$.

$\mathcal{A}.8$ The function $\left\{\varphi_{2t}\left(X^t,\alpha_0,\beta_0\right)+F_0\psi_t\left(X^t,\alpha_0\right)\right\}$ satisfies the CLT, i.e.

$$n^{-1/2}\sum_{t=1}^n\left(\varphi_{2t}\left(X^t,\alpha_0,\beta_0\right)+F_0\psi_t\left(X^t,\alpha_0\right)\right)\to^d N\left(0,J_0\right),$$

where

$$J_0\equiv\lim_{n\to\infty}Var\left(n^{-1/2}\sum_{t=1}^n\left(\varphi_{2t}\left(X^t,\alpha_0,\beta_0\right)+F_0\psi_t\left(X^t,\alpha_0\right)\right)\right)>0.$$

Assumption $\mathcal{A}.3$(ii) assumes that $\hat{\alpha}_n$ admits an asymptotic linear representation, which includes not only M-estimators but also other estimators such as GMM estimators.

Theorem A.1. *Under Assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, $\mathcal{A}.3$(i), $\mathcal{A}.4$(i)-(ii) and $\mathcal{A}.5$, $\hat{\beta}_n\to^P\beta_0$.*

Theorem A.2. *Under Assumptions $\mathcal{A}.1-\mathcal{A}.8$, $\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\to^d N\left(0,H_0^{-1}J_0H_0^{-1}\right)$.*

Theorems A.1 and A.2 are well known in the literature (see e.g. White (1994), Newey and McFadden (1994) and Wooldridge (1994)) and are only given here for completeness, but their proof is omitted for brevity.

Next, we provide a set of general conditions for bootstrap validity. Suppose that the bootstrap two-step M-estimator is defined as

$$\hat{\beta}_n^*=\arg\min_{\beta\in\mathcal{B}}Q_{2n}^*\left(\hat{\alpha}_n^*,\beta\right),$$

where $\hat{\alpha}_n^*$ is the first-step bootstrap analogue of $\hat{\alpha}_n$, and

$$Q_{2n}^*\left(\hat{\alpha}_n^*,\beta\right)\equiv n^{-1}\sum_{t=1}^n q_{2t}^*\left(X^{*t},\hat{\alpha}_n^*,\beta\right),$$

and where for each $\beta \in \mathcal{B}$, we let $q_{2t}^*\left(X^{*t}, \hat{\alpha}_n^*, \beta\right) = q_{2,\tau_t}\left(X^{\tau_t}, \hat{\alpha}_n^*, \beta\right)$ with τ_t denoting a set of indices chosen by the bootstrap. The first step bootstrap estimator $\hat{\alpha}_n^*$ is not necessarily an M-estimator. All we require in Assumption $\mathcal{B}^*$ below is that it has an asymptotic linear representation of the same type as $\hat{\alpha}_n$ but with $\psi_t\left(X^t, \alpha_0\right)$ replaced with $\psi_t^*\left(X^{*t}, \hat{\alpha}_n\right) = \psi_{\tau_t}\left(X^{\tau_t}, \hat{\alpha}_n\right)$. Thus, both $\hat{\alpha}_n^*$ and $\hat{\beta}_n^*$ depend on the same set of bootstrap indices $\{\tau_t\}$.

Assumption $\mathcal{B}^*$

$\mathcal{B}^*.1$ (i) $\hat{\alpha}_n^* - \hat{\alpha}_n \to^{P^*} 0$, in prob-$P$.
(ii) $\sqrt{n}\left(\hat{\alpha}_n^* - \hat{\alpha}_n\right) = n^{-1/2}\sum_{t=1}^n \psi_t^*\left(X^{*t}, \hat{\alpha}_n\right) + o_{P^*}(1)$, in prob-$P$.

$\mathcal{B}^*.2$ The functions $\left\{q_{2t}^*\left(X^{*t}, \alpha, \beta\right)\right\}$ satisfy a bootstrap ULLN on $\mathcal{A} \times \mathcal{B}$, i.e.

$$\sup_{\alpha,\beta} \left|Q_{2n}^*(\alpha, \beta) - Q_{2n}(\alpha, \beta)\right| \to^{P^*} 0,$$

in prob-P.

$\mathcal{B}^*.3$ The functions $\left\{\frac{\partial}{\partial \alpha'}\varphi_{2t}^*\left(X^{*t}, \alpha, \beta\right)\right\}$ and $\left\{\frac{\partial}{\partial \beta'}\varphi_{2t}^*\left(X^{*t}, \alpha, \beta\right)\right\}$ satisfy a bootstrap ULLN on $\mathcal{A}\times\mathcal{B}$, where $\varphi_{2t}^*\left(X^{*t}, \alpha, \beta\right) \equiv \frac{\partial}{\partial \beta} q_{2t}^*\left(X^{*t}, \alpha, \beta\right)$.

$\mathcal{B}^*.4$ $n^{-1/2}\sum_{t=1}^n \left(\varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n, \hat{\beta}_n\right) + F_0\psi_t^*\left(X^{*t}, \hat{\alpha}_n\right)\right) \to^{d^*} N(0, J_0)$, in prob-$P$, where

$$J_0 \equiv \lim_{n\to\infty} Var\left(n^{-1/2}\sum_{t=1}^n \left(\varphi_{2t}\left(X^t, \alpha_0, \beta_0\right) + F_0\psi_t\left(X^t, \alpha_0\right)\right)\right) > 0.$$

Assumption $\mathcal{B}^*$ imposes high level conditions on the bootstrap first step estimator and on the bootstrap second step objective function and its derivatives. These conditions can be verified for any particular bootstrap method used to obtain $\hat{\alpha}_n^*$ and $\hat{\beta}_n^*$, where $\hat{\beta}_n^*$ is a QMLE estimator and $\hat{\alpha}_n^*$ is any estimator admitting an asymptotic linear representation (as specified by Assumption $\mathcal{B}^*.2$). We verify these conditions for the two-step QMLE studied in Section 3.

Theorem A.3. *Suppose Assumptions $\mathcal{A}.1$, $\mathcal{A}.2$, $\mathcal{A}.3$(i), $\mathcal{A}.4$(i)-(ii) hold. If in addition Assumptions $\mathcal{B}^*.1$(i) and $\mathcal{B}^*.2$ are satisfied, then $\hat{\beta}_n^* - \hat{\beta}_n \to^{P^*} 0$, in prob-$P$.*

Theorem A.4. *Suppose Assumptions $\mathcal{A}.1 - \mathcal{A}.8$ hold. If in addition Assumptions $\mathcal{B}^*.1 - \mathcal{B}^*.4$ are satisfied, then $\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right) \to^{d^*} N\left(0, H_0^{-1}J_0H_0^{-1}\right)$, in prob-$P$.*

Theorems A.2 and A.4 imply that

$$\sup_{x\in\mathbb{R}^p} \left|P^*\left(\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right) \le x\right) - P\left(\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \le x\right)\right| \to^P 0,$$

as $n \to \infty$, thus justifying the use of the bootstrap distribution of $\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right)$ as a consistent estimator of the distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$.

A.3 Proofs of Theorems A.3, A.4, 4.1, 4.3 and 4.4

Proof of Theorem A.3. Let $\tilde{Q}_n(\beta) = Q_{2n}(\hat{\alpha}_n, \beta) = n^{-1}\sum_{t=1}^n q_{2t}\left(X^t, \hat{\alpha}_n, \beta\right)$. We apply Lemma A.2 of GW (2004) with $Q_n(\cdot, \theta) = \tilde{Q}_n(\beta)$. We can easily verify that $\tilde{Q}_n(\beta)$ satisfies the first part of this lemma, implying that $\hat{\beta}_n \to^P \beta_0$. Next, we verify that the function

$$\tilde{Q}_n^*(\beta) = Q_{2n}^*(\hat{\alpha}_n^*, \beta)$$

satisfies the second part of Lemma A.2. First, note that $\hat{\beta}_n^* = \arg\max_\beta \tilde{Q}_n^*(\beta)$, where $\tilde{Q}_n^*(\beta)$ satisfies the measurability and continuity assumptions given in particular Assumptions $\mathcal{A}.2$. Therefore, the result follows if we show that

$$\sup_{\beta\in\mathcal{B}}\left|\tilde{Q}_n^*(\beta) - \tilde{Q}_n(\beta)\right| \to^{P^*} 0, \text{ prob-}P.$$

To see that this is the case, note that

$$\begin{aligned}
\sup_{\beta\in\mathcal{B}}\left|\tilde{Q}_n^*(\beta) - \tilde{Q}_n(\beta)\right| &= \sup_{\beta\in\mathcal{B}}\left|Q_{2n}^*(\hat{\alpha}_n^*, \beta) - Q_{2n}(\hat{\alpha}_n, \beta)\right| \\
&\leq \sup_{\beta\in\mathcal{B}}\left|Q_{2n}^*(\hat{\alpha}_n^*, \beta) - Q_{2n}(\hat{\alpha}_n^*, \beta)\right| + \sup_{\beta\in\mathcal{B}}\left|Q_{2n}(\hat{\alpha}_n^*, \beta) - \bar{Q}_2(\hat{\alpha}_n^*, \beta)\right| \\
&\quad + \sup_{\beta\in\mathcal{B}}\left|Q_{2n}(\hat{\alpha}_n, \beta) - \bar{Q}_2(\hat{\alpha}_n, \beta)\right| + \sup_{\beta\in\mathcal{B}}\left|\bar{Q}_2(\hat{\alpha}_n^*, \beta) - \bar{Q}_2(\hat{\alpha}_n, \beta)\right| \\
&\leq \sup_{\alpha\in\mathcal{A},\beta\in\mathcal{B}}\left|Q_{2n}^*(\alpha, \beta) - Q_{2n}(\alpha, \beta)\right| + 2\sup_{\alpha\in\mathcal{A},\beta\in\mathcal{B}}\left|Q_{2n}(\alpha, \beta) - \bar{Q}_2(\alpha, \beta)\right| \\
&\quad + \sup_{\beta\in\mathcal{B}}\left|\bar{Q}_2(\hat{\alpha}_n^*, \beta) - \bar{Q}_2(\hat{\alpha}_n, \beta)\right|.
\end{aligned}$$

The first two terms are $o_{P^*}(1)$ and $o_P(1)$, respectively, given $\mathcal{B}^*.2$ and $\mathcal{A}.5$. The third term is $o_{P^*}(1)$ in prob-P, given the fact that $\bar{Q}_2(\alpha, \beta)$ is continuous on $\mathcal{A}\times\mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are compact subsets of finite dimensional Euclidean spaces, and the fact that that $\hat{\alpha}_n^* - \hat{\alpha}_n \to^{P^*} 0$, in prob-$P$ by Assumption $\mathcal{B}^*.1$.

Proof of Theorem A.4. By a mean value expansion of $n^{-1/2}\sum_{t=1}^n \varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n^*, \hat{\beta}_n^*\right)$ around $\hat{\beta}_n$,

$$0 = n^{-1/2}\sum_{t=1}^n \varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n^*, \hat{\beta}_n\right) + \left[n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\beta'}\varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n^*, \ddot{\beta}_n^*\right)\right]\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right),$$

where $\ddot{\beta}_n^*$ lies between $\hat{\beta}_n^*$ and $\hat{\beta}_n$. A second mean value expansion of $n^{-1/2}\sum_{t=1}^n \varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n^*, \hat{\beta}_n\right)$ around $\hat{\alpha}_n$ yields

$$\begin{aligned}
0 &= n^{-1/2}\sum_{t=1}^n \varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n, \hat{\beta}_n\right) + \left[n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\alpha'}\varphi_{2t}^*\left(X^{*t}, \ddot{\alpha}_n^*, \hat{\beta}_n\right)\right]\sqrt{n}\left(\hat{\alpha}_n^* - \hat{\alpha}_n\right) \\
&\quad + \left[n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\beta'}\varphi_{2t}^*\left(X^{*t}, \hat{\alpha}_n^*, \ddot{\beta}_n^*\right)\right]\sqrt{n}\left(\hat{\beta}_n^* - \hat{\beta}_n\right),
\end{aligned}$$

where $\ddot{\alpha}_n^*$ lies between $\hat{\alpha}_n^*$ and $\hat{\alpha}_n$. By a ULLN applied to $\frac{\partial}{\partial\alpha'}\varphi_{2t}^*\left(X^{*t}, \alpha, \beta\right)$ and $\frac{\partial}{\partial\beta'}\varphi_{2t}^*\left(X^{*t}, \alpha, \beta\right)$ (Assumption $\mathcal{B}^*.3$), we have that

$$n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\alpha'}\varphi_{2t}^*\left(X^{*t}, \ddot{\alpha}_n^*, \hat{\beta}_n\right) - n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\alpha'}\varphi_{2t}\left(X^t, \alpha_0, \beta_0\right) \to^{P^*} 0, \text{ in prob-}P,$$

which implies that

$$n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\alpha'}\varphi_{2t}^{*}\left(X^{*t},\ddot{\alpha}_{n}^{*},\hat{\beta}_{n}\right)\rightarrow^{P^{*}}F_{0},\text{ in prob-}P,$$

since $\hat{\alpha}_{n}^{*}\rightarrow^{P^{*}}\alpha_{0}$, $\hat{\beta}_{n}\rightarrow^{P}\beta_{0}$, and $n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\alpha'}\varphi_{2t}\left(X^{t},\alpha_{0},\beta_{0}\right)\rightarrow^{P}F_{0}$. Similarly,

$$n^{-1}\sum_{t=1}^{n}\frac{\partial}{\partial\beta'}\varphi_{2t}^{*}\left(X^{*t},\hat{\alpha}_{n}^{*},\ddot{\beta}_{n}^{*}\right)\rightarrow^{P^{*}}H_{0},\text{ in prob-}P,$$

since $\hat{\alpha}_{n}^{*}\rightarrow^{P^{*}}\alpha_{0}$ and $\hat{\beta}_{n}^{*}\rightarrow^{P^{*}}\beta_{0}$. It follows that

$$0=n^{-1/2}\sum_{t=1}^{n}\varphi_{2t}^{*}\left(X^{*t},\hat{\alpha}_{n},\hat{\beta}_{n}\right)+F_{0}\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\hat{\alpha}_{n}\right)+H_{0}\sqrt{n}\left(\hat{\beta}_{n}^{*}-\hat{\beta}_{n}\right)+o_{P^{*}}\left(1\right).$$

By Assumption $\mathcal{B}^{*}$.1(ii),

$$\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\hat{\alpha}_{n}\right)=n^{-1/2}\sum_{t=1}^{n}\psi_{t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)+o_{P^{*}}\left(1\right),$$

which implies that

$$0=n^{-1/2}\sum_{t=1}^{n}\varphi_{2t}^{*}\left(X^{*t},\hat{\alpha}_{n},\hat{\beta}_{n}\right)+F_{0}\left(n^{-1/2}\sum_{t=1}^{n}\psi_{t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)\right)+H_{0}\sqrt{n}\left(\hat{\beta}_{n}^{*}-\hat{\beta}_{n}\right)+o_{P^{*}}\left(1\right).$$

Hence,

$$\sqrt{n}\left(\hat{\beta}_{n}^{*}-\hat{\beta}_{n}\right)=-H_{0}^{-1}n^{-1/2}\sum_{t=1}^{n}\left(\varphi_{2t}^{*}\left(X^{*t},\hat{\alpha}_{n},\hat{\beta}_{n}\right)+F_{0}\psi_{t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)\right)+o_{P^{*}}\left(1\right).$$

The result now follows from Assumption $\mathcal{B}^{*}$.4.

Proof of Theorem 4.1. We verify that the high level conditions of Theorem A.4 are satisfied for the two-step QMLE under Assumption A as strengthened by Assumption B. In particular, we can show that Assumption $\mathcal{B}^{*}$.1(i) is satisfied for $\hat{\alpha}_{n}^{*}=\arg\max_{\alpha}Q_{1n}^{*}\left(\alpha\right)\equiv n^{-1}\sum_{t=1}^{n}\log f_{1t}^{*}\left(X^{*t},\alpha\right)$ by relying on GW (2004)'s Theorem 2.1 under Assumption A.1., A.6 and part (i) of Assumptions A.2-A.5 and A.7, A.8. Similarly, we can apply Theorem 2.2 of GW (2004) to conclude that $\mathcal{B}^{*}$.1(ii) is verified with $\psi_{t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)=-A_{0}^{-1}s_{1t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)$. To verify Assumption $\mathcal{B}^{*}$.2, we let $q_{2t}^{*}\left(X^{*t},\alpha,\beta\right)=-\log f_{2t}^{*}\left(X^{*t},\alpha,\beta\right)$ and apply Lemmas A.4 and A.5 of GW (2004). Assumptions A.4(ii) and A.5(ii) together with the requirement that $\ell_{n}=o\left(n\right)$ suffice to prove that $\mathcal{B}^{*}$.2 holds. $\mathcal{B}^{*}$.3 can be verified similarly by showing that a bootstrap ULLN applies to the derivatives of $s_{2t}^{*}\left(X^{*t},\alpha,\beta\right)$ with respect to α and β under A.4(ii) and A.5(ii) and the rate condition on the block size ℓ_{n}. Finally, to check that the bootstrap CLT (cf. Assumption $\mathcal{B}^{*}$.4) holds for $s_{t}^{*}\left(\hat{\alpha}_{n},\hat{\beta}_{n}\right)\equiv\varphi_{2t}^{*}\left(X^{*t},\hat{\alpha}_{n},\hat{\beta}_{n}\right)+F_{0}\psi_{t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)=-s_{2t}^{*}\left(X^{*t},\hat{\alpha}_{n},\hat{\beta}_{n}\right)+F_{0}A_{0}^{-1}s_{1t}^{*}\left(X^{*t},\hat{\alpha}_{n}\right)$ we proceed as in the proof of Theorem 2.2 of GW (2004). Specifically, we write

$$-n^{-1/2}\sum_{t=1}^{n}s_{t}^{*}\left(\hat{\alpha}_{n},\hat{\beta}_{n}\right)=n^{-1/2}\sum_{t=1}^{n}\left(s_{2t}^{*}\left(\hat{\alpha}_{n},\hat{\beta}_{n}\right)-F_{0}A_{0}^{-1}s_{1t}^{*}\left(\hat{\alpha}_{n}\right)\right)\equiv\xi_{1n}+\xi_{2n}+\xi_{3n}+\xi_{4n},$$

with

$$\begin{aligned}
\xi_{1n} &= n^{-1/2}\sum_{t=1}^{n}\left(\left(s_{2t}^{*}\left(\alpha_0,\beta_0\right)-F_0A_0^{-1}s_{1t}^{*}\left(\alpha_0\right)\right)-\left(s_{2t}\left(\alpha_0,\beta_0\right)-F_0A_0^{-1}s_{1t}\left(\alpha_0\right)\right)\right);\\
\xi_{2n} &= n^{-1/2}\sum_{t=1}^{n}\left(s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right)-s_{2t}\left(\alpha_0,\beta_0\right)\right)-F_0A_0^{-1}n^{-1/2}\sum_{t=1}^{n}\left(s_{1t}\left(\hat{\alpha}_n\right)-s_{1t}\left(\alpha_0\right)\right);\\
\xi_{3n} &= n^{-1/2}\sum_{t=1}^{n}\left(s_{2t}^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)-s_{2t}^{*}\left(\alpha_0,\beta_0\right)\right)-F_0A_0^{-1}n^{-1/2}\sum_{t=1}^{n}\left(s_{1t}^{*}\left(\hat{\alpha}_n\right)-s_{1t}^{*}\left(\alpha_0\right)\right);\\
\xi_{4n} &= n^{-1/2}\sum_{t=1}^{n}s_{2t}\left(\hat{\alpha}_n,\hat{\beta}_n\right)-F_0A_0^{-1}n^{-1/2}\sum_{t=1}^{n}s_{1t}\left(\hat{\alpha}_n\right).
\end{aligned}$$

By arguing exactly as in GW (2004), we can show that under Assumption A strengthened by Assumption B, $\xi_{1n}\to^{d^*}N\left(0,J_0\right)$, in prob-$P$, and $\xi_{2n}+\xi_{3n}=o_{P^*}\left(1\right)$ in prob-P, whereas $\xi_{4n}=o_P\left(1\right)$ by the first order conditions that define $\hat{\alpha}_n$ and $\hat{\beta}_n$.

Proof of Theorem 4.3. For some small $\delta>0$,

$$E^*\left|\sqrt{n}\left(\hat{\beta}_{1,n}^{*}-\hat{\beta}_n\right)\right|^{2+\delta}\leq\left\|\hat{H}_n^{-1}\right\|_1^{2+\delta}E^*\left|n^{-1/2}\sum_{t=1}^{n}s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)\right|^{2+\delta},$$

where $\|A\|_1$ is the spectral norm of a matrix A, i.e. $\|A\|_1^2=\max_{x\neq0}\frac{x'A'Ax}{x'x}$. Since $\hat{H}_n$ is a symmetric matrix, $\left\|\hat{H}_n^{-1}\right\|_1^{2+\delta}=\left(\lambda_{\min}^{-1}\left(\hat{H}_n\right)\right)^{2+\delta}=O_P\left(1\right)$ since $\lambda_{\min}\left(\hat{H}_n\right)\to^P\lambda_{\min}\left(H_0\right)\neq0$ by the assumption that H_0 is nonsingular. Thus, it suffices to show that $E^*\left|n^{-1/2}\sum_{t=1}^{n}s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)\right|^{2+\delta}=O_P\left(1\right)$. Using the definition of s_t^{*}, we can decompose this expectation as

$$\begin{aligned}
&E^*\left|n^{-1/2}\sum_{t=1}^{n}s_t^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)\right|^{2+\delta}\\
\leq\;&E^*\left|n^{-1/2}\sum_{t=1}^{n}s_{2t}^{*}\left(\hat{\alpha}_n,\hat{\beta}_n\right)\right|^{2+\delta}+\left\|\hat{F}_n\right\|^{2+\delta}\left\|\hat{A}_n^{-1}\right\|_1^{2+\delta}E^*\left|n^{-1/2}\sum_{t=1}^{n}s_{1t}^{*}\left(\hat{\alpha}_n\right)\right|^{2+\delta}.
\end{aligned}$$

Each of the bootstrap expectations on the RHS of the display can be shown to be $O_P\left(1\right)$ under our assumptions. The arguments are similar to those used by GW (2005). Take e.g. the second of these expectations. Adding and subtracting appropriately, we can bound it by I_1+I_2, where $I_1=2^{1+\delta}E^*\left|n^{-1/2}\sum_{t=1}^{n}s_{1t}^{*}\left(\alpha_0\right)\right|^{2+\delta}$ and $I_2=2^{1+\delta}E^*\left|n^{-1/2}\sum_{t=1}^{n}\left(s_{1t}^{*}\left(\hat{\alpha}_n\right)-s_{1t}^{*}\left(\alpha_0\right)\right)\right|^{2+\delta}$. Under Assumption B4′, $E\left(s_{1t}\left(\alpha_0\right)\right)=0$ and we can show that $\left\{s_{1t}\left(\alpha_0\right)\right\}$ is $L_{2+\delta}$-mixingale with bounded mixingale constants and absolutely summable coefficients given in particular the $L_{2+\delta}$-NED assumption on the score function $s_{1t}\left(\alpha\right)$. Hence, by Lemma A.1 of GW (2005), we have that $E\left(I_1\right)=O\left(1\right)+O\left(\left(\frac{\ell_n^2}{n}\right)^{(2+\delta)/2}\right)=O\left(1\right)$ since $\ell_n^2/n\to0$ by assumption. To show that $I_2=O_P\left(1\right)$, we rely on Assumption B5, the Lipschitz continuity assumption on $s_{1t}\left(\alpha\right)$. This assumption implies that

$$E^*\left|n^{-1/2}\sum_{t=1}^{n}\left(s_{1t}^{*}\left(\hat{\alpha}_n\right)-s_{1t}^{*}\left(\alpha_0\right)\right)\right|^{2+\delta}\leq\left(n^{-1}\sum_{t=1}^{n}E^*\left|L_{1t}^{*}\right|^{2+\delta}\right)\left|\sqrt{n}\left(\hat{\alpha}_n-\alpha_0\right)\right|^{2+\delta},$$

where $\left|\sqrt{n}\left(\hat{\alpha}_n-\alpha_0\right)\right|^{2+\delta}=O_P\left(1\right)$ and

$$n^{-1}\sum_{t=1}^{n}E^*\left|L_t^{*}\right|^{2+\delta}=n^{-1}\sum_{t=1}^{n}\left|L_{1t}\right|^{2+\delta}+O_P\left(\frac{\ell_n}{n}\right).$$

where $n^{-1}\sum_{t=1}^{n} E\left|L_{1t}\right|^{2+\delta} = O(1)$ under Assumption B5. The proof that

$$E^*\left|n^{-1/2}\sum_{t=1}^{n} s_{2t}^*\left(\hat{\alpha}_n, \hat{\beta}_n\right)\right|^{2+\delta} = O_P(1)$$

follows under similar arguments.

Proof of Theorem 4.4. The result follows from the triangle inequality if

$$\sup_n \mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right|^{2+\delta} < \infty \text{ and } \sup_n E\left|\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)\right|^{2+\delta} < \infty.$$

The moment condition on $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ holds by assumption. Then, the moment condition on $\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)$ follows by an argument similar to that used in Kato (2011). In particular, note that for any positive random variable Z and any $q \geq 1$, we can write $E\left|Z\right|^q = q\int_0^\infty t^{q-1}P(Z > t)\,dt$. Hence,

$$\mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right|^{2+\delta} = (2+\delta)\int_0^\infty t^{2+\delta-1}\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right| > t\right)dt.$$

We will show that $\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right| > t\right) \leq Kt^{-p}$ for $p > 2+\delta$ and some constant K. This will imply the result since

$$\mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right|^{2+\delta} \leq K\int_0^\infty t^{2+\delta-p-1}dt < \infty \text{ if } p > 2+\delta.$$

Let $\tilde{Q}_n(\beta) = Q_{2n}(\hat{\alpha}_n, \beta) = n^{-1}\sum_{t=1}^{n} q_{2t}\left(X^t, \hat{\alpha}_n, \beta\right)$, such that $q_{2t}\left(X^t, \hat{\alpha}_n, \beta\right) = \log f_{2t}\left(X^t, \hat{\alpha}_n, \beta\right)$. Note that $\hat{\beta}_n^* = \arg\max_\beta \tilde{Q}_n^*(\beta)$, where

$$\tilde{Q}_n^*(\beta) = Q_{2n}^*(\hat{\alpha}_n^*, \beta).$$

Partition the parameter space $\mathcal{B}$ into "shells" $S_{j,n} = \left\{\beta \in \mathcal{B} : 2^{j-1} < \left|\sqrt{n}(\beta - \beta_0)\right| \leq 2^j\right\}$ for any integer $j \geq 1$. If $\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right|$ is larger than 2^{j_0} for a given integer j_0, then $\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right|$ is in one of the shells $S_{j,n}$ with $j \geq j_0$. In that case, the supremum of the map $\beta \longmapsto \tilde{Q}_n^*(\beta) - \tilde{Q}_n^*(\beta_0)$ must be nonnegative by the definition of $\hat{\beta}_n^*$. This implies

$$\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^* - \beta_0\right)\right| > 2^{j_0}\right) \leq \sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}\left\{\tilde{Q}_n^*(\beta) - \tilde{Q}_n^*(\beta_0)\right\} \geq 0\right). \tag{8}$$

Next decompose $\tilde{Q}_n^*(\beta) - \tilde{Q}_n^*(\beta_0)$ as follows:

$$\begin{aligned}
\tilde{Q}_n^*(\beta) - \tilde{Q}_n^*(\beta_0) &= \left[Q_{2n}^*(\hat{\alpha}_n^*, \beta) - Q_{2n}^*(\hat{\alpha}_n^*, \beta_0)\right] - \left[Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0)\right] \\
&\quad + Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0) - E^*\left[Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0)\right] \\
&\quad + E^*\left[Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0)\right] - E\left(E^*\left[Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0)\right]\right) \\
&\quad + E\left[E^*\left(Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0)\right)\right] \\
&\equiv I_{2\text{-step},n}(\beta) + I_{1,n}(\beta) + I_{2,n}(\beta) + I_{3,n}(\beta).
\end{aligned}$$

Note that

$$\begin{aligned}
E^*\left(Q_{2n}^*(\alpha_0, \beta) - Q_{2n}^*(\alpha_0, \beta_0)\right) &= E^*\left(n^{-1}\sum_{t=1}^{n} q_{2t}^*(\alpha_0, \beta) - q_{2t}^*(\alpha_0, \beta_0)\right) \\
&= \sum_{t=1}^{n}\gamma_{nt}\left(q_{2t}(\alpha_0, \beta) - q_{2t}(\alpha_0, \beta_0)\right),
\end{aligned}$$

where the weighting function γ_{nt} is defined as

$$\gamma_{nt} = \begin{cases} \frac{t}{\ell_n(n-\ell_n+1)}, & \text{if} \quad t \in \{1,\ldots,\ell_n\} \\ \frac{1}{n-\ell_n+1}, & \text{if} \quad i \in \{\ell_n+1,\ldots,n-\ell_n\} \\ \frac{n-t+1}{\ell_n(n-\ell_n+1)}, & \text{if} \quad i \in \{n-\ell_n+1,\ldots,n\} \end{cases},$$

such that $\sum_{t=1}^{n}\gamma_{nt}=1$. It follows that

$$I_{3n}(\beta) = \sum_{t=1}^{n}\gamma_{nt}E\left(q_{2t}(\alpha_0,\beta)-q_{2t}(\alpha_0,\beta_0)\right) = \bar{Q}_2(\alpha_0,\beta)-\bar{Q}_2(\alpha_0,\beta_0),$$

given the time homogeneity of the moments $E(q_{2t}(\alpha,\beta))$ (which is part of Assumption B6(i)) and the fact that $\sum_{t=1}^{n}\gamma_{nt}=1$. By the quadratic behavior assumption, we can conclude that $-I_{3,n}(\beta) \geq K|\beta-\beta_0|^2 \geq K\frac{2^{2j-2}}{n}$ on $S_{j,n}$, for some $K>0$. Then, for each j the following inclusion holds

$$\begin{aligned} & \left\{\sup_{\beta\in S_{j,n}}\left\{\tilde{Q}_n^*(\beta)-\tilde{Q}_n^*(\beta_0)\right\}\geq 0\right\} \\ \subset & \left\{\sup_{\beta\in S_{j,n}}|I_{2\text{-step},n}(\beta)|+\sup_{\beta\in S_{j,n}}|I_{1,n}(\beta)|+\sup_{\beta\in S_{j,n}}|I_{2,n}(\beta)|\geq K\frac{2^{2j-2}}{n}\right\}. \end{aligned}$$

It follows that the right-hand side of (8) i.e., $\sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}\left\{\tilde{Q}_n^*(\beta)-\tilde{Q}_n^*(\beta_0)\right\}\geq 0\right)$ can be bounded by

$$\begin{aligned} & \sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}|I_{2\text{-step},n}(\beta)|\geq K\frac{2^{2(j-1)}}{n}\right) \\ & +\sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}|I_{1,n}(\beta)|\geq K\frac{2^{2(j-1)}}{n}\right)+\sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}|I_{2,n}(\beta)|\geq K\frac{2^{2(j-1)}}{n}\right). \end{aligned}$$

Thus, by Markov's inequality (with $p>2+\delta$) we have

$$\begin{aligned} & \sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}\left\{\tilde{Q}_n^*(\beta)-\tilde{Q}_n^*(\beta_0)\right\}\geq 0\right) \\ \leq & K\left[\begin{array}{c} \sum_{j=j_0}^{\infty}\left(\frac{2^{2(j-1)}}{n}\right)^{-p}\mathbb{E}\left(\sup_{\beta\in S_{j,n}}|I_{2\text{-step},n}(\beta)|^p\right) \\ +\sum_{j=j_0}^{\infty}\left(\frac{2^{2(j-1)}}{n}\right)^{-p}\mathbb{E}\left(\sup_{\beta\in S_{j,n}}|I_{1,n}(\beta)|^p\right)+\sum_{j=j_0}^{\infty}\left(\frac{2^{2(j-1)}}{n}\right)^{-p}\mathbb{E}\left(\sup_{\beta\in S_{j,n}}|I_{2,n}(\beta)|^p\right) \end{array}\right] \\ \leq & K\left[\begin{array}{c} \sum_{j=j_0}^{\infty}2^{-2pj}n^p\mathbb{E}\left(\sup_{\beta\in S_{j,n}}|I_{2\text{-step},n}(\beta)|^p\right) \\ +\sum_{j=j_0}^{\infty}2^{-2pj}n^p\mathbb{E}\left(\sup_{\beta\in S_{j,n}}|I_{1,n}(\beta)|^p\right)+\sum_{j=j_0}^{\infty}2^{-2pj}n^p\mathbb{E}\left(\sup_{\beta\in S_{j,n}}|I_{2,n}(\beta)|^p\right) \end{array}\right], \end{aligned}$$

where the constant K has changed from the first to second inequality. The crucial part of the proof is

to bound each expectation by $O\left(n^{-p}2^{pj}\right)$. This will imply that

$$\begin{aligned}
\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^*-\beta_0\right)\right|>2^{j_0}\right) &\leq \sum_{j=j_0}^{\infty}\mathbb{P}\left(\sup_{\beta\in S_{j,n}}\left\{\tilde{Q}_n^*(\beta)-\tilde{Q}_n^*(\beta_0)\right\}\geq 0\right)\leq K\sum_{j\geq j_0}2^{-pj}\\
&= \sum_{j\geq j_0}\left(\frac{1}{2}\right)^{pj}=(1/2)^{pj_0}+(1/2)^{p(j_0+1)}+\ldots\\
&= \left(\frac{1}{2}\right)^{pj_0}\underbrace{\left(1+(1/2)^p+(1/2)^{2p}+\ldots\right)}_{=\frac{1}{1-(1/2)^p}<K}\\
&\leq K2^{-pj_0}.
\end{aligned}$$

Since

$$\mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^*-\beta_0\right)\right|^{2+\delta}=p\int_0^{\infty}t^{2+\delta-1}\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^*-\beta_0\right)\right|>t\right)dt,$$

we can take the above result with $j_0=\log_2 t$. This implies

$$\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^*-\beta_0\right)\right|>t\right)\leq K2^{-p\log_2 t}=K2^{\log_2 t^{-p}}=Kt^{-p},$$

and since $p>2+\delta$, we can conclude

$$\begin{aligned}
\mathbb{E}\left|\sqrt{n}\left(\hat{\beta}_n^*-\beta_0\right)\right|^{2+\delta} &= p\int_0^{\infty}t^{2+\delta-1}\mathbb{P}\left(\left|\sqrt{n}\left(\hat{\beta}_n^*-\beta_0\right)\right|>t\right)dt\\
&\leq K\int_0^{\infty}t^{2+\delta-1}t^{-p}dt=K\int_0^{\infty}t^{-1-(p-2+\delta)}dt<\infty.
\end{aligned}$$

Bounding $E\left(\sup_{\beta\in S_{j,n}}\left|I_{2\text{-step},n}(\beta)\right|^p\right)$:

Recall that

$$\begin{aligned}
I_{2\text{-step},n}(\beta) &= \left[Q_{2n}^*(\hat{\alpha}_n^*,\beta)-Q_{2n}^*(\hat{\alpha}_n^*,\beta_0)\right]-\left[Q_{2n}^*(\alpha_0,\beta)-Q_{2n}^*(\alpha_0,\beta_0)\right]\\
&= \left[Q_{2n}^*(\hat{\alpha}_n^*,\beta)-Q_{2n}^*(\alpha_0,\beta)\right]-\left[Q_{2n}^*(\hat{\alpha}_n^*,\beta_0)-Q_{2n}^*(\alpha_0,\beta_0)\right]\\
&= n^{-1}\sum_{t=1}^{n}\left(q_{2t}^*(\hat{\alpha}_n^*,\beta)-q_{2t}^*(\alpha_0,\beta)\right)-n^{-1}\sum_{t=1}^{n}\left(q_{2t}^*(\hat{\alpha}_n^*,\beta_0)-q_{2t}^*(\alpha_0,\beta_0)\right)
\end{aligned}$$

By taking the Taylor series expansion of q_{2t} around $(\alpha,\beta)=(\alpha_0,\beta_0)$, we have

$$q_{2t}(\alpha,\beta)=q_{2t}(\alpha_0,\beta_0)+\frac{\partial}{\partial\alpha'}q_{2t}(\alpha_0,\beta_0)(\alpha-\alpha_0)+\frac{\partial}{\partial\beta'}q_{2t}(\alpha_0,\beta_0)(\beta-\beta_0)+R_2(\alpha,\beta),\qquad(9)$$

such that

$$R_2=\frac{1}{2!}\begin{bmatrix}(\alpha-\alpha_0)'\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}\left(\bar{\alpha},\bar{\beta}\right)(\alpha-\alpha_0)+(\beta-\beta_0)'\frac{\partial}{\partial\beta\partial\beta'}q_{2t}\left(\bar{\alpha},\bar{\beta}\right)(\beta-\beta_0)\\+(\alpha-\alpha_0)'\frac{\partial}{\partial\alpha\partial\beta'}q_{2t}\left(\bar{\alpha},\bar{\beta}\right)(\beta-\beta_0)+(\beta-\beta_0)'\frac{\partial}{\partial\beta\partial\alpha'}q_{2t}\left(\bar{\alpha},\bar{\beta}\right)(\alpha-\alpha_0)\end{bmatrix}$$

where $\bar{\alpha}$ lies between α and α_0 and $\bar{\beta}$ lies between β and β_0.

Then using (9), we can write

$$\begin{aligned}
q_{2t}^*(\hat{\alpha}_n^*,\beta)-q_{2t}^*(\alpha_0,\beta) &= \frac{\partial}{\partial\alpha'}q_{2t}^*(\alpha_0,\beta)(\hat{\alpha}_n^*-\alpha_0)\\
&\quad+\frac{1}{2!}(\hat{\alpha}_n^*-\alpha_0)'\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}^*(\bar{\alpha}_1,\beta)(\hat{\alpha}_n^*-\alpha_0)
\end{aligned}$$

where $\bar{\alpha}_1$ lies between $\hat{\alpha}_n^*$ and α_0. Similarly, we have

$$\begin{aligned} q_{2t}^*\left(\hat{\alpha}_n^*,\beta_0\right)-q_{2t}^*\left(\alpha_0,\beta_0\right) &= \frac{\partial}{\partial\alpha'}q_{2t}^*\left(\alpha_0,\beta_0\right)\left(\hat{\alpha}_n^*-\alpha_0\right) \\ &\quad +\frac{1}{2!}\left(\hat{\alpha}_n^*-\alpha_0\right)'\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}^*\left(\bar{\alpha}_2,\beta_0\right)\left(\hat{\alpha}_n^*-\alpha_0\right), \end{aligned}$$

where $\bar{\alpha}_2$ lies between $\hat{\alpha}_n^*$ and α. It follows that

$$\begin{aligned} I_{2\text{-step},n}\left(\beta\right) &= n^{-1}\sum_{t=1}^{n}\left(\frac{\partial}{\partial\alpha'}q_{2t}^*\left(\alpha_0,\beta\right)-\frac{\partial}{\partial\alpha'}q_{2t}^*\left(\alpha_0,\beta_0\right)\right)\left(\hat{\alpha}_n^*-\alpha_0\right) \\ &\quad +\frac{1}{2!}n^{-1}\sum_{t=1}^{n}\left(\hat{\alpha}_n^*-\alpha_0\right)'\left(\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}^*\left(\bar{\alpha}_1,\beta\right)-\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}^*\left(\bar{\alpha}_2,\beta_0\right)\right)\left(\hat{\alpha}_n^*-\alpha_0\right). \end{aligned}$$

Suppose that $\left\{\frac{\partial}{\partial\alpha'}q_{2t}\left(\alpha,\beta\right)\right\}$ and $\left\{\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}\left(\alpha,\beta\right)\right\}$ are Lipschitz continuous in (α,β) :

$$\left|\frac{\partial}{\partial\alpha'}q_{2t}\left(\alpha,\beta\right)-\frac{\partial}{\partial\alpha'}q_{2t}\left(\alpha_0,\beta_0\right)\right|\leq L_{1t}\left(X^t\right)\left(\left|\alpha-\alpha_0\right|+\left|\beta-\beta_0\right|\right),$$

and

$$\left|\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}\left(\alpha,\beta\right)-\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}\left(\alpha_0,\beta_0\right)\right|\leq L_{2t}\left(X^t\right)\left(\left|\alpha-\alpha_0\right|+\left|\beta-\beta_0\right|\right),$$

where the functions $L_{1t}\left(X^t\right)$ and $L_{2t}\left(X^t\right)$ do not depend on α nor β. Thus, we have

$$\left|\frac{\partial}{\partial\alpha'}q_{2t}^*\left(\alpha_0,\beta\right)-\frac{\partial}{\partial\alpha'}q_{2t}^*\left(\alpha_0,\beta_0\right)\right|\leq L_{1t}^*\left(\left|\beta-\beta_0\right|\right), \tag{10}$$

and similarly,

$$\begin{aligned} \left|\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}^*\left(\alpha_1,\beta\right)-\frac{\partial}{\partial\alpha\partial\alpha'}q_{2t}^*\left(\alpha_2,\beta_0\right)\right| &\leq L_{2t}^*\left(\left|\bar{\alpha}_1-\bar{\alpha}_2\right|+\left|\beta-\beta_0\right|\right) \\ &\leq L_{2t}^*\left(\left|\hat{\alpha}_n^*-\alpha_0\right|+\left|\beta-\beta_0\right|\right), \end{aligned} \tag{11}$$

where the last inequality follows because both $\bar{\alpha}_1$ and $\bar{\alpha}_2$ lie between $\hat{\alpha}_n^*$ and α_0. Therefore by the triangular inequality and using (10) and (11), we have

$$\begin{aligned} \left|I_{2\text{-step},n}\left(\beta\right)\right| &\leq n^{-1}\left(n^{-1}\sum_{t=1}^{n}L_{1t}^*\right)\left|\sqrt{n}\left(\hat{\alpha}_n^*-\alpha_0\right)\right|\left|\sqrt{n}\left(\beta-\beta_0\right)\right| \\ &\quad +n^{-3/2}\frac{1}{2!}\left(n^{-1}\sum_{t=1}^{n}L_{2t}^*\right)\left|\sqrt{n}\left(\hat{\alpha}_n^*-\alpha_0\right)\right|^2\left(\sqrt{n}\left|\hat{\alpha}_n^*-\alpha_0\right|+\left|\sqrt{n}\left(\beta-\beta_0\right)\right|\right). \end{aligned}$$

Hence, successive applications of the Hölder's inequality yields

$$
\begin{aligned}
& E\left(\sup_{\beta\in S_{j,n}}\left|I_{2\text{-step},n}\left(\beta\right)\right|^{p}\right)\\
\leq\ & Kn^{-p}\left(\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{\varepsilon p}\right)\right)^{\frac{1}{\varepsilon}}\left(\mathbb{E}\left(\left(\left(n^{-1}\sum_{t=1}^{n}L_{1t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p}\sup_{\beta\in S_{j,n}}\left|\sqrt{n}\left(\beta-\beta_{0}\right)\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)\right)\right)^{\frac{\varepsilon-1}{\varepsilon}}\\
& +Kn^{-\frac{3p}{2}}\left(\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{3\varepsilon p}\right)\right)^{\frac{1}{\varepsilon}}\left(\mathbb{E}\left(n^{-1}\sum_{t=1}^{n}L_{2t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p}\right)^{\frac{\varepsilon-1}{\varepsilon}}\\
& +Kn^{-\frac{3p}{2}}\left(\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{2\varepsilon p}\right)\right)^{\frac{1}{\varepsilon}}\left(\mathbb{E}\left(\left(n^{-1}\sum_{t=1}^{n}L_{2t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p}\sup_{\beta\in S_{j,n}}\left|\sqrt{n}\left(\beta-\beta_{0}\right)\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)\right)^{\frac{\varepsilon-1}{\varepsilon}}
\end{aligned}
$$

for some $\varepsilon>1$. Note that for $\beta\in S_{j,n}$, we have $\left|\sqrt{n}\left(\beta-\beta_{0}\right)\right|\leq 2^{j}$. This implies that

$$
\begin{aligned}
& E\left(\sup_{\beta\in S_{j,n}}\left|I_{2\text{-step},n}\left(\beta\right)\right|^{p}\right)\\
\leq\ & Kn^{-p}\left(\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{\varepsilon p}\right)\right)^{\frac{1}{\varepsilon}}2^{pj}\left(\mathbb{E}\left(\left(n^{-1}\sum_{t=1}^{n}L_{1t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p}\right)\right)^{\frac{\varepsilon-1}{\varepsilon}}\\
& +Kn^{-\frac{3p}{2}}\left(\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{3\varepsilon p}\right)\right)^{\frac{1}{\varepsilon}}\left(\mathbb{E}\left(n^{-1}\sum_{t=1}^{n}L_{2t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p}\right)^{\frac{\varepsilon-1}{\varepsilon}}\\
& +Kn^{-\frac{3p}{2}}\left(\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{2\varepsilon p}\right)\right)^{\frac{1}{\varepsilon}}2^{pj}\left(\mathbb{E}\left(\left(n^{-1}\sum_{t=1}^{n}L_{2t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p}\right)\right)^{\frac{\varepsilon-1}{\varepsilon}}
\end{aligned}
$$

Suppose we assume that $\mathbb{E}\left(\left|\sqrt{n}\left(\hat{\alpha}_{n}^{*}-\alpha_{0}\right)\right|^{3\varepsilon p}\right)<\infty$. If in addition we assume that $E\left(\left|L_{1t}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)<\infty$ and $E\left(\left|L_{2t}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)<\infty$, we can show that the expectations of average of the functions involving L_{1t}^{*} and L_{2t}^{*} are bounded. For instance,

$$
\begin{aligned}
\mathbb{E}\left(n^{-1}\sum_{t=1}^{n}L_{1t}^{*}\right)^{\frac{\varepsilon}{\varepsilon-1}p} &\leq Kn^{-1}\sum_{t=1}^{n}\mathbb{E}\left(\left|L_{1t}^{*}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)\\
&= Kn^{-1}\sum_{t=1}^{n}\left(E\left(E^{*}\left(\left|L_{1t}^{*}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)\right)\right)\\
&= KEE^{*}\left(n^{-1}\sum_{t=1}^{n}\left|L_{1t}^{*}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)\\
&= KE\left(\sum_{t=1}^{n}\gamma_{nt}\left|L_{1t}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)<\infty \text{ if } E\left(\left|L_{1t}\right|^{\frac{\varepsilon}{\varepsilon-1}p}\right)<\infty.
\end{aligned}
$$

Thus, under these assumptions

$$
\mathbb{E}\left(\sup_{\beta\in S_{j,n}}\left|I_{2\text{-step},n}\left(\beta\right)\right|^{p}\right)\leq Kn^{-p}2^{pj},
$$

which implies

$$\begin{aligned}\sum_{j=j_0}^{\infty} 2^{-2pj} n^p \mathbb{E}\left(\sup_{\beta\in S_{j,n}} |I_{2\text{-step},n}(\beta)|^p\right) &\leq K\sum_{j=j_0}^{\infty} 2^{-2pj} 2^{pj} \underbrace{n^p n^{-p}}_{=1} \\ &= K\sum_{j=j_0}^{\infty} 2^{-pj} \\ &\leq K 2^{-pj_0}.\end{aligned}$$

Bounding $\mathbb{E}\left(\sup_{\beta\in S_{j,n}} |I_{1,n}(\beta)|^p\right)$:

Note that by definition of $I_{1,n}(\beta)$, we have that

$$\begin{aligned}I_{1,n}(\beta) &= Q^*_{2n}(\alpha_0,\beta) - Q^*_{2n}(\alpha_0,\beta_0) - E^*\left[Q^*_{2n}(\alpha_0,\beta) - Q^*_{2n}(\alpha_0,\beta_0)\right] \\ &= n^{-1}\sum_{t=1}^{n}\left(q^*_{2t}(\alpha_0,\beta) - q^*_{2t}(\alpha_0,\beta_0)\right) - E^*\left(n^{-1}\sum_{t=1}^{n}\left(q^*_{2t}(\alpha_0,\beta) - q^*_{2t}(\alpha_0,\beta_0)\right)\right) \\ &\equiv n^{-1/2}\mathbb{G}^*_n\left(q_2(\alpha_0,\beta) - q_2(\alpha_0,\beta_0)\right),\end{aligned}$$

where for a class of functions $\mathcal{F} = \{f\}$, we define the empirical process $\mathbb{G}^*_n f$ as

$$\mathbb{G}^*_n f = n^{-1/2}\sum_{t=1}^{n}\left(f^*_t - E^* f^*_t\right).$$

Define the L^p norm of $\mathbb{G}^*_n f$ over $\mathcal{F}$ as

$$\left(\mathbb{E}\left|\mathbb{G}^*_n\right|^p_{\mathcal{F}}\right)^{1/p} = \left(\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\mathbb{G}^*_n f\right|\right)^p\right)^{1/p}.$$

With this notation,

$$\begin{aligned}\mathbb{E}\left(\sup_{\beta\in S_{j,n}} |I_{1,n}(\beta)|^p\right) &= \mathbb{E}\left(\sup_{\beta\in S_{j,n}}\left|n^{-1/2}\mathbb{G}^*_n\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right)\right|^p\right) \\ &= n^{-p/2}\mathbb{E}\left(\sup_{\beta\in S_{j,n}}\left|\mathbb{G}^*_n\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right)\right|^p\right) \\ &= n^{-p/2}\left\{\left(\mathbb{E}\left(\sup_{\beta\in S_{j,n}}\left|\mathbb{G}^*_n\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right)\right|^p\right)\right)^{1/p}\right\}^p \\ &= n^{-p/2}\left(\left(\mathbb{E}\left|\mathbb{G}^*_n\right|^p_{\mathcal{N}_\delta}\right)^{1/p}\right)^p,\end{aligned}$$

where we let $\mathcal{N}_\eta = \{q_2(\alpha_0,\beta) - q_2(\alpha_0,\beta_0) : |\beta-\beta_0| \leq \eta,\ (\alpha,\beta)\in\mathcal{A}\times\mathcal{B}\}$. Lemma A.1 shows that for any $\eta > 0$, $\left(\mathbb{E}\left|\mathbb{G}^*_n\right|^p_{\mathcal{N}_\eta}\right)^{1/p} \leq \eta$ holds under our assumptions. Thus, letting $\eta = \frac{2^j}{\sqrt{n}}$ yields $\left(\mathbb{E}\left|\mathbb{G}^*_n\right|^p_{\mathcal{N}_\eta}\right)^{1/p} \leq \left(\frac{2^j}{\sqrt{n}}\right)^p$, implying that

$$\mathbb{E}\left(\sup_{\beta\in S_{j,n}} |I_{1,n}(\beta)|^p\right) \leq n^{-p/2}\frac{2^{pj}}{n^{p/2}} = n^{-p}2^{pj}$$

It follows that

$$\sum_{j=j_0}^{\infty} 2^{-2pj} n^p \mathbb{E}\left(\sup_{\beta\in S_{j,n}} |I_{1,n}(\beta)|^p\right) \leq \sum_{j=j_0}^{\infty} 2^{-2pj} n^p n^{-p} 2^{pj} = \sum_{j=j_0}^{\infty} 2^{-pj} \leq K 2^{-pj_0},$$

as above.

Bounding $E\left(\sup_{\beta\in S_{j,n}} |I_{2,n}(\beta)|^p\right)$:

The argument is similar. By definition of $I_{2,n}(\beta)$, we have

$$\begin{aligned}
I_{2,n}(\beta) &= E^*\left[Q_{2n}^*(\alpha_0,\beta) - Q_{2n}^*(\alpha_0,\beta_0)\right] - E\left(E^*\left[Q_{2n}^*(\alpha_0,\beta) - Q_{2n}^*(\alpha_0,\beta_0)\right]\right) \\
&= n^{-1}\sum_{t=1}^{n} E^*\left(q_{2t}^*(\alpha_0,\beta) - q_{2t}^*(\alpha_0,\beta_0)\right) - n^{-1}\sum_{t=1}^{n} E\left(E^*\left(q_{2t}^*(\alpha_0,\beta) - q_{2t}^*(\alpha_0,\beta_0)\right)\right) \\
&= \sum_{t=1}^{n} \gamma_{nt}\left[\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right) - E\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right)\right] \\
&= n^{-1/2}\left(\sum_{t=1}^{n} \sqrt{n}\gamma_{nt}\left[\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right) - E\left(q_{2t}(\alpha_0,\beta) - q_{2t}(\alpha_0,\beta_0)\right)\right]\right) \\
&= n^{-1/2}\mathbb{G}_{n,\gamma}\left(q_2(\alpha_0,\beta) - q_2(\alpha_0,\beta_0)\right),
\end{aligned}$$

where we define the empirical process $\mathbb{G}_{n,\gamma}$ as

$$\mathbb{G}_{n,\gamma} f = \sum_{t=1}^{n} \sqrt{n}\gamma_{nt}\left(f_t - Ef_t\right),$$

with weights defined as above. Similarly, we define the L^p norm of $\mathbb{G}_{n,\gamma} f$ over $\mathcal{F} = \{f\}$ as

$$\left(E\left|\mathbb{G}_{n,\gamma}\right|_{\mathcal{F}}^p\right)^{1/p} = \left(E\left(\sup_{f\in\mathcal{F}} |\mathbb{G}_{n,\gamma} f|\right)^p\right)^{1/p}.$$

With this notation

$$\mathbb{E}\left(\sup_{\beta\in S_{j,n}} |I_{2,n}(\beta)|^p\right) = n^{-p/2}\left(\left(E\left|\mathbb{G}_{n,\gamma}\right|_{\mathcal{F}}^p\right)^{1/p}\right)^p.$$

It suffices to bound $\left(E\left|\mathbb{G}_{n,\gamma}\right|_{\mathcal{F}}^p\right)^{1/p}$. Assumption B6(ii) provides a bound on the L_p-norm of the empirical process $\mathbb{G}_n$, which differs from $\mathbb{G}_{n,\gamma}$ due to presence of the weights γ_{nt}. It is well known that these weights are introduced by the fact that the MBB puts less weight on the first and last ℓ observations in the sample. In particular, we can show that for any function f_t, the MBB expectation $E^*\left(\bar{f}_n^*\right) = \sum_{t=1}^n \gamma_{nt} f_t = n^{-1}\sum_{t=1}^n f_t + O_P\left(\frac{\ell}{n}\right)$. Using this insight, we can show that

$$\mathbb{G}_{n,\gamma} f = \frac{n}{n-\ell+1}\mathbb{G}_n f - \frac{n}{n-\ell+1}\mathbb{R}_{1n} f - \frac{n}{n-\ell+1}\mathbb{R}_{2n} f,$$

where

$$\begin{aligned}
\mathbb{R}_{1n} f &= \frac{1}{\sqrt{n}}\sum_{t=1}^{\ell}\left(1 - \frac{t}{\ell}\right)\left(f_t - Ef_t\right), \\
\mathbb{R}_{2n} f &= \frac{1}{\sqrt{n}}\sum_{t=1}^{\ell}\left(1 - \frac{t}{\ell}\right)\left(f_{n-t+1} - Ef_{n-t+1}\right).
\end{aligned}$$

By Minkowski's inequality,

$$\begin{aligned}\left(E\left|\mathbb{G}_{n,\gamma}\right|_{\mathcal{F}}^{p}\right)^{1/p} &\leq \frac{n}{n-\ell+1}\left\{\left(E\left|\mathbb{G}_{n}\right|_{\mathcal{F}}^{p}\right)^{1/p}+\left(E\left|\mathbb{R}_{1n}\right|_{\mathcal{F}}^{p}\right)^{1/p}+\left(E\left|\mathbb{R}_{2n}\right|_{\mathcal{F}}^{p}\right)^{1/p}\right\} \\ &\leq K\left(\left(E\left|\mathbb{G}_{n}\right|_{\mathcal{F}}^{p}\right)^{1/p}+\left(E\left|\mathbb{R}_{1n}\right|_{\mathcal{F}}^{p}\right)^{1/p}+\left(E\left|\mathbb{R}_{2n}\right|_{\mathcal{F}}^{p}\right)^{1/p}\right), \qquad (12)\end{aligned}$$

for some constant K since $\ell \to \infty$ such that $\ell = o\left(\sqrt{n}\right)$ under our assumptions. This implies that

$$\begin{aligned}\mathbb{E}\left(\sup_{\beta\in S_{j,n}}\left|I_{2,n}\left(\beta\right)\right|^{p}\right) &= n^{-p/2}\left(\left(E\left|\mathbb{G}_{n,\gamma}\right|_{\mathcal{F}}^{p}\right)^{1/p}\right)^{p} \\ &\leq Kn^{-p/2}\left(\left(E\left|\mathbb{G}_{n}\right|_{\mathcal{F}}^{p}\right)^{1/p}+\left(E\left|\mathbb{R}_{1n}\right|_{\mathcal{F}}^{p}\right)^{1/p}+\left(E\left|\mathbb{R}_{2n}\right|_{\mathcal{F}}^{p}\right)^{1/p}\right)^{p} \\ &\leq Kn^{-p/2}\left(\frac{2^{jp}}{n^{p/2}}+E\left|\mathbb{R}_{1n}\right|_{\mathcal{F}}^{p}+E\left|\mathbb{R}_{2n}\right|_{\mathcal{F}}^{p}\right),\end{aligned}$$

where we have used Assumption B6(ii) with $\eta = \frac{2^j}{\sqrt{n}}$ to bound $\left(E\left|\mathbb{G}_{n}\right|_{\mathcal{F}}^{p}\right)^{1/p}$. The remainder terms can be bounded by $O\left(\left(\frac{\ell}{\sqrt{n}}\right)^{p}\frac{2^{jp}}{n^{p}}\right)$ using the Lipschitz condition given in Assumption B6(iii), where the Lipschitz function for the log likelihood function $\{q_{2t}\left(\alpha,\beta\right)\}$ has a finite p^{th} order moment. Since $\ell = o\left(\sqrt{n}\right)$ by assumption, the contribution of the two remainder terms is smaller than that of the first term. We can then claim that

$$\mathbb{E}\left(\sup_{\beta\in S_{j,n}}\left|I_{2,n}\left(\beta\right)\right|^{p}\right) \leq Kn^{-p/2}\frac{2^{jp}}{n^{p/2}} = Kn^{-p}2^{jp},$$

and the proof follows as above.

A.4 Auxiliary lemmas used in the proof of Theorem 4.4

The main goal of this section is to show that a bootstrap version of the L_p maximal inequality stated in Assumption B6(iii) holds under our assumptions. In particular, we show that for some $p > 2+\delta$, $\left(\mathbb{E}\left|\mathbb{G}_{n}^{*}\right|_{\mathcal{N}_\eta}^{p}\right)^{1/p} \leq \eta$ holds when $\mathcal{N}_\eta$ is as defined in Assumption B6(iii) and $\mathbb{G}_n^*$ is defined as

$$\begin{aligned}&\mathbb{G}_{n}^{*}\left(q_{2}\left(\alpha_{0},\beta\right)-q_{2}\left(\alpha_{0},\beta_{0}\right)\right) \\ =\ & n^{-1}\sum_{t=1}^{n}\left(q_{2t}^{*}\left(\alpha_{0},\beta\right)-q_{2t}^{*}\left(\alpha_{0},\beta_{0}\right)\right)-E^{*}\left(n^{-1}\sum_{t=1}^{n}\left(q_{2t}^{*}\left(\alpha_{0},\beta\right)-q_{2t}^{*}\left(\alpha_{0},\beta_{0}\right)\right)\right),\end{aligned}$$

where $q_{2t}^{*}\left(\alpha,\beta\right)$ is a MBB version of $q_{2t}\left(\alpha,\beta\right) = \log f_{2t}\left(\alpha,\beta\right)$. This result is as follows.

Lemma A.1. *Suppose that Assumption B6(iii) holds, and assume that* $\{\log f_{2t}\left(\alpha,\beta\right)\}$ *satisfies a Lipschitz continuity condition on* $\mathcal{A}\times\mathcal{B}$*, a.s.-P, with Lipschitz functions* $\{L_t\}$ *such that* $E\left|L_t\right|^p < \infty$ *for* $p > 2+\delta$*, for some* $\delta > 0$*. Then,* $\left(\mathbb{E}\left|\mathbb{G}_{n}^{*}\right|_{\mathcal{N}_\eta}^{p}\right)^{1/p} \leq \eta$ *for any* $\eta > 0$*.*

To prove Lemma A.1, we rely on the following L_p multiplier inequality, which extends Lemma 4.1 of Praestgaard and Wellner (1993) by allowing for $p \geq 1$ rather than just $p = 1$.

To state this result, we need to introduce some notation. Recall that for a generic time series $\{X_t : t = 1,\ldots,n\}$, letting $k = \frac{n}{\ell}$ denote the number of blocks of size ℓ needed to define a MBB

sample of size n and letting $\{I_j : j = 1, \ldots, k\}$ be an i.i.d. uniform sequence of indices distributed on $\{1, \ldots, n - \ell + 1\}$ allows us to write the MBB average as

$$\bar{X}_n^* = n^{-1} \sum_{t=1}^{n} X_t^* = k^{-1} \sum_{j=1}^{k} \left(\ell^{-1} \sum_{t=1}^{\ell} X_{t+(j-1)\ell}^* \right) = k^{-1} \sum_{j=1}^{k} \left(\sum_{t=1}^{\ell} X_{t+I_j-1} \right) = n^{-1} \sum_{j=1}^{k} Z_{I_j}.$$

Another way to write this average is as follows. Let $N = n-\ell+1$, and let $\mathbf{W}_N = (W_1, \ldots, W_N)'$ denote a triangular array of weights whose distribution is the Multinomial $\left(k, \left(N^{-1}, \ldots, N^{-1}\right)\right)$ distribution[1]. Note that these are non-negative exchangeable random variables. We can then think of $\bar{X}_n^*$ as a weighted average of the block sums $Z_j = \sum_{t=1}^{\ell} X_{t+j-1}$, weighted by W_j :

$$\bar{X}_n^* = n^{-1} \sum_{j=1}^{N} W_j Z_j,$$

where W_j denotes the number of times the j^{th} block sum Z_j is drawn in the bootstrap sample. Note that if $\ell = 1$, then $N = k = n$, and this way of writing the bootstrap average is exactly the same as when studying the nonparametric i.i.d. bootstrap using the Multinomial distribution $\left(n, \left(n^{-1}, \ldots, n^{-1}\right)\right)$. Thus, our framework is an extension of the usual framework to the MBB. Our goal in Lemma A.1 is to bound the L_p moment of the bootstrap empirical process

$$\mathbb{G}_n^* f = n^{-1/2} \sum_{t=1}^{n} \left(f_t^* - E^* \left(f_t^* \right) \right).$$

With this new notation, we can write

$$\mathbb{G}_n^* f = n^{-1/2} \sum_{j=1}^{N} \left(W_j - E_W \left(W_j \right) \right) \left(\sum_{t=1}^{\ell} f_{t+j-1} \right),$$

where $E_W(\cdot)$ (and $P_W(\cdot)$) denotes expectation (and probability) with respect to the random vector $\mathbf{W}_N$ defined above. The L_p-multiplier we are about to state gives a bound on the L_p moments of averages defined as $n^{-1/2} \sum_{j=1}^{N} W_j Z_j$, where Z_j will play the role of the block sum $\sum_{t=1}^{\ell} f_{t+j-1}$ in our application.

To state this result, define the joint probability $\mathbb{P} = P \times P_W$, which we wrote before as $P \times P^*$, and let $\|W_1\|_{2,1} = \int_0^\infty \sqrt{P_W(W_1 \geq u)} du$. Some expressions below may be non-measurable; probability and expectation of these expressions are understood in terms of outer probability and outer expectation (see, e.g. van der Vaart and Wellner, 1996, p. 6). Application of Fubini's theorem to such expectations requires additional care. We assume that a measurability condition that restores the Fubini theorem is satisfied in all our applications below.

Lemma A.2. *Let $\mathbf{W}_N = (W_1, \ldots, W_N)'$ be an array of non-negative exchangeable random variables such that, for every N, $\|W_1\|_{2,1} = \int_0^\infty \sqrt{P_W(W_1 \geq u)} du < \infty$, and let R denote a random permutation uniformly distributed on Π_N, the set of permutations of $1, 2, \ldots, N$. Let $Z_1, \ldots, Z_N$ be a sequence of random elements such that $(\mathbf{W}_N, R)$ and $(Z_1, \ldots, Z_N)$ are independent, and write $\|Z_j\| = \sup_{h \in \mathcal{F}} |Z_j(h)|$.*

[1] For simplicity, we will drop the array notation and will write W_j rather than $W_{N,j}$. Similarly, we will omit the index n in the definition N_n.

Then for any N_0 such that $1 \leq N_0 < \infty$ and any $N > N_0$, the following inequality holds for any $p \geq 1$:

$$\left(\mathbb{E}\left\|n^{-1/2}\sum_{j=1}^{N} W_j Z_j\right\|^p\right)^{1/p} \leq \frac{N_0}{\sqrt{n}}\left(E_W\left|\max_{1\leq j\leq N} W_j\right|^p\right)^{1/p}\left(\frac{1}{N}\sum_{j=1}^{N} E\left\|Z_j\right\|^p\right)^{1/p} + \left\|W_1\right\|_{2,1}^{1/p}\cdot\left(E_{Z,R}\left(\max_{N_0<k\leq N}\left\|\frac{1}{\sqrt{k}}\sum_{i=N_0+1}^{k} Z_{R(i)}\right\|\right)^p\right)^{1/p},$$

where we let $E_{Z,R}(\cdot)$ denote the expectation with respect to $Z_1, \ldots, Z_N$ and R jointly.

This result extends Lemma 4.1 of Praestgaard and Wellner (1993) from $p = 1$ to $p \geq 1$. As in Praestgaard and Wellner (1993), we do not assume any particular dependence structure on the vector $(Z_1, \ldots, Z_N)$, the only assumption being that it is independent of the pair $(\mathbf{W}_N, R)$. This is in contrast with the L_p multiplier provided by Cheng (2014, p. 17), which assumes $Z_1, \ldots, Z_N$ to be i.i.d., while also allowing for any $p \geq 1$. The i.i.d. assumption on the random functions Z_j is too strong for our applications, where Z_j will be given by block sums of contributions to the log likelihood function. These are typically serially dependent in the time series context and this is the reason for given Lemma A.2, a result that might be of independent interest.

Next, we prove Lemma A.1 and then we prove Lemma A.2.

Proof of Lemma A.1 In the following '$\lesssim$' denote smaller than, up to an universal constant $K > 0$. Recalling the definition of $\mathbb{G}_n^* f$, where f is in the function class $\mathcal{N}_\eta$, and the property of the MBB weights, in particular, $\sum_{j=1}^{N} W_j = k$, implying that $E_W(W_j) = \frac{k}{N}$, we can rewrite $\mathbb{G}_n^* f$ as follows:

$$\begin{aligned}
\mathbb{G}_n^* f &= n^{-1/2}\sum_{j=1}^{N}\left(W_j - E_W(W_j)\right)\left(\sum_{t=1}^{\ell} f_{t+j-1}\right) \\
&= n^{-1/2}\sum_{j=1}^{N}\left(W_j - \frac{k}{N}\right)\left(\sum_{t=1}^{\ell} f_{t+j-1}\right), \text{ since } E_W(W_j) = \frac{k}{N} \\
&= n^{-1/2}\sum_{j=1}^{N}\left(W_j - \frac{k}{N}\right)\left[\left(\sum_{t=1}^{\ell} f_{t+j-1}\right) - E\left(\sum_{t=1}^{\ell} f_{t+j-1}\right)\right],
\end{aligned}$$

since $\sum_{t=1}^{N}\left(W_j - \frac{k}{N}\right) = 0$, and the expectation of $E\left(\sum_{t=1}^{\ell} f_{t+j-1}\right)$ is time invariant under Assumption B6(i). For $j = 1, 2, \ldots, N$, let

$$Y_j(f) = \sum_{t=1}^{\ell} f_{t+j-1} - E\left(\sum_{t=1}^{\ell} f_{t+j-1}\right) = \sum_{t=1}^{\ell}\left(f_{t+j-1} - E\left(f_{t+j-1}\right)\right). \tag{13}$$

With this notation, $\mathbb{G}_n^* f$ can be rewritten as

$$\mathbb{G}_n^* f = n^{-1/2}\sum_{j=1}^{N}\left(W_j - \frac{k}{N}\right) Y_j(f). \tag{14}$$

Our goal is to bound the L_p moment of the supremum of this empirical process. To do so, we follow the same arguments as in Cheng (2015, p. 19) to show that

$$\begin{aligned}
\left(\mathbb{E}\left\|\mathbb{G}_n^*\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} &= \left(\mathbb{E}\left(\sup_{f\in\mathcal{N}_\eta}\left|n^{-1/2}\sum_{t=1}^{N}\left(W_j - \frac{k}{N}\right) Y_j(f)\right|\right)^p\right)^{1/p} \\
&\lesssim 2\left(\mathbb{E}\left(\sup_{f\in\mathcal{N}_\eta}\left|n^{-1/2}\sum_{t=1}^{N} W_j Y_j(f)\right|\right)^p\right)^{1/p}.
\end{aligned} \tag{15}$$

Next, we apply the L_p multiplier inequality in Lemma A.2 (using (15)) with $Z_j = Y_j(f)$ and $\mathcal{F} = \mathcal{N}_\eta$. This yields

$$\begin{aligned}
\left(\mathbb{E}\left\|\mathbb{G}_n^*\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} &\lesssim \frac{N_0}{\sqrt{n}}\left(E_W\left|\max_{1\le j\le N} W_j\right|^p\right)^{1/p}\left(\frac{1}{N}\sum_{i=1}^N E\left\|Z_i\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} \\
&\quad + \left(\ell\left\|W_{N,1}\right\|_{2,1}\right)^{1/p}\left(\ell^{-1}E_{Z,R}\left(\max_{N_0<k\le N}\left\|\frac{1}{\sqrt{k}}\sum_{i=N_0+1}^k Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p} \\
&\lesssim N_0\frac{\ell}{\sqrt{n}}\left(E_W\left|\max_{1\le j\le N} W_j\right|^p\right)^{1/p}\left(\frac{1}{N\ell^p}\sum_{i=1}^N E\left\|Z_i\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} \\
&\quad + \left(\ell^{-1}E_{Z,R}\left(\max_{N_0<k\le N}\left\|\frac{1}{\sqrt{k}}\sum_{i=N_0+1}^k Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p} \\
&\lesssim \mathrm{I}+\mathrm{II}. \qquad (16)
\end{aligned}$$

for any $1 \le N_0 < \infty$ and $N > N_0$, (the second inequality follows because the MBB weight verifies the condition $\limsup_{N\to\infty}\ell\left\|W_{N,1}\right\|_{2,1} < \infty$, where $W_1 = W_{N,1}$). We first bound the first term in the preceding equation, then we bound the second term.

For the first term, note that

$$\begin{aligned}
&\frac{1}{N\ell^p}\sum_{i=1}^N E\left\|Z_i\right\|_{\mathcal{N}_\eta}^p \\
\le\ &\frac{1}{N\ell^p}\sum_{j=1}^N \ell^{p-1}\sum_{t=1}^\ell E\left\|\left(f_{t+j-1}-E\left(f_{t+j-1}\right)\right)\right\|_{\mathcal{N}_\eta}^p = n^{-1/2}\frac{\sqrt{n}}{N\ell}\sum_{j=1}^N\sum_{t=1}^\ell E\left\|\left(f_{t+j-1}-E\left(f_{t+j-1}\right)\right)\right\|_{\mathcal{N}_\eta}^p
\end{aligned}$$

from Minkowski's inequality. Using the same arguments as in the proof of Theorem 4.4 (see equation (12))), it follows that (and given Assumption B6(iii)),

$$\begin{aligned}
\left(\frac{1}{N\ell^p}\sum_{i=1}^N E\left\|Z_i\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} &\lesssim \left(\frac{1}{N\ell}\sum_{j=1}^N\sum_{t=1}^\ell E\left\|\left(f_{t+j-1}-E\left(f_{t+j-1}\right)\right)\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} \\
&= \left(EE^*\left(\frac{1}{n}\sum_{t=1}^n\left\|\left(f_{t+j-1}^*-E\left(f_{t+j-1}\right)\right)\right\|_{\mathcal{N}_\eta}^p\right)\right)^{1/p} \\
&\lesssim \left(\left(n^{-1}\sum_{t=1}^n E\left\|\left(f_t-E\left(f_t\right)\right)\right\|_{\mathcal{N}_\eta}^p\right)^{1/p}+\eta O\left(\frac{\ell}{\sqrt{n}}\right)\right), \qquad (17)
\end{aligned}$$

where the last term is asymptotically negligible given the condition $\ell = o(\sqrt{n})$. Next, we can show that

$$\left(n^{-1}\sum_{t=1}^n E\left\|\left(f_t-E\left(f_t\right)\right)\right\|_{\mathcal{N}_\eta}^p\right)^{1/p} \lesssim \left(E\left\|N_\eta\right\|^p\right)^{1/p}, \qquad (18)$$

where N_η is the envelope of the function class $\mathcal{N}_\eta$. Given the Lipschitz continuity assumption (cf.

Assumption B6(iv)), we can show that $(E\|N_\eta\|^p)^{1/p} \leq \eta$. This implies

$$\begin{aligned} & N_0 \frac{\ell}{\sqrt{n}} \left(E_W \left| \max_{1\leq j\leq N} W_j \right|^p \right)^{1/p} \left(\frac{1}{N\ell^p} \sum_{i=1}^N E\|Z_i\|^p_{\mathcal{N}_\eta} \right)^{1/p} \\ \lesssim & \underbrace{\left[\frac{\ell}{\sqrt{n}} \left(E_W \left| \max_{1\leq j\leq N} W_j \right|^p \right)^{1/p} \right]}_{o(1)} \underbrace{(E\|N_\eta\|^p)^{1/p}}_{\lesssim \eta} = o(\eta), \end{aligned}$$

provided the second factor is $o(1)$. Given that $\max_{1\leq j\leq N} W_j^p \geq 1$, $(E_W |\max_{1\leq j\leq N} W_j|^p)^{1/p} \leq E_W \left(\max_{1\leq j\leq N} W_j^p \right)$. Therefore, we have

$$\frac{\ell}{\sqrt{n}} \left(E_W \left| \max_{1\leq j\leq N} W_j \right|^p \right)^{1/p} \lesssim \underbrace{\sqrt{\frac{N}{n}}}_{\to 1} \frac{\ell}{\sqrt{N}} E_W \left(\max_{1\leq j\leq N} W_j^p \right).$$

Next, we appeal to Lemma 4.7 of Praestgaard and Wellner (1993) to show that

$$\frac{\ell}{\sqrt{N}} E_W \left(\max_{1\leq j\leq N} W_j^p \right) = o(1).$$

To do so, we verify that ℓW_1^p satisfies the necessary conditions of Lemma 4.7 of Praestgaard and Wellner (1993), i.e., the following two conditions

$$\limsup_{N\to\infty} \|\ell W_1^p\|_{2,1} < \infty, \tag{19}$$

and

$$\lim_{\lambda\to\infty} \limsup_{N\to\infty} \sup_{u\geq\lambda} u^2 P_W (\ell W_1^p > u) = 0, \tag{20}$$

where we recall that W_1 is a an element of a triangular array, i.e. $W_1 = W_{N,1}$. As argued by Cheng (2015), cf. his equation (29), a sufficient condition to obtain both conditions (19) and (20) is that

$$\limsup_{N\to\infty} E_W \left(\ell W_1^{(2+\varepsilon)p} \right) < \infty, \tag{21}$$

for some $\varepsilon > 0$, which in turn is implied by

$$\limsup_{N\to\infty} E_W \left(\ell W_1^5 \right) < \infty,$$

because for a small enough $\epsilon > 0$, we can always choose $p = 5/(2+\epsilon) > 2$. Using the property of multinomial distribution, we have

$$\begin{aligned} E_W \left(W_1^5 \right) &= \frac{k}{N_n} + 15\frac{k(k-1)}{N_n^2} + 25\frac{k(k-1)(k-2)}{N_n^3} + 10\frac{k(k-1)(k-2)(k-3)}{N_n^4} \\ &\quad + \frac{k(k-1)(k-2)(k-3)(k-4)}{N_n^5} \\ &= \frac{\frac{n}{\ell_n} N_n^4 + 15\frac{n}{\ell_n}\left(\frac{n}{\ell_n}-1\right) N_n^3 + 25\frac{n}{\ell_n}\left(\frac{n}{\ell_n}-1\right)\left(\frac{n}{\ell_n}-2\right) N_n^2}{N_n^5} \\ &\quad + \frac{10\frac{n}{\ell_n}\left(\frac{n}{\ell_n}-1\right)\left(\frac{n}{\ell_n}-2\right)\left(\frac{n}{\ell_n}-3\right) N_n}{N_n^5} \\ &\quad + \frac{\frac{n}{\ell_n}\left(\frac{n}{\ell_n}-1\right)\left(\frac{n}{\ell_n}-2\right)\left(\frac{n}{\ell_n}-3\right)\left(\frac{n}{\ell_n}-4\right)}{N_n^5}. \end{aligned}$$

Given the condition $\ell = o\left(\sqrt{n}\right)$, it follows that

$$\lim \sup_{N\to\infty} E_W\left(\ell W_1^5\right) = 1 < \infty.$$

We follow the same arguments as in Cheng (2015, p. 19) and write

$$\begin{aligned}
&\left(E_{Z,R}\left(\max_{N_0<k\leq N}\left\|\frac{1}{\sqrt{k}}\sum_{i=N_0+1}^{k} Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p} \\
\lesssim\ &\left(E_{Z,R}\left(\max_{N_0<k\leq N}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p} + \left(E_{Z,R}\left(\left\|\frac{1}{\sqrt{N_0}}\sum_{i=N_0+1}^{N} Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p} \\
\leq\ &2\left(E_{Z,R}\left(\max_{N_0\leq k\leq N}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p}
\end{aligned}$$

where the last inequality follows by the triangular inequality. Thus, the proof of Lemma A.1 is completed when

$$\mathrm{II} \lesssim \left(\ell^{-1}E_{Z,R}\left(\max_{N_0<k\leq N}\left\|\frac{1}{\sqrt{k}}\sum_{i=N_0+1}^{k} Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p},$$

holds for $p > 2+\delta$. Let

$$\widetilde{\mathbb{G}}_k = \frac{1}{\sqrt{k}}\sum_{i=1}^{k} Z_{R(i)},$$

for $N_0 \leq k \leq N$. It follows that when $k = N$, we have

$$\widetilde{\mathbb{G}}_N = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} Z_{R(i)},$$

Recall that for any positive random variable Y, the following holds

$$EY^q = \int_0^\infty qu^{q-1}P\left(Y > u\right)du,$$

for any $q > 0$. The Levy inequality (see e.g., proposition A.1.2 of van der Vaart and Wellner (1996)) implies that,

$$P\left(\max_{k\leq N}\left\|\widetilde{\mathbb{G}}_k\right\|_{\mathcal{N}_\eta} > \lambda\right) \leq KP\left(\left\|\widetilde{\mathbb{G}}_N\right\|_{\mathcal{N}_\eta} > \lambda\right), \tag{22}$$

for every $\lambda > 0$. Hence, we can deduce that

$$\mathrm{II} \lesssim \left(\ell^{-1}E_{Z,R}\left(\max_{N_0<k\leq N}\left\|\frac{1}{\sqrt{k}}\sum_{i=N_0+1}^{k} Z_{R(i)}\right\|_{\mathcal{N}_\eta}\right)^p\right)^{1/p} \lesssim K^{1/p}\left(\ell^{-1}E_{Z,R}\left\|\widetilde{\mathbb{G}}_N\right\|_{\mathcal{N}_\eta}^p\right)^{1/p}.$$

Proof of Lemma A.2 The proof follows closely that of Lemma 4.1 in Praestgaard and Wellner (1993). Define a random permutation S of $\{1,\ldots,N\}$ such that $W_{S(1)} \geq \ldots \geq W_{S(N)}$, and if $W_{S(t)} =$

$W_{S(t+1)}$ then $S(t) < S(t+1)$. Then, let R be a random permutation uniformly distributed on Π_N (i.e., the set of permutations of $1, 2, \ldots, N$) and independent of $(\mathbf{W}, S)$. Using the same arguments as in Praestgaard and Wellner (2003), and given the exchangeability of $\mathbf{W}_N$, we have that

$$\begin{aligned}
\left(\mathbb{E}\left\|n^{-1/2}\sum_{j=1}^{N} W_j Z_j\right\|^p\right)^{1/p} &= \left(\mathbb{E}\left\|n^{-1/2}\sum_{j=1}^{N} W_{(j)} Z_{R(j)}\right\|^p\right)^{1/p} \\
&\leq \left(\mathbb{E}\left\|n^{-1/2}\sum_{j=1}^{N_0} W_{(j)} Z_{R(j)}\right\|^p\right)^{1/p} + \left(\mathbb{E}\left\|n^{-1/2}\sum_{j=N_0+1}^{N} W_{(j)} Z_{R(j)}\right\|^p\right)^{1/p} \\
&\equiv I(N_0, N) + II(N_0, N).
\end{aligned}$$

where $W_{(j)} = W_{S(j)}$. We can bound the term $I(N_0, N)$ by

$$\begin{aligned}
\mathrm{I}(N_0, N) &\leq n^{-1/2}\sum_{j=1}^{N_0}\left(\mathbb{E}\left\|W_{(j)} Z_{R(j)}\right\|^p\right)^{1/p} \text{ by Minkowski's inequality} \\
&\leq n^{-1/2}\sum_{j=1}^{N_0}\left(E_W\left|W_{(j)}\right|^p\right)^{1/p}\left(\mathbb{E}\left\|Z_{R(j)}\right\|^p\right)^{1/p} \text{ by independence between } W \text{ and } R \\
&\leq \left(E_W\left(n^{-1/2}\max_{1\leq j\leq N} W_j\right)^p\right)^{1/p}\sum_{j=1}^{N_0}\left(\mathbb{E}\left\|Z_{R(j)}\right\|^p\right)^{1/p} \\
&= \left(E_W\left(n^{-1/2}\max_{1\leq j\leq N} W_j\right)^p\right)^{1/p}\sum_{j=1}^{N_0}\left(\frac{1}{N}\sum_{i=1}^{N} E\left\|Z_i\right\|^p\right)^{1/p} \text{ by the properties of } R \\
&\leq \frac{N_0}{\sqrt{n}}\left(E_W\left(\max_{1\leq j\leq N} W_j\right)^p\right)^{1/p}\left(\frac{1}{N}\sum_{i=1}^{N} E\left\|Z_i\right\|^p\right)^{1/p}.
\end{aligned}$$

Note in particular that

$$\begin{aligned}
\mathbb{E}\left\|Z_{R(j)}\right\|^p &= E_Z E_{R|Z}\left(\left\|Z_{R(j)}\right\|^p\right) \text{ by the LIE} \\
&= E_Z\left(\frac{1}{N}\sum_{i=1}^{N}\left\|Z_i\right\|^p\right) = \frac{1}{N}\sum_{i=1}^{N} E\left\|Z_i\right\|^p.
\end{aligned}$$

If $E\left\|Z_i\right\|^p$ does not depend on i, then this is equal to $E\left\|Z_1\right\|^p$ and we get that

$$\mathrm{I}(N_0, N) \leq \frac{N_0}{\sqrt{n}}\left(E_W\left(\max_{1\leq j\leq N} W_j\right)^p\right)^{1/p}\left(E\left\|Z_1\right\|^p\right)^{1/p},$$

which is what Cheng (2015) get.

Next, in order to bound the second term i.e., $\mathrm{II}(N_0, N)$, we follow Cheng (2015) and write

$$\sum_{j=N_0+1}^{N} W_{(j)} Z_{R(j)} = \sum_{j=N_0+1}^{N}\sqrt{j}\left(W_{(j)} - W_{(j+1)}\right) T_j,$$

where $T_j = j^{-1/2}\sum_{i=N_0+1}^{j} Z_{R(i)}$ and $W_{(N+1)} = 0$. Hence, following Cheng (2015),

$$\begin{aligned}
&\mathrm{II}\,(N_0, N)\\
=\;& \left(\mathbb{E}\left\|n^{-1/2}\sum_{j=N_0+1}^{N} W_{(j)} Z_{R(j)}\right\|^p\right)^{1/p}\\
=\;& \left(\mathbb{E}\left\|n^{-1/2}\sum_{j=N_0+1}^{N} \sqrt{j}\left(W_{(j)} - W_{(j+1)}\right) T_j\right\|^p\right)^{1/p}\\
\leq\;& \left(E_{Z,R}\left\|\max_{N_0<k\leq N}\|T_k\|\right\|^p\right)^{1/p} \cdot \left(E_W\left\|n^{-1/2}\sum_{j=N_0+1}^{N} \sqrt{j}\left(W_{(j)} - W_{(j+1)}\right)\right\|^p\right)^{1/p}\\
=\;& \left(E_{Z,R}\left\|\max_{N_0<j\leq N}\left|\frac{1}{\sqrt{k}}\sum_{j=N_0+1}^{k} Z_{R(j)}\right|\right\|^p\right)^{1/p} \cdot \left(E_W\left\|n^{-1/2}\sum_{j=N_0+1}^{N} \sqrt{j}\left(W_{(j)} - W_{(j+1)}\right)\right\|^p\right)^{1/p}.
\end{aligned}$$

Thus, the proof is completed if we can show that

$$\begin{aligned}
\left(E_W\left\|n^{-1/2}\sum_{j=N_0+1}^{N} \sqrt{j}\left(W_{(j)} - W_{(j+1)}\right)\right\|^p\right)^{1/p} &\leq n^{-1/2}\left(E_W\left\|\sum_{j=N_0+1}^{N} \sqrt{j}\left(W_{(j)} - W_{(j+1)}\right)\right\|^p\right)^{1/p}\\
&\leq n^{-1/2}\left(n^{1/2}\left\|W_1\right\|_{2,1}^{1/p}\right) \leq \left\|W_1\right\|_{2,1}^{1/p}
\end{aligned}$$

i.e., if we can show that

$$E_W\left\|\sum_{j=N_0+1}^{N} \sqrt{j}\left(W_{(j)} - W_{(j+1)}\right)\right\|^p \leq n^{p/2}\left\|W_1\right\|_{2,1}. \tag{23}$$

It is easy to see that the proof is completed by using exactly the same arguments as in Cheng (2015) (cf. the proof of their equation (43)).

References

[1] Andrews, D. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica* 70, 119-162.

[2] Armstrong, T. B., M. Bertanha and H. Hong (2014). A fast resample method for parametric and semiparametric models. *Journal of Econometrics*, 179, 128–133.

[3] Cattaneo, M.D., M. Jansson and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies*, 86, 1095-1122.

[4] Cochrane, J. H. (2001). *Asset Pricing*. Princeton University Press.

[5] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In: *Handbook of Econometrics*, VI, Elsevier Science.

[6] Chen, X., and Z. Liao (2015). Sieve semiparametric two-step GMM with weakly dependent data. *Journal of Econometrics*, 189, 163-186.

[7] Chen, X., O. Linton and I. van Keilegom. Estimation of Semiparametric Models When the Criterion Function is not Smooth, *Econometrica*, 2003, 71, 1591-1608.

[8] Cheng, G. (2015). Moment Consistency of the Exchangeably Weighted Bootstrap for Semiparametric M estimation. *Scandinavian Journal of Statistics*, 42, 665-684.

[9] Davidson, R. and J. MacKinnon (1999). Bootstrap testing in nonlinear models. *International Economic Review*, 40, 487-508.

[10] Engle, R. and K. Sheppard (2001). Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH, mimeo.

[11] Gallant, A.R. and H. White (1988) *A unified theory of estimation and inference for nonlinear dynamic models*. New York: Basil Blackwell.

[12] Ghosh, M., Parr, W.C., Singh, K. and G.J. Babu (1984). A note on bootstrapping the sample median. *Annals of Statistics*, 12, 1130-1135.

[13] Gonçalves, S. and R. de Jong. (2003). Consistency of the stationary bootstrap under weak moment conditions. *Economics Letters*, 81, 273-278.

[14] Gonçalves, S. and H. White (2002). The Bootstrap of the mean for dependent heterogeneous arrays, *Econometric Theory*, 2002, 18, 1367-1384.

[15] Gonçalves, S. and H. White (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics*, 119, 199-219.

[16] Gonçalves, S. and H. White (2005). Bootstrap standard error estimates for linear regressions, *Journal of the American Statistical Association* Vol. 100, No. 471, 970-979.

[17] Kato, K. (2011). A note on moment convergence of bootstrap M-estimators. *Statistics & Decisions* 28, 51-61.

[18] Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17, 1217-1241.

[19] Joe, H (1995). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94, 401-19.

[20] Liu, R. Y. and K. Singh (1992). Moving blocks jackknife and bootstrap capture weak dependence. In: LePage, R., Billiard, L. (Eds.), *Exploring the Limits of the Bootstrap*, Wiley, New York, 224-248.

[21] Nishiyama, Y. (2010). Moment convergence of M-estimators. *Stat. Neerlandica* 64, 505-507.

[22] Nelsen, R. B. (1999). *An Introduction to Copulas*. Lecture notes in statistics, Springer.

[23] Newey, W.K., McFadden, D. (1994). Large sample estimation and hypothesis testing. In: Engle, R. McFadden, D. (Eds.), Handbook of Econometrics, vol IV, North Holland.

[24] Oh, D.H. and A.J. Patton (2013). Simulated method of moments estimation for copula-based multivariate models. *Journal of the American Statistical Association,*. 108, 689-700.

[25] Patton, A. J., (2006). Estimation of Multivariate Models for Time Series of Possibly Different Lengths, *Journal of Applied Econometrics*, 21(2), 14-173 .

[26] Patton, A. J., (2012). Copula Methods for Forecasting Multivariate Time Series, *in Handbook of Economic Forecasting*, eds. G. Elliott and A. Timmermann, (Vol. 2), Oxford: Elsevier.

[27] Praestgaard, J., and J. A. Wellner (1993). Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability*, 2053-2086.

[28] Shao, J. (1992). Bootstrap variance estimators with truncation. *Statistics and Probability Letters*, 15, 95-101.

[29] van der Vaart, A. W., and J. A. Wellner (1996). Weak convergence and empirical processes: with applications to statistics, *Springer,* New York.

[30] White, H. (1994). Estimation, Inference and Specification Analysis. Econometric Society Monographs 22, Cambridge University Press: Cambridge.

[31] Wooldridge, J. (1994). Estimation and inference for dependent processes. In: Engle, R. McFadden, D. (Eds.), Handbook of Econometrics, vol IV, North Holland.