Econometric Theory, **35**, 2019, 111–141.
doi:10.1017/S0266466618000063

A SIMPLE ITERATIVE Z-ESTIMATOR FOR SEMIPARAMETRIC MODELS

DAVID T. FRAZIER
Monash University

We propose a new iterative estimation algorithm for use in semiparametric models where calculation of Z-estimators by conventional means is difficult or impossible. Unlike a Newton–Raphson approach, which makes use of the entire Hessian, this approach only uses curvature information associated with portions of the Hessian that are relatively easy to calculate. Consistency and asymptotic normality of estimators obtained from this algorithm are established under regularity conditions and an information dominance condition. Two specific examples, a quantile regression model with missing covariates and a GARCH-in-mean model with conditional mean of unknown functional form, demonstrate the applicability of the algorithm. This new approach can be interpreted as an extension of the maximization by parts estimation approach to semiparametric models.

1. INTRODUCTION

Many economic models depend on functions that are known up to a vector of finite-dimensional parameters and other functions that are unknown but which satisfy some smoothness conditions. The most basic scenario is one in which the finite-dimensional parameters are the parameters of interest, the unknown functions, also called infinite-dimensional parameters, are nuisance parameters and the two parameters are only linked through the likelihood function or some estimating equations.

However, there are many interesting cases where the two parameters cannot be separated, at which point the parameters are said to be *bundled* together. This terminology was originally used by Huang and Wellner (1997) to describe linear models under censoring but has since been used to describe general semiparametric models; see, e.g., Ding and Nan (2011) and Nan and Wellner (2013). In situations where the parameters are *bundled* together the infinite-dimensional parameters directly depend on the finite-dimensional parameters; specific examples include quantile regression models with missing covariates, GARCH models with unknown conditional mean, certain failure time models, and many models with (non)semiparametric generated covariates (see, e.g., Escanciano, Jacho-Chavez,

This article has benefited tremendously from the feedback provided by two anonymous referees and a co-editor. Additional thanks are given to Eric Renault, Mervyn Silvapulle, Saraswata Chaudhuri, Yanqin Fan, Dennis Kristensen, and Oliver Linton. Address correspondence to David Frazier, Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia; e-mail: david.frazier@monash.edu.

© Cambridge University Press 2018

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

and Lewbel, 2014; Mammen, Roth, and Schienle, 2016 for discussion of models with (non)semiparametrically generated covariates).

There is now a rich theory for semiparametric estimation in contexts where the parameters are *bundled* together; see, e.g., Newey (1994), Ichimura and Lee (2010), and Ding and Nan (2011) for M-estimators and Chen, Linton, and Van Keilegom (2003), Ai and Chen (2003), and Nan and Wellner (2013) for Z-estimators. However, standard estimation procedures for these semiparametric models often require directly solving, usually through some derivative-based procedure, for the zeros of a complicated vector of estimating equations and can therefore be computationally burdensome, which greatly limits the applicability of such models to practitioners.

If directly solving the estimating equations is complicated, a backfitting estimation approach can instead be employed, whereby the procedure iterates between estimates of the infinite-dimensional parameters and the finite-dimensional parameters. The backfitting approach severs the relationship between the finite and infinite-dimensional parameters, allowing estimation of these components to be carried out in-turn instead of simultaneously. The relative simplicity of the backfitting approach greatly contributes to its use in semiparametric and nonparametric statistics. However, backfitting generally requires under-smoothing to obtain precise estimates of the finite-dimensional parameters and even upon convergence the estimators can be inefficient (see Pastorello, Patilea, and Renault, 2003; Hu, Wang, and Carroll, 2004 for examples). Simulation evidence in Hu et al. (2004) also demonstrates that backfitting estimators may perform poorly even in simple semiparametric models.

The main contribution of this article is to propose a new iterative estimator for use in semiparametric models with bundled parameters. This new estimator is simpler to implement than standard estimators but does not suffer from the issues of backfitting estimators. We present this new estimator in the context of Z-estimation. In contrast to standard Z-estimators, this new procedure circumvents the need to directly solve the estimating equations by actively searching over portions of the equations that are simple to evaluate, while the more complicated portions are evaluated iteratively. Under regularity conditions and an identification condition particular to iterative estimators, often called an information dominance condition (see, e.g., Song, Fan, and Kalbfleisch, 2005; Fan, Pastorello, and Renault, 2015), this iterative approach produces estimators with the same asymptotic behavior as standard Z-estimators but at a greatly reduced computational cost.

This new estimation procedure can be seen as an extension of the Maximization by Parts (hereafter, MBP) estimation approach to semiparametric models. In a parametric setting with an additive log-likelihood function, where one portion of the log-likelihood is much simpler than the other portion, Song et al. (2005) propose the idea of MBP as a way to overcome computational difficulties associated with obtaining Maximum Likelihood estimates. MBP has been used in several fields of statistical modeling, including copula models (Zhang,

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Czado, and Min, 2011), multivariate volatility models (Amado and Terasvirta, 2013; Noureldin, Shephard, and Sheppard, 2014), and non-normal linear mixed-effect models (Song, Zhang, and Qu, 2007).

Subsequently, Fan et al. (2015) have extended the MBP approach of Song et al. (2005) to general (parametric) extremum estimators. In particular, Fan et al. (2015) note that MBP estimators can be developed so long as the first-order conditions for the extremum estimation problem can be written as the sum of two components, with one component of the first-order conditions being simpler to (numerically) evaluate than the other component.

The approach developed herein builds on Fan et al. (2015) to devise a computationally simple approach to estimation in semiparametric models with bundled parameters. While this approach in presented in the context of a general Z-estimator, if the estimating equations that define this Z-estimator are the first-order conditions of an extremum estimation problem, our approach can be viewed as an extension of Fan et al. (2015) to semiparametric models. (See Section 2.2 for more details).

The remainder of this article proceeds as follows. Section 2 details the general framework and presents the new estimation procedure. Section 3 applies this approach to a GARCH model with conditional mean of unknown functional form. Section 4 contains asymptotic results. For completeness, Section 5 considers an additional application of this procedure to a quantile regression model with missing covariates. Section 6 concludes. Technical results are gathered in the Appendix.

2. FRAMEWORK AND THE ALGORITHM

2.1. Framework

The framework we consider is similar to that of Newey (1994), Andrews (1994), Chen et al. (2003), Ichimura and Lee (2010), and Nan and Wellner (2013). Let $Z_1, \ldots, Z_N \in \mathcal{Z}$ be vector valued random variables, let $\theta \in \Theta \subset \mathbb{R}^p$ be a vector of finite-dimensional parameters and denote by $\eta : \mathcal{Z} \times \Theta \mapsto \mathbb{R}^q$, an element of $\mathcal{H} := \{(z,\theta) \mapsto \eta(z,\theta) \in \mathbb{R}^q : z \in \mathcal{Z},\ \theta \in \Theta\}$, the infinite-dimensional parameters. Throughout we state the infinite-dimensional parameter using the convention $\eta(\cdot,\theta)$, where "$\cdot$" denotes other arguments in η besides θ (for instance, a subset of Z). Denote by θ^0 and $\eta^0(\cdot,\theta^0)$ the 'true' unknown value of the parameters θ and $\eta(\cdot,\theta)$.

Given an observed sample $\{Z_i\}_{i=1}^N$, our goal is to estimate θ^0 by searching for the zeros of a p-dimensional vector of data dependent estimating equations. The sample estimating equations depend on both θ and $\eta(\cdot,\theta)$ and are denoted by $\Psi_N(\theta, \eta(\cdot,\theta))$. Given an estimate of $\eta(\cdot,\theta)$, say $\widehat{\eta}(\cdot,\theta)$, define the Z-estimator $\widehat{\theta}_N$ as the solution to $0 = \Psi_N(\theta, \widehat{\eta}(\cdot,\theta))$.

As an example of this general framework, consider estimating θ^0 using the scalar criterion function

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

$$M_N(\theta, \eta(\cdot,\theta)) = \frac{1}{N}\sum_{i=1}^{N} m(Z_i;\theta,\eta(\cdot,\theta)).$$

Assume, for simplicity, we have in our possession an estimator of the unknown function, denoted by $\widehat{\eta}(\cdot,\theta)$. Then, maximizing $M_N(\theta,\widehat{\eta}(\cdot,\theta))$ with respect to (hereafter, wrt) θ is equivalent to solving the following first-order conditions:

$$0 = \Psi_N(\theta,\widehat{\eta}(\cdot,\theta)) = \frac{\partial}{\partial\theta} M_N(\theta,\widehat{\eta}(\cdot,\theta)) = \frac{1}{N}\sum_{i=1}^{N} \dot{m}_1(Z_i;\theta,\widehat{\eta}(\cdot,\theta)) + \frac{1}{N}\sum_{i=1}^{N} \hat{\dot{\eta}}(\cdot,\theta)' \dot{m}_\eta(Z_i;\theta,\widehat{\eta}(\cdot,\theta))$$
$$= \Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta)) + \Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta)), \quad \textbf{(1)}$$

where $\dot{m}_1(Z_i;\theta,\eta(\cdot,\theta))$ denotes the derivative of $m(Z;\theta,\eta(\cdot,\theta))$ wrt the first occurrence of θ, $\dot{m}_\eta(Z;\theta,\eta(\cdot,\theta))$ denotes the Frechet derivative of $m(Z;\theta,\eta)$ wrt η, evaluated at $\eta(\cdot,\theta)$, and $\hat{\dot{\eta}}(\cdot,\theta)$ denotes an estimator of the $q \times p$ dimensional Frechet derivative of $\eta(\cdot,\theta)$ wrt θ, denoted by $\dot{\eta}(\cdot,\theta)$.[1]

From equation (1), we see that M-estimators in semiparametric models with *bundled* parameters can be cast as Z-estimators based on estimating equations that can be written as the sum of two components, $\Psi_{1,N}(\cdot)$ and $\Psi_{2,N}(\cdot)$. Moreover, it is critical to realize that one term in the first-order conditions (1) is relatively simple to evaluate, while the other term is more cumbersome to evaluate. In particular, $\Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta))$ only depends on the ordinary partial derivative wrt θ, and thus should be simpler to evaluate than $\Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta))$, which depends on the functional derivative of $\widehat{\eta}(\cdot,\theta)$.[2]

The following two examples discuss specific models where the estimating equations for θ are of the form (1).

Example

Consider the partially linear median regression model:

$$Y_i = X_{1i}'\theta^0 + \eta^0(X_{2i}) + \epsilon_i, \;\; \text{med}(\epsilon_i | X_{1i}, X_{2i}) = 0,$$

where Y_i is a response, X_{1i} is a $p \times 1$ vector of covariates, X_{2i} is a scalar covariate, ϵ_i is an i.i.d. error term, θ is a vector of unknown regression coefficients, and $\eta(\cdot)$ is an unknown function. For fixed $\theta \in \Theta$, define the function $\eta^0(X_{2i},\theta) = \text{med}(Y_i - X_{1i}'\theta | X_{2i})$ and note that $\eta^0(X_{2i}) = \eta^0(X_{2i},\theta^0)$. Using some smoothing method, $\eta^0(\cdot,\theta)$ can be estimated, for each θ, using $Y_i - X_{1i}'\theta$ and X_{2i}. Denote this estimator as $\widehat{\eta}(\cdot,\theta)$. For $Z_i = (Y_i, {X_{1i}'}', X_{2i}')'$ and $m(Z_i,\theta,\widehat{\eta}(\cdot,\theta)) = -|Y_i - X_{1i}'\theta - \widehat{\eta}(X_{2i},\theta)|$, van Keilgom and Carroll (2007) argue that an estimator of θ^0 can be obtained by solving

$$0 = \frac{1}{N}\sum_{i=1}^{N} X_{1,i}\left[2I\left(Y_i - X_{1i}'\theta - \widehat{\eta}(X_{2i},\theta)\right) - 1\right] + \frac{1}{N}\sum_{i=1}^{N}\left[2I\left(Y_i - X_{1i}'\theta - \widehat{\eta}(X_{2i},\theta)\right) - 1\right]\hat{\dot{\eta}}(X_{2i},\theta)$$
$$= \Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta)) + \Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta)),$$

where $I(\cdot)$ is the indicator function and $\hat{\dot{\eta}}(\cdot,\theta)$ is an estimator of $\dot{\eta}(\cdot,\theta)$. In this example, $\Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta))$ is relatively simple to evaluate, while $\Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta))$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

is more complicated since calculating this component requires us to evaluate $\widehat{\eta}(\cdot,\theta)$.

Example
Let $Z_i = (Y_i', X_i')'$ denote the observed data and let θ denote a vector of finite-dimensional parameters. In many semiparametric models in econometrics, the nonparametric component of the model will depend on a generated covariate. A generated covariate is a random vector $T(X_i,\theta)$, of fixed dimension, that depends on data X_i and parameters θ. Mammen et al. (2016) detail examples, and asymptotic theory, where the nonparametric component of the model is $\eta(\cdot,\theta) = E(Y|T(X,\theta) = \cdot)$ and where θ is estimated from the criterion function $M_N(\theta,\widehat{\eta}(\cdot,\theta)) = \frac{1}{N}\sum_{i=1}^{N} m(Z_i,\theta,\widehat{\eta}(\cdot,\theta))$, for some estimator $\widehat{\eta}(\cdot,\theta)$. Therefore, when $m(Z_i,\theta,\widehat{\eta}(\cdot,\theta))$ is smooth, a Z-estimator of θ can be obtained by solving estimating equations similar to equation (1).

From the above examples, we see that many semiparametric models will lead to Z-estimators that solve a vector of estimating equations $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ with a similar structure to equation (1); additional examples are given in Sections 3 and 5. That is, the estimating equations $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$, which define $\widehat{\theta}_N$, can often be written as

$$\Psi_N(\theta,\widehat{\eta}(\cdot,\theta)) = \Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta)) + \Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta)), \tag{2}$$

where one portion of $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ will be relatively easy to evaluate, while the other portion will be more cumbersome to evaluate. Without loss of generality, throughout the remainder we let $\Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta))$ denote the relatively simple portion of $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ and $\Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta))$ will denote the more complicated portion of $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$.

2.2. A Simple Z-Estimators for Bundled Parameters

Note that $\Psi_N(\theta,\eta(\cdot,\theta))$ depends on θ in several places. Therefore, calculating $\widehat{\theta}_N$ by directly solving $0 = \Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$, for a given estimator $\widehat{\eta}(\cdot,\theta)$, can be difficult in practice. If one were to use a Newton–Raphson (hereafter, NR) approach to calculate $\widehat{\theta}_N$ this would entail, at step $k \geq 1$, updating an initial estimator using

$$\hat{\theta}^{k+1} = \hat{\theta}^k - \left\{\partial\Psi_N(\hat{\theta}^k,\widehat{\eta}(\cdot,\hat{\theta}^k))/\partial\theta'\right\}^{-1}\left\{\Psi_{1,N}(\hat{\theta}^k,\widehat{\eta}(\cdot,\hat{\theta}^k)) + \Psi_{2,N}(\hat{\theta}^k,\widehat{\eta}(\cdot,\hat{\theta}^k))\right\}, \tag{3}$$

where $\partial\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))/\partial\theta'$ denotes the derivative of $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ wrt all occurrences of θ.[3] The presence of $\widehat{\eta}(\cdot,\theta)$ within $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ ensures that closed form formulas for $\partial\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))/\partial\theta'$ are most likely unavailable. Methods based on finite differences or Gauss–Newton type-algorithms are possible but can encounter numerical issues.[4]

Recall that, and with reference to equation (2), $\Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta))$ denotes the portion of $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ that is relatively easy to evaluate, while $\Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta))$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

denotes the portions that are more cumbersome to evaluate. Since $\Psi_{1,N}(\cdot)$ is relatively simple, solving $0 = \Psi_{1,N}(\theta, \widehat{\eta}(\cdot, \theta))$ should be relatively simple and an estimator, say $\widehat{\theta}_N^{(1)}$, could be obtained by solving $0 = \Psi_{1,N}(\theta, \widehat{\eta}(\cdot, \theta))$. Unfortunately, $\widehat{\theta}_N^{(1)}$ is at best inefficient relative to $\widehat{\theta}_N$ and at worst inconsistent. In most cases consistent estimation requires using the information about θ contained in $\Psi_{2,N}(\theta, \widehat{\eta}(\cdot, \theta))$. Moreover, from a practical standpoint, directly searching over the occurrence of θ in $\widehat{\eta}(\cdot, \theta)$ can be numerically cumbersome and should be avoided when possible.

Given the above issues, it would be useful to devise an estimation procedure that, one, uses the information contained in $\Psi_{2,N}(\theta, \widehat{\eta}(\cdot, \theta))$ without requiring us to directly search over the occurrences of θ in $\Psi_{2,N}(\theta, \widehat{\eta}(\cdot, \theta))$ and, two, does not require us to search over any occurrences of θ in $\widehat{\eta}(\cdot, \theta)$. An update of $\widehat{\theta}_N^{(1)}$ that adheres to these two requirements can be obtained by solving

$$\Psi_{1,N}(\theta, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(1)})) = -\Psi_{2,N}(\widehat{\theta}_N^{(1)}, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(1)})). \tag{4}$$

Iterating equation (4) produces the following estimation algorithm.

Algorithm 1.

Step 1. Obtain initial estimators $\widehat{\theta}_N^{(1)}, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(1)})$.

Step k+1. For $k \geq 1$, update $\widehat{\theta}_N^{(k)}$ by solving $\Psi_{1,N}(\theta, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)})) = -\Psi_{2,N}(\widehat{\theta}_N^{(k)}, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)}))$.

Remark. Assuming sufficient smoothness, **Step k+1** of Algorithm 1 can be replaced by a NR-type iteration associated with solving equation (4), which yields the alternative updating formula for **Step k+1**:

$$\widehat{\theta}_N^{(k+1)} = \widehat{\theta}_N^{(k)} - \left[\partial \Psi_{1,N}(\theta, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)}))/\partial \theta' \big|_{\theta = \widehat{\theta}_N^{(k)}}\right]^{-1} \Psi_N(\widehat{\theta}_N^{(k)}, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)})). \tag{5}$$

It is important to point out that, in contrast to a full NR procedure, the updating rule in (5) only requires calculating a portion of $\partial \Psi_N(\theta, \widehat{\eta}(\cdot, \theta))/\partial \theta'$.

Remark. Algorithm 1 resembles Algorithm 2.1 in Fan et al. (2015) (see their equation (2.7)) and can therefore be seen as a semiparametric extension of the MBP approach in Fan et al. (2015). In particular, if $\Psi_N(\theta, \widehat{\eta}(\cdot, \theta))$ represents the first-order conditions of an extremum criterion, as with the example of the M-estimator criterion $M_N(\theta, \widehat{\eta}(\cdot, \theta))$ given earlier, Algorithm 1 is the semiparametric extension of the MBP approach in Fan et al. (2015). The fundamental differences between these two approaches is that existing MBP approaches can only be applied to solve smooth estimating equations that depend on finite-dimensional parameters and hence are not applicable to the semiparametric setting of interest herein. However, given the similarities between the above algorithm and the MBP procedure in Fan et al. (2015), we interpret the above algorithm as an extension of the MBP approach in Fan et al. (2015) to semiparametric Z-estimation. In the sequel we refer to $\widehat{\theta}_N^{(k)}$ in Algorithm 1 as a *semiparametric* MBP (SP-MBP) estimator, while Algorithm 1 is referred to as the SP-MBP algorithm.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Remark. Let $\bar{\theta}_N[\cdot;\widehat{\eta}]$ be a function such that $\widehat{\theta}_N^{(k+1)}$ defined by **Step k+1** satisfies $\widehat{\theta}_N^{(k+1)} = \bar{\theta}_N[\widehat{\theta}_N^{(k)};\widehat{\eta}]$, for some possibly large value of k. Following Dominitz and Sherman (2005), we note that if $\bar{\theta}_N[\cdot;\widehat{\eta}]$ is an asymptotic contraction mapping then there exists a fixed-point that coincides with $\widehat{\theta}_N$, the solution to $\Psi_{1,N}(\theta,\widehat{\eta}(\cdot,\theta)) = -\Psi_{2,N}(\theta,\widehat{\eta}(\cdot,\theta))$, and $\widehat{\theta}_N^{(k)}$ converges to $\widehat{\theta}_N$ as $k \to \infty$. When θ^0 is a scalar, $\bar{\theta}_N[\theta_1;\widehat{\eta}]$ will be an asymptotic contraction mapping if the absolute value of its derivative, wrt θ and evaluated at θ^0, is less than one asymptotically. If θ^0 is a vector, $\bar{\theta}_N[\theta_1;\widehat{\eta}]$ will be an asymptotic contraction mapping if the norm of the derivative of $\bar{\theta}_N[\theta_1;\widehat{\eta}]$, evaluated at θ^0, is less than one asymptotically. This condition is called the *Information Dominance Condition* and is essentially a local-contraction mapping condition (Fan et al., 2015).

3. A GARCH-IN-MEAN EXAMPLE

In this section, we use a GARCH-in-mean model with unknown conditional mean to demonstrate the SP-MBP algorithm and to compare the finite-sample performance of this approach against existing iterative estimators.

GARCH models are a common means of modeling financial returns possessing constant or zero conditional mean. However, in certain cases the assumption of zero conditional mean is dubious. In these cases our focus of interest often shifts to analyzing the relationship between the conditional variance and conditional mean of excess returns, the so-called risk return tradeoff. There is a long history of such studies in the financial and economic literature with the most common model used in the analysis being the GARCH-in-mean (GARCH-M) model of Engle, Lilien, and Robins (1987). Parametric GARCH-M models assume that excess returns possess a nonzero conditional mean and can be modeled as

$$y_t = \mu + \gamma\, g(\sigma_t^2) + \epsilon_t,\ t = 1,\ldots,N, \tag{6}$$

for some known scalar function $g(\cdot)$, usually log or identity, and some GARCH conditional variance.

However, the model in (6) is restrictive since it requires that the relationship between conditional mean and conditional variance be linear. Linton and Perron (2003) propose a semiparametric version of (6), termed a semiparametric GARCH-M model, where

$$y_t = \eta(\sigma_t^2) + \epsilon_t,\ t = 1,\ldots,N, \tag{7}$$

and $\eta(\cdot)$ a smooth unknown function. In particular, Linton and Perron (2003) consider the case where σ_t^2 follows the exponential GARCH (EGARCH(p,q)) model of Nelson (1991)

$$\log(\sigma_t^2) = \omega + \sum_{i=1}^{p} \beta_i \log(\sigma_{t-i}^2) + \sum_{j=1}^{q} \alpha_j [(|\epsilon_{t-j}| - E|\epsilon_{t-j}|) + \lambda \epsilon_{t-j}].$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Using this model and a backfitting quasi-maximum likelihood (QML) estimation strategy, Linton and Perron (2003) find a nonlinear relationship between the conditional mean and variance of excess returns on the S&P500 index.

A similar semiparametric GARCH-M model was proposed by Christensen, Dahl, and Iglesias (2012), where the conditional variance specification is similar to the double-autoregressive model of Ling (2004):

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{8}$$

Using this form of the semiparametric GARCH-M model, Christensen et al. (2012) find a nonlinear relationship between the conditional mean and variance of excess returns.

Regardless of the parametric specification for the conditional variance, clearly $\eta(\sigma_t^2) := \eta(\sigma_t^2(\theta))$ and hence the parameters are *bundled* together. To maintain consistent notation, in the sequel we denote $\eta(\sigma_t^2)$ by $\eta(\sigma_t^2, \theta)$.

3.1. Parameter Estimation in the Semiparametric GARCH-M Model

Our goal is to estimate the parameters of the semiparametric GARCH-M model defined by equations (7) and (8). For $\theta = (\omega, \alpha, \beta)'$, denote the QML objective function as

$$M_N(\theta, \eta(\cdot, \theta)) = -\frac{1}{2}\sum_{t=1}^{N} \ln(\sigma_t^2) - \frac{1}{2}\sum_{t=1}^{N} \frac{(y_t - \eta(\sigma_t^2, \theta))^2}{\sigma_t^2}.$$

Christensen et al. (2012) propose a two-step strategy to estimate θ^0 and $\eta^0(\cdot, \theta)$: first, fix θ and maximize a locally smoothed version of $M_N(\theta, \eta(\cdot, \theta))$ wrt η to obtain the estimator $\widehat{\eta}(\cdot, \theta)$;[5] subsequently, form $M_N(\theta, \widehat{\eta}(\cdot, \theta))$ and maximize the resulting quantity wrt θ. However, because $\eta(\cdot)$ depends on $\sigma_t^2 = \sigma_t^2(\theta)$, $\widehat{\eta}(\cdot, \theta)$ cannot be estimated for fixed θ. To alleviate this issue, Christensen et al. (2012) implement the above two-step approach through the following iterative steps, which are carried out till convergence is achieved.

Step 1. Obtain initial estimators $\widetilde{\theta}_N^{(1)}, \widetilde{\eta}(\cdot, \widetilde{\theta}_N^{(1)})$.

Step k+1. Update $\widetilde{\theta}_N^{(k)}$ by maximizing $M_N(\theta, \widetilde{\eta}(\cdot, \widetilde{\theta}_N^{(k)}))$, where $\widetilde{\eta}(\cdot, \widetilde{\theta}_N^{(k)})$ is obtained by maximizing a locally smoothed version of $M_N(\theta, \eta)$, wrt η, at the point $\theta = \widetilde{\theta}_N^{(k)}$.

Christensen et al. (2012) claim the above estimator of θ^0 converges to the quasi-maximum likelihood estimate (QMLE) $\widehat{\theta}_N$, defined as the solution to

$$0 = \sum_{t=1}^{N} \left(\frac{1}{\hat{\sigma}_t^2(\theta)} - \frac{\tilde{\epsilon}_t^2(\theta)}{(\hat{\sigma}_t^2(\theta))^2} \right) \frac{\partial \hat{\sigma}_t^2(\theta)}{\partial \theta} - \sum_{t=1}^{N} \frac{\tilde{\epsilon}_t(\theta)}{\hat{\sigma}_t^2(\theta)} \dot{\tilde{\eta}}(\hat{\sigma}_t^2, \theta), \tag{9}$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

where $\tilde{\epsilon}_t(\theta^*) = y_t - \widetilde{\eta}(\hat{\sigma}_t^2(\theta^*), \theta^*)$, $\hat{\sigma}_t^2(\theta^*) = \omega^* + \alpha^* y_{t-1}^2 + \beta^* \hat{\sigma}_{t-1}^2(\theta^*)$ is the conditional variance estimate calculated at the point $\theta = \theta^*$, and $\dot{\eta}(\cdot, \theta)$ denotes the Frechet derivative of $\eta(\cdot, \theta)$ wrt θ. Letting

$$\Psi_{1,N}(\theta, \eta(\cdot, \theta)) = \sum_{t=1}^{N} \left(\frac{1}{\sigma_t^2(\theta)} - \frac{\epsilon_t^2(\theta)}{(\sigma_t^2(\theta))^2} \right) \frac{\partial \sigma_t^2(\theta)}{\partial \theta}, \quad \Psi_{2,N}(\theta, \eta(\cdot, \theta)) = \sum_{i=1}^{N} \frac{\epsilon_t(\theta)}{\sigma_t^2(\theta)} \dot{\eta}(\sigma_t^2, \theta),$$

we arrive at the general specification considered throughout the article. The term $\Psi_{1,N}(\theta, \eta(\cdot, \theta))$ is essentially the quasi-score from a parametric GARCH model and the term $\Psi_{2,N}(\theta, \eta(\cdot, \theta))$ captures the change in the conditional mean wrt a change in the conditional variance parameters.

While the algorithm of Christensen et al. (2012) is simple to implement, it is clear that $\widetilde{\theta}_N^{(k)}$ solves, at each iteration, a set of estimating equations that differ from (9). In particular, $\widetilde{\theta}_N^{(k+1)}$ solves

$$0 = \Psi_{1,N}(\theta, \widetilde{\eta}(\cdot, \widetilde{\theta}_N^{(k)})) = \sum_{t=1}^{N} \left(\frac{1}{\tilde{\sigma}_t^{2(k)}} - \frac{\left(\tilde{\epsilon}_t^{(k)}\right)^2}{(\tilde{\sigma}_t^{2(k)})^2} \right) \frac{\partial \tilde{\sigma}_t^{2(k)}}{\partial \theta}, \tag{10}$$

where $\tilde{\epsilon}_t^{(k)} = \tilde{\epsilon}_t(\widetilde{\theta}_N^{(k)})$, $\tilde{\sigma}_t^{2(k)} = \hat{\sigma}_t^2(\widetilde{\theta}_N^{(k)})$ and so there is generally no reason to suspect that $\widehat{\theta}_N$, the solution to (9), and $\widetilde{\theta}_N^{(k+1)}$, the solution to (10), will be asymptotically equivalent as $k \to \infty$. From this analysis, we see that the estimator proposed in Christensen et al. (2012) is actually a backfitting estimator and therefore inherits the potential issues associated with backfitting estimators (see, e.g., Pastorello et al., 2003; Hu et al., 2004 for a discussion of these issues).

While the approach of Christensen et al. (2012) delivers a feasible estimator, indeed, this type of backfitting approach is the only existing implementation for semiparametric GARCH-M models in the literature, we expect this approach to perform worse than an estimator that utilizes all of the information contained in (9).

The SP-MBP algorithm can deliver an estimator that is no more computationally involved than the estimator of Christensen et al. (2012), but that utilizes all of the information contained in (9). To see this, note that, the **(k+1)**-th step of the SP-MBP and backfitting approaches solve (respectively)

$$\Psi_{1N}(\theta, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)})) = -\Psi_{2N}(\widehat{\theta}_N^{(k)}, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)})) \text{ (SP-MBP)}$$
$$\Psi_{1,N}(\theta, \widetilde{\eta}(\cdot, \widetilde{\theta}_N^{(k)})) = 0 \text{ (backfitting)}.$$

Moreover, while the backfitting estimator solves $\Psi_{1,N}(\theta, \widetilde{\eta}(\cdot, \widetilde{\theta}_N^{(k)})) = 0$, the SP-MBP estimator only requires solving $\Psi_{1,N}(\theta, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)})) = -C$, for $C = \Psi_{2,N}(\widehat{\theta}_N^{(k)}, \widehat{\eta}(\cdot, \widehat{\theta}_N^{(k)}))$. We can therefore interpret the SP-MBP estimator as a correction of the backfitting estimator that utilizes a recentered vector of estimating equations. In addition, as the number of iterations diverge, this recentering utilizes

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

all of the available information in the estimating equations and yields an estimator as efficient as the Z-estimator $\widehat{\theta}_N$ that solves (9).

Precise implementation of the SP-MBP procedure defines $\widehat{\theta}_N^{(k+1)}$ as the solution to

$$0 = \sum_{t=1}^{N} \left(\frac{1}{\hat{\sigma}_t^{2(k+1)}} - \frac{(\hat{\epsilon}_t^{(k)})^2}{\left(\hat{\sigma}_t^{2(k+1)}\right)^2} \right) \frac{\partial \hat{\sigma}_t^{2(k+1)}}{\partial \theta} - \sum_{t=1}^{N} \frac{\hat{\epsilon}_t^{(k)}}{\hat{\sigma}^{2(k)}} \hat{\eta}(\hat{\sigma}_t^{2(k)}, \widehat{\theta}_N^{(k)}), \tag{11}$$

where $\hat{\epsilon}_t^{(k)} = \hat{\epsilon}_t(\widehat{\theta}_N^{(k)})$ and $\hat{\sigma}_t^{2(k)} = \hat{\sigma}_t^2(\widehat{\theta}_N^{(k)})$. Obtaining $\widehat{\theta}_N^{(k+1)}$ using a NR procedure is simple since the second-derivatives used in the calculations are known in closed form and are given by (see Fiorentini, Calzolari, and Panattoni, 1996 for details)

$$\ddot{\Psi}_{1,N}(\theta, \widehat{\eta}(\cdot, \theta)) = \frac{1}{2} \sum_{t=1}^{N} \left\{ \left[\frac{\hat{\epsilon}_t^2(\theta)}{\sigma_t^2} - 1 \right] \left\{ \frac{1}{\sigma_t^2} \frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'} - \frac{1}{(\sigma_t^2)^2} \frac{\partial \sigma_t^2}{\partial \theta} \frac{\partial \sigma_t^2}{\partial \theta'} \right\} - \frac{1}{(\sigma_t^2)^2} \frac{\partial \sigma_t^2}{\partial \theta} \frac{\partial \sigma_t^2}{\partial \theta'} \frac{\hat{\epsilon}_t^2(\theta)}{\sigma_t^2} \right\},$$

where, for $z_t = (1, y_{t-1}^2, \sigma_{t-1}^2)'$,

$$\frac{\partial \sigma_t^2}{\partial \theta} = z_t + \beta \frac{\partial \sigma_{t-1}^2}{\partial \theta},$$
$$\frac{\partial^2 \sigma_t^2}{\partial \theta \partial \theta'} = \frac{\partial z_t}{\partial \theta} + \frac{\partial \sigma_{t-1}^2}{\partial \theta} \frac{\partial \beta}{\partial \theta'} + \beta \frac{\partial \sigma_{t-1}^2}{\partial \theta \partial \theta'},$$

and where

$$\frac{\partial z_t}{\partial \theta} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{\partial \sigma_{t-1}^2}{\partial \omega} & \frac{\partial \sigma_{t-1}^2}{\partial \alpha} & \frac{\partial \sigma_{t-1}^2}{\partial \beta} \end{pmatrix}, \quad \frac{\partial \sigma_{t-1}^2}{\partial \theta} \frac{\partial \beta}{\partial \theta'} = \frac{\partial z_t}{\partial \theta'}.$$

3.2. Simulation Experiments

We analyze the finite sample properties of the SP-MBP estimator and compare its performance against the backfitting approach of Christensen et al. (2012). We consider two separate data generating processes, denoted A and B, given by[6]

A: $y_t = \sigma_t^2 + 0.5 \sin(10\sigma_t^2) + \sigma_t \epsilon_t$, $\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2$, with $\beta = 0.68$;

B: $y_t = 0.5\sigma_t^2 + 0.1 \sin(0.5 + 20\sigma_t^2) + \sigma_t \epsilon_t$, $\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2$, with $\beta = 0.80$;

where ϵ_t are i.i.d. standard normal and the parameters (ω, γ) are fixed at $(0.01, 0.10)$. For both estimation methods the unknown function $\eta(\cdot)$ is approximated using quadratic splines.

To define the spline approximation for $\eta(\cdot)$ let $[a, b]$ be the range of $\sigma_t^2 := \sigma_t^2(\theta)$. Let the knots $\{t_1, \ldots, t_{R_N}\}$ partition $[a, b]$, where the total number of knots

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

TABLE 1. True parameters under specification A: $\theta^0 = (\omega^0, \alpha^0, \beta^0)' = (0.01, 0.1, 0.68)'$ and $\eta(\sigma^2) = \sigma^2 + 0.5\sin(10\sigma^2)$. SP-MBP: SP-MBP estimator, BF: backfitting estimator

	$N = 500$	ω	α	β	$N = 1{,}000$	ω	α	β
SP-MBP	BIAS	0.0013	0.0009	−0.0161	BIAS	0.0004	0.0007	−0.0073
	STD	0.0071	0.0195	0.0576	STD	0.0053	0.0143	0.0421
	MAE	0.0044	0.0141	0.0406	MAE	0.0035	0.0106	0.0302
	RMSE	0.0073	0.0195	0.0598	RMSE	0.0053	0.0143	0.0427
	$N = 500$	ω	α	β	$N = 1{,}000$	ω	α	β
BF	BIAS	0.0043	0.0003	−0.0328	BIAS	0.0016	0.0008	−0.0138
	STD	0.0126	0.0227	0.0869	STD	0.0065	0.0156	0.0478
	MAE	0.0081	0.0170	0.0630	MAE	0.0049	0.0119	0.0377
	RMSE	0.0133	0.0227	0.0928	RMSE	0.0067	0.0156	0.0497

satisfies $R_N = O(N^\nu)$, $\max_{1\le j\le R_N+1} |t_j - t_{j-1}| = O(N^{-\nu})$ and $\nu \in (0, 1/2)$. Given R_N knots, the quadratic spline approximation of $\eta(\sigma_t^2, \theta)$ is given by

$$\eta(\sigma_t^2, \theta) = \sum_{j=0}^{2} \gamma_j (\sigma_t^2)^j + \sum_{k=1}^{R_N} \gamma_{2+k} (\sigma_t^2 - t_k)^2 I(\sigma_t^2 \ge t_k),$$

where $I(\cdot)$ is the indicator function. The estimator $\widehat{\eta}(\sigma^2, \theta)$ is then obtained by estimating the corresponding coefficients γ_j, $j = 0, \ldots, R_N$, which, from the definition of $M_N(\theta, \eta)$, only requires simple least-squares calculations.

Lastly, for both estimators, SP-MBP and backfitting, we choose the knots $\{t_1, \ldots, t_{R_N}\}$ to partition the percentiles of the estimated σ_t^2. The specific number of knots R_N is taken to be the floor of $N^{3/20}$.

For specifications A, B we consider two sample sizes, $N = 500$ and $N = 1{,}000$, with 1,000 replications across both designs. The results for the estimates of θ^0 are given in Tables 1 and 2, which detail the bias (BIAS), standard deviation (STD), mean absolute error (MAE), and root mean squared error (RMSE) across the different sample sizes and specifications. For both estimators the numerical

TABLE 2. True parameters under specification B: $\theta^0 = (\omega^0, \alpha^0, \beta^0)' = (0.01, 0.1, 0.8)'$ and $\eta(\sigma^2) = 0.5\sigma^2 + 0.1\sin(0.5 + 20\sigma^2)$. Table entries are the same as in Table 1

	$N = 500$	ω	α	β	$N = 1{,}000$	ω	α	β
SP-MBP	BIAS	0.0028	0.0015	−0.0322	BIAS	0.0012	0.0006	−0.0143
	STD	0.0082	0.0369	0.0919	STD	0.0052	0.0246	0.0620
	MAE	0.0048	0.0257	0.0580	MAE	0.0034	0.0180	0.0414
	RMSE	0.0086	0.0369	0.0974	RMSE	0.0054	0.0246	0.0636
	$N = 500$	ω	α	β	$N = 1{,}000$	ω	α	β
BF	BIAS	0.0083	0.0106	−0.0892	BIAS	0.0042	0.0077	−0.0473
	STD	0.0165	0.0439	0.1567	STD	0.0112	0.0298	0.1104
	MAE	0.0101	0.0324	0.1103	MAE	0.0059	0.0222	0.0693
	RMSE	0.0184	0.0451	0.1803	RMSE	0.0119	0.0307	0.1201

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

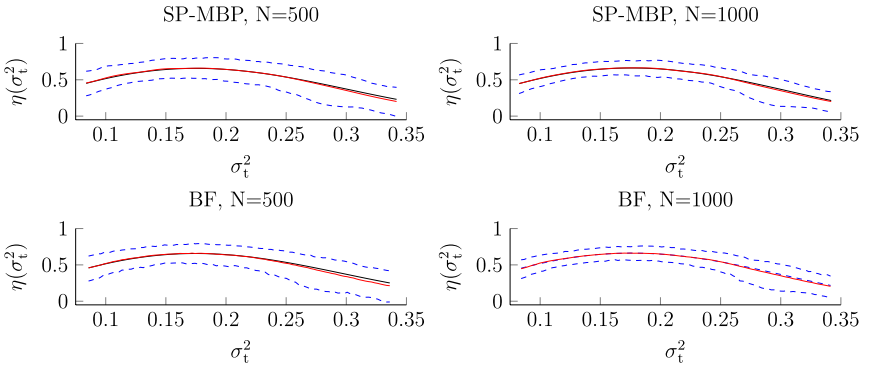


FIGURE 1. Conditional mean estimates under specification A. The black line is the true function, the red line is the mean estimate and the blue dashed lines are corresponding 0.025 and 0.975 quantiles (across the Monte Carlo replications). SP-MBP: SP-MBP estimator, BF: backfitting estimator.

tolerance for convergence was $1.0e^{-05}$ and the convergence criteria was chosen to be $|\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)}|$, where $|\cdot|$ is the Euclidean norm.

The results in Tables 1 and 2 demonstrate that the SP-MBP approach yields more precise estimates with smaller bias than those obtained from backfitting (i.e., the estimator of Christensen et al., 2012). As alluded to earlier, this result is due to the recentering of the SP-MBP estimating equations at every iteration.

Figures 1 and 2 present the resulting conditional mean estimates for both estimators across all four Monte Carlo designs. These figures demonstrate that both estimation procedures yield precise estimates of the unknown conditional mean across the different designs.

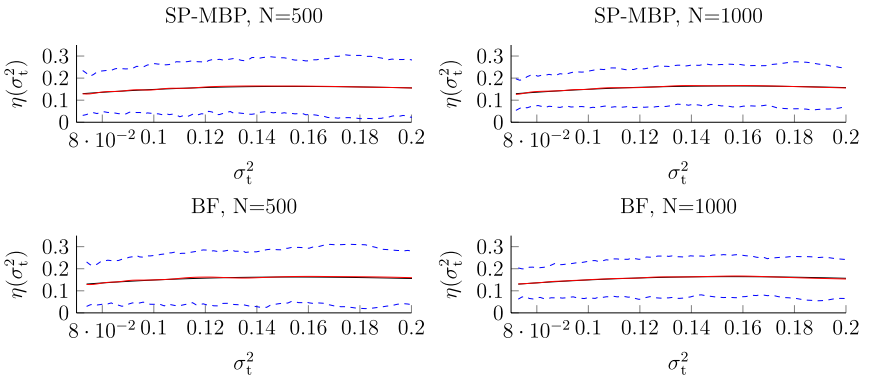


FIGURE 2. Conditional mean estimates under specification B. The plot is the same as in Figure 1.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

TABLE 3. Computing time comparison between SP-MBP and backfitting. The table entry is the ratio of average computing time, over the replications, for the SP-MBP and backfitting approaches. Ratios larger than unity indicate that SP-MBP has a larger computational time than backfitting. N is the sample size and Spec. refers to specification A or B in the simulation design

(N, Spec.)	(500, A)	(1,000, A)	(500, B)	(1,000, B)
SP-MBP/BF	1.431445	1.775611	1.46305	1.852346

To measure the computational cost of both estimation methods, Table 3 reports the relative computational time required for both estimation approaches over the different simulation designs. The results indicate that SP-MBP is slightly more computationally costly than backfitting. This result is not surprising given that SP-MBP must calculate, at each iteration, a recentering term, which backfitting is not required to calculate. However, it should also be noted that the SP-MBP estimator converges quickly, with the average number of iterations required to obtain convergence being between nine and twelve across the different specifications. Therefore, while the SP-MBP approach delivers better performance than backfitting, it does so at a slightly higher computational cost.

In total, these results demonstrate that the SP-MBP algorithm yields accurate estimates for both the finite and infinite-dimensional components in the semiparametric GARCH-M model at a cost that is not significantly higher than that of the backfitting estimator.

4. ASYMPTOTIC THEORY

This section develops the asymptotic properties of the SP-MBP estimator $\widehat{\theta}_N^{(k)}$. Let $|\cdot|$ denote the Euclidean norm. For $\mathcal{H}$ a class of functions from $\mathcal{Z} \times \Theta \to \mathbb{R}^q$, define $\|\cdot\|_{\mathcal{H}}$ as the sup-norm over Θ and $\mathcal{Z}$; i.e., for $\eta \in \mathcal{H}$, $\|\eta\|_{\mathcal{H}} = \sup_{z\in\mathcal{Z}} \sup_{\theta\in\Theta} |\eta(z,\theta)|$. Let θ^0 denote the true value of θ and $\eta^0(\cdot,\theta)$ the true functional form of η.

To simplify the discussion, we define a sample 'score' function associated with the SP-MBP algorithm

$$S_N(\theta,\theta_1;\eta) = \Psi_{1,N}(\theta,\eta(\cdot,\theta_1)) + \Psi_{2,N}(\theta_1,\eta(\cdot,\theta_1)).$$

The notion of θ and θ_1 in $S_N(\theta,\theta_1;\eta)$ is taken with reference to **Step k+1** of the SP-MBP algorithm (i.e., Algorithm 1): θ denotes the occurrences over which the algorithm directly searches and θ_1 denotes the occurrences over which the algorithm iterates. Using $S_N(\theta,\theta_1;\widehat{\eta})$ we can equivalently define $\widehat{\theta}_N^{(k)}$ as $\widehat{\theta}_N^{(k)} = \bar{\theta}_N[\widehat{\theta}_N^{(k-1)};\widehat{\eta}]$, where

$$\bar{\theta}_N[\theta_1;\eta] = \arg\min_{\theta\in\Theta} |S_N(\theta,\theta_1;\eta)|.$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

4.1. Consistency

Consistency of $\widehat{\theta}_N^{(k)}$ for θ^0 (as $N \to \infty$) requires the existence and regularity of limit counterparts to $S_N(\theta, \theta_1; \eta)$ and $\bar{\theta}_N[\theta_1; \eta]$. To this end, define

$$S(\theta, \theta_1; \eta) = \operatorname*{plim}_{N\to\infty} S_N(\theta, \theta_1; \eta) = \Psi_1(\theta, \eta(\cdot, \theta_1)) + \Psi_2(\theta_1, \eta(\cdot, \theta_1)).$$

Likewise, for fixed $\theta_1 \in \Theta$ and some $\eta \in \mathcal{H}$, define the limit counterpart of $\bar{\theta}_N[\theta_1; \eta]$ as

$$\bar{\theta}[\theta_1; \eta] = \arg\min_{\theta\in\Theta} |S(\theta, \theta_1; \eta)|.$$

To establish consistency, we impose the following assumptions.

Assumption A. (1) $\|\widehat{\eta}(\cdot, \theta) - \eta^0(\cdot, \theta)\|_{\mathcal{H}} = o_p(1)$; (2) Uniformly for all $\theta_1 \in \Theta$, $|S_N(\bar{\theta}_N[\theta_1; \widehat{\eta}], \theta_1; \widehat{\eta})| \leq \inf_{\theta\in\Theta} |S_N(\theta, \theta_1; \widehat{\eta})| + o_p(1)$; (3) $S(\theta, \theta_1; \eta)$ exists and is continuous in (θ, θ_1, η); (4) For all positive sequences $\delta_N = o(1)$: $\sup_{\theta\in\Theta, \theta_1\in\Theta, \|\eta-\eta^0\|_{\mathcal{H}}\leq\delta_N} |S_N(\theta, \theta_1; \eta) - S(\theta, \theta_1; \eta^0)| = o_p(1)$; (5) $\Theta \subset \mathbb{R}^p$ is compact and $\theta^0 \in \Theta$ uniquely satisfies $S(\theta^0, \theta^0; \eta^0) = 0$.

Assumption B. The map $\bar{\theta}[\cdot; \eta^0] : \Theta \to \Theta$ exists and satisfies: (1) For any $\theta_1 \in \Theta$, $\theta \mapsto |S(\theta, \theta_1; \eta^0)|$ has unique minimum $\bar{\theta}[\theta_1; \eta^0]$; (2) $\theta_1 \mapsto \bar{\theta}[\theta_1; \eta^0]$ is continuous in θ_1 and θ^0 is a fixed-point of $\bar{\theta}[\cdot; \eta^0]$.

Remark. Assumption **A** is similar to those found elsewhere in the literature on semiparametric Z-estimation; see, e.g., Theorem 1 in Chen et al. (2003). Assumption **B** is specifically needed to deal with the iterative formulation of the estimator presented in this article and is similar to assumptions in Fan et al. (2015).

Since $\widehat{\theta}_N^{(k)}$ is obtained through an iterative procedure, proving consistency of $\widehat{\theta}_N^{(k)}$ for θ^0 (as $N \to \infty$) requires different arguments than those used to prove consistency of standard Z-estimators. However, note that, by the triangle inequality,

$$|\widehat{\theta}_N^{(k)} - \theta^0| \leq |\bar{\theta}_N[\widehat{\theta}_N^{(k-1)}; \widehat{\eta}] - \bar{\theta}[\widehat{\theta}_N^{(k-1)}; \eta^0]| + |\bar{\theta}[\widehat{\theta}_N^{(k-1)}; \eta^0] - \theta^0| \tag{12}$$

and consistency will follow if both terms on the RHS of (12) are $o_p(1)$. Let $\widehat{\theta}_N^{(1)}$ be the starting value of the algorithm. For now, assume that the first term on the RHS of equation (12) is $o_p(1)$. We then have the following two cases for $\widehat{\theta}_N^{(1)}$ to consider:

(1) $|\widehat{\theta}_N^{(1)} - \theta^0| = o_p(1)$. Since $\widehat{\theta}_N^{(1)}$ is consistent as $N \to \infty$, under general regularity conditions $\widehat{\theta}_N^{(k)}$ will be consistent for $k > 1$ and (12) will be satisfied.

(2) $|\widehat{\theta}_N^{(1)} - \theta^0| = O_p(1)$. If $\bar{\theta}[\theta_1; \eta^0]$ is a contraction mapping with Lipschitz constant $c \in (0, 1)$ and fixed-point θ^0, then

$$|\widehat{\theta}_N^{(k)} - \theta^0| \leq |\bar{\theta}_N[\widehat{\theta}_N^{(k-1)}; \widehat{\eta}] - \bar{\theta}[\widehat{\theta}_N^{(k-1)}; \eta^0]| + |\bar{\theta}[\widehat{\theta}_N^{(k-1)}; \eta^0] - \bar{\theta}[\theta^0; \eta^0]| \leq o_p(1) + c|\widehat{\theta}_N^{(k-1)} - \theta^0|.$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Iterating the above relationship yields (see the proof of Proposition 1 for more details)

$$|\widehat{\theta}_N^{(k)} - \theta^0| \le o_p(1) + c|\widehat{\theta}_N^{(k-1)} - \theta^0| \le \cdots \le o_p(1) + c^{k-1}|\widehat{\theta}_N^{(1)} - \theta^0|,$$

which will converge to zero in probability if $k \to \infty$ as $N \to \infty$.

The above intuition leads to the following assumption.

Assumption C. One of the following is satisfied: (C.1) $|\widehat{\theta}_N^{(1)} - \theta^0| = o_p(1)$; (C.2) for all $\theta_1, \theta_2 \in \Theta$ and for some $c \in (0,1)$, $|\bar{\theta}[\theta_1; \eta^0] - \bar{\theta}[\theta_2; \eta^0]| \le c|\theta_1 - \theta_2|$.

Remark. Assumption **C** requires that the initial estimator used in the algorithm, $\widehat{\theta}_N^{(1)}$, is consistent or $\bar{\theta}[\cdot; \eta^0]$ is a contraction mapping. In the quantile regression example considered in Section 5 a consistent initial value can be obtained, whereas in the GARCH-M example, Section 3, no such initial estimator exists in practice.

Even if $\widehat{\theta}_N^{(1)}$ is not consistent, under the contraction assumption, $\widehat{\theta}_N^{(k)}$ would be consistent if it were feasible to conduct an infinite number of iterations. While such a choice for k is infeasible in practice, consistency can still be achieved by allowing the number of iterations, k, to diverge as the sample size, N, diverges. Define $\{k_N\}_{N \ge 1}$ to be a sequence of positive integers linking the number of iterations, k, and the sample size, N, such that $k_N \to \infty$ as $N \to \infty$.

PROPOSITION 1. *If Assumptions **A** and **B** hold, then as $N \to \infty$:*

(*i*) *under Assumption **C.1**, $|\widehat{\theta}_N^{(k)} - \theta^0| = o_p(1)$ for any $k \ge 1$;*

(*ii*) *under Assumption **C.2**, $|\widehat{\theta}_N^{(k)} - \theta^0| = o_p(1)$ for $k = k_N \to \infty$ as $N \to \infty$.*

4.2. Asymptotic Distribution

Deriving the asymptotic distribution of $\sqrt{N}(\widehat{\theta}_N^{(k)} - \theta^0)$ requires us to first discuss the derivatives of $S(\theta, \theta_1; \eta) = \Psi_1(\theta, \eta(\cdot, \theta_1)) + \Psi_2(\theta_1, \eta(\cdot, \theta_1))$. For $\delta > 0$, define $\Theta_\delta := \{\theta \in \Theta : |\theta - \theta^0| \le \delta\}$ and $\mathcal{H}_\delta := \{\eta \in \mathcal{H} : \|\eta - \eta^0\|_{\mathcal{H}} \le \delta\}$. For $(\theta, \theta_1, \eta) \in \Theta_\delta \times \Theta_\delta \times \mathcal{H}_\delta$, we denote the usual derivative of $S(\theta, \theta_1; \eta)$ wrt θ as $\dot{S}_1(\theta, \theta_1; \eta)$. For $(\theta, \theta_1) \in \Theta_\delta \times \Theta_\delta$, we say that $S(\theta, \theta_1; \eta)$ is Frechet differentiable at $\eta \in \mathcal{H}_\delta$ if there exists a continuous linear map $\dot{S}_\eta(\theta, \theta_1; \eta)$ such that for all $\{h_N \subset \mathcal{H}\}$ with $\|h_N\|_{\mathcal{H}} \to 0$ and $\eta + h_N \in \mathcal{H}$, for all $N \ge 1$,

$$|S(\theta, \theta_1; \eta + h_N) - S(\theta, \theta_1; \eta) - \dot{S}_\eta(\theta, \theta_1; h_N)| = o(\|h_N\|_{\mathcal{H}}).$$

The Frechet derivative of $\eta(\cdot, \theta_1)$ wrt θ_1 can be defined in a similar manner. Denote the $q \times p$ matrix of Frechet derivatives of $\eta(\cdot, \theta_1)$ wrt θ_1, and evaluated at $\theta_1 \in \Theta_\delta$, by $\dot{\eta}(\cdot, \theta_1)$. Lastly, define the Frechet derivative of $S(\theta, \theta; \eta^0)$ wrt all the occurrences of θ as $\dot{S}_\theta(\theta, \theta; \eta^0)$. For simplicity, when derivatives of $S(\cdot)$ are evaluated at $(\theta^0, \theta^0, \eta^0(\cdot, \theta^0))$ we disregard their dependence on these values; i.e., in the sequel $\dot{S}_1 = \dot{S}_1(\theta^0, \theta^0; \eta^0)$ and $\dot{S}_\eta = \dot{S}_\eta(\theta^0, \theta^0; \eta^0)$.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

We can now state the following assumption.

Assumption D. For $\|\cdot\|_\rho$ the spectral norm: $0 < \|[-\dot{S}_1]^{-1}\{\dot{S}_\theta - \dot{S}_1\}\|_\rho < 1$.

Remark. Assumption **D** is often called the *information dominance condition* and can be interpreted as a requirement that the information about θ^0 contained in $\dot{S}_1$ be greater than the information about θ^0 contained in $\dot{S}_\theta - \dot{S}_1$. Assumption **D** is actually implied by Assumption **C.2** and thus can be treated as a local version of **C.2**. Therefore, if **C.2** is satisfied Assumption **D** is redundant. However, Assumption **D** is needed in situations where the contraction mapping condition is only satisfied in a neighborhood of θ^0 but there exists an initially consistent starting value $\widehat{\theta}_N^{(1)}$.

To derive the asymptotic distribution of $\sqrt{N}(\widehat{\theta}_N^{(k)} - \theta^0)$ we impose the following assumptions.

Assumption E. (1) For some $\beta > 0$, $\|\widehat{\eta}(\cdot,\theta) - \eta^0(\cdot,\theta)\|_{\mathcal{H}} = O_p(N^{-\beta})$;
(2) For $k = k_N$, and $k_N \to \infty$ as $N \to \infty$, $|S_N(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta})| = o_p(N^{-1/2})$;
(3) For all positive sequences $\delta_N = o(1)$:

$$\sup_{(\theta,\theta_1,\eta)\in\Theta_{\delta_N}\times\Theta_{\delta_N}\times\mathcal{H}_{\delta_N}} \sqrt{N}|S_N(\theta,\theta_1;\eta) - S(\theta,\theta_1;\eta) - S_N(\theta^0,\theta^0;\eta^0)| = o_p(1);$$

(4) $S(\theta,\theta_1;\eta)$ is Frechet differentiable at $(\theta^0,\theta^0,\eta^0)$ with derivatives $\dot{S}_1, \dot{S}_\theta$, and $\dot{S}_\eta$ a continuous linear functional, and with each term continuous at $(\theta^0,\theta^0,\eta^0(\cdot,\theta^0))$. For some $\alpha > 1$ satisfying $\beta\alpha > 1/2$ and for $(\theta,\theta_1,\eta) \in \Theta_\delta \times \Theta_\delta \times \mathcal{H}_\delta$:

$$\begin{aligned}\big|S(\theta,\theta_1;\eta) - S(\theta^0,\theta^0;\eta^0) - \dot{S}_1(\theta-\theta^0) - \{\dot{S}_\theta - \dot{S}_1\}(\theta_1-\theta^0) - \dot{S}_\eta[\eta(\cdot,\theta^0) - \eta^0(\cdot,\theta^0)]\big| \\ = o(|\theta-\theta^0|) + o(|\theta_1-\theta^0|) + O(\|\eta-\eta^0\|_{\mathcal{H}}^{\alpha});\end{aligned}$$

(5) (i) $\sqrt{N}\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0) - \eta^0(\cdot,\theta^0)] = O_P(1)$; (ii) $\sqrt{N}S_N(\theta^0,\theta^0;\eta^0) = O_p(1)$;
(iii) For some matrix V, $\sqrt{N}\left\{S_N(\theta^0,\theta^0;\eta^0) + \dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0) - \eta^0(\cdot,\theta^0)]\right\} \xrightarrow{d} \mathcal{N}(0,V)$;
(6) $-\dot{S}_\theta$ is positive definite and $[-\dot{S}_1]^{-1}$ exists;
(7) $\theta^0 \in \text{int}(\Theta)$ uniquely satisfies $S(\theta^0,\theta^0;\eta^0) = 0$.

Remark. Assumption **E** is similar to those found elsewhere in the literature on semiparametric Z-estimation (see, e.g., Chen et al., 2003; Chen, 2007; Nan and Wellner, 2013) and ensures that $\widehat{\theta}_N$, the solution to $0 = S_N(\theta,\theta;\widehat{\eta})$, is asymptotically normal.

PROPOSITION 2. *If Assumptions **A**, **B** and **E** are satisfied, then as $N \to \infty$:*

(i) $\sqrt{N}(\widehat{\theta}_N - \theta^0) \to_d \mathcal{N}(0, \dot{S}_\theta^{-1} V \dot{S}_\theta^{-1})$;

*(ii) for $k_N/\ln(N) \to \infty$ and $k = k_N$, under Assumptions **C** and **D**,* $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N) = o_p(1)$.

Remark. Asymptotic equivalence between $\widehat{\theta}_N^{(k)}$ and $\widehat{\theta}_N$ requires that $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)}) = o_p(1)$, which in turn requires that k diverge faster than $\ln(N)$. This rate

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

requirement is needed since the asymptotic distribution of $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)})$ can be expressed in terms of $\Sigma^{k-1}\sqrt{N}(\widehat{\theta}_N^{(1)} - \theta^0)$, regardless of whether or not $\widehat{\theta}_N^{(1)}$ is consistent, and where we recall that Assumption **C.1** only guarantees $|\widehat{\theta}_N^{(1)} - \theta^0| = o_p(1)$ (with no explicit rate). Therefore, the term $\Sigma^{k-1}\sqrt{N}(\widehat{\theta}_N^{(1)} - \theta^0)$ will be $o_p(1)$ so long as $\Sigma^{k-1}\sqrt{N} \to 0$. Under Assumptions **C.1** and **D** (or Assumption **C.2**), $\|\Sigma\|_\rho < 1$ and choosing a sequence of iterations $k = k_N$, where $\frac{k_N}{\ln(N)} \to \infty$, ensures that $\Sigma^{k-1}\sqrt{N} \to 0$ as $N \to \infty$.

If one is willing to assume $\sqrt{N}(\widehat{\theta}_N^{(1)} - \theta^0) = O_p(1)$, then similar arguments to those in the proof of Proposition 2 show that $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N) = o_p(1)$, as $N \to \infty$, so long as $k \to \infty$ and with no particular relation to N.

Remark. By Proposition 2 part (ii), $\widehat{\theta}_N^{(k)}$ is asymptotically equivalent to $\widehat{\theta}_N$. Therefore, estimating the variance of $\sqrt{N}(\widehat{\theta}_N^{(k)} - \theta^0)$, for large k, can be carried out as if one was going to estimate the asymptotic variance of $\sqrt{N}(\widehat{\theta}_N - \theta^0)$. For a discussion on estimating the asymptotic variance of $\sqrt{N}(\widehat{\theta}_N - \theta^0)$ we refer the interested reader to Ackerberg, Chen, and Hahn (2012).

5. ADDITIONAL EXAMPLE: DOUBLY ROBUST MOMENT CONDITIONS

The following example shows that estimation in problems with missing data can be cast as Z-estimation in a semiparametric model with bundled parameters. A specific quantile regression model with missing covariates illustrates the SP-MBP algorithm, i.e., Algorithm 1, in this context.

Assume we are interested in estimating the parameter governing the just-identified p-dimensional estimating functions $g(Z;\theta) = (g_1(Z;\theta), \ldots, g_p(Z;\theta))'$, where $Z = (W', X')'$, W is a scalar response variable and X is a (vector valued) covariate. Suppose that there exists a unique θ^0 satisfying $\mathbb{E}[g(Z;\theta)] = 0$ and our goal is estimation of θ^0. We observe an i.i.d. data set $\{Z_i = (W_i', X_i')'\}_{i=1}^N$ where the covariate X_i is observed with missingness. The missingness in X_i follows a missing-at-random (MAR) pattern: let D_i be a missingness indicator such that $D_i = 1$ if X_i is observed and $D_i = 0$ else, we say that X_i is MAR if $\Pr(D_i = 1|Z_i) = \Pr(D_i = 1|W_i)$; i.e., the missingness depends only on the observed data (W_i) and not the missing data (X_i). See Wooldridge (2007) and Graham, Pinto, and Egel (2012) for discussion of the MAR assumption and its applicability in economics.

Under an MAR missingness pattern, estimation of θ^0 from the complete case sample units $D_i g(Z_i;\theta)$ is generally insufficient to deliver a consistent estimator of θ^0. However, consistent and efficient estimates of θ^0 can be obtained by a doubly robust (DR) estimation strategy where models are specified for $p^0(w) = \Pr(D = 1|W = w)$, often called the propensity score, and $\eta^0(w,\theta) = \mathbb{E}[g(Z;\theta)|W = w]$, the conditional expectation. Let $p(w)$ and $\eta(w,\theta)$ denote the models for $p^0(w)$ and $\eta^0(w,\theta)$. DR estimation then proceeds by using the

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

observed data $\{W_i, D_i, D_i X_i\}$ and the models $p(w)$, $\eta(w,\theta)$ to estimate the population estimating equations

$$\mathbb{E}\left[\frac{D}{p(W)}g(Z;\theta)\right]+\mathbb{E}\left[\frac{D-p(W)}{p(W)}\eta(W,\theta)\right].$$

When the models $p(\cdot)$ and $\eta(\cdot,\theta)$ are correct, DR estimators based on a finite sample version of the above will deliver a consistent and efficient estimator of θ^0; for a discussion on DR estimators, see, e.g., Robins, Rotnitzky, and Zhao (1994), Scharfstein, Rotnitzky, and Robins (1999), and Chen, Hong, and Tarozzi (2008).

DR parameter estimates appear in a number of important settings, such as, linear regression models with missing covariates, quantile regression models with missing covariates and treatment effect models. It is most common for $p(\cdot)$ and $\eta(\cdot,\theta)$ to be parametric models of the unknown functions $p^0(\cdot)$ and $\eta^0(\cdot,\theta)$. Then, so long as one of the models is correctly specified, DR estimators will deliver a consistent estimator for θ^0. However, if either $p(\cdot)$ or $\eta(\cdot,\theta)$ are misspecified, the resulting DR estimator will not be efficient. Following Roth and Firpo (2016), an alternative to popular parametric DR approaches is to consider a nonparametric DR approach where we model $p^0(\cdot)$ and $\eta^0(\cdot,\theta)$ nonparametrically in a first-stage and then estimate θ^0. For $\widehat{\eta}(\cdot,\theta)$ (respectively, $\widehat{p}(\cdot)$) an estimator of $\eta^0(\cdot,\theta)$ (respectively, $p^0(\cdot)$), a consistent and efficient estimator of θ^0 can be obtained by solving:

$$0=\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))=\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{\widehat{p}(W_i)}g(Z_i;\theta)+\frac{D_i-\widehat{p}(W_i)}{\widehat{p}(W_i)}\widehat{\eta}(W_i,\theta).$$

Clearly, estimation of θ^0 from $\Psi_N(\theta,\widehat{\eta}(\cdot,\theta))$ requires solving for the zeros of estimating equations that satisfy the general decomposition given in equation (2). Also, one portion of the estimating equations is much simpler than the other portion; i.e., $(D_i/\widehat{p}(W_i))g(Z_i;\theta)$ is simpler than $(D_i-\widehat{p}(W_i)/\widehat{p}(W_i))\widehat{\eta}(W_i,\theta)$. Moreover, note that the estimated conditional expectation $\widehat{\eta}(W_i,\theta)$ depends directly on θ through $g(Z_i;\theta)$.[7] Thus, in an attempt to obtain a consistent and efficient estimator, we are left with parameters θ and $\eta(\cdot,\theta)$ that cannot be separated and are therefore *bundled* together.

5.1. Quantile Regression with Missing Covariates

Consider the following quantile regression (QR) model

$$Y_i=X_{1i}'\theta_1(\tau)+X_{2i}'\theta_2(\tau)+\epsilon_i,\ i=1,\ldots,N. \tag{13}$$

Y is a response, $X=(X_1',X_2')'$ is a $p\times 1$ vector of covariates, ϵ is an i.i.d. error term satisfying $\Pr(\epsilon_i\leq 0)=\tau$ for the τ-th quantile, and $\theta(\tau)=(\theta_1(\tau),\theta_2(\tau))$ are regression coefficients that are allowed to vary across quantiles. For simplicity assume $\{Y_i,X_i\}_{i=1}^N$ is an i.i.d sample from $\{Y,X\}$.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

The response variable Y_i and covariate X_{1i} are always observed but the covariates X_{2i} can be missing. Define $D_i = 1$ if X_{2i} is missing and $D_i = 0$ else. Let $W_i = (Y_i, X'_{1i})'$. The X_{2i}'s are assumed to be missing at random (MAR) with propensity score $p(W_i) = \Pr(D_i = 1|Y_i, X_{1i})$.

If there is no missing data the QR estimator of $\theta(\tau)$ can be obtained as the solution to

$$\frac{1}{N}\sum_{i=1}^{N} X_i(\tau - I(\epsilon_i(\theta) \leq 0)) = 0, \tag{14}$$

where $I(\cdot)$ is the indicator function and $\epsilon(\theta) = Y - X'\theta$. However, when X_{2i} is missing the solution to equation (14) is generally biased and inconsistent for some true θ^0. Many methods exist to deal with this issue, such as, inverse probability weighting (IPW) (Horvitz and Thompson, 1952), the estimating equation projection (EEP) method (Zhou, Wan, and Wang, 2008), or the nonparametric augmented IPW (AIPW) method (Xue, 2009; Roth and Firpo, 2016; Chen, Wan, and Zhou, 2015). In this section, we focus our attention on the nonparametric AIPW estimator, hereafter simply referred to as the AIPW estimator.

The AIPW estimator combines nonparametric estimation of the propensity score $p(W_i)$ and the conditional expectation of the estimating equations in equation (14). For $\widehat{p}(W_i)$ denoting the estimated propensity score, the AIPW estimator can be defined as the (unique) zero of the sample criterion function

$$\Psi_N(\theta, \widehat{\eta}(\cdot,\theta)) = \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\widehat{p}(W_i)}(X_i(\tau - I(\epsilon_i(\theta) \leq 0)) - \frac{(D_i - \widehat{p}(W_i))}{\widehat{p}(W_i)})\widehat{\eta}(W_i,\theta), \tag{15}$$

where $\widehat{\eta}(w,\theta)$ is a nonparametric estimator of $\eta(w,\theta) = E[D_i(X_i(\tau - I(\epsilon_i(\theta) \leq 0)))|W = w]$. In this example we focus on estimation of $\eta(\cdot,\theta)$ and $p(W_i)$ using kernel methods and so

$$\widehat{\eta}(w,\theta) = \sum_{j=1}^{N} D_j X_j(\tau - I(\epsilon_j(\theta) \leq 0))\mathcal{K}_h(w - W_j) \Big/ \left\{\sum_{j=1}^{N} D_j \mathcal{K}_h(w - W_j)\right\},$$

$$\widehat{p}(w) = \sum_{j=1}^{N} D_j L_b(W_j - w) \Big/ \max\left\{1, \sum_{i=1}^{N} D_j L_b(W_j - w)\right\},$$

where $\mathcal{K}_h(\cdot)$, $L_b(\cdot)$ are kernel functions and h, b are bandwidths.

Regardless of the nonparametric method chosen, comparing (15) with the general structure presented in Section 2 we have that

$$\Psi_{1,N}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\widehat{p}(W_i)}(X_i(\tau - I(\epsilon_i(\theta) \leq 0)), \Psi_{2,N}(\theta, \widehat{\eta}(\cdot,\theta)) = \frac{1}{N}\sum_{i=1}^{N} \frac{(D_i - \widehat{p}(W_i))}{\widehat{p}(W_i)})\widehat{\eta}(W_i,\theta),$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

and $\Psi_{1,N}(\theta)$ is of a much simpler form than $\Psi_{2,N}(\theta, \widehat{\eta}(Z,\theta))$. Clearly, the above estimating equations have the same general structure considered throughout this article.

The sample estimating equations in (15) can be solved directly to estimate θ^0; however, there exist two difficulties in this approach: one, the occurrences of θ within $\widehat{\eta}(\cdot,\theta)$ can lead to numerical instabilities within the optimization; two, we must employ numerical methods that can handle nonsmooth estimating equations.

As noted in Chen et al. (2015), directly estimating θ^0 from (15) is difficult since $\Psi_N(\theta, \widehat{\eta}(\cdot,\theta))$ is nonsmooth in θ. Chen et al. (2015) propose an iterative procedure based on the MM algorithm[8] of Lange, Hunter, and Yang (2000) as a means of simplifying the computations. The MM algorithm works by optimizing a smooth surrogate function in lieu of the actual objective function. For quantile regression Hunter and Lange (2000) propose the surrogate function:

$$\zeta_\tau^\epsilon(\epsilon|\epsilon_1) = \frac{1}{4}\left[\frac{\epsilon^2}{(\epsilon+|\epsilon_1|)} + (4\tau-2)\epsilon + c\right],$$

which is a surrogate for $\rho_\tau^\epsilon(\epsilon) = \rho_\tau(\epsilon) - \frac{\epsilon}{2}\ln(\xi+|\epsilon|)$, where $\rho_\tau(\epsilon)$ is the usual quantile check function, $\xi > 0$ is a constant and c is a constant such that $\zeta_\tau^\epsilon(\epsilon_1|\epsilon_1) = \rho_\tau^\epsilon(\epsilon_1)$.

While the above surrogate function is useful for quantile regression, estimating θ from the sample criterion function in (15) means we cannot directly use the surrogate function approach. Instead, an estimation procedure based on the derivative, wrt θ, of the surrogate function $\zeta_i^\epsilon(\epsilon(\theta)|\epsilon_1)$ and given by $\frac{\partial\zeta_i^\epsilon(\epsilon(\theta)|\epsilon_1)}{\partial\theta} = \frac{1}{4}\left\{\frac{X_i(\epsilon_i(\theta))}{(\epsilon_i(\theta)+|\epsilon_{1i}|)^2} - X_i(4\tau-2) - \frac{2X_i(\epsilon_i(\theta))}{(\epsilon_i(\theta)+|\epsilon_{1i}|)}\right\}$ can be implemented. Applying the MM principle to the criterion in (15) and given a starting value $\widetilde{\theta}_N^{(1)}$, Chen et al. (2015) propose to update the estimate of θ iteratively by solving

$$0 = \frac{1}{N}\sum_{i=1}^N \frac{D_i}{\widehat{p}(W_i)}\frac{\partial\zeta_i^\epsilon(\epsilon(\theta)|\widetilde{\epsilon}^{(k)})}{\partial\theta} - \frac{1}{N}\sum_{i=1}^N \frac{D_i-\widehat{p}(W_i)}{\widehat{p}(W_i)}\frac{\sum_{j=1}^N D_j\frac{\partial\zeta_j^\epsilon(\epsilon(\theta)|\widetilde{\epsilon}^{(k)}))}{\partial\theta}\mathcal{K}_h(W_j-W_j)}{\sum_{j=1}^N D_j\mathcal{K}_h(W_j-W_j)}, \tag{16}$$

where $\widetilde{\epsilon}^{(k)} = \epsilon(\widetilde{\theta}_N^{(k)})$. equation (16) could be solved using, for instance, a NR approach.

The estimation procedure of Chen et al. (2015) is overly complicated since it treats the left-hand-side (LHS) and right-hand-side (RHS) of (16) as equivalent in their complexity. As an alternative to solving (16), we apply the MM principle of Hunter and Lange (2000) as a means of implementing the SP-MBP estimator to solve (15). Using this idea and given a starting value $\widehat{\theta}_N^{(1)}$, at the k+1-th step ($k>1$), the SP-MBP estimator obtains $\widehat{\theta}_N^{(k+1)}$ by solving

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

$$0 = \frac{1}{N}\sum_{i=1}^{N} \frac{D_i}{\widehat{p}(W_i)} \frac{\partial \zeta_i^{\epsilon}(\epsilon(\theta)|\widehat{\epsilon}^{(k)})}{\partial \theta} - \frac{1}{N}\sum_{i=1}^{N} \frac{D_i - \widehat{p}(W_i)}{\widehat{p}(W_i)} \frac{\sum_{j=1}^{N} D_j \frac{\partial \zeta_j^{\epsilon}(\widehat{\epsilon}^{(k)}|\widehat{\epsilon}^{(k)}))}{\partial \theta} \mathcal{K}_h(W_j - W_j)}{\sum_{j=1}^{N} D_j \mathcal{K}_h(W_j - W_j)}, \tag{17}$$

where $\widehat{\epsilon}^{(k)} = \epsilon(\widehat{\theta}_N^{(k)})$. Equation (17) can be solved using a NR based approach, which is particularly useful since the only second-derivative required to implement this procedure is

$$\frac{1}{N}\sum_{i}^{N} \frac{D_i}{\widehat{p}(W_i)} \frac{\partial^2 \zeta_i^{\epsilon}(\epsilon(\theta)|\epsilon(\widehat{\theta}_N^{(k)}))}{\partial \theta \partial \theta'}|_{\theta=\widehat{\theta}_N^{(k)}}, \text{ where}$$

$$\frac{\partial^2 \zeta_i^{\epsilon}(\epsilon(\theta)|\epsilon_1)}{\partial \theta \partial \theta'} = \frac{X_i X_i'}{2}\left\{ \frac{1}{(\epsilon_i(\theta) + |\epsilon_{1i}|)} - \frac{2\epsilon_i(\theta)}{(\epsilon_i(\theta) + |\epsilon_{1i}|)^2} + \frac{\epsilon_i(\theta)^2}{(\epsilon_i(\theta) + |\epsilon_{1i}|)^3} \right\}.$$

This approach, while similar to the estimator proposed in Chen et al. (2015), is simpler in at least one key aspect. Besides estimation of $\widehat{p}(w)$, implementation of the procedure outlined in Chen et al. (2015) using a NR approach requires calculating, at every iteration, two nonparametric components: one associated with the RHS of (16) and one associated with its derivative, wrt θ;[9] a NR approach based on solving equation (17) only requires calculating one nonparametric component, the one associated with the LHS of equation (17).

It is interesting to point out that in this example a simple initially consistent starting value can be obtained using nonparametric IPW. Implementing the algorithm proposed in this article, when such a starting value exists, is no more difficult than implementing the nonparametric IPW estimator since the right hand side of (17) is a constant. Therefore, up to the calculation of $\widehat{\eta}(w, \widehat{\theta}_N^{(k)})$, each iteration of the SP-MBP estimator is no more difficult than an iteration of the nonparametric IPW estimator.

5.2. Simulation Experiments

To analyze the finite sample properties of the proposed AIPW estimator obtained by solving (17), hereafter referred to as the iterative AIPW estimator, we conduct a simulation study similar to the design in Chen et al. (2015). We consider outcomes generated from

$$Y_i = \theta_0 + X_{1i}\theta_1 + X_{2i}\theta_2 + \epsilon_i,$$

where $(\theta_0, \theta_1, \theta_2)' = (1, -2, 1)'$, X_{1i} is distributed $\chi^2(1) - 1$, X_{2i} is distributed $N(1, 1)$ and X_{1i}, X_{2i} independent. The error ϵ_i is standard normal and the τ-th conditional quantile of Y_i is given by

$$Q_\tau(Y_i|X_i) = \theta_0(\tau) + X_{1i}\theta_1 + X_{2i}\theta_2,$$

where $\theta_0(\tau) = \theta_0 + Q_\tau(\epsilon_i)$. For the simulation we consider three quantiles $\tau \in \{0.25, 0.5, 0.75\}$.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Letting $W = (X_1, Y)'$, we consider four difference specifications for the propensity score:

(S1) $p(x_1, y) = \exp(\gamma' W^2)/(1 + \exp((\gamma' W)^2))$, for $\gamma = (\sqrt{(2.5)}, 0)'$;
(S2) $p(x_1, y) = \exp(\gamma' W)/(1 + \exp(\gamma' W))$, for $\gamma = (1, 1)'$;
(S3) $p(x_1, y) = \exp(\sin(\gamma' W - 1))/(1 + \exp(\sin(\gamma' W - 1)))$, for $\gamma = (1, 1)'$;
(S4) $p(x_1, y) = 1/4 + \arctan(\gamma' W 0.^2)/\pi$, for $\gamma = (1, 1)'$.

All specifications set the missingness on X_2, with the missingness under (S1) not depending on Y. Specification (S1) and (S2) are similar to those in the design of Chen et al. (2015), while (S3) and (S4) feature elsewhere in the literature. Under the four different missingness specifications missing data account for about 21% of the sample under (S1), 32% under (S2), 50% under (S3) and 38% under (S4).

Within the Monte Carlo design we compare the nonparametric IPW estimator and the iterative AIPW estimator. For nonparametric IPW, estimation of $p(x_1, y)$

TABLE 4. Comparison of IPW and iterative AIPW estimators for the $\tau = 0.5$ quantile. S1–S4 refer to the propensity score specifications, which can be found in the main text. BIAS is the mean bias, MAE is mean absolute error, RMSE is root mean squared error, STD is the empirical standard deviation and COV is the Monte Carlo coverage. The true parameter values are $\theta_1 = 1, \theta_2 = -2, \theta_3 = 1$. BIAS, MAE, RMSE have been multiplied by 100 for scaling purposes

	$\tau = 0.5$		IPW	$\tau = 0.5$		AIPW
S1	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
BIAS	−0.2699	0.1518	−0.4499	−0.2173	0.0979	−0.4346
MAE	5.0817	3.2017	5.3338	5.0921	3.2067	5.3369
RMSE	6.8473	4.0444	7.2495	6.8545	4.0473	7.2557
COV	4.1	5.6	4	4.1	5.7	3.9
STD	0.0685	0.0404	0.0724	0.0685	0.0405	0.0725
S2	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
BIAS	3.1065	3.0516	−2.9979	3.1427	2.9735	−2.9841
MAE	8.4745	10.9690	7.4610	8.4890	10.9388	7.475
RMSE	10.9557	14.6378	9.6730	10.9633	14.6182	9.6872
COV	5	4	4.9	4.9	4	4.9
STD	0.1051	0.1432	0.0920	0.1051	0.1432	0.0922
S3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
BIAS	−1.1872	0.6767	−2.4216	−1.1383	0.6071	−2.4120
MAE	7.4115	5.1928	7.8559	7.3981	5.1910	7.8562
RMSE	13.8812	8.2827	11.0578	13.8767	8.2836	11.0692
COV	1.7	2.5	3.3	1.7	2.3	3.5
STD	0.1384	0.0826	0.1079	0.1384	0.0827	0.1081
S4	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
BIAS	−0.6619	0.1122	−1.0793	−0.6079	0.0527	−1.0722
MAE	5.9954	3.6741	5.8307	6.0058	3.6847	5.8318
RMSE	7.9209	4.6549	7.4778	7.9339	4.6618	7.4810
COV	3.7	4.9	5.2	3.7	4.9	5.1
STD	0.0790	0.0466	0.0740	0.0791	0.0466	0.0741

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

is based on kernel smoothing, while for the iterative AIPW estimator kernel smoothing is used to estimate both $p(x_1, y)$ and $\eta(w, \theta)$. For both procedures the kernel function was chosen standard normal and the bandwidth was taken to be $1.02\sigma_W N^{-1/(4+d)}$, where d is the dimension of W and $\sigma_W = \sigma_{X_1} * \sigma_Y$.

We consider a sample size of $N = 500$ observations and 1,000 Monte Carlo replications across the four designs (S1)–(S4). The results for the $\tau = 0.5$ quantile are reported in Table 4.[10] The estimators are compared on the basis of mean bias (BIAS), mean absolute error (MAE), root mean squared error (RMSE), Monte Carlo coverage (COV), and Monte Carlo standard deviation (STD).

In general the results demonstrate that the iterative AIPW estimator consistently yields estimates that display smaller bias than the IPW estimator. While the Monte Carlo bias is generally smaller for the iterative AIPW estimator, the IPW approach often yields estimators with slightly smaller MAE and RMSE. The Monte Carlo coverages between the iterative AIPW method and the IPW procedure are similar and are also close to the 95% level. The Monte Carlo standard errors between the two methods are also similar. This yields convincing evidence that the iterative AIPW estimator loses little in terms of precision, despite the fact that it does not use all available second-derivative information to estimate the parameters.

Across the four specifications for the propensity score the average number of iterations required to obtain convergence of the iterative AIPW procedure are as follows (for $\tau = 0.5$): twenty-one iterations under (S1), twenty iterations under (S2), twenty-four under (S3), and twenty-two under (S4).

6. CONCLUSION

We propose a new iterative estimator for use in semiparametric models. Several examples illustrate the broad range of applications for this approach and the simplicity with which it can be implemented. Monte Carlo evidence from two examples demonstrates that the estimator behaves well in finite samples and compares favorably to existing semiparametric estimators.

Similar to the MBP algorithms of Song et al. (2005) and Fan et al. (2015), the procedure developed herein relies on a contraction mapping condition. The required contraction mapping condition can be represented, at least locally, through an *information dominance condition*. Under the *information dominance condition*, and additional technical conditions, the estimator derived from this new procedure is consistent and asymptotically normal with the same asymptotic distribution as standard Z-estimators.

NOTES

1. To keep the statement of the example as simple as possible, we forgo a precise definition of the functional derivatives until Section 4.

2. Similar to M-estimation with bundled parameters, GMM estimators based on an over-identified vector of sample moments, say $\sum_{i=1}^{N} g(Z_i, \theta, \widehat{\eta}(\cdot, \theta))/N$, also admit first-order conditions with the

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

structure in equation (1). Moreover, these first-order conditions would also be such that one term in the first-order conditions would be simpler to evaluate than the other term.

3. A careful description of these derivatives is given in Section 4.

4. Finite difference methods can be computationally costly and estimation algorithms incorporating these numerical approximations can encounter difficulties in invertibility and/or positive definiteness at update values (Song et al., 2005; Fan et al., 2015). Gauss–Newton type algorithms are not required to calculate the full Hessian matrix, but can encounter issues when the sample size is small or the dimension of θ is large, Song et al. (2005).

5. Christensen et al. (2012) utilize kernel smoothing to obtain $\widehat{\eta}(\cdot, \theta)$, however, other methods such as local-polynomials or splines could also be used.

6. The following data generating processes are the same as those analyzed in Christensen et al. (2012).

7. With reference to the general setup for $\eta(\cdot, \theta)$ in the previous section, in this example we have $q = p$. We also note that a similar structure for the estimating equations would maintain if $g(Z; \theta)$ over-identifies θ. However, we focus on the just-identified case for simplicity.

8. MM stands for majorize in minimization and minorize in maximization.

9. Calculating this derivative can be time consuming if the sample size is large or if the dimension of W_i is large.

10. Results for $\tau = 0.25$ and $\tau = 0.75$ quantiles have also been obtained and are similar to those obtained for the $\tau = 0.5$ quantile. These additional results are not reported for brevity but are available from the author upon request.

REFERENCES

Ackerberg, D., X. Chen, & J. Hahn (2012) A practical asymptotic variance estimator for two-step semiparametric estimators. *The Review of Economics and Statistics* 94(2), 481–498.

Ai, C. & X. Chen (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* 71(6), 1795–1843.

Amado, C. & T. Terasvirta (2013) Modelling volatility by variance decomposition. *Journal of Econometrics* 175(2), 142–153.

Andrews, D.W.K. (1994) Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* 62(1), 43–72.

Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models. In J.J. Heckman & E.E. Learner (eds.), *Handbook of Econometrics*, vol. 6, Part B, pp. 5549–5632. Elsevier.

Chen, X., H. Hong, & A. Tarozzi (2008) Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics* 36, 808–843.

Chen, X., O. Linton, & I. Van Keilegom (2003) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 71(5), 1591–1608.

Chen, X., A.T.K. Wan, & Y. Zhou (2015) Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association* 110(510), 723–741.

Christensen, B.J., C.M. Dahl, & E.M. Iglesias (2012) Semiparametric inference in a garch-in-mean model. *Journal of Econometrics* 167(2), 458–472.

Ding, Y. & B. Nan (2011) A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *The Annals of Statistics* 39(6), 3032–3061.

Dominitz, J. & R.P. Sherman (2005) Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory* 21(4), 838–863.

Engle, R.F., D.M. Lilien, & R.P. Robins (1987) Estimating time varying risk premia in the term structure: The arch-m model. *Econometrica* 55(2), 391–407.

Escanciano, J.C., D.T. Jacho-Chavez, & A. Lewbel (2014) Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing. *Journal of Econometrics* 178(0), 426–443.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Fan, Y., S. Pastorello, & E. Renault (2015) Maximization by parts in extremum estimation. *The Econometrics Journal* 18(2), 147–171.

Fiorentini, G., G. Calzolari, & L. Panattoni (1996) Analytic derivatives and the computation of garch estimates. *Journal of Applied Econometrics* 11(4), 399–417.

Graham, B.S., C.C. Pinto, & D. Egel (2012) Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies* 79, 1053–1079.

Horvitz, D.G. & D.J. Thompson (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260), 663–685.

Hu, Z., N. Wang, & R.J. Carroll (2004) Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika* 91(2), 251–262.

Huang, J. & J. Wellner (1997) Interval censored survival data: A review of recent progress. In D. Lin & T. Fleming (eds.), *Proceedings of the First Seattle Symposium in Biostatistics*. Lecture Notes in Statistics, pp. 123–169. Springer.

Hunter, D.R. & K. Lange (2000) Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics* 9(1), 60–77.

Ichimura, H. & S. Lee (2010) Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics* 159(2), 252–266.

Lange, K., D.R. Hunter, & I. Yang (2000) Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics* 9(1), 1–20.

Ling, S. (2004) Estimation and testing stationarity for double-autoregressive models. *Journal of the Royal Statistical Society Series B* (*Methodology*) 66(1), 63–78.

Linton, O. & B. Perron (2003) The shape of the risk premium. *Journal of Business & Economic Statistics* 21(3), 354–367.

Mammen, E., C. Roth, & M. Schienle (2016) Semiparametric estimation with generated covariates. *Econometric Theory* 32, 1140–1177.

Nan, B. & J.A. Wellner (2013) A general semiparametric z-estimator for case-cohort studies. *Statistical Sinica* 23(23), 1155–1180.

Nelson, D.B. (1991) Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 347–370.

Newey, W.K. (1994) The asymptotic variance of semiparametric estimators. *Econometrica* 62(6), 1349–1382.

Noureldin, D., N. Shephard, & K. Sheppard (2014) Multivariate rotated arch models. *Journal of Econometrics* 179(1), 16–30.

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

Pastorello, S., V. Patilea, & E. Renault (2003) Iterative and recursive estimation in structural nonadaptive models [with comments, rejoinder]. *Journal of Business & Economic Statistics* 21(4), 449–482.

Robins, M., A. Rotnitzky, & L. Zhao (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* 427, 846–866.

Roth, C. & S. Firpo (2016) Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. Working paper.

Scharfstein, D.O., A. Rotnitzky, & J.M. Robins (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94, 1096–1146.

Song, P.X.K., Y. Fan, & J. Kalbfleisch (2005) Maximization by parts in likelihood inference [with comments, rejoinder]. *Journal of the American Statistical Association* 100(472), 1145–1167.

Song, P., P. Zhang, & A. Qu (2007) Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica* 17(3), 929–943.

van Keilgom, I. & R.J. Carroll (2007) Backfitting versus profiling in general criterion functions. *Statistica Sinica* (17), 797–816.

Wooldridge, J.M. (2007) Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141(2), 1281–1301.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Xue, L. (2009) Empirical likelihood confidence intervals for response mean with data missing at random. *Scandinavian Journal of Statistics* 36(4), 671–685.
Zhang, R., C. Czado, & A. Min (2011) Efficient maximum likelihood estimation of copula based meta-distributions. *Computational Statistics and Data Analysis* 55(3), 1196–1214.
Zhou, Y., A.T.K. Wan, & X. Wang (2008) Estimating equations inference with missing data. *Journal of the American Statistical Association* 103(483), 1187–1199.

APPENDIX A: Technical Results

This appendix contains the proofs of Propositions 1 and 2, and lemmas used to establish these results. To simplify the notation used in this appendix, define

$$\bar{\theta}_N[\theta_1] = \bar{\theta}_N[\theta_1; \widehat{\eta}] \text{ and } \bar{\theta}[\theta_1] = \bar{\theta}[\theta_1; \eta^0].$$

A.1. Useful Lemmas

We first introduce a few lemmas that will be used to help prove Propositions 1 and 2.

LEMMA A.1. *Under Assumptions* ***A*** *and* ***B***: $\sup_{\theta_1 \in \Theta} |\bar{\theta}_N[\theta_1] - \bar{\theta}[\theta_1]| = o_p(1)$.

Proof. The proof is similar to Proposition 1 of Pastorello et al. (2003) and Theorem 1 of Chen et al. (2003), which is itself based on Corollary 3.2 of Pakes and Pollard (1989).

Define $Q[\theta, \theta_1; \eta] = |S(\theta, \theta_1; \eta)|$ and $Q_N[\theta, \theta_1; \eta] = |S_N(\theta, \theta_1; \eta)|$, where $S(\theta, \theta_1; \eta)$ and $S_N(\theta, \theta_1; \eta)$ were defined in Section 4. Let $\Theta^2 = \Theta \times \Theta$. Define

$$\epsilon^*(\Theta^2) = \inf_{(\theta, \theta_1) \in \Theta^2} \left\{ Q[\theta, \theta_1; \eta^0] - Q[\bar{\theta}[\theta_1], \theta_1; \eta^0] \right\}$$

and note that $\epsilon^*(\Theta^2)$ exists and, for $V \subset \Theta^2$, $\epsilon^*(V) \geq 0$. Define

$$V_\delta = \left\{ (\theta, \theta_1) \in \Theta^2 : |\theta - \bar{\theta}[\theta_1]| \geq \delta \right\}.$$

By the compactness of Θ and the continuity of $\theta_1 \mapsto \bar{\theta}[\theta_1]$, V_δ is compact and, for any $\delta > 0$,

$$\epsilon^*(V_\delta) = \inf_{(\theta, \theta_1) \in V_\delta} \left\{ Q[\theta, \theta_1; \eta^0] - Q[\bar{\theta}[\theta_1], \theta_1; \eta^0] \right\} > 0. \tag{A.1}$$

Fix $\delta > 0$. Equation (A.1) then implies that there exists some ϵ^*, with $0 < \epsilon^* < \epsilon^*(V_\delta)$, such that

$$\mathrm{P}\left(\sup_{\theta_1 \in \Theta} |\bar{\theta}_N[\theta_1] - \bar{\theta}[\theta_1]| \geq \delta \right) \leq \mathrm{P}\left(\sup_{\theta_1 \in \Theta} \left\{ Q[\bar{\theta}_N[\theta_1], \theta_1; \eta^0] - Q[\bar{\theta}[\theta_1], \theta_1; \eta^0] \right\} \geq \epsilon^* \right).$$

The result follows if

$$\lim_{N \to \infty} \mathrm{P}\left(\sup_{\theta_1 \in \Theta} \left\{ Q[\bar{\theta}_N[\theta_1], \theta_1; \eta^0] - Q[\bar{\theta}[\theta_1], \theta_1; \eta^0] \right\} \geq \epsilon^* \right) = 0.$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Note that, for any $\theta_1 \in \Theta$, by Assumption (A.2)

$$Q_N[\bar{\theta}_N[\theta_1],\theta_1;\widehat{\eta}] \le Q_N[\bar{\theta}[\theta_1],\theta_1;\widehat{\eta}] + o_p(1). \tag{A.2}$$

Let $\gamma > 0$. By Assumptions (A.1) and (A.4)

$$\lim_{N\to\infty} \mathrm{P}\left(\sup_{\theta_1\in\Theta} \left\{ Q[\bar{\theta}_N[\theta_1],\theta_1;\eta^0] - Q_N[\bar{\theta}_N[\theta_1],\theta_1;\widehat{\eta}] \right\} < \gamma/2 \right) = 1, \tag{A.3}$$

$$\lim_{N\to\infty} \mathrm{P}\left(\sup_{\theta_1\in\Theta} \left\{ Q_N[\bar{\theta}[\theta_1],\theta_1;\widehat{\eta}] - Q[\bar{\theta}[\theta_1],\theta_1;\eta^0] \right\} < \gamma/2 \right) = 1. \tag{A.4}$$

From equations (A.2)–(A.4) we can conclude

$$\lim_{N\to\infty} \mathrm{P}\left(\sup_{\theta_1\in\Theta} \left\{ Q[\bar{\theta}_N[\theta_1],\theta_1;\eta^0] - Q[\bar{\theta}[\theta_1],\theta_1;\eta^0] \right\} < \gamma \right) = 1. \tag{A.5}$$

Taking $\gamma = \epsilon^*$ yields the result. ■

LEMMA A.2. *Assumption* ***E*** *is satisfied and* $|\widehat{\theta}_N^{(k)} - \theta^0| = o_p(1)$. *If* $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)}) = o_p(1)$ *as* $N \to \infty$ *and* $k = k_N \to \infty$, *then*

$$S(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta}) = O_p(N^{-1/2}).$$

Proof. From consistency of $\widehat{\theta}_N^{(k)}$ there exists some positive sequence $\delta_N = o(1)$ such that $|\widehat{\theta}_N^{(k)} - \theta^0| \le \delta_N$ and $|\widehat{\theta}_N^{(k-1)} - \theta^0| \le \delta_N$, with probability approaching 1, and by Assumptions (E.1) and (E.3)

$$\sqrt{N}\left| S_N(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta}) - S(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta}) - S_N(\theta^0, \theta^0; \eta^0) \right| = o_p(1).$$

Applying the inequality $-|a| + |b| - |c| \le |a - b - c|$ to the above yields

$$\begin{aligned} -\sqrt{N}|S_N(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta})| + \sqrt{N}|S(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta})| - \sqrt{N}|S_N(\theta^0, \theta^0; \eta^0)| \\ \le \sqrt{N}\left| S_N(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta}) - S(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta}) - S_N(\theta^0, \theta^0; \eta^0) \right| \\ = o_p(1). \end{aligned}$$

By Assumptions (E.2) and (E.5) (part (ii)), rearranging terms yields (as $N \to \infty$ and $k \to \infty$)

$$\begin{aligned} \sqrt{N}|S(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta})| &\le o_p(1) + \sqrt{N}|S_N(\widehat{\theta}_N^{(k)}, \widehat{\theta}_N^{(k-1)}; \widehat{\eta})| + \sqrt{N}|S_N(\theta^0, \theta^0; \eta^0)| \\ &= o_p(1) + o_p(1) + O_p(1). \end{aligned}$$

■

LEMMA A.3. *Assumption* ***E*** *is satisfied and* $|\widehat{\theta}_N^{(k)} - \theta^0| = o_p(1)$. *If* $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)}) = o_p(1)$ *as* $N \to \infty$ *and* $k = k_N \to \infty$, *then*

$$|\widehat{\theta}_N^{(k)} - \theta^0| = O_p(N^{-1/2}).$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Proof. From consistency of $\widehat{\theta}_N^{(k)}$ and $\widehat{\eta}(\cdot,\theta)$, application of (E.4) yields

$$\begin{aligned}\sqrt{N}\left|S(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta})-S(\theta^0,\theta^0;\eta^0)-\dot{S}_1(\widehat{\theta}_N^{(k)}-\theta^0)-[\dot{S}_\theta-\dot{S}_1](\widehat{\theta}_N^{(k-1)}-\theta^0)-\dot{S}_\eta\left[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)\right]\right| \\ = o_p\left(\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|+\sqrt{N}|\widehat{\theta}_N^{(k-1)}-\theta^0|\right)+O_p\left(\sqrt{N}\|\widehat{\eta}(\cdot,\theta)-\eta^0(\cdot,\theta)\|_{\mathcal{H}}^{\alpha}\right) \\ = o_p\left(1+\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|+\sqrt{N}|\widehat{\theta}_N^{(k-1)}-\theta^0|\right),\end{aligned}$$

where the last line follows from (E.1) and (E.4), since $O_p\left(\sqrt{N}\|\widehat{\eta}(\cdot,\theta)-\eta^0(\cdot,\theta)\|_{\mathcal{H}}^{\alpha}\right)=o_p(1)$. By (E.7), $S(\theta^0,\theta^0;\eta^0)=0$. Then, applying the inequality $-|a|+|b|-|c|\leq|a-b-c|$ to the above yields

$$\begin{aligned}&-\sqrt{N}|S(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta})|+\sqrt{N}|\dot{S}_1(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N^{(k-1)})+\dot{S}_\theta(\widehat{\theta}_N^{(k-1)}-\theta^0)|-\sqrt{N}\left|\dot{S}_\eta\left[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)\right]\right| \\ &\leq o_p\left(1+\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|+\sqrt{N}|\widehat{\theta}_N^{(k-1)}-\theta^0|\right).\end{aligned}$$

By assumption $\sqrt{N}(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N^{(k-1)})=o_p(1)$ (as $N\to\infty$ and $k\to\infty$), and

$$\begin{aligned}&-\sqrt{N}|S(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta})|+\sqrt{N}|\dot{S}_\theta(\widehat{\theta}_N^{(k)}-\theta^0)|-\sqrt{N}\left|\dot{S}_\eta\left[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)\right]\right| \\ &\leq o_p\left(1+\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|\right). \qquad \textbf{(A.6)}\end{aligned}$$

Using the result of Lemma A.2 and (E.5) (part (i)) we can rewrite equation (A.6) as

$$-O_p(1)+\sqrt{N}|\dot{S}_\theta(\widehat{\theta}_N^{(k)}-\theta^0)|-O_p(1)\leq o_p\left(1+\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|\right). \qquad \textbf{(A.7)}$$

By (E.6), $\dot{S}_\theta$ is full rank and there exists some $c_1>0$ such that

$$|\dot{S}_\theta(\widehat{\theta}_N^{(k)}-\theta^0)|\geq c_1|\widehat{\theta}_N^{(k)}-\theta^0|.$$

Applying this result to equation (A.7) and rearranging terms gives

$$\begin{aligned}O_p(1)\geq& c_1\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|-o_p\left(\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|\right)-O_p(1)-o_p(1) \\ =&\left[O_p(1)-o_p(1)\right]\sqrt{N}|\widehat{\theta}_N^{(k)}-\theta^0|-O_p(1)\end{aligned}$$

and the result follows. ■

A.2. Proof of Proposition 1

Recall that $\widehat{\theta}_N^{(k)}=\bar{\theta}_N[\widehat{\theta}_N^{(k-1)}]$ and $\theta^0=\bar{\theta}[\theta^0]$. The triangle inequality yields

$$\begin{aligned}|\widehat{\theta}_N^{(k)}-\theta^0|&\leq|\bar{\theta}_N[\widehat{\theta}_N^{(k-1)}]-\bar{\theta}[\widehat{\theta}_N^{(k-1)}]|+|\bar{\theta}[\widehat{\theta}_N^{(k-1)}]-\bar{\theta}[\theta^0]| \\ &\leq\sup_{\theta_1\in\Theta}|\bar{\theta}_N[\theta_1]-\bar{\theta}[\theta_1]|+|\bar{\theta}[\widehat{\theta}_N^{(k-1)}]-\bar{\theta}[\theta^0]|. \qquad \textbf{(A.8)}\end{aligned}$$

We now consider the two cases in the result: (1), $|\widehat{\theta}_N^{(1)}-\theta^0|=o_p(1)$; (2), $|\widehat{\theta}_N^{(1)}-\theta^0|=O_p(1)$.

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Case (1). By Lemma A.1, $\sup_{\theta_1\in\Theta}|\bar{\theta}_N[\theta_1]-\bar{\theta}[\theta_1]| = o_p(1)$. By continuity of $\bar{\theta}[\cdot]$ and Assumption (C.1), we have $|\bar{\theta}[\widehat{\theta}_N^{(1)}]-\bar{\theta}[\theta^0]| = o_p(1)$ as $N\to\infty$. Taking $k=2$ in equation (A.8) delivers

$$|\widehat{\theta}_N^{(2)}-\theta^0| \le \sup_{\theta_1\in\Theta}|\bar{\theta}_N[\theta_1]-\bar{\theta}[\theta_1]| + |\bar{\theta}[\widehat{\theta}_N^{(1)}]-\bar{\theta}[\theta^0]| = o_p(1).$$

Therefore, $\widehat{\theta}_N^{(2)}$ is consistent and by continuity of $\bar{\theta}[\cdot]$ so is each subsequent iteration.

Case (2). By the contraction hypothesis (C.2), for some $c\in(0,1)$, $|\bar{\theta}[\widehat{\theta}_N^{(k-1)}]-\bar{\theta}[\theta^0]| \le c|\widehat{\theta}_N^{(k-1)}-\theta^0|$. Applying this hypothesis to equation (A.8) and iterating yields

$$\begin{aligned}
|\widehat{\theta}_N^{(k)}-\theta^0| &\le \sup_{\theta_1\in\Theta}|\bar{\theta}_N[\theta_1]-\bar{\theta}[\theta_1]| + c|\widehat{\theta}_N^{(k-1)}-\theta^0| \\
&\le (1+c)\sup_{\theta_1\in\Theta}|\bar{\theta}_N[\theta_1]-\bar{\theta}[\theta_1]| + c^2|\widehat{\theta}_N^{(k-2)}-\theta^0| \\
&\vdots \\
&\le \left(\sup_{\theta_1\in\Theta}|\bar{\theta}_N[\theta_1]-\bar{\theta}[\theta_1]|\right)\sum_{j=0}^{k-1}c^j + c^{k-1}|\widehat{\theta}_N^{(1)}-\theta^0| = o_p(1)\frac{1-c^{k-1}}{1-c} + c^{k-1}|\widehat{\theta}_N^{(1)}-\theta^0|,
\end{aligned}$$

where the $o_p(1)$ term follows from Lemma A.1 as $N\to\infty$. Choosing a sequence of iterations $k=k_N$ such that $k_N\to\infty$ as $N\to\infty$, we conclude that $c^{k-1}|\widehat{\theta}_N^{(1)}-\theta^0| = o_p(1)$ as $N\to\infty$. ■

A.3. Proof of Proposition 2

To simplify notation, when derivatives are evaluated at $(\theta^0,\theta^0,\eta^0)$ we drop this dependence from the expressions; e.g., $\dot{S}_1 = \dot{S}_1(\theta^0,\theta^0;\eta^0)$ and $\dot{S}_\theta = \dot{S}_\theta(\theta^0,\theta^0;\eta^0)$.

Proof of (i). Result (i) of Proposition 2 is similar to Theorem 2 of Chen et al. (2003) and Lemma 2.2 of Nan and Wellner (2013). Therefore, the proof is not repeated for the sake of brevity.

Proof of (ii). To prove Result (ii) of Proposition 2, we break the proof into two parts: part (1) shows that $\sqrt{N}(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N^{(k-1)}) = o_p(1)$ as $N\to\infty$, with $k=k_N$ as given in the result; part (2) uses $\sqrt{N}(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N^{(k-1)}) = o_p(1)$ to deduce $\sqrt{N}(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N) = o_p(1)$, for $k=k_N$ as in the result.

Part (1). Demonstration of $\sqrt{N}(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N^{(k-1)}) = o_p(1)$.

From consistency of $\widehat{\theta}_N^{(k)}$, $\widehat{\theta}_N^{(k-1)}$ and Assumptions (E.2) and (E.3),

$$\begin{aligned}
o_p(N^{-1/2}) &= \left[S_N(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta}) - S(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta}) - S_N(\theta^0,\theta^0;\eta^0)\right] \\
&= o_p(N^{-1/2}) - \left[S(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta}) + S_N(\theta^0,\theta^0;\eta^0)\right]. \qquad \textbf{(A.9)}
\end{aligned}$$

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

Expanding (A.9) using Assumption (E.4) and consistency of $\widehat{\theta}_N^{(k)}$, $\widehat{\theta}_N^{(k-1)}$:

$$\begin{aligned} S_N(\theta^0,\theta^0;\eta^0)+S(\widehat{\theta}_N^{(k)},\widehat{\theta}_N^{(k-1)};\widehat{\eta}) &= S_N(\theta^0,\theta^0;\eta^0)+S(\theta^0,\theta^0;\eta^0)+\dot{S}_1(\widehat{\theta}_N^{(k)}-\theta^0) \\ &\quad +[\dot{S}_\theta-\dot{S}_1](\widehat{\theta}_N^{(k-1)}-\theta^0)+\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)] \\ &\quad +o_p(|\widehat{\theta}_N^{(k)}-\theta^0|)+o_p(|\widehat{\theta}_N^{(k-1)}-\theta^0|)+O_p(\|\widehat{\eta}-\eta^0\|_{\mathcal{H}}^{\alpha}) \\ &= o_p(N^{-1/2}). \end{aligned} \tag{A.10}$$

By Assumption (E.1), $\|\widehat{\eta}-\eta^0\|_{\mathcal{H}} = O_p(N^{-\beta})$, so that for $\alpha > 1$ and $\alpha\beta > 1/2$, $O_p(\sqrt{N}\|\widehat{\eta}-\eta^0\|_{\mathcal{H}}^{\alpha}) = o_p(1)$. Therefore, from equation (A.10), Assumptions (E.6) and (E.7), and disregarding higher-order terms,

$$(\widehat{\theta}_N^{(k)}-\theta^0) = [-\dot{S}_1]^{-1}\left\{S_N(\theta^0,\theta^0;\eta^0)+\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)]+[\dot{S}_\theta-\dot{S}_1](\widehat{\theta}_N^{(k-1)}-\theta^0)+o_p(N^{-1/2})\right\}$$
$$(\widehat{\theta}_N^{(k-1)}-\theta^0) = [-\dot{S}_1]^{-1}\left\{S_N(\theta^0,\theta^0;\eta^0)+\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)]+[\dot{S}_\theta-\dot{S}_1](\widehat{\theta}_N^{(k-2)}-\theta^0)+o_p(N^{-1/2})\right\}.$$

Define $\Sigma := [-\dot{S}_1]^{-1}[\dot{S}_\theta-\dot{S}_1]$. Iterating the above relationships yields

$$(\widehat{\theta}_N^{(k)}-\theta^0) = \sum_{j=0}^{k-2}\Sigma^j[-\dot{S}_1]^{-1}\left\{S_N(\theta^0,\theta^0;\eta^0)+\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)]\right\}+\Sigma^{k-1}(\widehat{\theta}_N^{(1)}-\theta^0)+o_p(N^{-1/2}),$$
$$(\widehat{\theta}_N^{(k-1)}-\theta^0) = \sum_{j=0}^{k-3}\Sigma^j[-\dot{S}_1]^{-1}\left\{S_N(\theta^0,\theta^0;\eta^0)+\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)]\right\}+\Sigma^{k-2}(\widehat{\theta}_N^{(1)}-\theta^0)+o_p(N^{-1/2}).$$

Subtracting the two equations and multiplying by $\sqrt{N}$ yields

$$\begin{aligned} \sqrt{N}(\widehat{\theta}_N^{(k)}-\widehat{\theta}_N^{(k-1)}) &= \Sigma^{k-2}[-\dot{S}_1]^{-1}\sqrt{N}\left\{S_N(\theta^0,\theta^0;\eta^0)+\dot{S}_\eta[\widehat{\eta}(\cdot,\theta^0)-\eta^0(\cdot,\theta^0)]\right\} \\ &\quad +\Sigma^{k-2}(\Sigma-I_p)\sqrt{N}(\widehat{\theta}_N^{(1)}-\theta^0)+o_p(1) \\ &= \Sigma^{k-2}O_p(1)+\Sigma^{k-2}(\Sigma-I_p)\sqrt{N}(\widehat{\theta}_N^{(1)}-\theta^0)+o_p(1), \end{aligned} \tag{A.11}$$

where the last line follows by Assumption (E.5).

The result follows if the RHS of equation (A.11) is $o_p(1)$. Consider the following two cases for $|\widehat{\theta}_N^{(1)}-\theta^0|$: (1), $|\widehat{\theta}_N^{(1)}-\theta^0| = o_p(1)$ and (2), $|\widehat{\theta}_N^{(1)}-\theta^0| = O_p(1)$.

Case (1). From Assumption **D**, $0 < \|\Sigma\|_\rho < 1$, with $\|\cdot\|_\rho$ the spectral norm. Thus, the first term in equation (A.11), $\Sigma^{k-2}O_p(1)$, satisfies $\Sigma^{k-2}O_p(1) = o_p(1)$ as $k \to \infty$. However, even though $|\widehat{\theta}_N^{(1)}-\theta^0| = o_p(1)$ (by Assumption (C.1)), the term $\sqrt{N}\Sigma^{k-2}(\Sigma - I_p)(\widehat{\theta}_N^{(1)}-\theta^0)$ may not be $o_p(1)$ even if $k \to \infty$. However, we can control this term by appropriately choosing the number of iterations $k = k_N$, such that $k_N \to \infty$ as $N \to \infty$.

Case (2). Assumption (C.2) implies that $\|\Sigma\|_\rho < 1$. For $k \to \infty$ the first term in equation (A.11) satisfies $\Sigma^{k-2}O_p(1) = o_p(1)$. Therefore, similar to Case (1), we must control the second term in equation (A.11) by appropriately choosing $k = k_N$.

To control $\sqrt{N}\Sigma^{k-2}(\Sigma - I_p)(\widehat{\theta}_N^{(1)}-\theta^0)$ in equation (A.11) (under either case), first note that $\|\Sigma^{k-2}\|_\rho \leq \|\Sigma\|_\rho^{k-2}$ and $\|\Sigma\|_\rho < 1$. The result will then follow (under either case) if we can choose k such that $\sqrt{N}\|\Sigma\|_\rho^{k-2} \to 0$. Take $k = k_N$, where k_N satisfies

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press

$\frac{k_N}{(-0.5\ln(N)/\ln(\|\Sigma\|_\rho))} \to \infty$.[11] It then follows that $\sqrt{N}\|\Sigma\|_\rho^{k-2} \to 0$ as $N \to \infty$. Consequently, for $k = k_N$ as above, $\sqrt{N}\Sigma^{k-2}(\Sigma - I_p)(\widehat{\theta}_N^{(1)} - \theta^0) = o_p(1)$ under both cases and

$$\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)}) = o_p(1).$$

Part (2). Demonstration of $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N) = o_p(1)$.

Since $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N^{(k-1)}) = o_p(1)$ as $N \to \infty$, for $k = k_N$ as stated in the result, we replace $\widehat{\theta}_N^{(k-1)}$ in equation (A.10) with $\widehat{\theta}_N^{(k)}$ to obtain

$$o_p(N^{-1/2}) = S_N(\theta^0, \theta^0; \eta^0) + S(\theta^0, \theta^0; \eta^0) + \dot{S}_\theta(\widehat{\theta}_N^{(k)} - \theta^0) + \dot{S}_\eta[\widehat{\eta}(\cdot, \theta^0) - \eta^0(\cdot, \theta^0)] + o_p(|\widehat{\theta}_N^{(k)} - \theta^0|) + O_p(\|\widehat{\eta} - \eta^0\|_{\mathcal{H}}^\alpha).$$

By Assumption (E.1) and (E.4), $O_p(\sqrt{N}\|\widehat{\eta} - \eta^0\|_{\mathcal{H}}^\alpha) = o_p(1)$ and by Assumption (E.7) $0 = S(\theta^0, \theta^0; \eta^0)$. Likewise, by Lemma A.3, $o_p(\sqrt{N}|\widehat{\theta}_N^{(k)} - \theta^0|) = o_p(1)$ for $k = k_N$ as stated in the result. Applying these conclusions,

$$o_p(N^{-1/2}) = S_N(\theta^0, \theta^0; \eta^0) + \dot{S}_\eta[\widehat{\eta}(\cdot, \theta^0) - \eta^0(\cdot, \theta^0)] + \dot{S}_\theta(\widehat{\theta}_N^{(k)} - \theta^0) + o_p(N^{-1/2}).$$

From Assumption (E.6) and algebra, we can conclude $\sqrt{N}(\widehat{\theta}_N^{(k)} - \widehat{\theta}_N) = o_p(1)$ as $N \to \infty$. ■

https://doi.org/10.1017/S0266466618000063 Published online by Cambridge University Press