EFFICIENCY BOUNDS FOR SEMIPARAMETRIC MODELS WITH SINGULAR SCORE FUNCTIONS

PROSPER DOVONON AND YVES F. ATCHADÉ

(Dec. 2015)

Abstract. This paper is concerned with asymptotic efficiency bounds for the estimation of the finite dimension parameter $\theta \in \mathbb{R}^p$ of semiparametric models that have singular score function for θ at the true value $\theta_\star$. The resulting singularity of the matrix of Fisher information means that the standard bound derived by Begun et. al. ([1]) for $\theta - \theta_\star$ is not defined. We study the case of single rank deficiency of the score and focus on the case where the derivative of the root density in the direction of the last parameter component, θ_2, is nil while the derivatives in the $p-1$ other directions, θ_1, are linearly independent. We then distinguish two cases: (i) The second derivative of the root density in the direction of θ_2 and the first derivative in the direction of θ_1 are linearly independent and (ii) The second derivative of the root density in the direction of θ_2 is also nil but the third derivative in θ_2 is linearly independent of the first derivative in the direction of θ_1. We show that in both cases, efficiency bounds can be obtained for the estimation of $\kappa_j(\theta) = (\theta_1 - \theta_{\star 1}, (\theta_2 - \theta_{\star 2})^j)$, with $j = 2$ and 3, respectively and argue that an estimator $\hat{\theta}$ is efficient if $\kappa_j(\hat{\theta})$ reaches its bound. We provide the bounds in form of convolution and asymptotic minimax theorems. For case (i), we propose a transformation of the Gaussian variable that appears in our convolution theorem to account for the restricted set of values of $\kappa_2(\theta)$. This transformation effectively gives the efficiency bound for the estimation of $\kappa_2(\theta)$ in the model configuration (i). We apply these results to locally under-identified moment condition models and show that the generalized method of moment (GMM) estimator using $V_\star^{-1}$ as weighting matrix, where $V_\star$ is the variance of the estimating function, is optimal even in these non standard settings.

Keywords: Efficient estimation, semiparametric models, singular score, moment condition models, under-identification.

1. Introduction

Efficiency bounds for parameter estimation is a cornerstone in statistical inference. Such bounds set up a benchmark that helps assess whether a proposed estimator makes use of all the information that a sample can carry regarding the parameter of interest. Fundamental results in the form of convolution and asymptotic minimax theorem have been developed that are useful to derive these bounds in (a) parametric models (see e.g. [21, 23]); (b) non-parametric models (see e.g. [25, 2, 3, 27, 35]) and (c) semiparametric models ([33, 6, 1, 32, 9]).

Consider a semiparametric models[1] with a finite-dimensional parameter of interest $\theta \in \mathbb{R}^p$, and an infinite-dimensional nuisance parameter u which is a member of some large functional class. Begun, Hall, Huang, and Wellner ([1]), henceforth denoted BHHW, have shown that the asymptotic lower

This research is partially supported by the Fonds de Recherche du Québec - Société et Culture (FRQSC), and by the National Science Foundation grant DMS 1228164 and SES 1229261.

P. Dovonon: Concordia University, 1455 de Maisonneuve Blvd. West Montreal, Quebec H3G 1M8, Canada. *Email address:* prosper.dovonon@concordia.ca.

Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, MI 48109, United States. *E-mail address:* yvesa@umich.edu.

[1]We refer to ([29]) for a more precise characterization of semiparametric models.

bound for the estimation of θ under standard conditions is the inverse of the asymptotic Fisher information matrix. The existence of this bound requires that the Fisher information matrix be nonsingular at the truth. Even though this condition is fulfilled in many applications, there are some instances where it fails. Following Bickel ([6]), we shall refer to such parameter values as irregular. All the examples introduced in Section 3 feature true parameter values that are irregular. But, in spite of their irregular nature, it is shown that these parameters can still be consistently estimated. This motivates us to explore efficient estimation in such a context.

This paper is concerned with the efficient estimation of θ, the parametric component of a semi-parametric model, when the score function in the direction of θ is degenerate at the truth. This degeneracy implies that the Fisher information matrix is singular at the true value. We focuses on the case where the variance of the score function is of rank $p-1$ at the truth, with p the size of θ. In particular, we assume that the score in the direction of the first $p-1$ components of θ, say θ_1, are linearly independent while the score in the direction of the last component θ_2 is equal to 0. It is worth mentioning that the general rank $p-1$ setting fits into this configuration up to a rotation of the parameter space. Efficiency shall then be studied in the resulting system of coordinates in the light of the approach that we expose in this paper.

We build on the work of BHHW who rely on Hellinger differentiability of root-density function $f(\theta,u,\cdot)$ to obtain a proper characterization of the set of limit experiments over which the bound is derived. As we show, when the score function is degenerate, higher order approximation of the root density is needed to get a relevant set of limit experiments[2]. In particular, if (i) $\nabla^{(2)}_{\theta_2} f(\theta_\star,u_\star,\cdot)$ is not linearly dependent with $\nabla_{\theta_1} f(\theta_\star,u_\star,\cdot)$, we rely on second-order approximation through second-order Hellinger differentiability of $(\theta,u,\cdot)$ and if (ii) $\nabla^{(2)}_{\theta_2} f(\theta_\star,u_\star,\cdot)=0$ but $\nabla^{(3)}_{\theta_2} f(\theta_\star,u_\star,\cdot)$ is not linearly dependent of $\nabla_{\theta_1} f(\theta_\star,u_\star,\cdot)$, we rely on third-order approximation of the root-density; with $(\theta_\star,u_\star)$ denoting the true value of (θ,u).

This approach gives rise to a polynomial function $\kappa_\ell(\theta)$ for which efficiency bounds are derived in the form of convolution and minimax theorems. Specifically, $\kappa_\ell(\theta)=(\theta_1-\theta_{\star 1},(\theta_2-\theta_{\star 2})^\ell)$, with $\ell=2$ in case (i) and $\ell=3$ in case (ii). Our convolution theorems have the same flavor as the standard results for the estimation of $\theta-\theta_\star$ with the difference that the score in the direction of θ_2, in the form $\nabla_{\theta_2} f(\theta_\star,u_\star,\cdot)$, is replaced by $\frac{1}{2}\nabla^{(2)}_{\theta_2} f(\theta_\star,u_\star,\cdot)$ in case (i) and $\frac{1}{3!}\nabla^{(3)}_{\theta_2} f(\theta_\star,u_\star,\cdot)$ in case (ii). Since under (i) or (ii), the Fisher information matrix is singular and the lower bound for the estimation of $\theta-\theta_\star$ as derived by BHHW is not defined, we claim that an estimator $\hat\theta$ of $\theta_\star$ is efficient if $\kappa_\ell(\hat\theta)$, properly scaled, reaches the efficiency bound that we derive for $\kappa_\ell(\theta)$.

However, the parameter function $\kappa_\ell(\theta)$ with $\ell=2$ (in case (i)) has its last component that is nonnegative and nil at the truth and this exposes some admissibility issue. In effect, the convolution theorem that we establish gives the best Gaussian asymptotic approximation of any regular estimator

[2]([30]) uses a similar strategy for maximum likelihood estimation of finite-dimensional parameters with degenerate score functions

of $\kappa_\ell(\theta)$. But, a Gaussian approximation can only be a poor representation of $\kappa_2(\hat{\theta})$ since their supports do not coincide. One possibility of solving this problem is to adopt a Bayesian approach that incorporates the information on the support as a prior. Bickel ([5]) has used such an approach to derive bounds for the mean in a fully parametric and simple problem where data are generated from normal distribution with known variance and bounded mean. Solutions of this nature to general problems are not available in the literature. We rely on a different approach to incorporate this information on the support of $\kappa_2(\hat{\theta})$. Letting $Z_\star = (Z_{\star 1}, Z_{\star 2})$ be the best Gaussian asymptotic approximation of $\kappa_2(\hat{\theta})$, $\tilde{Z}_{\star 2} = Z_{\star 2}\mathbb{I}(Z_{\star 2} \geq 0)$, where $\mathbb{I}(\cdot)$ is the usual indicator function, is a better approximation to the last component $(\hat{\theta}_2 - \theta_{\star 2})^2$ of $\kappa_2(\hat{\theta})$. If $Z_{\star 1}$ and $Z_{\star 2}$ were independent, a natural more efficient approximation for $\kappa_2(\hat{\theta})$ would be $(Z_{\star 1}, \tilde{Z}_{\star 2})$. Allowing for dependence, we rely on a projection argument. Let $aZ_{\star 2}$ be the projection of $Z_{\star 1}$ on the span of $Z_{\star 2}$, and let $bU = Z_{\star 1} - aZ_{\star,2}$ be the residual, so that $Z_\star = (aZ_{\star 2} + bU, Z_{\star 2})$, where $Z_{\star,2}$ and U are independent Gaussian random variables, and the expressions of a, b are given in Section 4. We then define the best asymptotic approximation of any regular estimator of $\kappa_2(\theta)$ that incorporate the support information as $\tilde{Z}_\star = (a\tilde{Z}_{\star 2} + bU, \tilde{Z}_{\star 2})$. And we define the semiparametric information bound for estimating $\kappa_2(\theta)$ as $\mathsf{Var}(\tilde{Z}_\star)$.

Even though these bounds can be applied to any semiparametric model satisfying either (i) or (ii), our main motivation comes from moment condition models as reflected in our examples in Section 3. As we show in Section 2, these models can be represented as semiparametric models that depend not only on the parameter of interest θ but also on a nuisance parameter u that lies in a Hilbert space. Efficiency bound for the estimation of θ has been derived by Chamberlain ([8]) under the condition of first order local identification, i.e. the Jacobian of the estimating moment function at the true parameter value $\theta_\star$ is of full column-rank. We show that this corresponds in the semiparametric setting to nonsingular score function for θ at $(\theta_\star, u_\star)$. However, several papers have highlighted the possibility of failure of the first order local identification while higher order local identification is ensured (see e.g. [31, 24, 14, 15, 12, 13]). Higher order local identification refers to cases where the moment condition model is uniquely solved by $\theta_\star$ but more than a linear expansion of the moment function around $\theta_\star$ is needed to yield an approximation that uniquely determines $\theta_\star$. Examples 3.1, 3.2 and 3.3 in Section 3 show moment condition models with first order local identification failure. In particular, the conditional heteroskedasticity and skewness co-features in Example 3.3 are solution of moment condition models that are identified locally at the second and third order, respectively.

We show that the local behaviour of the moment function at $\theta_\star$ is quite connected to that of the implicit semiparametric density function that it defines. In particular, if the first derivative of the moment function in certain direction is nil, so is the derivative of the density in that direction. Also, if in addition, the second derivative of the moment function is nil, the same is true for the density function. Hence, depending on the local under-identification pattern of the moment condition model, the implicit semiparametric model satisfies (i) or (ii) and the bounds that we previously derived can be applied to these moment condition models in our examples.

Dovonon and Hall ([13]) have derived the asymptotic distribution of the generalized method of moment (GMM) estimator when the Jacobian matrix of the moment function is of rank $p-1$ while local identification is ensured at the second order. We show that when the weighting matrix is set to $V_\star^{-1}$, with $V_\star$ being the variance of the estimating function evaluated at the true value $\theta_\star$, the GMM estimator $\hat{\theta}$ is optimal in the sense that $\kappa_2(\hat{\theta}) = (\hat{\theta}_1 - \theta_{\star 1}, (\hat{\theta}_2 - \theta_{\star 2})^2)$, properly scaled, is asymptotically distributed as $\tilde{Z}_\star$.

We also derive the asymptotic distribution of the GMM estimator when the rank of the Jacobian matrix of the moment function at the truth is $p-1$ and the first two derivatives in the direction of θ_2 is nil while local identification is ensured at the third order. We show again that when $V_\star^{-1}$ is used as weighting matrix, the GMM estimator is efficient in the sense given above. That is, $\kappa_3(\hat{\theta}) = (\hat{\theta}_1 - \theta_{\star 1}, (\hat{\theta}_2 - \theta_{\star 2})^3)$, properly scaled, is asymptotically distributed as $\tilde{Z}_\star$. These results show that the well established optimality of the GMM estimator (using $V_\star^{-1}$ as weighting matrix) in standard models carries over to non standard models where the Jacobian matrix is rank deficient.

The rest of the paper is organized as follows: In Section 2, we introduce the moment condition models and derive the implicit semiparametric models that they induce. Applying the standard method of BHHW, we derive a lower bound for the parameter of interest using this implicit model. Our results confirm those of Chamberlain ([8]) namely that the GMM estimator with $V_\star^{-1}$ is efficient. Section 3 gives examples of moment condition models in which the true parameter value is not locally identified at the first order but rather at second or third order. Section 4 introduces our approach to derive efficiency bounds for semiparametric models with singular score function whereas Section 5 applies these results to moment condition models and establishes the efficiency of GMM estimator with $V_\star^{-1}$ as weighting matrix even in these non standard settings. We close the paper with some remarks in Section 6. Lengthy proofs are placed in the Appendix.

2. Semiparametric representation of moment equality models

The main motivation behind this work is the efficient estimation of parameters in moment equation models. We consider moment equality models describing data through some moment equations up to an unknown finite dimensional parameter $\theta \in \mathbb{R}^p$. Extending a result by Chamberlain ([8]), we first show that a moment equation model implicitly induces a semiparametric model that represents the distribution of the data up to θ and an infinite dimensional nuisance parameter u that lies in a Hilbert space. This semiparametric representation then provides a framework within which efficiency bounds for the estimation of θ can be derived. Although the semiparametric efficiency bounds subsequently derived can be applied more broadly, we shall mostly be concerned with their applications to moment equality models. As a result, we devote this section to a brief introduction to moment equality models, and their representation as semiparametric models.

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed $\mathbb{R}^k$-valued random variables with probability distribution $P_\star$. We write $L^2(P_\star)$ to denote the Hilbert space $L^2(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), P_\star)$ of

real-valued functions on $\mathbb{R}^k$. Assume that we are given a function ψ which maps $\mathbb{R}^p \times \mathbb{R}^k$ into $\mathbb{R}^q$ with the restriction on $P_\star$ taking the form of moment condition model:

$$\mathbb{E}_\star(\psi(\theta_\star, X)) \equiv \int \psi(\theta_\star, x) P_\star(dx) = 0, \tag{1}$$

where $\theta_\star$ is some point in $\mathbb{R}^p$ ($p \leq q$).

For $i \geq 0$, and for any $x \in \mathbb{R}^k$, we will use the notation $\nabla_\theta^{(i)}\psi(\theta, x)$ to denote the i-th order differential of the map $u \mapsto \psi(u, x)$ evaluated at θ (with the convention that $\nabla^{(0)}\psi(\theta, x) = \psi(\theta, x)$), and $\|\nabla_\theta^{(i)}\psi(\theta, x)\|$ will denote the operator norm of the differential. We make the following assumption:

Assumption 1. (1.1) *There exists a neighborhood Θ of $\theta_\star$, a $L^2(P_\star)$-neighborhood $\mathcal{N}$ of $f_\star \equiv 1$, a finite constant C, an integer $r \geq 1$, such that for $P_\star$-almost all $x \in \mathbb{R}^k$, $u \mapsto \psi(u, x)$ is r-times continuously differentiable on Θ, and for all $f \in \mathcal{N}$,*

$$\int \sup_{\theta \in \Theta} \|\nabla_\theta^{(i)}\psi(\theta, x)\| f^2(x) P_\star(dx) \leq C,$$

for $i = 0, \ldots, r$.

(1.2) *The matrix $V_\star \equiv \mathsf{Var}_\star(\psi(\theta_\star, X)) = \int \psi(\theta_\star, x)\psi'(\theta_\star, x) P_\star(dx)$ is positive definite.*

Remark 1. *Notice that the moment condition equation (1) implies that $\int \|\psi(\theta_\star, x)\| f_\star^2(x) P_\star(dx) < \infty$, a slightly stronger version of which is the integrability condition imposed in Assumption (1.1) when $i = 0$. This condition is needed for the function $\theta \mapsto \int \nabla_\theta^{(i)}\psi(\theta, x) f^2(x) P_\star(dx)$ to be well behaved.*

A commonly used estimator in the moment equation model (1) is the GMM estimator defined as

$$\hat{\theta}_{GMM} = \arg\min_{\theta \in \Theta} \bar{\psi}'(\theta) V_n \bar{\psi}'(\theta), \tag{2}$$

where $\bar{\psi}(\theta) = \frac{1}{n}\sum_{i=1}^n \psi(\theta, x_i)$ and $V_n \in \mathbb{R}^{q \times q}$ is a positive definite matrix. We are then naturally led to the question of whether the GMM estimator is efficient, and this question constitutes the main practical question addressed in this work.

To proceed, we introduce some notation that we carry throughout the paper. We equip the Hilbert space $L^2(P_\star)$ with the inner product $\langle u, v \rangle = \int u(x) v(x) P_\star(dx) = \mathbb{E}_\star(u(X)v(X))$. More generally, for $u_0 : \mathbb{R}^k \to \mathbb{R}$, $u : \mathbb{R}^k \to \mathbb{R}^{s \times r}$, and $v : \mathbb{R}^k \to \mathbb{R}^{p \times r}$, we set $\langle u, v \rangle \equiv \mathbb{E}_\star(u(X)v(X)')$, $\langle u_0, u \rangle \equiv \mathbb{E}_\star(u_0(X)u(X)')$, and $\langle u, u_0 \rangle \equiv \mathbb{E}_\star(u_0(X)u(X))$. Notice with these definitions that $\langle u, v \rangle' = \langle v, u \rangle$.

Let $\phi^0(x) = \left(\phi_1^0(x), \phi_2^0(x), \ldots, \phi_q^0(x)\right)' \equiv V_\star^{-1/2}\psi(\theta_\star, x)$ and $\phi_{q+1}(x) = 1$. We also define for $\theta \in \Theta$, $\psi_\theta(x) \equiv \psi(\theta, x)$, and $\phi_\theta^0(x) \equiv V_\star^{-1/2}\psi_\theta(x)$. We further introduce $\bar{\phi}^0 \equiv (1, \psi_{\theta_\star}' V_\star^{-1/2})' = (\phi_{q+1}, \{\phi^0\}')'$, and $\bar{\phi}_\theta^0 \equiv (1, \psi_\theta' V_\star^{-1/2})' = (\phi_{q+1}, \{\phi_\theta^0\}')'$.

Clearly, under the moment condition, the q components of ϕ^0 and ϕ_{q+1} are $q+1$ orthonormal vectors of $L^2(P_\star)$. Since $L^2(P_\star)$ is separable, we complete $\{\phi^0, \phi_{q+1}\}$ to have an orthonormal basis $\{\phi_j : j \geq 1\}$ of $L^2(P_\star)$. We denote by $\mathcal{E}$ the (closed) subspace of $L^2(P_\star)$ generated by the orthonormal

family $\{\phi_k : k \geq q+2\}$. Now, we consider the map $\mathcal{M} : \mathbb{R}^p \times \mathcal{E} \times L^2(P_\star) \to L^2(P_\star)$ defined by:

$$\mathcal{M}(\theta, u, f) \equiv \frac{1}{2} \left\langle f^2, \phi_\theta^0 \right\rangle \phi^0 + \frac{1}{2} \left(\int f^2(x) P_\star(dx) - 1 \right) \phi_{q+1} + \sum_{j=q+2}^{\infty} \langle \phi_j, f - u \rangle \phi_j. \tag{3}$$

Lemma 2.1. *Assume Assumption 1. Then $\mathcal{M}$ is r-times continuously differentiable on $\Theta \times \mathcal{E} \times \mathcal{N}$ and for all $(\theta, u, f) \in \Theta \times \mathcal{E} \times \mathcal{N}$, $\delta \in \mathbb{R}^p$, $k \in \mathcal{E}$, and $h \in L^2(P_\star)$,*

$$\nabla_\theta \mathcal{M}(\theta, u, f) \cdot \delta = \frac{1}{2} \delta' \left\langle f^2, \nabla_\theta \bar{\phi}_\theta^0 \right\rangle \bar{\phi}^0,$$

$$\nabla_u \mathcal{M}(\theta, u, f) \cdot k = -k, \qquad \text{and ,}$$

$$\nabla_f \mathcal{M}(\theta, u, f) \cdot h = \left\langle fh, \bar{\phi}_\theta^0 \right\rangle \bar{\phi}^0 + \sum_{j \geq q+2} \langle \phi_j, h \rangle \phi_j.$$

Proof. We write $\mathcal{M}$ as $\mathcal{M} = \mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3$, where $\mathcal{M}_1(\theta, u, f) = \frac{1}{2} \left\langle f^2, \phi_\theta^0 \right\rangle \phi^0$, $\mathcal{M}_2(\theta, u, f) = \frac{1}{2} \left(\int f^2(x) P_\star(dx) - 1 \right) \phi_{q+1}$, and $\mathcal{M}_3(\theta, u, f) = \sum_{j=q+2}^{\infty} \langle \phi_j, f - u \rangle \phi_j$, so that it is enough to establish the desired properties for each of these functions $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. $\mathcal{M}_3$ is a linear map and is trivially of class C^∞. $\mathcal{M}_2$ is quadratic, hence also of class C^∞. By Assumption 1, and standard results for exchanging integral and derivatives, it is straightforward to check that $\mathcal{M}_1$ is of class C^r. Hence the result. The expressions of the partial derivatives are straightforward to derive. □

The following lemma sets up the moment condition model (1) as a parametric model suitably indexed.

Lemma 2.2. *If $\theta_\star$ satisfies (1), and Assumption 1 holds for some $r \geq 1$, then there exists a neighborhood $\mathcal{V}$ of $(\theta_\star, u_\star)$ in $\mathbb{R}^p \times \mathcal{E}$, where $u_\star$ denotes the zero element of $\mathcal{E}$, a family $\{f(\theta, u, \cdot) : (\theta, u) \in \mathcal{V}\}$ of measurable functions on $\mathbb{R}^k$, such that $f(\theta_\star, u_\star, \cdot) \equiv 1$, and for all $(\theta, u) \in \mathcal{V}$,*

$$\int \psi(\theta, x) f^2(\theta, u, x) P_\star(dx) = 0, \quad \int f^2(\theta, u, x) P_\star(dx) = 1.$$

Furthermore, the map $(\theta, u) \mapsto f(\theta, u, \cdot)$ is r times differentiable and its first partial derivatives are given by

$$\forall h \in \mathcal{E}, \ \nabla_u f(\theta, u, \cdot) \cdot h = h - \left\langle f_{\theta,u} h, \bar{\phi}_\theta^0 \right\rangle \left\langle f_{\theta,u} \bar{\phi}^0, \bar{\phi}_\theta^0 \right\rangle^{-1} \bar{\phi}^0,$$

and $\forall w \in \mathbb{R}^p$,

$$\nabla_\theta f(\theta, u, \cdot) \cdot w = -\frac{1}{2} w' \left\langle f_{\theta,u}^2, \nabla_\theta \bar{\phi}_\theta^0 \right\rangle \left\langle f_{\theta,u} \bar{\phi}^0, \bar{\phi}_\theta^0 \right\rangle^{-1} \bar{\phi}^0.$$

In particular, $\nabla_\theta f(\theta, u, \cdot)$ evaluated at $(\theta_\star, u_\star)$ is $\nabla_\theta f(\theta_\star, u_\star, \cdot) = -\frac{1}{2} \Gamma' V_\star^{-1/2} \phi^0$, where

$$\Gamma \equiv \mathbb{E}_\star \left(\nabla_\theta \psi(\theta_\star, X) \right).$$

Remark 2. *For convenience in the notation we will at times write $f(\theta, u, \cdot)$ as $f_{\theta,u}$, and similarly for its derivatives.*

Lemma 2.2 shows that the moment condition (1), under Assumption 1, implicitly defines a semiparametric model $\{f^2(\theta,u,x)P_\star(dx),\ (\theta,u)\in\mathcal{V}\}$. This result is an extension of Lemma 1 of [8] which establishes a similar result for the case where the random variable X has finite support.

Perhaps, one of the most practical interests of Lemma 2.2 is the possibility it offers to study the asymptotic efficiency of estimating $\theta_\star$ through the induced semiparametric model $\{f^2(\theta,u,\cdot):(\theta,u)\in\mathcal{V}\}$. BHHW has developed a general methodology for deriving such bounds. However, and as noted above, their theory applies only to models that have a non-degenerate score at the true value; that is $\left\langle\nabla_\theta f_{(\theta_\star,u_\star)}(\cdot),\nabla_\theta f_{(\theta_\star,u_\star)}(\cdot)\right\rangle_\mu$ is non-singular[3]. From the last conclusion of Lemma 2.2, this condition is equivalent to Γ having full-column rank—the so-called first order local identification condition. This means that the standard method elaborated by BHHW does not apply to derive efficiency bounds for moment condition models that are not first-order identified. Examples of such models are given in next section.

Before moving to non-standard models, we first highlight how BHHW can be applied to the induced semiparametric model to get the efficiency bound for first-order locally identified (standard) moment condition models. In doing so, we also introduce concepts that will appear throughout the rest of the paper.

The local asymptotic normality (LAN) property of the sequence of experiments under consideration is essential in deriving the asymptotic efficiency bound through the standard techniques. In our case, a sequence of experiments is determined by any sequence $\{(\theta_n,u_n):n\in\mathbb{N}\}$ of elements of $\mathcal{V}$ (where $\mathcal{V}$ is the neighborhood of $(\theta_\star,0)$ obtained in Lemma 2.2) which in turn determines $f_n(\cdot)\equiv f(\theta_n,u_n,\cdot)$; the square of which is equal to the sequence of probability densities with respect to $P_\star$. As shown by BHHW, for the sequence of experiments determined by $\{(\theta_n,u_n):n\in\mathbb{N}\}$ to be LAN at $(\theta_\star,u_\star)$, it suffices that:

$$\|\sqrt{n}(f_n-f_\star)-\alpha\|_{L^2(P_\star)}\to 0\quad\text{as}\quad n\to\infty, \tag{4}$$

for some $\alpha\in L^2(P_\star)$, where $f_\star(\cdot)=f(\theta_\star,u_\star,\cdot)\equiv 1$.

The asymptotic efficiency bound is obtained as a function of the norm of one of such α's, the determination of which requires the characterization of the subset H_1 of $L^2(P_\star)$ of eligible values for α. Minimally, H_1 is determined by the sequence of experiments (θ_n,u_n) of interest and the local behaviour of the map $(\theta,u)\mapsto f(\theta,u,\cdot)$ in the neighborhood of $(\theta_\star,u_\star)$. The Hellinger differentiability property is considered for the *root density* function f.

Definition 1. ***(Hellinger differentiability of** f**).*** *The function $f=f(\theta,u,\cdot)$ is said to be first order Hellinger-differentiable at $(\theta,u)\in\mathcal{V}$ if there exists a function $\rho_\theta\in L^2(P_\star)$ and a bounded linear*

[3]The score function in the direction of θ is given by $\nabla_\theta\ln(f^2_{(\theta,u)}(\cdot))$ which amounts to $2\nabla_\theta f_{(\theta,u)}(\cdot)/f_{(\theta,u)}(\cdot)$.

operator $A : L^2(P_\star) \to L^2(P_\star)$ *such that, with* $f_n \equiv f(\theta_n, u_n, \cdot)$,

$$\frac{\|f_n - f - \{\rho_\theta \cdot (\theta_n - \theta) + A(u_n - u)\}\|_{L^2(P_\star)}}{\|\theta_n - \theta\| + \|u_n - u\|_{L^2(P_\star)}} \to 0 \quad as \quad n \to \infty,$$

for all sequences $\theta_n \to \theta$ *and* $u_n \to u$ *and* $(\theta_n, u_n) \in \mathcal{V}$ *for all* $n \geq 1$.

Remark 3. *Frechet differentiability is a sufficient condition for Hellinger differentiability. In that respect, the implicit semiparametric model defined by* $f(\theta, u, \cdot)$ *is Hellinger differentiable at any* $(\theta, u) \in \mathcal{V}$ *under Assumption 1. In this case,* ρ_θ *is simply* $\nabla_\theta f_{(\theta,u)}(\cdot)$. *The score function at* (θ, u) *in the direction of* θ *is* $2\rho_\theta(\cdot)/f_{(\theta,u)}(\cdot)$.

We now characterize the sequences of experiments of interest. Let $\theta_0 \in \mathbb{R}^p$, $\eta \in \mathbb{R}^p$ and let R_n be a diagonal (p,p)-matrix, with diagonal elements depending solely on n and diverging to ∞ as $n \to \infty$. Let $\Theta_1(\theta_0, \eta)$ denote the collection of all sequences $\{\theta_n\}_{n\geq 1}$ such that

$$R_n(\theta_n - \theta_0) - \eta \to 0, \quad as \quad n \to \infty,$$

and $\Theta_1(\theta_0) = \bigcup\{\Theta_1(\theta_0, \eta) : \eta \in \mathbb{R}^p\}$. Similarly, let $\mathcal{C}_1(u_0, \beta)$ ($\beta \in \mathcal{E}$) denote the collection of all sequences $\{u_n\}_{n\geq 1}$ with each $u_n \in \mathcal{U}$, the projection of $\mathcal{V}$ on $\mathcal{E}$, such that

$$\|\sqrt{n}(u_n - u_0) - \beta\|_{L^2(P_\star)} \to 0 \quad as \quad n \to \infty.$$

Let

$$\mathcal{B}_1(u_0) = \left\{\beta \in \mathcal{E} : \|\sqrt{n}(u_n - u_0) - \beta\|_{L^2(P_\star)} \to 0 \quad \text{as} \quad n \to \infty \text{ for some sequence } (u_n)_{n\geq 1} \text{ with all } u_n \in \mathcal{E}\right\}$$

and $\mathcal{C}_1(u_0) \equiv \bigcup_{\beta \in \mathcal{B}_1(u_0)} \mathcal{C}_1(u_0, \beta)$.

Under the assumption of Hellinger differentiability of f at $(\theta_\star, u_\star)$, Proposition 2.1 of BHHW establishes that, for sequences of experiments belonging to $\Theta_1(\theta_\star, \eta) \times \mathcal{C}_1(u_\star, \beta)$, α (as defined in (4)) is given by $\alpha = \rho_{\theta_\star} \cdot \eta + A_\star \beta$. More generally, when sequences of experiments are considered to be in $\Theta_1(\theta_\star) \times \mathcal{C}_1(u_\star)$, the corresponding collection of α's is given by the subset $H_1(\theta_\star, u_\star)$ of $L^2(P_\star)$ defined by:

$$H_1(\theta_\star, u_\star) = \left\{\alpha \in L^2(P_\star) : \alpha = \rho_{\theta_\star} \cdot \eta + A_\star \beta \text{ for some } \eta \in \mathbb{R}^p, \beta \in \mathcal{B}_1(u_\star)\right\}.$$

In the light of Lemma 2.1 of BHHW, as far as the LAN properties of the model of interest are concerned, we can index f_n either by $(\theta_n, u_n)_n \in \Theta_1(\theta_\star) \times \mathcal{C}_1(u_\star)$ or by $(\eta, \beta) \in \mathbb{R}^p \times \mathcal{B}_1(u_\star)$ or equivalently by $\alpha \in H_1(\theta_\star, u_\star)$. We make this explicit in Theorem 2.2 for instance where the notation $(f_{n,\alpha})_n$ for the sequence $(f_n)_n$ refers to $\alpha \in H_1(\theta_\star, u_\star)$ such that $\sqrt{n}(f_n - f_\star) \to \alpha$ in $L^2(P_\star)$ as $n \to \infty$.

The convolution result that follows next applies to sequences of estimators $\hat{\theta}_n$ that are regular.

Definition 2. *An estimator* $\hat{\theta}_n$ *of* θ_0 *is* R_n*-regular at* $f^2(\cdot) = f^2(\theta_0, u_0, \cdot)$ *if, for every sequence* $(f_n)_{n\geq 1}$, *with* $f_n(\cdot) \equiv f(\theta_n, u_n, \cdot)$ *and* $(\theta_n, u_n)_{n\geq 1} \in \Theta_1(\theta_0) \times \mathcal{C}_1(u_0)$, $R_n(\hat{\theta}_n - \theta_n)$ *converges in distribution under* f_n^2 *to* S *that depends only on* f^2, *i.e. only on* θ_0 *and* u_0.

The following result gives a convolution decomposition of the asymptotic distribution of any regular estimator of $\theta_\star$ of the moment condition model (1).

Theorem 2.1. *Let $R_{1n} = \sqrt{n} I_p$ and let $\hat{\theta}_n$ be an estimator of $\theta_\star$ R_{1n}-regular at $f^2_\star = f^2(\theta_\star, u_\star, \cdot)$ with limit distribution $S_\star$ and Γ be defined as in Lemma 2.2. Assume that Assumption 1 holds, and that Rank $(\Gamma) = p$. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{d} S_\star \stackrel{d}{=} Z_\star + W, \tag{5}$$

where $Z_\star \sim N(0, I^{-1}_{\star(1)})$ independent of the random vector W and $I_{\star(1)} = \Gamma' V_\star^{-1} \Gamma$.

We also have an asymptotic minimax optimality result for general class of loss functions that we state below. Let $\ell : \mathbb{R}^p \to \mathbb{R}_+$ be a loss function that is subconvex (i.e. $\{x : \ell(x) \leq y\}$ is closed, convex, and symmetric for every $y \geq 0$).

Theorem 2.2. *Under Assumption 1, if ℓ is subconvex, $\hat{\theta}_n$ is a measurable sequence of estimators of $\theta_\star$ and Rank $(\Gamma) = p$, then:*

$$\sup_{I \subset H} \liminf_{n \to \infty} \sup_{\alpha \in I} \mathbb{E}_{f^2_{n,\alpha}} \ell(n^{1/2}(\hat{\theta}_n - \theta_n)) \geq \mathbb{E}\ell(Z_\star),$$

where $Z_\star$ is defined in (5)*. The first supremum is taken over all finite subsets I of $H \equiv H_1(\theta_\star, u_\star)$.*

Proof. Follows readily from Theorem 3.11.5 of [34], page 417. □

This theorem is a simple application of the minimax theorem 3.11.5 of [34]. The measurability condition can be replaced by asymptotic measurability of the sequence of estimator $\hat{\theta}_n$. In this case, the first expectation in the conclusion of the theorem is taken with respect to the inner probability measure.

2.1. **Implications of these results for the GMM estimator.** The consequence of the above two theorems is that any regular estimator of $\theta_\star$ has an asymptotic variance that is at least as large as $I^{-1}_{\star(1)}$. Therefore, $I^{-1}_{\star(1)}$ stands for the efficiency bound for estimating $\theta_\star$ from the moment condition-based model (1). A similar result has been established by [8], using a different approach to ours. This result shows in particular that the GMM estimator $\hat{\theta}_{GMM}$ defined in (2) is asymptotically efficient under standard conditions, if V_n is a sequence of symmetric positive definite matrices that converges in probability to $V_\star^{-1}$. Indeed in this case, the estimator has $(\Gamma' V_\star^{-1} \Gamma)^{-1}$ as asymptotic variance.

The bounds provided by the convolution and the minimax theorems above are defined only if Γ is of full column rank. This is the so-called first order local identification condition at $\theta_\star$ for the moment condition model (1). Rank deficiency of Γ implies that $I_{\star(1)}$ is singular, therefore, $Z_\star$ is not a proper Gaussian variable. We explore in the next sections how these results are altered when first order local identification fails at the true value of the parameter of interest.

3. Some examples of moment equation models with rank deficient matrix Γ

The following examples illustrate the configuration where the moment condition model is solved at a certain value $\theta_\star$ that is unique solution in the parameter space but at which the Jacobian matrix Γ is rank deficient.

3.1. **A toy example.** Consider $y_i \sim iid(0,1)$, $x_i \sim iid(0,1)$ and x_i independent of y_i $(i = 1, \ldots, n)$ described by the moment condition model:

$$m(\theta) \equiv \mathbb{E}((y_i - \theta x_i)^2 - 1) = 0, \quad \theta \in \mathbb{R}.$$

Clearly, the moment function $m(\theta) = \theta^2$ and the moment condition model identifies the true parameter value $\theta_\star = 0$. However, the first derivative of m evaluated at $\theta_\star$ is nil meaning that the full rank condition fails in this model.

3.2. **An example by Rotnitzky, Cox, Bottai and Robins ([30]).** Suppose that $Y_i = (W_i, X_i)$, $i = 1, \ldots, n$ are independent random variables and conditionally on X_i,

$$W_i = e^{-\theta X_i} - \sum_{k=0}^{s-1} \frac{(-1)^k}{k!} \theta^k X_i^k + \varepsilon_i$$

with $\mathbb{E}(\varepsilon_i | X_i) = 0$. Take $s = 2$ and assume that $X_i \sim N(0,1)$ and $\mathbb{E}(W_i) = 0$. Y_i can then be described by the moment condition

$$m(\theta) \equiv \mathbb{E}(\varepsilon_i) = 0.$$

It is not hard to see that $m(\theta) = -e^{\theta^2/2} + 1$ so that this moment condition model identifies $\theta_\star = 0$. But, clearly the first order local identification condition fails since $\partial m(\theta_\star)/\partial\theta = 0$. One may want to rely on more moment restrictions to restore local identification but the problem typically persists. Consider the moment condition model of the form:

$$\mathbb{E}(h(X_i)\varepsilon_i) = 0,$$

with $h(X_i)$ being any $\mathbb{R}^q$-valued function of X_i that meets the integrability requirements with one component being a constant function. We can show that this moment condition also identifies $\theta_\star = 0$ but the first order local identification condition for $\theta_\star$ fails.

3.3. **Volatility and skewness co-features in asset returns.** Let $r_t \equiv (r_{1t}, r_{2t})', t = 1, \ldots$ be a bivariate stationary process of two stock returns. Let $\mathfrak{F}_t$ be an increasing filtration of the information available on the market up to t (including past returns). The process (r_{it}) has a conditionally heteroskedastic feature if $\mathsf{Var}(r_{it}|\mathfrak{F}_{t-1})$ is time variant. This is the consequence of the so-called volatility clustering feature that is a well-known stylized fact for stock returns. Similarly, this process has dynamic asymmetry feature if $\mathbb{E}(r_{it}^3|\mathfrak{F}_t)$ is time variant. This is also a well-documented feature for stock returns. (See e.g. [19, 20, 17, 11]).

If these two assets share a *common conditionally heteroskedastic factor*, they can be represented as:

$$r_t = \Lambda f_t + u_t, \tag{6}$$

with $\Lambda = (\lambda_1, \lambda_2)' \in \mathbb{R}^2$, f_t a common factor that is a $\mathbb{R}$-valued process such that $\mathbb{E}(f_t|\mathfrak{F}_{t-1}) = 0$ and $\mathsf{Var}(f_t|\mathfrak{F}_{t-1})$ time variant. $u_t = (u_{1t}, u_{2t})'$ is the vector of idiosyncratic shocks satisfying: $\mathbb{E}(u_t|\mathfrak{F}_{t-1}) = 0$, $\mathsf{Var}(u_t|\mathfrak{F}_{t-1}) = \Omega$, constant and $\mathsf{Cov}(u_t, f_t|\mathfrak{F}_{t-1}) = 0$.

Such factor structure is appealing for multivariate volatility modeling. The conditional heteroskedasticity in returns passes by the common factor to which the idiosyncratic shocks are conditionally uncorrelated. This shocks are not conditionally heteroskedastic. (See [10, 22, 18, 15] for more details on these models.)

This factor structure can be tested by observing that it implies the existence of a linear combination of the returns that offsets the conditionally heteroskedastic feature. That is, there exists a so-called co-feature vector $(\theta_1, \theta_2) \neq (0, 0)$ such that:

$$\mathsf{Var}(\theta_1 r_{1t} + \theta_2 r_{2t}|\mathfrak{F}_{t-1}) = cst. \tag{7}$$

For identification purpose the co-feature is determined uniquely by setting e.g. $\theta_1 = 1$ (see [15] for more details). Using a vector of instrument $(1, z_t')'$ belonging to $\mathfrak{F}_t$, the conditional moment restriction (7) implies an unconditional moment restriction

$$m_1(\theta) \equiv \mathbb{E}\left\{(z_{t-1} - \mathbb{E}(z_{t-1})[(r_{1t} + \theta r_{2t})^2 - \mathbb{E}(r_{1t} + \theta r_{2t})^2)]\right\} = 0. \tag{8}$$

Dovonon and Renault ([15]) propose a test for common conditionally heteroskedastic features based on this moment conditional model. They establish in particular that if the factor structure is correct, (8) identifies the co-feature θ (which can therefore be consistently estimated) but the first order local identification condition does not hold as they show that

$$\frac{\partial m_1}{\partial \theta}(\theta_\star) = 0, \quad with \quad \theta_\star = -\lambda_1/\lambda_2. \tag{9}$$

Common factors in skewness can also be evaluated in asset returns that all have time varying conditional third moment through a similar factor structure to (6) with $\mathbb{E}(f_t|\mathfrak{F}_{t-1}) = 0$ and $s_{t-1} \equiv \mathbb{E}(f_t^3|\mathfrak{F}_{t-1})$ time variant. $u_t = (u_{1t}, u_{2t})'$ is the vector of idiosyncratic shocks satisfying: $\mathbb{E}(u_t|\mathfrak{F}_{t-1}) = 0$, $\mathbb{E}(Vec(u_t u_t')u_t'|\mathfrak{F}_{t-1}) = s$, constant, and $\mathsf{Cov}(Vec(u_t u_t'), f_t|\mathfrak{F}_{t-1}) = 0$ and $\mathsf{Cov}(u_t, f_t^2|\mathfrak{F}_{t-1}) = 0$.

In the same spirit as for volatility, the skewness co-feature is determined by

$$\mathbb{E}((r_{1t} + \theta r_{2t})^3|\mathfrak{F}_{t-1}) = cst \tag{10}$$

which implies the unconditional moment condition:

$$m_2(\theta) \equiv \mathbb{E}\left\{(z_{t-1} - \mathbb{E}(z_{t-1})[(r_{1t} + \theta r_{2t})^3 - \mathbb{E}(r_{1t} + \theta r_{2t})^3)]\right\} = 0. \tag{11}$$

Note that nothing guarantees that the skewness co-feature is the same as that of volatility since common features may exist in volatility and not in skewness and vice versa. The moment condition

(11) is useful to test whether there is a common factor in skewness and also to consistently estimate the skewness co-feature. Actually, it is not hard to see that

$$m_2(\theta) = (\lambda_1 + \lambda_2\theta)^3 \mathsf{Cov}(z_{t-1}, s_{t-1})$$

so that if there is at least one component of z_t that is correlated with s_t, $m_2(\theta) = 0$ identifies $\theta_\star = -\lambda_1/\lambda_2$ that solves the moment restrictions. But, we can also see that:

$$\frac{\partial m_2}{\partial \theta}(\theta_\star) = 0, \quad \text{and} \quad \frac{\partial^2 m_2}{\partial \theta^2}(\theta_\star) = 0, \quad \text{whereas} \quad \frac{\partial^3 m_2}{\partial \theta^3}(\theta_\star) = 6\lambda_2^3 \mathsf{Cov}(z_{t-1}, s_{t-1}) \neq 0. \tag{12}$$

Both (9) and (12) show that estimating co-(volatility or skewness)-features in the framework of factor models lead to models that are not first order locally identified. In the case of co-skewness-features, the relevant moment condition models may even have second-order derivatives that are zero at the true value.

4. Efficiency bound for semiparametric models with singular score

Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d) $\mathfrak{X}$-valued random variables with density function $f^2(\theta_1, \theta_2, u, \cdot)$ with respect to a sigma-finite measure μ on a measurable space $(\mathfrak{X}, \mathscr{C})$ where θ_1, $\theta_2 \in \mathbb{R}$ and u is a measurable function on $(\mathfrak{Y}, \mathscr{D})$, a measurable space equipped with a sigma-finite measure ν. Let $L^2(\mu)$ and $L^2(\nu)$ denote $L^2(\mathfrak{X}, \mathscr{C})$ and $L^2(\mathfrak{Y}, \mathscr{D})$, respectively. We assume that $u \in L^2(\nu)$. By definition, $f \in L^2(\mu)$ and $\|f\|_\mu = 1$.

Our goal is to derive the efficiency bound for estimating (θ_1, θ_2) while u is treated as a nuisance parameter. We consider the standard case where f is differentiable at the true value $(\theta_\star, u_\star) = (\theta_{\star 1}, \theta_{\star 2}, u_\star)$ but we depart from the standard settings by considering that the score function vanishes in the direction of θ_2 at $(\theta_\star, u_\star)$, that is

$$\nabla_{\theta_2} f(\theta_\star, u_\star, \cdot) \equiv 0. \tag{13}$$

This singularity implies in particular that the Fisher information matrix for estimating $(\theta_{\star 1}, \theta_{\star 2})$ is singular and the efficiency bounds for its estimation cannot be derived using the standard approaches. This non-suitability carries over to the search of bounds in either direction (θ_1 or θ_2). For instance, from the results of BHHW, if $u_\star$ is known, a bound for the estimation of $\theta_{\star 1}$ is simply the inverse of 4 times the squared $L^2(\mu)$-norm of the regression residual of $\nabla_{\theta_1} f(\theta_\star, u_\star, \cdot)$ on $\nabla_{\theta_2} f(\theta_\star, u_\star, \cdot)$. Under (13), this is the inverse of 4 times the squared $L^2(\mu)$-norm of $\nabla_{\theta_1} f(\theta_\star, u_\star, \cdot)$. This corresponds to the efficiency bound for estimating $\theta_{\star 1}$ if $\theta_{\star 2}$ were actually known. Intuitively, such a bound would not be sharp as it would not be reachable by any regular estimator of $\theta_\star$.

The standard treatment of efficiency bounds derivation is based on first order approximation of f. The function f is assumed first order Hellinger differentiable at $(\theta_\star, u_\star)$ (that is f is Frechet-differentiable at $(\theta_\star, u_\star)$) and, thanks to the linear independence of the vector of the components of $\nabla_\theta f(\theta_\star, u_\star, \cdot)$, this first order approximation of f is enough to establish the mapping of the sequences

of experiments indexed by $(\theta_n, u_n) \in \Theta_1(\theta_\star, \eta) \times \mathcal{C}_1(u_\star, \beta)$ into the space $H_1(\theta_\star, u_\star)$[4] which is big enough to allow for the study of the local efficiency of a large class of estimators. However, when there is linear dependence of scores, $H_1(\theta_\star, u_\star)$ is not big enough to get general results such as Theorem 3.1 of BHHW. In fact, $I_\star$ as defined in that theorem is nil and $Z_\star$ as defined in their convolution theorem is not a proper Gaussian random variable.

A natural way to explore larger sets of limit experiments consists of exploring higher order approximations of f. This leads us to introduce the notion of second (or higher) order Hellinger differentiability. In what follows, we consider $\theta_2 \in \mathbb{R}$ since this is the case where easily interpretable results are possible, and $\theta_1 \in \mathbb{R}^k, k \geq 1$. We set k to 1 without loss of generality so that typically, $\theta \in \mathbb{R} \times \mathbb{R}$.

Definition 3. *$f(\theta, u, \cdot)$ is said to be second-order Hellinger-differentiable at $(\theta, u) \in \mathbb{R} \times \mathbb{R} \times L^2(\nu)$ if there exists: $\rho_\theta = (\rho_{\theta_1} \ \rho_{\theta_2})$ with $\rho_\theta \in L^2(\mu) \times L^2(\mu)$, a bounded linear operator $A : L^2(\nu) \to L^2(\mu)$; $\rho_{\theta\theta} = (\rho_{\theta_i\theta_j})_{ij} : 1 \leq i, j \leq 2$, with $\rho_{\theta_i\theta_j} \in L^2(\mu)$, $\forall i, j$; a continuous bilinear operator $B : L^2(\nu) \times L^2(\nu) \to L^2(\mu)$; and two continuous bilinear operators $C_1, C_2 : L^2(\nu) \times \mathbb{R} \to L^2(\mu)$ such that, for all sequences $\theta_n \to \theta$ and $u_n \to u$ in $L^2(\nu)$,*

$$\frac{\|f_n - f - \xi(\theta_n - \theta, u_n - u)\|_\mu}{(\|\theta_n - \theta\| + \|u_n - u\|_\nu)^2} \to 0 \quad as \quad n \to \infty,$$

with $f_n \equiv f(\theta_n, u_n, \cdot)$, and

$$\begin{aligned} \xi(\theta_n - \theta, u_n - u) \ \equiv \ & \rho_\theta \cdot (\theta_n - \theta) + A(u_n - u) + \tfrac{1}{2}(\theta_n - \theta)' \rho_{\theta\theta} (\theta_n - \theta) \\ & + \tfrac{1}{2} B(u_n - u, u_n - u) + C_1(u_n - u, \theta_{1n} - \theta_1) + C_2(u_n - u, \theta_{2n} - \theta_2). \end{aligned} \tag{14}$$

Remark 4. *If $f(\theta, u, \cdot)$ is twice differentiable at (θ, u), then $f(\theta, u, \cdot)$ is second-order Hellinger-differentiable at (θ, u). This follows from the Taylor formula. In this case, ρ_θ and $\rho_{\theta\theta}$ are the first and second partial derivatives of f with respect to θ at (θ, u), A and B are the first and second partial derivatives of f with respect to u at (θ, u), and C is the second partial derivative of f with respect to u and θ at (θ, u).*

Under second-order Hellinger differentiability, even if ρ_{θ_2} vanishes, so long as $\rho_{\theta_2\theta_2}$ does not vanish and is linearly independent of ρ_{θ_1}, it will be possible to suitably enlarge the set of experiments beyond the standard one. In doing so, the new sequences of experiments will allow the determination of relevant efficiency bounds. In some problems (see Example 3.3), there is a possibility that both ρ_{θ_2} and $\rho_{\theta_2\theta_2}$ vanish. In such situations, higher order Hellinger differentiability would rather be considered.

We will first study the case where, at $(\theta_\star, u_\star)$, $\rho_{\theta_2} = 0$ but ρ_{θ_1} and $\rho_{\theta_2\theta_2}$ are linearly independent. This case is encapsulated in Assumption 2 below. We will follow this by the case where third-order Hellinger differentiability is required; in which case, we will assume that $\rho_{\theta_2} = \rho_{\theta_2\theta_2} = 0$ and linear independence of ρ_{θ_1} and $\rho^{(3)}_{\theta_2}$, the third derivative of f with respect to θ_2.

[4]The sets Θ_1, $\mathcal{C}_1$ and H_1 are defined as in Section 2 but with the spaces introduced in the current section.

Assumption 2. *f is second-order Hellinger differentiable at $(\theta_\star, u_\star)$ where $\rho_{\theta_2} = 0$ and ρ_{θ_1} and $\rho_{\theta_2\theta_2}$ are linearly independent.*

In the framework of Assumption 2, the standard sequences of experiment determined by the root-n-rate of convergence as introduced through $R_{1n} = \sqrt{n}I_p$ in the previous section would not be relevant for our theory of efficiency. This is because under the assumption $\rho_{\theta_2} = 0$, the rate $\sqrt{n}$ is no longer typical for estimators of $\theta_\star$ as illustrated by the following result. For the sake of simplicity, we assume that f depends only on θ_2 (θ_1 and u are supposed known or absent from the model).

Lemma 4.1. *Assume that $X_1, \ldots, X_n$ are iid $\mathfrak{X}$-valued random variables with density function $f^2(\theta_2, \cdot)$ with respect to a sigma-finite measure μ on a measurable space $(\mathfrak{X}, \mathscr{C})$. Assume that f is Hellinger differentiable at $\theta_{\star 2}$ and $\rho_{\theta_2} \equiv \nabla_{\theta_2} f(\theta_{\star 2}, \cdot) = 0$. Then, there is no $\sqrt{n}$-regular estimator for $\theta_{\star 2}$.*

This result complements Theorem 1(ii) of [7] who has provided a different proof than ours. In spite of this, it is still possible to estimate consistently θ_2 but typically at a slower rate. We know from [30] that, under Assumption 2, with $u_\star$ known or nonexistent, the maximum likelihood estimator of θ_2 is $n^{1/4}$-consistent and that of $\theta_{\star 1}$ is $\sqrt{n}$-consistent. Therefore, it makes sense to explore efficiency properties in the family of experiments that are indexed by θ_1 and θ_2 that lie in a $\sqrt{n}$-shrinking neighborhood of $\theta_{\star 1}$ and $n^{1/4}$-neighborhood of $\theta_{\star 2}$, respectively. Let R_{2n} be the diagonal $(2,2)$-matrix with diagonal elements $\sqrt{n}$ and $n^{1/4}$, respectively. For $\eta \in \mathbb{R}^2$, let $\Theta_2(\theta, \eta)$ be the collection of all sequences $\{\theta_n\}_{n\geq 1}$ such that:

$$R_{2n}(\theta_n - \theta) - \eta \to 0, \quad \text{as} \quad n \to \infty$$

and let $\Theta_2(\theta) = \bigcup\{\Theta_2(\theta, \eta) : \eta \in \mathbb{R}^2\}$. We let $\mathcal{C}_2(u, \beta)$ be defined analogously to $\mathcal{C}_1(u, \beta)$ in Section 2 but with sequences $\{u_n\}_{n\geq 1}$ having elements in $L^2(\nu)$ and $\beta \in L^2(\nu)$. We also define $\mathcal{B}_2(u)$ similarly to $\mathcal{B}_1(u)$ and $\mathcal{C}_2(u) = \bigcup_{\beta \in \mathcal{B}_2(u)} \mathcal{C}_2(u, \beta)$.

Proposition 4.1. *Suppose that f is second-order Hellinger-differentiable at $(\theta, u) \in \mathbb{R}^2 \times L^2(\nu)$ and that $\rho_{\theta_2} = 0$. Let $\{(\theta_n, u_n)\}_{n\geq 1} \in \Theta_2(\theta, \eta) \times \mathcal{C}_2(u, \beta)$ for some $\eta \in \mathbb{R}^2$ and $\beta \in L^2(\nu)$. Then, with $f_n \equiv f(\theta_n, u_n, \cdot)$ and $f \equiv f(\theta, u, \cdot)$,*

$$\|\sqrt{n}(f_n - f) - \alpha\|_\mu \to 0 \quad as \quad n \to \infty, \tag{15}$$

where $\alpha \in L^2(\mu)$ is given by:

$$\alpha = \eta_1 \rho_{\theta_1} + \frac{1}{2}\eta_2^2 \rho_{\theta_2\theta_2} + A\beta.$$

Proof. Sketch: Write the second-order Hellinger-differentiability definition for f_n, (θ_n, u_n). Use the triangle inequality to conclude. □

This proposition is analogue to Proposition 2.1 of BHHW and characterizes the limits of experiments indexed by sequences in $\Theta_2(\theta, \eta) \times \mathcal{C}_2(u, \beta)$. The main difference with BHHW is that the linear term

$\eta_2\rho_{\theta_2}$ which can be considered as the score in the direction of θ_2 is replaced by a quadratic term in η_2: $\eta_2^2\rho_{\theta_2\theta_2}$. This linear term vanishes because the score of the model f_n in the direction of θ_2 vanishes at $\theta_{\star 2}$. The second-order quadratic term does not drop out because $\Theta_2(\theta,\eta)$ includes $n^{1/4}$-neighborhoods of θ_2 which are large enough as to make information from second-order expansion count in the determination of the limit experiments.

For $\{f_n\}_{n\geq 1}$ and f defined as in Proposition 4.1, the following lemma establishes the local asymptotic normality of the local likelihood ratio L_n and the contiguity result useful to derive our convolution theorem.

Let

$$L_n = \log\{\prod_{i=1}^{n}[f_n^2(X_i)/f^2(X_i)]\}.$$

Lemma 4.2. *If f_n and f as in Proposition 4.1 satisfy* (15)*, then, for every $\varepsilon > 0$,*

$$P_f\{|L_n - 2n^{-1/2}\sum_{i=1}^{n}\alpha(X_i)/f(X_i) + \sigma^2/2| > \varepsilon\} \to 0 \quad as \quad n \to \infty$$

where $\sigma^2 = 4\|\alpha\|_\mu^2$. Thus, under P_f,

$$L_n \xrightarrow{d} N(-\sigma^2/2, \sigma^2) \quad as \quad n \to \infty$$

and the sequences $\{\prod_{i=1}^n f_n^2(x_i)\}$ and $\{\prod_{i=1}^n f^2(x_i)\}$ are contiguous.

We refer to BHHW and the references therein for the proof of this lemma. In the light of Proposition 4.1 and Lemma 4.2, as far as the LAN property of the sequences of experiments considered is of concern, we can either index these experiments by sequences $\{(\theta_n, u_n)\}_{n\geq 1} \in \Theta_2(\theta_\star) \times \mathcal{C}_2(u_\star)$, by their limits: $(\eta,\beta) \in \mathbb{R}^2 \times \mathcal{B}_2(u_\star)$, or, alternatively, by $\alpha \in H_2(\theta_\star, u_\star)$:

$$H_2(\theta_\star, u_\star) = \left\{\alpha \in L^2(\mu) : \alpha = \eta_1\rho_{\theta_1} + \frac{1}{2}\eta_2^2\rho_{\theta_2\theta_2} + A\beta; (\eta_1,\eta_2) \in \mathbb{R}^2, \beta \in \mathcal{B}_2(u_\star)\right\}.$$

In preparation for our convolution theorem, we introduce the notion of regular estimator in the context of Assumption 2. It is natural to consider estimators $\hat{\theta}$ of $\theta_\star$ that are R_{2n}-regular at $f_\star^2 = f^2(\theta_\star, u_\star, \cdot)$ in the sense that, for every sequence $\{f_n = f(\theta_n, u_n, \cdot)\}_{n\geq 1}$ with $\{(\theta_n, u_n)\}_{n\geq 1} \in \Theta_2(\theta_\star) \times \mathcal{C}_2(u_\star)$, $R_{2n}(\hat{\theta} - \theta)$ converges in distribution under f_n^2 to S that depends only on $f_\star^2$. But, since $\rho_{\theta_{\star 2}} = 0$, the Fisher information is nil in the direction of $\theta_{\star 2}$ making the quest for efficiency bound for $\theta - \theta_\star$ rather difficult even in the family of R_{2n}-regular estimators; because of the singular of the Fisher information matrix. Nevertheless, a natural function of θ_2 that can be estimated at the standard $\sqrt{n}$-rate is $t_2(\theta_2) = (\theta_2 - \theta_{\star 2})^2$. Instead of searching for efficiency bound on $\theta - \theta_\star$, we will rather derive bounds for the estimation of $\kappa_2(\theta) = (\theta_1 - \theta_{\star 1}, t_2(\theta_2))$ which can be dealt with using some existing framework upon some further elaboration. The bound that we will derive for $\kappa_2(\theta)$ has some connection with the Bhattacharyya bound ([4]) in the same way the standard asymptotic bounds are connected to the Cramer-Rao bound.

We say that $(\hat{\theta}_{1n}, \hat{t}_{2n})$ is a $\sqrt{n}$-regular estimator of $(\theta_1, t_2(\theta_2))$ at $f_\star^2 = f^2(\theta_\star, u_\star, \cdot)$ if for every sequence $\{f_n = f(\theta_n, u_n, \cdot)\}_{n\geq 1}$ with $\{(\theta_n, u_n)\}_{n\geq 1} \in \Theta_2(\theta_\star) \times \mathcal{C}_2(u_\star)$, $\sqrt{n}(\hat{\theta}_{1n} - \theta_{1n}, \hat{t}_{2n} - t_2(\theta_{2n}))$ converges in distribution (under f_n^2) to S that depends only on $f_\star^2$, i.e. only on $\theta_\star$ and $u_\star$.

Toward the statement of the convolution result, we make the following assumption:

Assumption 3. *$\mathcal{B}_2(u_\star)$ is a subspace of $L^2(\nu)$.*

Let the orthogonal projections of ρ_{θ_1} and $\frac{1}{2}\rho_{\theta_2\theta_2}$ onto $\{A\beta : \beta \in \mathcal{B}_2(u_\star)\}$ be given by $A\beta_1^\star$ and $A\beta_2^\star$, respectively. Assumption 3 guarantees the existence of $\beta_1^\star$ and $\beta_2^\star$ in $\mathcal{B}_2(u_\star)$ such that $\rho_{\theta_1} - A\beta_1^\star \perp A\beta$ and $\frac{1}{2}\rho_{\theta_2\theta_2} - A\beta_2^\star \perp A\beta$, for all $\beta \in \mathcal{B}_2(u_\star)$. Let

$$s_\star = \begin{pmatrix} \rho_{\theta_1} - A\beta_1^\star \\ \frac{1}{2}\rho_{\theta_2\theta_2} - A\beta_2^\star \end{pmatrix} \quad \textit{and} \quad I_{\star(2)} = 4\langle s_\star, s_\star \rangle_\mu.$$

We have the following:

Theorem 4.1. *Suppose that $(\hat{\theta}_{1n}, \hat{t}_{2n})'$ is an estimator of $(\theta_{\star 1}, t_2(\theta_{\star 2}))$ that is $\sqrt{n}$-regular at $f_\star^2 = f(\theta_\star, u_\star, \cdot)$ with limit distribution S under $f_\star^2$, i.e. $\sqrt{n}(\hat{\theta}_{1n} - \theta_{\star 1}, \hat{t}_{2n} - t_2(\theta_{\star 2})) \xrightarrow{d} S$, under $f_\star^2$. Suppose, in addition that Assumption 3 and the conclusion of Proposition 4.1 hold at $f_\star$ with α as specified and that $I_{\star(2)}$ is nonsingular. Then*

$$S \stackrel{d}{=} Z_\star + W, \tag{16}$$

where $Z_\star \sim N\left(0, I_{\star(2)}^{-1}\right)$; with $Z_\star$ and W independent.

Proof. See Appendix. □

This result is similar to that of BHHW but with ρ_{θ_2} replaced by $\frac{1}{2}\rho_{\theta_2\theta_2}$. As it turns out, as the standard score in the direction of θ_2 (ρ_{θ_2}) vanishes at $(\theta_\star, u_\star)$, the second-order derivative $\rho_{\theta_2\theta_2}$ now plays the role of score in the definition of minimum achievable variance. The condition $I_{\star(2)}$ nonsingular implies that $\rho_{\theta_2\theta_2}$ does not vanish and even more, that the functions ρ_{θ_1} and $\rho_{\theta_2\theta_2}$ are linearly independent. This is an essential condition to obtain estimators of $\theta_{\star 2}$ that have optimal rate $n^{1/4}$.

This result shows in particular that any $\sqrt{n}$-regular estimator of $(\theta_{\star 1}, t_2(\theta_{\star 2}))$ must have an asymptotic variance that is at least as large as $I_{\star(2)}^{-1}$. Let

$$\begin{pmatrix} (I_{\star(2)})_{11} & (I_{\star(2)})_{12} \\ (I_{\star(2)})_{21} & (I_{\star(2)})_{22} \end{pmatrix}$$

denote the partition of $I_{\star(2)}$ along the dimensions of the two components θ_1 and θ_2. As a result, the minimum variance of any such estimator of $\theta_{\star 1}$ is given by

$$\left((I_{\star(2)})_{11} - \frac{1}{(I_{\star(2)})_{22}}(I_{\star(2)})_{12}(I_{\star(2)})_{21}\right)^{-1},$$

and that of $t_2(\theta_{\star 2})$ is

$$\left((I_{\star(2)})_{22} - (I_{\star(2)})_{21}(I_{\star(2)})_{11}^{-1}(I_{\star(2)})_{12}\right)^{-1}.$$

However, it is important to mention that under the conditions of the theorem, $\hat{t}_{2n}$ consistently estimates $(\theta - \theta_{\star 2})^2$ which is a nonnegative quantity with true value 0 lying on the boundary of the parameter set. One would therefore expect $\hat{t}_{2n}$ to be nonnegative for admissibility purpose. This prior information is not taken into account in deriving the convolution result above. Such prior may be more suitable to incorporate in Bayesian frameworks (see [5]) but, to the best of our knowledge, no Bayesian theory exists to deal with this problem.

One would expect that an efficiency bound for estimating $t_2(\theta_{\star 2})$ in the family of regular estimators that account for this information to be smaller than $\mathsf{Var}(Z_{\star 2}) = \left((I_{\star(2)})_{22} - (I_{\star(2)})_{21}(I_{\star(2)})_{11}^{-1}(I_{\star(2)})_{12}\right)^{-1}$. For the same reason, the efficiency bound for regular estimators of $\theta_{1\star}$ that account for the range of $t_2(\theta_{\star 2})$ shall be at most, as large as $\mathsf{Var}(Z_{\star 1})$. We proceed as follows in an attempt to insert this prior information into the derived bound. Since $t_2(\theta_2) \geq 0$ with true value at 0, it is reasonable to consider that $\sqrt{n}(\hat{t}_{2n} - t_2(\theta_{\star 2}))$ is better approximated by $\tilde{Z}_{\star 2} \stackrel{\text{def}}{=} Z_{\star 2}\mathbb{I}(Z_{\star 2} \geq 0)$. Let $\begin{pmatrix} aZ_{\star 2} \\ Z_{\star 2} \end{pmatrix}$ be the projection of $Z_\star = \begin{pmatrix} Z_{\star 1} \\ Z_{\star 2} \end{pmatrix}$ on $Z_{\star 2}$, and let $\begin{pmatrix} bU \\ 0 \end{pmatrix}$, with $U \sim \mathbf{N}(0, I_{p-1})$, be the linear projection of $Z_\star$ on the space orthogonal to $Z_{\star 2}$, so that

$$Z_\star = \begin{pmatrix} aZ_{\star 2} + bU \\ Z_{\star 2} \end{pmatrix}. \tag{17}$$

By construction, $Z_{\star 2}$ and U are independent Gaussian random variables, and from the joint distribution of $Z_\star$, we have

$$a = \frac{1}{(I_{\star(2)}^{-1})_{22}}(I_{\star(2)}^{-1})_{12}, \quad \text{and} \quad b = \left((I_{\star(2)}^{-1})_{11} - \frac{1}{(I_{\star(2)}^{-1})_{22}}(I_{\star(2)}^{-1})_{12}(I_{\star(2)}^{-1})_{21}\right)^{1/2}.$$

Replacing $Z_{\star 2}$ by $\tilde{Z}_{\star 2}$ in (17) we define

$$\tilde{Z}_\star \stackrel{\text{def}}{=} \begin{pmatrix} aZ_{\star 2}\mathbb{I}(Z_{\star 2} \geq 0) + bU \\ Z_{\star 2}\mathbb{I}(Z_{\star 2} \geq 0) \end{pmatrix}. \tag{18}$$

And we define $\mathsf{Var}(\tilde{Z}_\star)$ as the minimum asymptotic variance of $\sqrt{n}(\hat{\theta}_{1n} - \theta_{\star 1}, \hat{t}_{2n} - t_2(\theta_{\star 2}))$ in presence of the nonnegativity constraint. Note that by construction $\mathsf{Var}(\tilde{Z}_\star) \preceq \mathsf{Var}(Z_\star)$. In spite of the fact that this bound is not derived directly from the convolution theorem, it will prove useful in explaining the behavior of the GMM estimator under first-order local identification failure as we shall see in next section.

We now turn our attention to efficiency bounds when the first two derivatives of the density function vanish in some direction at the true parameter value. The asset returns' skewness co-feature model (11) in Example 3.3 is such a case as we shall see in next section. Again, assume that the model is parameterized by (θ, u) and $\theta \in \mathbb{R} \times \mathbb{R}$ with first and second derivatives, ρ_{θ_2} and $\rho_{\theta_2\theta_2}$ at $(\theta_\star, u_\star)$ both nil. The existence of the bound derived in Theorem 4.1 requires the linear independence of ρ_{θ_1} and $\rho_{\theta_2\theta_2}$). This bound is therefore not applicable in this case.

Actually, we can show along the lines of Lemma 4.1 that θ_2 cannot be estimated by any $n^{1/4}$-regular estimator. Estimation results for this framework are available in Rotnitzky et al. (2000), albeit in a parametric framework with finite dimension parameter. They show that θ_2 can be consistently estimated and under linear independence of ρ_{θ_1} and $\rho^{(3)}_{\theta_2}$ at the true parameter value ($\rho^{(3)}_{\theta_2}$ standing for the third derivative of f in the direction of θ_2), the rate of convergence of the maximum likelihood estimator of θ_1 is $\sqrt{n}$ and that of θ_2 is $n^{1/6}$.

Following the same approach as in the configuration of Assumption 2, we will consider sequences of experiments indexed by sequences of parameters that are in a $\sqrt{n}$-shrinking-neighborhood of $\theta_{\star 1}$ and a $n^{1/6}$-shrinking-neighborhood of $\theta_{\star 2}$, respectively. We let R_{3n} be the diagonal (2,2)-matrix with diagonal elements $\sqrt{n}$ and $n^{1/6}$. The sequences of parameters are $\{\theta_n\}_{n\geq 1}$ such that:

$$R_{3n}(\theta_n - \theta) - \eta \to 0, \quad \text{as} \quad n \to \infty$$

($\eta \in \mathbb{R}^2$) and are collected in the set $\Theta_3(\theta, \eta)$. We let $\Theta_3(\theta) = \bigcup\{\Theta_3(\theta, \eta) : \eta \in \mathbb{R}^2\}$ and $\mathcal{C}_3(u, \beta)$, $\mathcal{B}_3(u)$ and $\mathcal{C}_3(u)$ be defined similarly to $\mathcal{C}_2(u, \beta)$ and $\mathcal{C}_1(u, \beta)$; $\mathcal{B}_2(u)$ and $\mathcal{B}_1(u)$; and $\mathcal{C}_2(u)$ and $\mathcal{C}_1(u)$, respectively.

By analogy to the previous case, we will need that f is third-order Hellinger differentiable at $(\theta_\star, u_\star)$. A formal definition can be stated along the lines of Definition 3. Third order-Hellinger differentiability is guaranteed by third-order Frechet differentiability. We make the following assumption that summarizes the framework under study:

Assumption 4. *f is third-order Hellinger differentiable at $(\theta_\star, u_\star)$ where $\rho_{\theta_2} = \rho_{\theta_2\theta_2} = 0$ and ρ_{θ_1} and $\rho^{(3)}_{\theta_2}$ are linearly independent.*

Under Assumption 4, it is not hard to derive the limits of $\sqrt{n}(f_n - f)$, where the sequence of experiments f_n are properly indexed:

Proposition 4.2. *Suppose that f is third-order Hellinger differentiable at $(\theta, u) \in \mathbb{R}^2 \times L^2(\nu)$ and that $\rho_{\theta_2} = 0$ and $\rho_{\theta_2\theta_2} = 0$. Let $\{(\theta_n, u_n)\}_{n\geq 1} \in \Theta_3(\theta, \eta) \times \mathcal{C}_3(u, \beta)$ for some $\eta \in \mathbb{R}^2$ and $\beta \in L^2(\nu)$. Then, with $f_n \equiv f(\theta_n, u_n, \cdot)$ and $f \equiv f(\theta, u, \cdot)$, (15) holds with $\alpha \in L^2(\mu)$ given by:*

$$\alpha = \eta_1 \rho_{\theta_1} + \frac{1}{6}\eta_2^3 \rho^{(3)}_{\theta_2} + A\beta. \tag{19}$$

Proof. Sketch: Write the third-order Hellinger-differentiability definition for f_n, (θ_n, u_n) and make use of successive applications of the triangle inequality to conclude. □

As previously seen, the conclusion of this proposition is sufficient to deduce the LAN property of the likelihood ratio of the experiments described here as established by Lemma 4.2 with α given by (19).

The natural function of the parameter that we consider for which an asymptotic efficiency bound is derived is $\kappa_3(\theta) = (\theta_1, t_3(\theta_2))$, with $t_3(\theta_2) = (\theta_2 - \theta_{\star 2})^3$. We claim that $(\hat{\theta}_{1n}, \hat{t}_{3n})$ is a $\sqrt{n}$-regular estimator of $(\theta_1, t_3(\theta_2))$ at $f^2_\star = f^2(\theta_\star, u_\star, \cdot)$ if for every sequence $\{f_n = f(\theta_n, u_n, \cdot)\}_{n\geq 1}$ with

$\{(\theta_n, u_n)\}_{n\geq 1} \in \Theta_3(\theta_\star) \times \mathcal{C}_3(u_\star)$, $\sqrt{n}(\hat{\theta}_{1n} - \theta_{1n}, \hat{t}_{3n} - t_3(\theta_{2n}))$ converges in distribution (under f_n^2) to S that depends only on $f_\star^2$, i.e. only on $\theta_\star$ and $u_\star$.

As in Theorem 4.1, the convolution theorem here requires that:

Assumption 5. *$\mathcal{B}_3(u_\star)$ is a subspace of $L^2(\nu)$.*

Let $I_{\star(3)}$ be the same as $I_{\star(2)}$ of Theorem 4.1 but with $\frac{1}{6}\rho_{\theta_2}^{(3)}$ and $\mathcal{B}_3(u_\star)$ replacing $\frac{1}{2}\rho_{\theta_2\theta_2}$ and $\mathcal{B}_2(u_\star)$, respectively. We have:

Theorem 4.2. *Suppose that $(\hat{\theta}_{1n}, \hat{t}_{3n})'$ is an estimator of $(\theta_{\star 1}, t_3(\theta_{\star 2}))$ that is $\sqrt{n}$-regular at $f_\star^2 = f^2(\theta_\star, u_\star, \cdot)$ with limit distribution S under $f_\star^2$, i.e. $\sqrt{n}(\hat{\theta}_{1n} - \theta_{\star 1}, \hat{t}_{3n} - t_3(\theta_{\star 2})) \xrightarrow{d} S$ under $f_\star^2$. Suppose, in addition that Assumption 5 and the conclusion of Proposition 4.2 hold at $f_\star$ with α given by (19) and that $I_{\star(3)}$ is nonsingular. Then*

$$S \stackrel{d}{=} Z_\star + W, \tag{20}$$

where $Z_\star \sim N(0, I_{\star(3)}^{-1})$; with $Z_\star$ and W independent.

We do not provide the proof of this result since it is similar to that of Theorem 4.1. As expected, as the first and second derivatives vanish at the true parameter value in the direction of θ_2, the third-order derivative kicks in to replace the score that appears in the standard case. It is worth mentioning that in this case where the transformation estimated is $t_3(\theta_2) = (\theta_2 - \theta_{\star 2})^3$, Gaussian estimators are admissible since the true value of $t_3(\theta_{\star 2}) = 0$ is interior to the range of $t_3(\theta_2)$ which is the whole real line $\mathbb{R}$. In this case, no prior information is useful as opposed to the previous case where $t_2(\theta_2)$ is nonnegative with true value on the boundary.

The following asymptotic minimax result also holds for the estimation of $(\theta_{\star 1}, t_3(\theta_{\star 2}))$:

Theorem 4.3. *Suppose that the conclusion of Proposition 4.2 and Assumption 5 hold, $I_{\star(3)}$ is non-singular, $\hat{\kappa}_n = (\hat{\theta}_{1n}, \hat{t}_{3n})'$ is a measurable estimator of $\kappa_3(\theta_\star) \equiv (\theta_{\star 1}, t_3(\theta_{\star 2}))$ and that ℓ is a subconvex function. Then*

$$\sup_{I \subset H} \liminf_{n\to\infty} \sup_{\alpha \in I} \mathbb{E}_{f_{n,\alpha}^2} \ell(\sqrt{n}(\hat{\kappa}_n - \kappa_3(\theta_n))) \geq \mathbb{E}\ell(Z_\star),$$

where $Z_\star$ is defined as in (20). The first supremum is taken over all finite subsets I of $H \equiv H_3(\theta_\star, u_\star)$, where $H_3(\theta, u)$ is defined similarly to $H_2(\theta, u)$ with $\frac{1}{2}\eta_2^2\rho_{\theta_2\theta_2}$ replaced by $\frac{1}{6}\eta_2^3\rho_{\theta_2}^{(3)}$ and $\mathcal{B}_2(u)$ by $\mathcal{B}_3(u)$.

Proof. Follows readily from Theorem 3.11.5 of [34]. □

We end this section by the following remarks:

Remark 5. *If the nonparametric component $u_\star$ is known, the results in Theorems 4.1, 4.2 and 4.3 hold with $\beta_1^\star = \beta_2^\star = 0$. Also, if $(\rho_{\theta_1}, \rho_{\theta_2\theta_2})$ is orthogonal to $\mathcal{B}_2(u_\star)$ in the context of Theorem 4.1 (or $(\rho_{\theta_1}, \rho_{\theta_2}^{(3)})$ is orthogonal to $\mathcal{B}_3(u_\star)$ in the context of Theorem 4.2), $\beta_1^\star$ and $\beta_2^\star$ are nil and we get the same bounds for the estimation of $\theta_\star$ as if $u_\star$ were known. These conditions give the possibility to have sequences of estimators $\hat{\theta}_n$ of $\theta_\star$ that are adaptive to the nonparametric direction.*

Remark 6. *Our results can easily be extended to semiparametric models in which all derivatives at* $(\theta_\star, u_\star)$ *in the direction of* θ_2 *are nil up to* $j-1$, $j \geq 2$. *In this case, efficiency bound can be obtained for the estimation of* $(\theta_{\star 1}, t_j(\theta_{\star 2}))$, *with* $t_j(\theta_2) = (\theta_2 - \theta_{\star 2})^j$. *Under similar assumptions to those maintained in Theorem 4.1, the conclusion of that theorem holds with* $I_{\star(2)}$ *replaced by* $I_{\star(j)}$; *this latter is analogue to* $I_{\star(2)}$ *but with* $\frac{1}{2}\rho_{\theta_2\theta_2}$ *replaced by* $\frac{1}{j!}\rho^{(j)}_{\theta_2}$.

Of course, when j *is even, one has to be aware of the prior information that* $t_j(\theta_2)$ *is nonnegative with true value at 0. This information can be integrated to the efficiency bound determination in line with the approach suggested following Theorem 4.1.*

At this stage, it is worth recalling that asymptotic efficiency bound for the estimation of $\theta - \theta_\star$ cannot be obtained through existing techniques under Assumption 2 or 4. Under these assumptions, bounds for $\kappa_j(\theta) = (\theta_1 - \theta_{\star 1}, (\theta_2 - \theta_{\star 2})^j)$ are derived in this paper. It makes sense then to explore efficiency of an estimator $\hat{\theta}$ of θ through the efficiency of $\kappa_j(\hat{\theta})$. We can therefore claim that an estimator $\hat{\theta}$ of θ is efficient in the context of Assumption 2 if $\kappa_2(\hat{\theta})$ reaches the efficiency bounds derived for $\kappa_2(\theta)$ by (18) and in the context of Assumption 4 if $\kappa_3(\hat{\theta})$ reaches the efficiency bounds derived by Theorems 4.2 and 4.3.

It is also worth mentioning the possibility of the score function ρ_θ to be singular without vanishing in a particular direction. Such configuration has not been explicitly studied in this paper. However, it is not hard to see that such a model can be re-parameterized through a change of coordinate system such that, so long as the score degenerates in a single direction, we have ρ_{η_1} is not degenerate and $\rho_{\eta_2} = 0$, with $(\eta_1, \eta_2) \in \mathbb{R}^k \times \mathbb{R}$ corresponding to θ in the new coordinate system. Efficient estimation of such models can then be explored by our method in this new model re-parameterization.

5. Application to locally under-identified moment equality models

This section derives efficiency bounds for the estimation of θ in locally under-identified moment condition models and investigates whether the two-step GMM reaches those bounds?

5.1. **Efficiency bounds.**

We have seen in Lemma 2.2 that the moment condition model can be represented locally as a semiparametric model $\{f_{\theta,u},\ (\theta,u) \in \mathcal{V}\}$. As shown by [8] (see also Theorem 2.1 and Theorem 2.2 above), this representation can be used to derived the semiparametric efficiency bound for estimating θ, under the assumption that the moment condition model is first order identifiable. One important consequence of this analysis is the conclusion that the GMM estimator is efficient, since it reaches the semiparametric efficiency bound.

When the first-order identifiability assumption breaks down, the general results (Theorem 4.1 and 4.2) derived above can be used to obtain the semiparametric efficiency bound of θ. We specialize these results to moment condition models, and compare the semiparametric bound to the asymptotic variance of the GMM estimator.

We focus on the case where $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{p-1} \times \mathbb{R}$. The following lemma which actually is a mere consequence of Lemma 2.2 highlights some local properties of the implicit family of density induced by a moment condition model when this latter fails the first order local identification condition. We use the notation $D \stackrel{\text{def}}{=} \mathbb{E}\left[\nabla_{\theta_1}\psi_{\theta_\star}(X)\right]$, $G_i \stackrel{\text{def}}{=} \mathbb{E}\left[\nabla^{(i)}_{\theta_2}\psi_{\theta_\star}(X)\right]$ $(i \geq 1)$, as well as the notation of Section 2.

Lemma 5.1. *Assume Assumption 1, and let $\{f_{\theta,u},\ (\theta,u) \in \mathcal{V}\}$ be the semiparametric model defined by the moment condition model, as obtained in Lemma 2.2.*

(1) *If Assumption 1 holds then*

$$\nabla_{\theta_1} f_{(\theta_\star,0)} = -\frac{1}{2} D' V_\star^{-1/2} \phi^0, \quad \textit{and} \quad \nabla_{\theta_2} f_{(\theta_\star,0)} = -\frac{1}{2} G_1' V_\star^{-1/2} \phi^0.$$

In particular, if $G_1 = 0$, then $\nabla_{\theta_2} f_{(\theta_\star,0)} = 0$.

(2) *If Assumption 1 holds with $r \geq 2$, and $\nabla_{\theta_2} f_{(\theta_\star,0)} = 0$, then*

$$\nabla^{(2)}_{\theta_2} f_{(\theta_\star,0)} = -\frac{1}{2} G_2' V_\star^{-1/2} \phi^0.$$

(3) *If Assumption 1 holds with $r \geq 3$, and $\nabla_{\theta_2} f_{(\theta_\star,0)} = 0$, $\nabla^{(2)}_{\theta_2} f_{(\theta_\star,0)} = 0$, then*

$$\nabla^{(3)}_{\theta_2} f_{(\theta_\star,0)} = -\frac{1}{2} G_3' V_\star^{-1/2} \phi^0.$$

Using this result, we can apply Theorem 4.1 and 4.2 to the moment equation model. Set $t_j(\theta_2) = (\theta_2 - \theta_{\star 2})^j$.

Corollary 5.1. *Suppose that Assumption 1 holds with some $r \geq 2$, with $G_1 = 0$, and $\mathsf{Rank}(D, G_2) = p$. Let $(\hat{\theta}_{1n}, \hat{t}_{2n})'$ be a $\sqrt{n}$-regular estimator of $(\theta_{\star 1}, t_2(\theta_{\star 2}))$ at $f^2_\star = f^2(\theta_\star, u_\star, \cdot)$ with limit distribution S under $f^2_\star$. Then*

$$S \stackrel{d}{=} Z_{\star(2)} + W, \tag{21}$$

where $Z_{\star(2)} \sim N\left(0, I^{-1}_{\star(2)}\right)$ independent of W, and

$$I_{\star(2)} = \begin{pmatrix} D' V_\star^{-1} D & \frac{1}{2} D' V_\star^{-1} G_2 \\ \frac{1}{2} G_2' V_\star^{-1} D & \frac{1}{4} G_2' V_\star^{-1} G_2 \end{pmatrix}.$$

Proof. □

If $\hat{\theta}_n$ is an estimator of $\theta_\star$ that is such that $\kappa_2(\hat{\theta}_n)$ is $\sqrt{n}$-regular of $\kappa_2(\theta_\star)$, this corollary suggests that $Z_{\star(2)}$ is the Gaussian that best approximates $\sqrt{n}\kappa_2(\hat{\theta}_n)$ asymptotically. Note some similarities between this optimal Gaussian approximation and that obtained in the standard case for $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ as studied in Section 2. In particular, the asymptotic variance of the latter is $(\Gamma' V_\star^{-1} \Gamma)^{-1}$ while the variance of the former can be written $\left(\Gamma_{(2)}' V_\star^{-1} \Gamma_{(2)}\right)^{-1}$; where $\Gamma_{(2)} = \begin{pmatrix} D & \frac{1}{2} G_2 \end{pmatrix}$ is of the same form as Γ except for its last column which replaces the first derivative of the moment function in the direction of θ_2 by half of its second derivatives (see Theorem 2.1).

However, as discussed following Theorem 4.1, approximating $\sqrt{n}\kappa_2(\hat{\theta}_n)$ by a Gaussian variate can only lead to a poor approximation because of the non negativeness of the last component making $Z_{\star(2)}$ a *naive* approximation. A better approximation that uses the information on the support of the last component is given by $\tilde{Z}_{\star(2)}$ which we define similarly to $\tilde{Z}_\star$ in Equation (18) using $I_{\star(2)}$ as given in the corollary.

We now consider the case where $G_1 = G_2 = 0$, and θ is identifies only at the third order.

Corollary 5.2. *Suppose that Assumption 1 holds with some $r \geq 2$, with $G_1 = 0$, $G_2 = 0$, and $\mathsf{Rank}(D, G_3) = p$. Let $(\hat{\theta}_{1n}, \hat{t}_{3n})'$ be a $\sqrt{n}$-regular estimator of $(\theta_{\star 1}, t_3(\theta_{\star 2}))$ at $f^2_\star = f^2(\theta_\star, u_\star, \cdot)$ with limit distribution S under $f^2_\star$. Then*

$$S \overset{d}{=} Z_{\star(3)} + W, \tag{22}$$

where $Z_{\star(3)} \sim N\left(0, I^{-1}_{\star(3)}\right)$ independent of W, and

$$I_{\star(3)} = \begin{pmatrix} D'V_\star^{-1}D & \frac{1}{6}D'V_\star^{-1}G_3 \\ \frac{1}{6}G_3'V_\star^{-1}D & \frac{1}{36}G_3'V_\star^{-1}G_3 \end{pmatrix}.$$

Proof. □

From this result, if $\hat{\theta}_n$ is an estimator of θ such that $\kappa_3(\hat{\theta}_n)$ is a $\sqrt{n}$-regular estimator of $\kappa_3(\theta_\star)$, the best Gaussian asymptotic approximation of $\sqrt{n}\kappa_3(\hat{\theta}_n)$ is $Z_{\star(3)}$. Note once again the similarity between the variance of $Z_{\star(3)}$ given by $\left(\Gamma_{(3)}'V_\star^{-1}\Gamma_{(3)}\right)^{-1}$, with $\Gamma_{(3)} = \begin{pmatrix} D & \frac{1}{6}G_3 \end{pmatrix}$ and that of the best Gaussian approximation of $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ in the standard setting. Unlike the previous result, there is no support restriction for $\sqrt{n}\kappa_3(\hat{\theta}_n)$ so that a Gaussian approximation is admissible.

5.2. **Efficiency of the GMM estimator.** In this section, we show that the GMM estimator is asymptotically efficient if the sequence of weighting matrices V_n converges in probability to $V = V_\star^{-1}$. As we have reviewed in Section 2 confirming the work of Chamberlain ([8]), the GMM estimator using such sequence of weighting matrices, the so-called efficient GMM, is efficient in standard models i.e. those without first-order local identification issue. Our findings suggest that the efficiency property of the efficient GMM is immune to local identification issues in the sense that the function $\kappa_j(\theta)$ of the parameter is efficiently estimated by that function of the GMM estimator.

Let $\hat{\theta}$ be the GMM estimator as defined by (2) and consider the same parameter partition $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{p-1} \times \mathbb{R}$ as in Section 5.1. Let $\theta_\star$ be the unique parameter value that solves (1). Assume further that the moment condition function is sufficiently smooth around $\theta_\star$ and $G_1 = 0$ while D is full column-rank $p-1$ so that the model is first order locally non identified. We will distinguish two cases of local identification patterns:

(i) Second-order identification[5]: the matrix $\begin{pmatrix} D & G_2 \end{pmatrix}$ has full column-rank p,

(ii) Third-order identification: $G_2 = 0$ and the matrix $\begin{pmatrix} D & G_3 \end{pmatrix}$ has full column-rank p.

[5]The reader can refer to Dovonon and Renault (2009) for a more general specification of the second-order local identification condition and its characterizations.

We show that in case (i), $\sqrt{n}\kappa_2(\hat{\theta})$ is asymptotically distributed as $\tilde{Z}_{\star(2)}$ as introduced in (18) with $\tilde{Z}_{\star(2)}$ given by Corollary 5.1 whereas in case (ii), $\sqrt{n}\kappa_3(\hat{\theta})$ is asymptotically distributed as $Z_{\star(3)}$ given by Corollary 5.2.

The asymptotic distribution of $\sqrt{n}\kappa_2(\hat{\theta})$ has been derived by Dovonon and Hall ([13]) under conditions including (i).

Letting the probability limit V of V_n be equal to $V_\star^{-1}$, and $M = I_q - V_\star^{-1/2} D(D'V_\star^{-1}D)^{-1}D'V_\star^{-1/2}$, they establish that:

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_1 - \theta_{\star 1} \\ (\hat{\theta}_2 - \theta_{\star 2})^2 \end{pmatrix} \xrightarrow{d} \mathbb{S}_{(2)} = \begin{pmatrix} C\tilde{\mathbb{Z}}_0 + CV_\star^{-1/2} G_2 \mathbb{V}/2 \\ \mathbb{V} \end{pmatrix}, \tag{23}$$

with

$$C = -(D'V_\star^{-1}D)^{-1}D'V_\star^{-1/2}, \quad \mathbb{V} = -2\frac{\mathbb{Z}\mathbb{I}(\mathbb{Z}<0)}{\sigma_{G_2}}, \quad \sigma_{G_2} = G_2'V_\star^{-1/2}MV_\star^{-1/2}G_2,$$

$$\mathbb{Z} = G_2'V_\star^{-1/2}M\tilde{\mathbb{Z}}_0, \qquad \tilde{\mathbb{Z}}_0 \sim N(0, I_q),$$

and $\mathbb{I}(\cdot)$ is the usual indicator function.

The asymptotic distribution of $\sqrt{n}\kappa_3(\hat{\theta})$ under (ii) is given by Theorem A.1 in Appendix. Again, letting the probability limit V of V_n be equal to $V_\star^{-1}$, we obtain that:

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_1 - \theta_{\star 1} \\ (\hat{\theta}_2 - \theta_{0,2})^3 \end{pmatrix} \xrightarrow{d} \mathbb{S}_{(3)} = \Delta\tilde{\mathbb{Z}}_0,$$

with:

$$\Delta = \begin{pmatrix} -(D'V_\star^{-1}D)^{-1}D'V_\star^{-1/2}\left(I_q - V_\star^{-1/2}G_3G_3'V_\star^{-1/2}M/\sigma_{G_3}\right) \\ -6G_3'V_\star^{-1/2}M/\sigma_{G_3} \end{pmatrix}, \text{ and } \sigma_{G_3} = G_3'V_\star^{-1/2}MV_\star^{-1/2}G_3.$$

The next result establishes the asymptotic efficiency of the GMM estimator using weighting matrix with probability limit $V_\star^{-1}$ in the context of local identification patterns (i) and (ii):

Proposition 5.1. *(a)* $\mathbb{S}_{(2)} \overset{d}{=} \tilde{Z}_{\star(2)}$ *and (a)* $\mathbb{S}_{(3)} \overset{d}{=} Z_{\star(3)}$.

Part (a) of Proposition 5.1 shows in particular that under Assumption 1, the local identification pattern (i) and other regularity conditions[6] that guarantee that $\sqrt{n}\kappa_2(\hat{\theta})$ converges in distribution, the GMM estimator defined by (2) is asymptotically efficient in the sense that $\sqrt{n}\kappa_2(\hat{\theta})$ reaches the semiparametric efficiency bound derived in the previous section.

Whereas Part (b) shows that under Assumption 1, the local identification pattern (ii) and the assumptions of Theorem A.1, the GMM estimator defined by (2) is asymptotically efficient in the sense that $\sqrt{n}\kappa_3(\hat{\theta})$ reaches the semiparametric efficiency bound derived in the previous section.

[6]We refer to Dovonon and Hall (2015, Theorem 1) for an explicit account of these conditions.

6. Concluding remarks

We have developed in this paper an efficiency theory in semiparametric models where the score function is degenerate at the true value. To avoid cumbersome technical details, we have focused on the case where the degeneracy occurs in only one direction of the parameter space (θ_2), and partial derivatives of the root density up to order ℓ ($\ell = 2, 3$) is needed to form a non-degenerate pseudo-score function at the true value. In this setting, we have shown that the question of efficient estimation is well-posed if one focuses on the quantity $\kappa_\ell(\theta) = (\theta_1 - \theta_{\star 1}, (\theta_2 - \theta_{\star 2})^\ell)$, and we have derived the corresponding asymptotic efficiency bound. The case where $\ell = 2$ has raised an interesting phenomenon whereby the semiparametric bound produced by the convolution theorem of the model can in fact be improved by utilizing the support information of the parameter. In such cases, and using a projection argument, we have proposed a new efficiency bound, that differs from the variance in the convolution theorem, and appropriately accounts for the support information. We have then proceeded to apply these results to under-identified moment condition models. For such models, we have shown that when the weighting matrix is set to $V_\star^{-1}$, the GMM estimator $\hat{\theta}$ is optimal in the sense that $\sqrt{n}\kappa_\ell(\hat{\theta})$ (for $\ell = 2$ or 3) converges to a distribution with a covariance matrix given by the proposed efficiency bound.

This work tackles the problem of efficient estimation in statistical models with degenerate Fisher information. One interesting direction for future work is further exploration of the case $\ell = 2$, in particular whether is it possible to incorporate the parameter restrictions directly in a convolution theorems, as opposed to the projection argument used in this work. Another possible direction for future work is the extension of the results of this paper to more general pattern of degeneracy of the score function.

Appendix A. Asymptotic distribution of GMM under third-order identification

We let the GMM estimator $\hat{\theta}$ be defined as in (2) and consider the parameter partition $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{p-1} \times \mathbb{R}$. We derive the asymptotic distribution of $\hat{\theta}$ under third-order local identification maintaining the following assumptions:

Assumption 6. (a) *The data sample is given by* $\{x_i : i = 1, \ldots, n\}$*, a sequence of i.i.d. random variables with values in* $\mathbb{R}^k$*.*
(b) $\mathbb{E}(\psi(x, \theta)) = 0 \Leftrightarrow \theta = \theta_\star$.
(c) $\theta_\star \in \Theta$ *compact.*
(d) $\psi(x, \theta)$ *is continuous at each* $\theta \in \Theta$ *with probability one.*
(e) $\mathbb{E}\left(\sup_{\theta \in \Theta} \|\psi(x, \theta)\|\right) < \infty$.

Assumption 7. (a) $\mathbb{E}\left(\frac{\partial \psi}{\partial \theta_2}(x, \theta_\star)\right) = 0, \quad \mathbb{E}\left(\frac{\partial^2 \psi}{\partial \theta_2^2}(x, \theta_\star)\right) = 0.$
(b) *Rank* $\begin{pmatrix} D & G_3 \end{pmatrix} = p$*, with* $D = \mathbb{E}\left(\frac{\partial \psi}{\partial \theta_1'}(\theta_\star, x)\right)$ *and* $G_3 = \mathbb{E}\left(\frac{\partial^3 \psi}{\partial \theta_2^3}(x, \theta_\star)\right)$.

Assumption 8. (a) $\psi(x, \theta)$ *is three time continuously differentiable in a neighborhood* $\mathcal{N}$ *of* $\theta_\star$ *with probability approaching one.*
(b) $\mathbb{E}\left\{\max\left(\sup_{\theta \in \mathcal{N}} \left\|\frac{\partial \psi}{\partial \theta_1'}(x, \theta)\right\|, \sup_{\theta \in \mathcal{N}} \left\|\frac{\partial^2 \psi}{\partial \theta \partial \theta'}(x, \theta)\right\|, \sup_{\theta \in \mathcal{N}} \left\|\frac{\partial^3 \psi}{\partial \theta_2^3}(x, \theta)\right\|\right)\right\} < \infty.$
(c) $\sqrt{n}\bar{\psi}(\theta_\star) \xrightarrow{d} \mathbb{Z}_0, \quad$ *with* $\quad \mathbb{Z}_0 \sim N(0, V_\star)$.

(d) $V_n = V + o_P(1)$, *where* V *is a symmetric positive definite matrix,*
$\frac{\partial \bar{\psi}}{\partial \theta_2}(\theta_\star) = O_P(n^{-1/2})$, $\frac{\partial^2 \bar{\psi}}{\partial \theta_2^2}(\theta_\star) = O_P(n^{-1/2})$.

We have the following result; the proof of which can be found in the proofs' section of this appendix.

Theorem A.1. *If Assumptions 6, 7, 8 hold and* $\theta_\star$ *is interior to the parameter set* Θ*, then:*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_1 - \theta_{\star 1} \\ (\hat{\theta}_2 - \theta_{\star 2})^3 \end{pmatrix} \xrightarrow{d} \Delta V^{1/2}\mathbb{Z}_0,$$

with

$$\Delta = \begin{pmatrix} -(D'VD)^{-1}D'V^{1/2}\left(I_q - V^{1/2}G_3G_3'V^{1/2}M/\sigma_{G_3}\right) \\ -6G_3'V^{1/2}M/\sigma_{G_3} \end{pmatrix},\ M = I_q - V^{1/2}D(D'VD)^{-1}D'V^{1/2}$$

and $\sigma_{G_3} = G_3'V^{1/2}MV^{1/2}G_3$.

Appendix B. Proofs

Proof of Lemma 2.2 Let $u_\star = 0_{L^2(P_\star)}$ and $f_\star = 1$. We have

$$\mathcal{M}(\theta_\star, u_\star, f_\star) = 0$$

and

$$\nabla_f \mathcal{M}(\theta_\star, u_\star, f_\star)\cdot h = \left\langle h, \bar{\phi}^0\right\rangle \bar{\phi}^0 + \sum_{j\geq q+2}\langle \phi_j, h\rangle \phi_j,$$

$h \in L^2(P_\star)$ which is an isomorphism of $L^2(P_\star)$. Therefore, by the implicit function theorem, there exists a class C^r function $f : \mathcal{V} \to \mathcal{U}$ defined on some neighborhood $\mathcal{V}$ of $(\theta_\star, u_\star)$ to some neighborhood $\mathcal{U}$ of $f_\star$ such that $f(\theta_\star, u_\star, \cdot) = f_\star(\cdot)$ and for all $(\theta, u) \in \mathcal{V}$,

$$\mathcal{M}(\theta, u, f(\theta, u, \cdot)) = 0.$$

In particular, for all $(\theta, u) \in \mathcal{V}$,

$$\int \psi'(\theta, x) f^2(\theta, u, x) P_\star(dx) V_\star^{-1/2} = 0, \ \textit{and}\ \int f^2(\theta, u, x) P_\star(dx) = 1.$$

The first result follows since $V_\star$ is nonsingular.

The derivatives of the functional $f(\theta, u, \cdot)$ are obtained applying the usual formulas:

$$\nabla_u f(\theta, u, \cdot)\cdot h = -\left(\nabla_f \mathcal{M}(\theta, u, f)\right)^{-1} \circ \left(\nabla_u \mathcal{M}(\theta, u, f)\cdot h\right), \quad \forall h \in \mathcal{E}$$

and

$$\nabla_\theta f(\theta, u, \cdot)\cdot w = -\left(\nabla_f \mathcal{M}(\theta, u, f)\right)^{-1} \circ \left(\frac{\partial \mathcal{M}}{\partial \theta'}(\theta, u, f)\cdot w\right), \quad \forall w \in \mathbb{R}^p.$$

To obtain explicit formulas we first derive the expression of $\left(\nabla_f \mathcal{M}(\theta, u, f)\right)^{-1}$. Notice that $\left\langle \bar{\phi}^0_{\theta_\star}, f_{\theta_\star, u_\star}\bar{\phi}^0\right\rangle = I_{q+1}$. Hence by the inverse application theorem, there exists a neighborhood of $(\theta_\star, u_\star)$ that we take without any loss of generality as $\mathcal{V}$ such that the matrix $\left\langle \bar{\phi}^0_\theta, f_{\theta,u}\bar{\phi}^0\right\rangle$ is also invertible for all $(\theta, u) \in \mathcal{V}$.

Now, for $h = \sum_{j\geq 1} a_i\phi_i$, suppose that $\nabla \mathcal{M}_f(\theta, u, f)\cdot h = v = \left\langle fh, \bar{\phi}^0_\theta\right\rangle \bar{\phi}^0 + \sum_{j\geq q+1}\langle \phi_j, h\rangle \phi_j$. Then for $j \geq q+2$, $a_j = \langle h, \phi_j\rangle = \langle v, \phi_j\rangle$, and hence

$$v = \left\langle f\sum_{j=1}^{q+1} a_j\phi_j, \bar{\phi}^0_\theta\right\rangle \bar{\phi}^0 + \left\langle f\sum_{j\geq q+2} a_j\phi_j, \bar{\phi}^0_\theta\right\rangle \bar{\phi}^0 + \sum_{j\geq q+2}\langle u, \phi_j\rangle \phi_j.$$

Setting $A = (a_{q+1}, a_1, \ldots, a_q)'$, this translates into

$$\left\langle \bar{\phi}^0_\theta, f\bar{\phi}^0\right\rangle A = \left\langle \bar{\phi}^0, v\right\rangle - \left\langle \bar{\phi}^0_\theta, f\sum_{j\geq q+2}\langle v, \phi_j\rangle \phi_j\right\rangle.$$

Hence

$$\left(\nabla_f \mathcal{M}(\theta, u, f)\right)^{-1}\cdot v = \left(\left\langle v, \bar{\phi}^0\right\rangle - \left\langle f\sum_{j\geq q+2}\langle v, \phi_j\rangle \phi_j, \bar{\phi}^0_\theta\right\rangle\right)\left\langle f\bar{\phi}^0, \bar{\phi}^0_\theta\right\rangle^{-1}\bar{\phi}^0 + \sum_{j\geq q+2}\langle v, \phi_j\rangle \phi_j.$$

Using the expression above we obtain:

$$\nabla_u f(\theta,u,\cdot)\cdot h = h - \left\langle f_{\theta,u}h, \bar{\phi}^0_\theta\right\rangle \left\langle f_{\theta,u}\bar{\phi}^0, \bar{\phi}^0_\theta\right\rangle^{-1}\bar{\phi}^0, \quad \forall h\in\mathcal{E}$$

$$\nabla_\theta f_{\theta,u}\cdot w = -\frac{1}{2}w'\left\langle f^2_{\theta,u}, \nabla_\theta\bar{\phi}^0_\theta\right\rangle \left\langle f_{\theta,u}\bar{\phi}^0, \bar{\phi}^0_\theta\right\rangle^{-1}\bar{\phi}^0.$$

□

Proof of Theorem 2.1 We verify that the implicit parametric model $\{f(\theta,u):(\theta,u)\in\mathcal{V}\}$ induced by the moment condition model (1) satisfies the conditions of Theorem 3.1 of BHHW and the conclusion follows readily. These conditions are the following:

(1) The set

$$\mathcal{B}(u_\star)=\left\{\beta\in\mathcal{E}:\|\sqrt{n}(u_n-u_\star)-\beta\|_{L^2(P_\star)}\to 0,\ as\ n\to\infty, \text{for some sequence } (u_n) \text{ with all } u_n\in\mathcal{E}\right\}$$

is a subspace of $\mathcal{E}$.

(2) Let $\rho_{\theta_\star}=\nabla_\theta f(\theta_\star,u_\star,\cdot)$ and $A=\nabla_u f(\theta_\star,u_\star,\cdot)$.

(2.i) $\rho_{\theta_\star}\in(L^2(P_\star))^p$,

(2.ii) A is a bounded operator from $\mathcal{E}$ to $L^2(P_\star)$,

(2.iii) Hellinger differentiability of f at $(\theta_\star,u_\star)$:

$$\frac{\|f(\theta_n,u_n)-f(\theta_\star,u_\star)-\{\rho'_{\theta_\star}(\theta_n-\theta_\star)+A(u_n-u_\star)\}\|_{L^2(P_\star)}}{\|\theta_n-\theta_\star\|+\|u_n-u_\star\|_{L^2(P_\star)}}\to 0\ as\ n\to\infty$$

for all sequences $\theta_n\to\theta_\star$ and $u_n\to u_\star$ in $L^2(P_\star)$, where $u_n\in\mathcal{E}$ for all $n\geq 1$.

Let us check these conditions:

(1) Choosing $u_n=u_\star$ for all $n\geq 1$, we can see that $u_\star\in\mathcal{B}(u_\star)$ which is nonempty as it contains $0_{L^2(P_\star)}$. Let $\beta_1,\beta_2\in\mathcal{B}(u_\star)$. Then there exists two sequences $(u_{1n})_{n\geq1}$ and $(u_{2n})_{n\geq1}$ all in $\mathcal{E}$ such that $\sqrt{n}(u_{in}-u_\star)-\beta_i\to 0$ in $L^2(P_\star)$ as $n\to\infty$ $(i=1,2)$. For any $\alpha_1,\alpha_2\in\mathbb{R}$, by the triangle inequality, we can see that $\sqrt{n}(\alpha_1u_{1n}+\alpha_2u_{2n}-u_\star)-(\alpha_1\beta_1+\alpha_2\beta_2)\to 0$ in $L^2(P_\star)$, recalling that $u_\star=0_{L^2(P_\star)}$.

(2.i) From Lemma 2.2, $\rho_{\theta_\star}(\cdot)=-\frac{1}{2}\Gamma' V_\star^{-1/2}\phi^0(\cdot)$. Thus (2.i) is satisfied thanks to the condition (1.2) of Assumption 1.

(2.ii) From Lemma 2.2, it is not hard to see that $A\equiv\nabla_u f(\theta_\star,u_\star,\cdot)=Id_{\mathcal{E}}$ which is a linear continuous map from $\mathcal{E}$ in $L^2(P_\star)$. As such, A is a bounded operator.

(2.iii) Follows immediately from the fact that $(\theta,u)\mapsto f(\theta,u)$ is Frechet-differentiable at $(\theta_\star,u_\star)$.

We can now apply Theorem 3.1 of BHHW and deduce (5) with $I_{\star(1)}$ given by:

$$I_{\star(1)}=4\int(\rho_{\theta_\star}(x)-(A\beta_\star)(x))(\rho_{\theta_\star}(x)-(A\beta_\star)(x))'P_\star(dx),$$

where $\beta_\star\in\mathcal{B}(u_\star):\rho_{\theta_\star}-A\beta_\star\perp A\beta$ for all $\beta\in\mathcal{B}(u_\star)$. But, since $A=Id_{\mathcal{E}}$, $\forall\beta\in\mathcal{E}$, $A\beta=\sum_{j=q+2}^{\infty}\langle\phi_j,\beta\rangle\phi_j$. Hence, $\forall\beta\in\mathcal{B}(u_\star)$, $A\beta\perp\rho_{\theta_\star}$. As a result, $\beta_\star=0$. Thus

$$I_{\star(1)}=4\int\rho_{\theta_\star}(x)\rho'_{\theta_\star}(x)P_\star(dx)=\Gamma' V_\star^{-1}\Gamma.$$

□

Proof of Lemma 5.1. We have established in Lemma 2.2 that under Assumption 1,

$$\nabla_\theta f(\theta,u,\cdot)\cdot w=-\frac{1}{2}w'\left\langle f^2_{\theta,u},\nabla_\theta\bar{\phi}^0_\theta\right\rangle\left\langle f_{\theta,u}\bar{\phi}^0,\bar{\phi}^0_\theta\right\rangle^{-1}\bar{\phi}^0.$$

This gives Part (1). Straightforward differentiation implies that $\forall w_1,w_2\in\mathbb{R}^p$, $(\theta,u)\in\mathcal{V}$,

$$\begin{aligned}\nabla^{(2)}_\theta f_{\theta,u}\cdot(w_1,w_2)=&-\frac{1}{2}w_1'\left[\left\langle f^2_{\theta,u},\nabla^{(2)}_\theta\bar{\phi}^0_\theta\cdot w_2\right\rangle+2\left\langle f_{\theta,u}\left(\nabla_\theta f_{\theta,u}\cdot w_2\right),\nabla_\theta\bar{\phi}^0_\theta\right\rangle\right]\left\langle f_{\theta,u}\bar{\phi}^0,\bar{\phi}^0_\theta\right\rangle^{-1}\bar{\phi}^0\\&+\frac{1}{2}w_1'\left\langle f^2_{\theta,u},\nabla_\theta\bar{\phi}^0_\theta\right\rangle\left\langle f_{\theta,u}\bar{\phi}^0,\bar{\phi}^0_\theta\right\rangle^{-1}\left[\left\langle\left(\nabla_\theta f_{\theta,u}\cdot w_2\right)\bar{\phi}^0,\bar{\phi}^0_\theta\right\rangle+\left\langle f_{\theta,u}\bar{\phi}^0,\nabla_\theta\bar{\phi}^0_\theta\cdot w_2\right\rangle\right]\left\langle f_{\theta,u}\bar{\phi}^0,\bar{\phi}^0_\theta\right\rangle^{-1}\bar{\phi}^0.\end{aligned}$$

This readily gives Part (2). Part (3) follows similarly. □

Proof of Lemma 4.1. Let $\eta_2 \in \mathbb{R}$ and $\{\theta_{2n}\}$ a sequence of real numbers such that: $\varepsilon_n \equiv \sqrt{n}(\theta_{2n}-\theta_{\star 2})-\eta_2 \to 0$ as $n \to \infty$. Let $f_n = f(\theta_{2n},\cdot)$ and $f_\star = f(\theta_{\star 2},\cdot)$. By the Hellinger differentiability of f at $\theta_{\star 2}$, we can apply Proposition 2.1 of BHHW (without the nonparametric component) and obtain that

$$\sqrt{n}(f_n - f_\star) \to 0,$$

in $L^2(\mu)$ since $\rho_{\theta_2} = 0$ at $\theta_{\star 2}$. We can therefore deduce from their Lemma 2.1 that:

$\forall \eta_2 \in \mathbb{R}, L_n = \log\{\prod_{i=1}^n [f_n^2(X_i)/f_\star^2(X_i)]\} \to 0$, in probability (both under f_n and $f_\star$) since α (in this lemma) is equal to 0 here.

By the regularity assumption, we have:

$$\left(\sqrt{n}(\hat\theta_{2n}-\theta_{2n}), L_n\right) \xrightarrow{d_{f_n}} (S,0) \quad \textit{and} \quad \left(\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}), L_n\right) \xrightarrow{d_{f_\star}} (S,0).$$

Let us consider the characteristic function of $\sqrt{n}(\hat\theta_{2n}-\theta_{2n})$ at $w \in \mathbb{R}$. We have:

$$\begin{aligned}
\mathbb{E}_{f_n} \exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{2n})) &= \mathbb{E}_{f_n} \exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}) - iw(\varepsilon_n+\eta_2)) \\
&= \mathbb{E}_{f_n} \exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}) - iw\eta_2) + o(1) \\
&= \mathbb{E}_{f_\star} \exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}) + L_n - iw\eta_2) + o(1).
\end{aligned}$$

By the almost sure representation theorem, $\left(\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}), L_n\right) \to (S,0)$, almost surely in some probability space. Hence, $\exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}) + L_n - iw\eta_2)$ converges almost surely to $\exp(iwS)\exp(-iw\eta_2)$ in that space. The fact that $\mathbb{E}_{f_\star}|\exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}) + L_n - iw\eta_2)| = \mathbb{E}_{f_\star}\exp(L_n) = 1 = \mathbb{E}|\exp(iwS)\exp(-iw\eta_2)|$ guarantees uniform integrability. Hence,

$$\mathbb{E}_{f_\star} \exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{\star 2}) + L_n - iw\eta_2) \to \mathbb{E}\left(\exp(iwS)\exp(-iw\eta_2)\right),$$

while

$$\mathbb{E}_{f_n} \exp(iw\sqrt{n}(\hat\theta_{2n}-\theta_{2n})) \to \mathbb{E}\exp(iwS).$$

Hence,

$$\mathbb{E}\exp(iwS) = \exp(-iw\eta_2)\mathbb{E}\exp(iuS) : \forall w, \eta_2 \in \mathbb{R}.$$

Thus $\forall w, \eta_2 \in \mathbb{R}, \exp(-iw\eta_2) = 1$ which establishes the contradiction. □

Proof of Theorem 4.1. This is essentially an adaptation of the proof of Theorem 3.1 of BHHW. Let $S_n = \sqrt{n}\begin{pmatrix} \hat\theta_{1n}-\theta_{1n} \\ \hat t_{2n} - t_2(\theta_{2n}) \end{pmatrix}$. The characteristic function of S_n under f_n is:

$$\begin{aligned}
\mathbb{E}_{f_n}\left(\exp\{iw'S_n\}\right) &= \mathbb{E}_{f_n}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{1n}) + iw_2\sqrt{n}(\hat t_{2n}-t_2(\theta_{2n}))\right)\right) \\
&= \mathbb{E}_{f_n}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{\star 1}+\theta_{\star 1}-\theta_{1n}) + iw_2\sqrt{n}(\hat t_{2n}-t_2(\theta_{2n}))\right)\right) \\
&= \mathbb{E}_{f_n}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{\star 1}) - iw_1\sqrt{n}(\theta_{1n}-\theta_{\star 1})) + iw_2\sqrt{n}(\hat t_{2n}-(\theta_{2n}-\theta_{\star 2})^2)\right)\right).
\end{aligned}$$

But, $\theta_{1n} = \theta_{\star 1} + (\eta_1+\varepsilon_{1n})/\sqrt{n}$ and $\theta_{2n} = \theta_{\star 2} + (\eta_2+\varepsilon_{2n})/n^{1/4}$ with ε_{1n}, $\varepsilon_{2n} \to 0$ as $n \to \infty$. Then,

$$\begin{aligned}
\mathbb{E}_{f_n}\left(\exp\{iw'S_n\}\right) &= \mathbb{E}_{f_n}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{\star 1}) - iw_1(\varepsilon_{1n}+\eta_1)\right) \times \exp\left(iw_2\sqrt{n}\hat t_{2n} - iw_2(\varepsilon_{2n}+\eta_2)^2\right)\right) \\
&= \mathbb{E}_{f_n}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{\star 1}) - iw_1\eta_1\right) \times \exp\left(iw_2\sqrt{n}\hat t_{2n} - iw_2\eta_2^2\right)\right) + o(1) \\
&= \mathbb{E}_{f_n}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{\star 1}) + iw_2\sqrt{n}\hat t_{2n} - iw_1\eta_1 - iw_2\eta_2^2\right)\right) + o(1) \\
&= \mathbb{E}_{f_\star}\left(\exp\left(iw_1\sqrt{n}(\hat\theta_{1n}-\theta_{\star 1}) + iw_2\sqrt{n}\hat t_{2n} - iw_1\eta_1 - iw_2\eta_2^2 + L_n\right)\right) + o(1).
\end{aligned}$$

This holds for any $\alpha \in H_2(\theta_\star, u_\star)$. We choose α defined with η_1 and η_2 considered so far but with β free. Using the fact that $\mathcal{B}_1(u_\star)$ is a space, we can write $\alpha = \eta_1(\rho_{\theta_1} - A\beta_1) + \eta_2^2(\frac{1}{2}\rho_{\theta_2\theta_2} - A\beta_2) \equiv \eta_1\alpha_1(\beta_1) + \eta_2^2\alpha_2(\beta_2)$, with

$\beta_1, \beta_2 \in \mathcal{B}_1(u_\star)$. Let

$$I(\beta_1,\beta_2) = 4\begin{pmatrix} \|\alpha_1(\beta_1)\|_\mu^2 & \langle \alpha_1(\beta_1), \alpha_2(\beta_2)\rangle_\mu \\ \\ \langle \alpha_1(\beta_1), \alpha_2(\beta_2)\rangle_\mu & \|\alpha_2(\beta_2)\|_\mu^2 \end{pmatrix}.$$

From Lemma 4.2 and under $f_\star = f(\theta_\star, u_\star, \cdot)$, the random vector $\Big(\sqrt{n}(\hat{\theta}_{1n} - \theta_{\star 1}), \sqrt{n}\hat{t}_{2n}, 2n^{-1/2}\sum_{\iota=1}^n \alpha(X_\iota)/f_\star(X_\iota)\Big)$ converges weakly coordinate-wise to $(S_1, S_2, \delta' Z)$; $Z \sim N(0, I(\beta_1,\beta_2))$, $\delta' = (\eta_1 \ \eta_2^2)$.

By Prohorov's theorem, there is a subsequence of $\Big(\sqrt{n}(\hat{\theta}_{1n} - \theta_{\star 1}), \sqrt{n}\hat{t}_{2n}, L_n\Big)$ that converges weakly under $f_\star$ to $\big(S_1, S_2, \delta' Z - \frac{1}{2}\delta' I(\beta_1,\beta_2)\delta\big)$.

By the regularity assumption, the characteristic function

$$\mathbb{E}_{f_n}\Big(\exp iw_1\sqrt{n}(\hat{\theta}_{1n} - \theta_{1n}) + iw_2\sqrt{n}(\hat{t}_{2n} - t_2(\theta_{2n}))\Big)$$

converges to $\mathbb{E}\exp(i(w_1 S_1 + w_2 S_2))$. By similar argument as in BHHW, we can claim that:

$$\mathbb{E}_{f_\star}\exp\Big(iw_1\sqrt{n}(\hat{\theta}_{1n} - \theta_{\star 1}) + iw_2\sqrt{n}\hat{t}_{2n} - iw_1\eta_1 - iw_2\eta_2^2 + L_n\Big)$$

converges to

$$\mathbb{E}\exp\left(iw_1 S_1 + iw_2 S_2 + \delta' Z\right)\exp\left(-\frac{1}{2}\delta' I(\beta_1,\beta_2)\delta - iw_1\eta_1 - iw_2\eta_2^2\right).$$

Letting $S = (S_1, S_2)'$ and $\phi_1(v,w) = \mathbb{E}\exp(iv'S + iw'Z)$, we have:

$$\phi_1(w,0) = \mathbb{E}\exp\left(iw_1 S_1 + iw_2 S_2 + \delta' Z\right)\exp\left(-\frac{1}{2}\delta' I(\beta_1,\beta_2)\delta - iw_1\eta_1 - iw_2\eta_2^2\right).$$

Since the right hand side of this equation is analytic in (η_1,η_2) and constant for each $(\eta_1,\eta_2) \in \mathbb{R}^2$, it is also constant for all (η_1,η_2) complex. The choice $\begin{pmatrix} \eta_1 \\ \eta_2^2 \end{pmatrix} = -iI^{-1}(\beta_1,\beta_2)\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ yields:

$$\phi_1(w,0) = \mathbb{E}\exp iw'\left(S - I^{-1}(\beta_1,\beta_2)Z\right) \times \exp\left(-\frac{1}{2}w' I^{-1}(\beta_1,\beta_2)w\right).$$

This is a factorization into the characteristic function of $W = S - Z_0$ and Z_0 with W and Z_0 independent and $Z_0 = I^{-1}(\beta_1,\beta_2)Z \sim N(0, I^{-1}(\beta_1,\beta_2))$. This conclusion is true for any $\beta_1, \beta_2 \in \mathcal{B}_1(u_\star)$.

The relevant bound is obtained by choosing β_1, β_2 so that $I^{-1}(\beta_1,\beta_2)$ is maximum. We now show that this maximum is reached at $(\beta_1^\star, \beta_2^\star)$ defined by:

$$\beta_1^\star = \arg\min_\beta \langle \rho_{\theta_1} - A\beta, \rho_{\theta_1} - A\beta\rangle_\mu \quad \textit{and} \quad \beta_2^\star = \arg\min_\beta \left\langle \frac{1}{2}\rho_{\theta_2\theta_2} - A\beta, \frac{1}{2}\rho_{\theta_2\theta_2} - A\beta \right\rangle_\mu.$$

For this, we show that

$$\forall w \in \mathbb{R}^2, \quad w'(I(\beta_1^\star,\beta_2^\star) - I(\beta_1,\beta_2))w \le 0.$$

First, note that, for all $\beta \in \mathcal{B}_1(u_\star)$, $\langle \rho_{\theta_1} - A\beta_1^\star, A\beta\rangle = 0$ and $\langle \frac{1}{2}\rho_{\theta_2\theta_2} - A\beta_2^\star, A\beta\rangle = 0$. Using this, it is not hard to see that:

$$\forall w \in \mathbb{R}^2, \quad w'(I(\beta_1^\star,\beta_2^\star) - I(\beta_1,\beta_2))w = -\langle w_1 A(\beta_1 - \beta_1^\star) + w_2 A(\beta_2 - \beta_2^\star), w_1 A(\beta_1 - \beta_1^\star) + w_2 A(\beta_2 - \beta_2^\star)\rangle \le 0,$$

and we conclude. □

Proof of Proposition 5.1. (a) We have to show that

$$\mathbb{S}_{(2)} \overset{d}{=} \begin{pmatrix} AZ_{\star 2}\mathbb{I}(Z_{\star 2} \ge 0) + BU \\ Z_{\star 2}\mathbb{I}(Z_{\star 2} \ge 0) \end{pmatrix},$$

with $Z_{\star 2} \sim N(0, I^{22}_{\star(2)})$ independent of $U \sim N(0, I_{p-1})$, $A = \frac{I^{12}_{\star(2)}}{I^{22}_{\star(2)}}$, $B = \left(I^{11}_{\star(2)} - \frac{I^{12}_{\star(2)}\cdot I^{21}_{\star(2)}}{I^{22}_{\star(2)}}\right)^{1/2}$, with $I^{ij}_{\star(2)}$, $i,j = 1,2$ are the entries of $I^{-1}_{\star(2)}$, with

$$I_{\star(2)} = \begin{pmatrix} D'V_\star^{-1}D & \frac{1}{2}D'V_\star^{-1}G_2 \\ \frac{1}{2}G_2'V_\star^{-1}D & \frac{1}{4}G_2'V_\star^{-1}G_2 \end{pmatrix}.$$

Using the formula for the inverse partitioned matrix (see e.g. [26], p.11), we have:

$$\begin{aligned} I^{11}_{\star(2)} &= (D'V_\star^{-1}D)^{-1} + (D'V_\star^{-1}D)^{-1}D'V_\star^{-1}G_2G_2'V_\star^{-1}D(D'V_\star^{-1}D)^{-1}/\sigma_{G_2}, \\ I^{12}_{\star(2)} &= -2(D'V_\star^{-1}D)^{-1}D'V_\star^{-1}G_2/\sigma_{G_2}, \\ I^{21}_{\star(2)} &= (I^{12}_{\star(2)})', \\ I^{22}_{\star(2)} &= 4/\sigma_{G_2}. \end{aligned} \tag{A.1}$$

Considering the distribution of $\mathbb{S}_{(2)}$ given by (23), we can see that: $B\tilde{\mathbb{Z}}_0$ and $\mathbb{Z}$ are linear function of the Gaussian $\tilde{\mathbb{Z}}_0$ with covariance nil. Hence they are independent. As a result, we can claim that:

$$\mathbb{S}_{(2)} \stackrel{d}{=} \begin{pmatrix} CV_\star^{-1/2}G_2\mathbb{V}/2 + (D'V_\star^{-1}D)^{-1/2}\mathbb{U}_0 \\ \mathbb{V} \end{pmatrix},$$

with $\mathbb{U}_0 \sim N(0, I_{p-1})$ independent of $\mathbb{Z}$. Let $U = \mathbb{U}_0$ and $Z_{\star 2} = -\frac{2\mathbb{Z}}{\sigma_{G_2}}$. Then,

$$\mathbb{S} \stackrel{d}{=} \begin{pmatrix} CV_\star^{-1/2}G_2Z_{\star 2}\mathbb{I}(Z_{\star 2} \geq 0) + (D'V_\star^{-1}D)^{-1/2}U \\ Z_{\star 2}\mathbb{I}(Z_{\star 2} \geq 0) \end{pmatrix}.$$

To conclude, it suffices to show that:

$$\mathsf{Var}(Z_{\star 2}) = I^{22}_{\star(2)}, \quad CV_\star^{-1/2}G_2 = 2 \times \frac{I^{12}_{\star(2)}}{I^{22}_{\star(2)}}, \quad \text{and} \quad (D'V_\star^{-1}D)^{-1} = I^{11}_{\star(2)} - I^{12}_{\star(2)} \cdot I^{21}_{\star(2)}/I^{22}_{\star(2)}.$$

This can be done easily using (A.1) and the fact that $\mathsf{Var}(\mathbb{Z}) = \sigma_{G_2}$.

To establish (b), we just have to show that $\Delta\Delta'$ is equal to $I^{-1}_{\star(3)}$, with

$$I_{\star(3)} = \begin{pmatrix} D'V_\star^{-1}D & \frac{1}{6}D'V_\star^{-1}G_3 \\ \frac{1}{6}G_3'V_\star^{-1}D & \frac{1}{36}G_3'V_\star^{-1}G_3 \end{pmatrix}.$$

Let $I^{ij}_{\star(3)}$, $i,j = 1,2$ be the entries of $I^{-1}_{\star(3)}$. Using again the formula for inverse of partitioned matrix, we get after some straightforward calculations, we get:

$$\begin{aligned} I^{11}_{\star(3)} &= (D'V_\star^{-1}D)^{-1} + (D'V_\star^{-1}D)^{-1}D'V_\star^{-1}G_3G_3'V_\star^{-1}D(D'V_\star^{-1}D)^{-1}/\sigma_{G_3}, \\ I^{12}_{\star(3)} &= -6(D'V_\star^{-1}D)^{-1}D'V_\star^{-1}G_3/\sigma_{G_3}, \\ I^{21}_{\star(3)} &= (I^{12}_{\star(3)})', \\ I^{22}_{\star(3)} &= 36/\sigma_{G_3}. \end{aligned}$$

By a straightforward expansion of the terms in $\Delta\Delta'$ and using the fact that $MV_\star^{-1/2}D = 0$, the expected result becomes transparent. □

Proof of Theorem A.1. The consistency of the GMM estimator is established by Newey and McFadden (1994) under Assumption 6. Towards the asymptotic distribution, we first derive the asymptotic order of magnitude of convergence of $\hat{\theta}_1 - \theta_{\star 1}$ and $\hat{\theta}_2 - \theta_{\star 2}$. For this, we do a first order mean-value expansion of $\theta_1 \to \bar{\psi}(\theta_1, \hat{\theta}_2)$ and then a third order expansion of $\theta_2 \to \bar{\psi}(\theta_{\star_1}, \theta_2)$. This gives:

$$\bar{\psi}(\hat{\theta}) = \bar{\psi}(\theta_\star) + \frac{\partial\bar{\psi}}{\partial\theta_1'}(\bar{\theta}_1, \hat{\theta}_2)(\hat{\theta}_1 - \theta_{\star 1}) + \frac{\partial\bar{\psi}}{\partial\theta_2}(\theta_\star)(\hat{\theta}_2 - \theta_{\star 2}) + \frac{1}{2}\frac{\partial^2\bar{\psi}}{\partial\theta_2^2}(\theta_\star)(\hat{\theta}_2 - \theta_{\star 2})^2 + \frac{1}{6}\frac{\partial^3\bar{\psi}}{\partial\theta_2^3}(\theta_{\star 1}, \bar{\theta}_2)(\hat{\theta}_2 - \theta_{\star 2})^3,$$

where $\bar{\theta}_1 \in (\theta_{\star 1}, \hat{\theta}_1)$ and $\bar{\theta}_2 \in (\theta_{\star 2}, \hat{\theta}_2)$ and may differ from row to row. Hence,

$$\bar{\psi}(\hat{\theta}) = \bar{\psi}(\theta_\star) + \bar{D}(\hat{\theta}_1 - \theta_{\star 1}) + \frac{1}{6}\bar{G}_3(\hat{\theta}_2 - \theta_{\star 2})^3 + o_P(n^{-1/2}), \tag{A.2}$$

with $\bar{D} = \frac{\partial \bar{\psi}}{\partial \theta_1'}(\bar{\theta}_1, \hat{\theta}_2)$ and $\bar{G}_3 = \frac{\partial^3 \bar{\psi}}{\partial \theta_2^3}(\theta_{\star 1}, \bar{\theta}_2)$. Pre-multiplying this equality by $\bar{D}'V_n$ and solving in $\hat{\theta}_1 - \theta_{\star 1}$, we have:

$$(\hat{\theta}_1 - \theta_{\star 1}) = (\bar{D}'V_n\bar{D})^{-1}\bar{D}'V_n\left(\bar{\psi}(\hat{\theta}) - \bar{\psi}(\theta_\star) - \frac{1}{6}\bar{G}_3(\hat{\theta}_2 - \theta_{\star 2})^3\right) + o_P(n^{-1/2}). \tag{A.3}$$

Plugging this into (A.2), we get:

$$\bar{\psi}(\hat{\theta}) = \bar{\psi}(\theta_\star) + \bar{D}(\bar{D}'V_n\bar{D})^{-1}\bar{D}'V_n(\bar{\psi}(\hat{\theta}) - \bar{\psi}(\theta_\star)) - \frac{1}{6}V_n^{-1/2}\bar{M}V_n^{1/2}\bar{G}_3(\hat{\theta}_2 - \theta_{\star 2})^3 + o_P(n^{-1/2}),$$

with $\bar{M} = I_q - V_n^{1/2}\bar{D}(\bar{D}'V_n\bar{D})^{-1}\bar{D}'V_n^{1/2}$.

Hence,

$$\bar{\psi}'(\hat{\theta})V_n\bar{\psi}(\hat{\theta}) = \bar{\psi}'(\theta_\star)V_n\bar{\psi}(\theta_\star) + \tfrac{1}{36}\bar{G}_3'V_n^{1/2}\bar{M}V_n^{1/2}\bar{G}_3(\hat{\theta}_2 - \theta_{\star 2})^6 + (\hat{\theta}_2 - \theta_{\star 2})^3 O_P(n^{-1/2}) + O_P(n^{-1}). \tag{A.4}$$

The orders of magnitude in (A.4) follow from the fact that $\bar{M}$ converges in probability to M and therefore is $O_P(1)$ and the fact that $\bar{\psi}(\hat{\theta}) = O_P(n^{-1/2})$. This latter comes from the fact that $\bar{\psi}'(\hat{\theta})V_n\bar{\psi}(\hat{\theta}) \leq \bar{\psi}'(\theta_\star)V_n\bar{\psi}(\theta_\star)$ (by definition of GMM estimator). Since V_n converges in probability to V symmetric positive definite, we can claim that $\bar{\psi}(\hat{\theta}) = O_P(n^{-1/2})$ as it is bounded by $\bar{\psi}(\theta_\star)$ which is $O_P(n^{-1/2})$. Again, by the definition of the GMM estimator, the right hand side of (A.4) is less or equal to $\bar{\psi}'(\theta_\star)V_n\bar{\psi}(\theta_\star)$ and this gives:

$$\frac{1}{36}G_3'V^{1/2}MV^{1/2}G_3 n(\hat{\theta}_2 - \theta_{\star 2})^6 + o_P(1)n(\hat{\theta}_2 - \theta_{\star 2})^6 \leq O_P(1) + \sqrt{n}(\hat{\theta}_2 - \theta_{\star 2})^3 O_P(1) \tag{A.5}$$

Thanks to Assumption 7(b) and the fact that V is nonsingular, $MV^{1/2}G_3 \neq 0$. As a result, $G_3'V^{1/2}MV^{1/2}G_3 \neq 0$ which is sufficient to deduce from (A.5) that $n(\hat{\theta}_2 - \theta_{\star 2})^6 = O_P(1)$; or equivalently that $n^{1/6}(\hat{\theta}_2 - \theta_{\star 2}) = O_P(1)$. We obtain $\hat{\theta}_1 - \theta_{\star 1} = O_P(n^{-1/2})$ from (A.3).

Using these orders of magnitude, we can write that:

$$\bar{\psi}(\hat{\theta}) = \bar{\psi}(\theta_\star) + D(\hat{\theta}_1 - \theta_{\star 1}) + \frac{1}{6}G_3(\hat{\theta}_2 - \theta_{\star 2})^3 + o_P(n^{-1/2}) \tag{A.6}$$

The first order condition for $\hat{\theta}$ in the direction of θ_1 is:

$$\frac{\partial \bar{\psi}'}{\partial \theta_1}V_n\bar{\psi}(\hat{\theta}) = 0.$$

Using (A.6), this implies that

$$D'V\bar{\psi}(\theta_\star) + D'VD(\hat{\theta}_1 - \theta_{\star 1}) + \frac{1}{6}D'VG_3(\hat{\theta}_2 - \theta_{\star 2})^3 = o_P(n^{-1/2}).$$

Therefore, we have

$$(\hat{\theta}_1 - \theta_{\star 1}) = -(D'VD)^{-1}D'V\left(\bar{\psi}(\theta_\star) + \frac{1}{6}G_3(\hat{\theta}_2 - \theta_{\star 2})^3\right) + o_P(n^{-1/2}). \tag{A.7}$$

Plugging this in (A.6), we have:

$$\bar{\psi}(\hat{\theta}) = V^{-1/2}MV^{1/2}\left(\bar{\psi}(\theta_\star) + \frac{1}{6}G_3(\hat{\theta}_2 - \theta_{\star 2})^3\right) + o_P(n^{-1/2}). \tag{A.8}$$

The first order condition for $\hat{\theta}$ in the direction of θ_2 is:

$$\frac{\partial \bar{\psi}'}{\partial \theta_2}(\hat{\theta})V_n\bar{\psi}(\hat{\theta}) = 0.$$

Note that

$$\frac{\partial \bar{\psi}}{\partial \theta_2}(\hat{\theta}) = \frac{\partial \bar{\psi}}{\partial \theta_2}(\theta_\star) + \frac{\partial^2 \bar{\psi}}{\partial \theta_2 \partial \theta'}(\theta_\star)(\hat{\theta} - \theta_\star) + \frac{1}{2}\frac{\partial^3 \bar{\psi}}{\partial \theta_2^3}(\bar{\theta})(\hat{\theta}_2 - \theta_{\star 2})^2 + o_P(n^{-1/2}) = G_3(\hat{\theta}_2 - \theta_{\star 2})^2 + o_P(n^{-1/3}).$$

From this and using (A.8), the first order condition gives:

$$n^{1/3}(\hat{\theta}_2 - \theta_{\star 2})^2\left(G_3'V^{1/2}MV^{1/2}\left(\sqrt{n}\bar{\psi}(\theta_\star) + \frac{1}{6}G_3\sqrt{n}(\hat{\theta}_2 - \theta_{\star 2})^3\right)\right) = o_P(1).$$

Hence,

$$n^{1/3}(\hat{\theta}_2 - \theta_{\star 2})^2 = o_P(1) \quad \text{or} \quad G_3' V^{1/2} M V^{1/2} \left(\sqrt{n}\bar{\psi}(\theta_\star) + \frac{1}{6} G_3 \sqrt{n}(\hat{\theta}_2 - \theta_{\star 2})^3 \right) = o_P(1).$$

That is

$$\sqrt{n}(\hat{\theta}_2 - \theta_{\star 2})^3 = o_P(1) \quad \text{or} \quad \sqrt{n}(\hat{\theta}_2 - \theta_{\star 2})^3 = -6 \frac{G_3' V^{1/2} M V^{1/2}}{G_3' V^{1/2} M V^{1/2} G_3} \sqrt{n}\bar{\psi}(\theta_\star) + o_P(1).$$

Using (A.8) to obtain $n\bar{\psi}'(\hat{\theta}) V_n \bar{\psi}(\hat{\theta})$ and plugging in either of these two values of $\sqrt{n}(\hat{\theta}_p - \theta_{\star p})^3$, we can see that the minimum is actually reached by the second quantity. Plugging this expression into (A.7), we get:

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_1 - \theta_{\star 1} \\ (\hat{\theta}_2 - \theta_{\star 2})^3 \end{pmatrix} = \Delta V^{1/2} \sqrt{n}\bar{\psi}(\theta_\star) + o_P(1),$$

with $\Delta = \begin{pmatrix} -(D'VD)^{-1} D' V^{1/2} \left(I_q - V^{1/2} G_3 G_3' V^{1/2} M / \sigma_{G_3} \right) \\ -6 G_3' V^{1/2} M / \sigma_{G_3} \end{pmatrix}$ and the result follows. □

References

[1] Begun, J. M., W. J. Hall, W.-M. Huang and J. Wellner, 1983. "Information and asymptotic efficiency in parametric-nonparametric models," *Annals of Statistics*, 11, 432-452.

[2] Beran, R., 1977. "Estimating a distribution function," *Annals of Statistics*, 5, 400-404.

[3] Beran, R., 1978. "An efficient and robust adaptive estimator of location," *Annals of Statistics*, 6, 292-313.

[4] Bhattacharyya, A., 1946. "On some analogues to the amount of information and their uses in statistical estimation," *Sankhya*, 8, 1-14, 201-208, 277-280.

[5] Bickel, P., 1981. "Minimax estimation of the mean of a normal distribution when the parameter space is restricted," *Annals of Statistics*, 9, 1301-1309.

[6] Bickel, P., 1982. "On Adaptive Estimation," *Annals of Statistics*, 10, 647-671.

[7] Chamberlain, G. 1986. "Asymptotic efficiency in semiparametric models with censoring," *Journal of Econometrics*, 32, 189-218.

[8] Chamberlain, G. 1987. "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[9] Dalalyan, A. S., G. K. Golubev and A. B. Tsybakov, 2006. "Penalized maximum likelihood and semiparametric second-order efficiency," *Annals of Statistics*, 34, 169201.

[10] Diebold F. and M. Nerlove, 1989. "The dynamics of exchange rate volatility: a multivariate latent factor ARCH model," *Journal of Applied Econometrics*, 4, 121.

[11] Dovonon P., 2013. "Conditionally heteroskedastic factor models with skewness and leverage effects," *Journal of Applied Econometrics*, 28, 1110-1137.

[12] Dovonon, P. and S. Gonçalves, 2015. "Bootstrapping the GMM overidentification test under first-order underidentification," *Working Paper*, Concordia University and Western University.

[13] Dovonon, P. and A. R. Hall, 2015. "The asymptotic properties of GMM and indirect inference under second-order identification," *Working Paper*, Concordia University and University of Manchester.

[14] Dovonon, P. and E. Renault, 2009. "GMM overidentification test with first order underidentification," *Working Paper*, Concordia University and Brown University.

[15] Dovonon, P. and E. Renault, 2013. "Testing for common conditionally heteroskedastic factors," *Econometrica*, 81, 2561-2586.

[16] Engle, R. F. and S. Kozicki, 1993. "Testing for common features," *Journal of Business & Economic Statistics*, 11(4), 369-395.

[17] Engle, R. F. and A. Mistry, 2007. "Priced risk and asymmetric volatility in the cross-section of skewness," *Working paper*, NYU.

[18] Fiorentini G., E. Sentana, N. Shephard, 2004. "Likelihood-based estimation of generalised ARCH structures," *Econometrica*, 72, 1481-1517.

[19] Harvey C. R., A. Siddique, 1999. "Autoregressive conditional skewness," *Journal of Financial and Quantitative Analysis*, 34, 465-487.

[20] Harvey, C. R., A. Siddique, 2000. "Conditional skewness in asset pricing tests," *Journal of Finance*, 55, 1263-1295.

[21] Hájek, J., 1972. "Local asymptotic minimax and admissibility in estimation," *Proc. Sixth Berkeley Symp. Math. Statist. and Probab.*, 1, 245-261. University of California Press, Berkeley.

[22] King, M. A., E. Sentana, S. B. Wadhwani, 1994. "Volatility and links between national stock markets," *Econometrica* 62, 901-933.

[23] LeCam, L., 1972. "Limits experiments," *Proc. Sixth Berkeley Symp. Math. Statist. and Probab.*, 1, 175-194. University of California Press, Berkeley.

[24] Lee, L.-F. and A. Chesher, 1986. "Specification testing when score test statistics are identically zero," *Journal of Econometrics*, 31, 121-149.

[25] Levit, B. Y., 1975. "On the efficiency of a class of non-parametric estimates," *Theory Probab. Appl.* 20, 723-740.

[26] Magnus, J. R., and Neudecker, H., 1988. "Matrix Differential Calculus With Applications in Statistics and Econometrics," *Wiley*, Chichester.

[27] Millar, P. W., 1979. "Asymptotic minimax theorems for the sample distribution function," *Z. Wahrsch. verw. Gebiete* 48, 233-252.

[28] Newey, K. W. and D. McFadden, 1994. "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, IV, Edited by R.F. Engle and D. L. McFadden, 2112-2245.

[29] Powell, J. L., 1994. "Estimation of semiparametric," in *Handbook of Econometrics*, IV, Edited by R.F.

Engle and D. L. McFadden, 2443-2521.
[30] Rotnitzky, A., D. R. Cox, M. Bottai and J. Robins, 2000. "Likelihood-based inference with singular information matrix," *Bernoulli*, 6(2), 243-284.
[31] Sargan, J. D., 1983. "Identification and lack of identification," *Econometrica*, 51, 1605-1633.
[32] Schick, A., 1986. "On asymptotically efficient estimation in semiparametric models," *Annals of Statistics* 14, 1139-1151.
[33] Stein, C., 1956. "Efficient nonparametric testing and estimation," *Proc. Third Berkeley Symp. Math. Statist. and Probab.* 1, 187-195. University of California Press, Berkeley.
[34] van der Vaart, A. W., and J. A. Wellner, 1996. "Weak convergence and empirical processes with application to statistics," *Springer-Verlag*, New York.
[35] Wellner, J. A., 1982. "Asymptotic optimality of the product-limit estimator," *Annals of Statistics*, 10, 595-602.