Efficiency bound under identifiability constraints in semiparametric models

Patrice Bertail, Mélanie Zetlaoui

Efficiency bound under identifiability constraints in semiparametric models

Patrice Bertail[1] and Mélanie Zetlaoui[1]

[1]MODAL'X UMR 9023, UPL, University Paris Nanterre, 200, Ave de la République, Nanterre, 92001 cedex, France.

Contributing authors: patrice.bertail@gmail.com; melanie.zetlaoui@gmail.com;

Abstract

The purpose of this work is to define an adequate efficiency bound in some models presenting some identification problems. We show how it is possible to define such bounds in some regular semi-parametric models (in the sense of Le Cam) when an identifying constraint is available, despite the degeneracy of the information matrix. We establish a new convolution theorem in this context. We illustrate the computation of the information bound for some standard identifiability constraints, in some interesting models, including probit, single-index, Anova models. We also show how a two-step procedure still based on a preliminary estimator satisfying approximately the constraint, allows us to obtain an efficient estimator of the parameters.

Keywords: semiparametric models, efficiency bounds, unidentifiable parameters, Bahadur efficiency

MSC Classification: 35A01 , 65L10 , 65L12 , 65L20 , 65L70

1 Introduction

The problem of asymptotic efficiency is an old statistical problem to which Bahadur, Fischer, Cramér, and Rao have brought important contributions. In particular, Bahadur (1960), (see also Shen, 2001) has shown that the notion of Bahadur efficiency is strongly related to Fisher information in identifiable regular parametric models. However, in many parametric and semiparametric

models, the parameters of interest are not directly identifiable and can only be estimated under some well-chosen identifiability constraints. Unless a clear reparametrization of the problem is easy to implement, this makes the computation and the notion of efficiency difficult to quantify. For instance, even in some very simple models (with multidimensional parameters), the Fisher Information matrix may be degenerate (not full rank) making the definition of an efficiency bound (in all different senses) or a Cramér-Rao bound quite problematic: recall that Rothenberg (1971) has shown that weak local identification of the parameter of interest is equivalent to the non-degeneracy of the Fisher Information matrix in parametric models. In some parametric models, it is possible to compute such a bound by a correct reparametrization of the original (non-identifiable) parameters. Consider for instance the case of a probit model, in which we constrain the variance of the latent model to be equal to 1, then, in that case, the identification problem is easily handled and the efficiency bound may be computed only over the remaining parameters. However, reparametrization may be more complex in many semiparametric models, and computation of the corresponding efficiency bound is even more problematic. This problem appears in many semiparametric models including the following models.

- Single index models (and their generalization, multiple index model) or slice inverse regression (see Ma & Zhu, 2013). The parameters are generally identifiable only under some normalizing constraints (for instance the $L2$ norm of the parameter vector say $\|\beta\|_2$ is fixed to one) and a sign constraint indicating the direction of at least one component. A spherical reparametrization as suggested in the work of Ma and Zhu (2013) may be difficult to implement in practice to get both an efficient estimator and an explicit computation of the information matrix.
- Factor analysis and non-negative matrix factorization: the same problems appear here as in single-index models. The direction of the cone to which the data belongs should satisfy some constraints (see **?**).
- Simultaneous equations problems in econometric theory: identifiability conditions are also generally necessary but reparametrization of the model in terms of structural equations with identifiable parameters is not always easy.
- Mixture models: identifiability constraints are also necessary and identification becomes quite intricate for semi-parametric models.
- Neural network models: Fukumizu (1996) has obtained conditions for the non-degeneracy of the Fisher Information matrix which essentially amount to eliminating a layer(s) of the network (that is putting some parameters equal to 0). See also Fukumizu (2003) for local conic reparametrization of the model in this framework.

The purpose of this paper is to propose a general framework for computing efficiency bounds in some non-identifiable semiparametric models when some identifiability constraints are available. Despite some obvious connections, our problem is different from the problem of obtaining efficiency bounds in an

identifiable model, under some constraints on the parameters, as studied in Stoica and Ng (1998) for parametric models, and in Susyanto and Klaassen (2017) and Klaassen and Susyanto (2019) for semiparametric models. In our case, the Fisher information of the original model is not invertible. One may think that the approach of these authors may be extended by using a generalized inverse of the Fisher information. However, this is not true as will be shown in a simple example. Despite this fact, some ideas are quite similar: in particular, the main idea in a parametric model is to project the score function in the space orthogonal to the gradients of the constraints. The same techniques are applied to semiparametric models, extending arguments from Bickel, Klaassen, Ritov, and Wellner (1998), Klaassen and Susyanto (2019) and Kosorok (2008). The efficiency bound that we propose is itself degenerate and depends on the identifying constraint but enables one to recover the right bounds for any reparametrization (satisfying the constraints). We also show that the usual two-step procedure considered in Cheng (2013) based on a preliminary estimator of the parameters (satisfying approximately the constraints up to order $O(n^{-1/2})$ and an adequate estimator of the efficient score function allows us to get efficient estimators of the parameters.

2 Efficiency bounds under identifiability constraints

2.1 Background: Theory and Concepts

Consider a parametric model $\mathbb{P}_\Theta = \left\{P_\theta,\ \theta \in \Theta \subset \mathbb{R}^k\right\}$ which is supposed to be dominated by a measure μ. The space $\mathbb{R}^k$ is endowed with a generic L_2 norm $\|.\|$. The set Θ is assumed to be an open set. Put $p_\theta = \frac{dP_\theta}{d\mu}$. As in the theory of Le Cam (see Le Cam, 1986) we use the reparametrization

$$\begin{aligned} s\ &: \mathbb{R}^k \to L_2(\mu) \\ &\theta \mapsto s(\theta) = \sqrt{p_\theta}, \end{aligned} \tag{1}$$

which allows to consider $s(\theta) = \sqrt{p_\theta}$ as an element of an Hilbert space. In this space, we recall the definition of quadratic differentiability (see Le Cam, 1986).

The model is said to be quadratically differentiable at $\theta \in \Theta$, if there exists a gradient $\dot{l}_\theta \in \mathbb{R}^k$ such that

$$\int \left(p_{\theta+h}^{\frac{1}{2}} - p_\theta^{\frac{1}{2}} - \frac{1}{2}h^t \dot{l}_\theta p_\theta^{\frac{1}{2}}\right)^2 d\mu = o(\|h\|^2) \quad as\ h \to 0.$$

Recall that differentiability in quadratic mean is essentially a smoothness assumption allowing for some singularities in the model, which ensures that the classical properties of maximum likelihood still hold. For example the Laplace distribution with $p_\theta(x) = \frac{1}{2}\exp(-\mid x-\theta \mid)$ is differentiable in quadratic mean with gradient $\dot{l}_\theta$ given by $\dot{l}_\theta(x) = sign(x-\theta)$.

In a quadratically differentiable model, the Fisher information matrix is defined by

$$I(\theta) = E(\dot{l}_\theta \ \dot{l}_\theta{}^t).$$

When a parametric model is not identifiable this matrix is not full rank (see Rothenberg, 1971) so that the Fréchet-Darmois-Cramér-Rao information bound can not be defined as the inverse of the matrix. In the same way, Bahadur efficiency can not be obtained because of the degeneracy of Fisher Information.

Definition 1 In the following, we say that a parametric model $\mathbb{P}_\Theta$ is regular (in Le Cam sense) if for any $\theta \in \Theta$, p_θ is quadratically differentiable and the Fisher information exists and is positive semidefinite.

Assume now that $\theta \in \Theta \subset \mathbb{R}^k$ is not identifiable but that we have l identifiability constraints of the form $G(\theta) = 0$, where $G : \mathbb{R}^k \to \mathbb{R}^l$ is some measurable function. This means that there exists a transformation $\phi : \mathbb{R}^k \to \mathbb{R}^{k-l}$ such that $\phi = \phi(\theta)$ is identifiable. However, such transformation may not be explicit and sometimes difficult to exhibit.

Define the gradient matrix of the constraints in the set of the real matrices of size (l, k), $\mathcal{M}_{l,\,k}(\mathbb{R})$, by $\dot{G}_\theta = \frac{dG_\theta}{d\theta^t}$. It is assumed to have full rank l. Introduce U_θ in $\mathcal{M}_{k,\,k-l}(\mathbb{R})$ a matrix such that $U_\theta^t U_\theta = I_{k-l}$, whose columns form an orthonormal basis for the null space of $\dot{G}_\theta$ that is such that

$$\dot{G}_\theta U_\theta = 0.$$

We now recall the notion of local regular estimators (see Bickel et al., 1998).

Definition 2 Let T_n be a $\sqrt{n} - consistent$ estimator of θ. Define a sequence of values $\theta_n = \theta + h/\sqrt{n}$, $h \in \mathbb{C}^k$, endowed with the $L2$ norm and such that, for n large enough, $\theta_n \in \Theta$. The statistics T_n is said to be locally regular at θ with limiting distribution Z iff, for any h such that $\|h\| \leq M$, where M is a positive constant, we have

$$\sqrt{n}(T_n - \theta_n) \underset{n\to\infty}{\overset{L_{P_{\theta_n}}}{\longrightarrow}} Z,$$

where $\overset{L_{P_{\theta_n}}}{\longrightarrow}$ means convergence in law under the distribution P_{θ_n} of the observations.

Intuitively, it means that the statistics T_n is asymptotically locally robust, in that a little change in the parameter $\theta_n = \theta + h/\sqrt{n}$ will not change the limiting distribution Z of T_n when it is correctly centered by θ_n .

2.2 Convolution theorem under identifiability constraints in parametric models

2.2.1 Convolution theorem

The Hajek-Le Cam convolution theorem (see van der Vaart, 2000,Theorem 25.20, and Le Cam, 1986) is an important tool for defining the notion of efficient estimators. The following result extends this theorem to parameters that are not identifiable, up to a replacement of the inverse of Fisher Information bound by the bound

$$\mathcal{B}_\theta = U_\theta (U_\theta^t I(\theta) U_\theta)^{-1} U_\theta^t.$$

The form of this bound may be understood as the generalized inverse of the variance of a projected score. Let us introduce the orthogonal projection matrix onto the columns of U_θ in $\mathcal{M}_{k,\,k}(\mathbb{R})$ given by

$$P_U = U_\theta (U_\theta^t U_\theta)^{-1} U_\theta^t = U_\theta U_\theta^t.$$

Then the projected score $P_U \dot{l_\theta}$ has variance

$$V(P_U \dot{l_\theta}) = U_\theta U_\theta^t I(\theta) U_\theta U_\theta^t,$$

which plays the role of the new Fisher information under the given constraints. Then it is easy to see that the generalized inverse of this quantity is given by $\mathcal{B}_\theta$. Indeed, using the fact that $U_\theta^t U_\theta = I_{k-l}$, we have

$$\begin{aligned}
\mathcal{B}_\theta V(P_U \dot{l_\theta}) \mathcal{B}_\theta &= U_\theta (U_\theta^t I(\theta) U_\theta)^{-1} U_\theta^t [U_\theta U_\theta^t I(\theta) U_\theta U_\theta^t] U_\theta (U_\theta^t I(\theta) U_\theta)^{-1} U_\theta^t \\
&= U_\theta (U_\theta^t I(\theta) U_\theta)^{-1} (U_\theta^t I(\theta) U_\theta)(U_\theta^t I(\theta) U_\theta)^{-1} U_\theta^t \\
&= U_\theta (U_\theta^t I(\theta) U_\theta)^{-1} U_\theta^t = \mathcal{B}_\theta
\end{aligned}$$

and

$$\begin{aligned}
V(P_U \dot{l_\theta}) \mathcal{B}_\theta V(P_U \dot{l_\theta}) &= U_\theta U_\theta^t I(\theta) U_\theta U_\theta^t [U_\theta (U_\theta^t I(\theta) U_\theta)^{-1} U_\theta^t] U_\theta U_\theta^t I(\theta) U_\theta U_\theta^t \\
&= U_\theta (U_\theta^t I(\theta) U_\theta)(U_\theta^t I(\theta) U_\theta)^{-1} (U_\theta^t I(\theta) U_\theta) U_\theta^t \\
&= U_\theta U_\theta^t I(\theta) U_\theta U_\theta^t = V(P_U l_\theta).
\end{aligned}$$

In the unbiased case, this formula also covers the identifiable case with some constraints given by the function G on the parameter. Such constrained parameters have been considered in Stoica and Ng (1998), Susyanto and Klaassen (2017) and Klaassen and Susyanto (2019) and their expressions can used directly in that case when $V(P_U \dot{l_\theta})$ is of full rank. Our proof covers unidentified parametric models. In our case, since the original parameter is not identifiable, the matrix $V(P_U \dot{l_\theta})$ is not directly invertible (as assumed in the paper of Stoica & Ng, 1998).

For reasons that will become clearer later, the quantity $\mathcal{B}_\theta \dot{l_\theta}$ is called the efficient score and we have as in the usual identifiable parametric case

$$V(\mathcal{B}_\theta \dot{l_\theta}) = \mathcal{B}_\theta.$$

Lemma 1 *Consider a regular parametric model* $\mathbb{P}_\Theta = \left\{P_\theta,\ \theta \in \Theta \subset \mathbb{R}^k\right\}$, *which is quadratically differentiable in* $\theta \in \Theta$. *The function* $G : \mathbb{R}^k \to \mathbb{R}^l$ *defining the constraints is assumed to be differentiable in* θ. *Let* T_n *be a locally regular estimator of* θ. *Then, we have the following Hajek-Le Cam convolution theorem, under the identifiability constraints,*

$$\begin{pmatrix} \sqrt{n}\,(T_n - \theta) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{B}_\theta \dot{l_\theta}\,(X_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{B}_\theta \dot{l}\,(X_i) \end{pmatrix} \underset{n\to\infty}{\overset{L_{P_\theta}}{\longrightarrow}} \begin{pmatrix} \Delta_{P_\theta} \\ Z_\theta \end{pmatrix},$$

where Z_θ *is distributed as a centered normal distribution* $\mathcal{N}(0, \mathcal{B}_\theta)$ *with variance-covariance matrix* $\mathcal{B}_\theta$ *and is independent from the asymptotic distribution* Δ_{P_θ}.

Proof We follow the arguments of Bickel et al. (1998), Theorem 1 p. 24, with slight modifications to take into account the identifiability constraints. Define the couple of r.v.'s

$$(R_n, V_n) = (\sqrt{n}\,(T_n - \theta)\,, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} P_U \dot{l_\theta}\,(X_i))$$

and denote (R, V) its joint limit (for eventually a particular sub-sequence that we denote by n by a slight abuse of notation, see a similar construction in Bickel et al., 1998). By the CLT, the limiting distribution of the second component is $V = N(0, P_U I(\theta) P_U)$. Now define, for $h \in \mathbb{C}^k$,

$$W_n(h) = \sum^{n} \log(p_{\theta + \frac{P_U\, h}{\sqrt{n}}}\,(X_i)) - \frac{1}{\sqrt{n}} \sum^{n} log(p_\theta\,(X_i))$$

with limiting distribution $h^t V - \frac{1}{2} h^t P_U I(\theta) P_U h$.
Notice that under $P_{\theta + \frac{P_U\, h}{\sqrt{n}}}$ we have $R_n = R + P_U h + o_P(1)$ it follows that

$$E_{P_{\theta + \frac{P_U\, h}{\sqrt{n}}}} \exp\left(ia^t R_n\right) \underset{n\to\infty}{\to} E(\exp(ia^t R)) \exp(ia^t P_U h).$$

But by contiguity, we have as well

$$E_{P_{\theta + \frac{P_U\, h}{\sqrt{n}}}} \exp(ia^t R_n) = E_{P_\theta} \exp(ia^t R_n + W_n(h)) + o(1).$$

It follows that, by identifying these two expressions at the limit that we have

$$E_{P_\theta} \exp(ia^t R + h^t V - \frac{1}{2} h^t P_U I(\theta) P_U h) = E(\exp(ia^t R)) \exp(ia^t P_U h),$$

yielding

$$E_{P_\theta} \exp(ia^t R + h^t V) = E(\exp(ia^t R)) \exp(ia^t P_U h + \frac{1}{2} h^t P_U I(\theta) P_U h).$$

For any $b \in \mathbb{R}^k$, put $h^t = -i(a-b)^t\mathcal{B}_\theta$ then, using the fact that $P_U\mathcal{B}_\theta = \mathcal{B}_\theta$ and $\mathcal{B}_\theta I(\theta)\mathcal{B}_\theta = \mathcal{B}_\theta$, we get

$$\begin{aligned} &E_{P_\theta}\exp(ia^tR - i(a-b)^t\mathcal{B}_\theta V) \\ &= E_{P_\theta}\exp(ia^t(R-\mathcal{B}_\theta V) + ib^t\mathcal{B}_\theta V) \end{aligned}$$

$$\begin{aligned} &= E(\exp(ia^tR))\exp\left(a^tP_U\mathcal{B}_\theta(a-b) - \frac{1}{2}(a-b)^t\mathcal{B}_\theta P_U I(\theta)P_U\mathcal{B}_\theta(a-b)\right) \\ &= E(\exp(ia^tR))\exp\left(a^t\mathcal{B}_\theta(a-b) - \frac{1}{2}(a-b)^t\mathcal{B}_\theta(a-b)\right). \end{aligned}$$

It follows that, for $b=0$,

$$E_{P_\theta}\exp(ia^t(R-\mathcal{B}_\theta V) = E(\exp(ia^tR))\exp\left(\frac{1}{2}a^t\mathcal{B}_\theta a\right),$$

yielding

$$\begin{aligned} E_{P_\theta}\exp(ia^t(R-\mathcal{B}_\theta V) + ib^t\mathcal{B}_\theta V) &= E_{P_\theta}\exp(ia^t(R-\mathcal{B}_\theta V)\exp\left(-\frac{1}{2}a^t\mathcal{B}_\theta a\right) \\ &\exp\left(a^t\mathcal{B}_\theta(a-b) - \frac{1}{2}(a-b)^t\mathcal{B}_\theta(a-b)\right) \\ &= E_{P_\theta}\exp(ia^t(R-\mathcal{B}_\theta V)\exp(-\frac{1}{2}b^t\mathcal{B}_\theta b). \end{aligned}$$

Since this is the limiting characteristic function of $(R_n - \mathcal{B}_\theta V_n, \mathcal{B}_\theta V_n)$, we conclude that these two components are independent by the factorization of the characteristic function and that the second component is Gaussian. From this, using again the fact that $\mathcal{B}_\theta P_U = \mathcal{B}_\theta$, we get the convolution theorem.

Notice that $V_{P_\theta}(\sqrt{n}\,(T_n-\theta) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{B}_\theta \dot{l_\theta}(X_i)) = Var(\Delta_{P_\theta})$ and by independence we get

$$\lim_{n\to\infty} V_{P_\theta}(\sqrt{n}\,(T_n-\theta)) \geq V_{P_\theta}(Z_\theta) = \mathcal{B}_\theta I(\theta)\mathcal{B}_\theta^t = \mathcal{B}_\theta.$$

It follows that the smallest variance of any regular estimator of T_n is given by $\mathcal{B}_\theta$ and T_n is efficient iff T_n is asymptotically a.s. linear with

$$\sqrt{n}\,(T_n-\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{B}_\theta \dot{l_\theta}(X_i) + o(1).$$

□

Remark 1 If we have an identifiable reparametrization $\phi = \phi(\theta)$ then we can choose U_θ to be the orthogonalized version of $\dot{\phi}(\theta)^t$ in $\mathcal{M}_{(k,k-l)}$.

Lemma 2 *The efficiency bound for θ under the identifiability constraint G is given by*

$$\mathcal{B}_\theta = U_\theta(U_\theta^t I(\theta)U_\theta)^{-1}U_\theta^t,$$

where $I(\theta)$ is the Fisher information of the unconstrained likelihood. .

Proof The result follows immediately from the convolution theorem given before. Indeed the smallest variance of T_n is obtained when $Var(\Delta_{P_\theta}) = 0$ and in that case the variance of the estimator becomes $\mathcal{B}_\theta$. □

2.2.2 Examples

Example 1 Consider a simple Probit model. For this, consider Y with Bernoulli distribution $B(1, \Phi(\frac{\mu}{\sigma}))$, $\theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}$, where Φ is the cdf of the standard normal distribution. Only $\frac{\mu}{\sigma}$ and functions of this quantity are identifiable in this model. One may propose the identifiability constraints $\sigma = 1$ but also $\mu = 1$. This is a toy example that illustrates the behavior of the tools we introduced before.

The log-likelihood is given by

$$l_\theta(y) = \log(p_\theta(y)) = y \log \Phi(\frac{\mu}{\sigma}) + (1-y) \log(1 - \Phi(\frac{\mu}{\sigma}))$$

so that the gradient or score function is given by

$$\begin{aligned} \dot{l}_\theta &= \begin{pmatrix} \frac{y}{\sigma} \frac{\phi(\frac{\mu}{\sigma})}{\Phi(\frac{\mu}{\sigma})} - \frac{(1-y)}{\sigma} \frac{\phi(\frac{\mu}{\sigma})}{1-\Phi(\frac{\mu}{\sigma})} \\ -\frac{y\mu}{\sigma^2} \frac{\phi(\frac{\mu}{\sigma})}{\Phi(\frac{\mu}{\sigma})} + \frac{(1-y)\mu}{\sigma^2} \frac{\phi(\frac{\mu}{\sigma})}{1-\Phi(\frac{\mu}{\sigma})} \end{pmatrix} \\ &= \frac{\phi(\frac{\mu}{\sigma})(y - \Phi(\frac{\mu}{\sigma}))}{\sigma \Phi(\frac{\mu}{\sigma})(1 - \Phi(\frac{\mu}{\sigma}))} \begin{pmatrix} 1 \\ -\frac{\mu}{\sigma} \end{pmatrix}, \end{aligned}$$

yielding

$$I(\theta) = \frac{1}{\sigma^2} \frac{\phi(\frac{\mu}{\sigma})^2}{\Phi(\frac{\mu}{\sigma})(1 - \Phi(\frac{\mu}{\sigma}))} \begin{pmatrix} 1 & -\frac{\mu}{\sigma} \\ -\frac{\mu}{\sigma} & \frac{\mu^2}{\sigma^2} \end{pmatrix}.$$

Notice that $I(\theta)$ is obviously of rank 1. The generalized inverse of this matrix can be obtained by singular value decomposition and is given by

$$I(\theta)^- = \sigma^2 \frac{\Phi(\frac{\mu}{\sigma})(1 - \Phi(\frac{\mu}{\sigma}))}{\phi(\frac{\mu}{\sigma})^2 (1 + \frac{\mu^2}{\sigma^2})^2} \begin{pmatrix} 1 & -\frac{\mu}{\sigma} \\ -\frac{\mu}{\sigma} & \frac{\mu^2}{\sigma^2} \end{pmatrix}.$$

Under the identifiability constraint $G(\theta) = \sigma - 1 = 0$ we get $\dot{G}_\theta = (0, 1)$ and we have

$$U_\theta = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

so that the efficiency bound is given by a straightforward calculus, by

$$\mathcal{B}_\theta = \frac{\Phi(\mu)(1 - \Phi(\mu))}{\phi(\mu)^2 (1 + \mu^2)^2} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The upper left corner clearly gives the right bound for μ (which is identifiable now) and since $\sigma = 1$ is constant we clearly have the bound for σ equal to 0. Notice that clearly, the information will depend on the given constraint. Notice also that the expression appearing in Stoica and Ng (1998) or Susyanto & Klaassen, 2017 and Klaassen & Susyanto, 2019) are only valid when $I(\theta)$ is full rank. It does not hold if one replaces the inverse with the generalized inverse, as could be expected. Indeed, the expression given in Stoica and Ng (1998) would become

$$\mathcal{I}_\theta^{-1} = I(\theta)^- - I(\theta)^- \dot{G}_\theta^t (\dot{G}_\theta \; I(\theta)^- \dot{G}_\theta^t)^{-1} \dot{G}_\theta I(\theta)^-.$$

A straightforward calculation shows in this example that the right bound is given by

$$\mathcal{I}_\theta^{-1} = I(\theta)^- - \frac{1}{\mu^2} \frac{\Phi(\mu)(1 - \Phi(\mu))}{\phi(\mu)^2 (1 + \mu^2)^2} \begin{pmatrix} 1 & -\mu \\ -\mu & \mu^2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu \\ -\mu & \mu^2 \end{pmatrix}$$

$$= \frac{\Phi(\mu)(1-\Phi(\mu))}{\phi(\mu)^2(1+\mu^2)^2}\begin{pmatrix} 1 & -\mu \\ -\mu & \mu^2 \end{pmatrix} - \frac{\Phi(\mu)(1-\Phi(\mu))}{\phi(\mu)^2(1+\mu^2)^2}\begin{pmatrix} 1 & -\mu \\ -\mu & \mu^2 \end{pmatrix}$$
$$= 0.$$

Example 2 We consider the following artificial linear model in order to explain later, how the information matrix for a single index model is defined :

$$Y = cX^t\beta + \varepsilon \ , \ \theta = \begin{pmatrix} c \\ \beta \end{pmatrix}.$$

The residual ε is assumed to have known density f with $E_f(\varepsilon \mid X) = 0$ and $V_f(\varepsilon \mid X) = \sigma^2 < \infty$. The X_i are i.i.d with density g.

The identifiability condition $c = 1$ clearly leads to the usual linear regression for which we can compute directly the Fisher information for β. However, we may be rather interested, as in single-index models, by the direction of β assuming that $\|\beta\|^2 = 1$ and $c > 0$.

The likelihood of the model is given by

$$\Pi_{i=1}^n f(y_i - cx_i^t\beta)g(x_i).$$

It follows that the individual score is given by

$$\dot{l}_\theta(y \mid x) = \begin{pmatrix} x^{\,t}\beta \\ cx \end{pmatrix}\frac{f'(y-cx^t\beta)}{f(y-cx^t\beta)}$$

$$V(\dot{l}_\theta(y \mid x)) = E_f\left(\frac{f'(\varepsilon)}{f(\varepsilon)}\right)^2 E\left(\begin{pmatrix} x^t\beta \\ cx \end{pmatrix}\begin{pmatrix} \beta^t x & cx^t \end{pmatrix}\right)$$
$$= E_f\left(\frac{f'(\varepsilon)}{f(\varepsilon)}\right)^2 \begin{pmatrix} E(x^t\beta\beta^t x) & cE(x^t\beta x^t) \\ cE(x\beta^t x) & c^2E(xx^t) \end{pmatrix}$$
$$= E_f\left(\frac{f'(\varepsilon)}{f(\varepsilon)}\right)^2 \begin{pmatrix} \beta^t E(xx^t)\beta & c\beta^t E(xx^t) \\ cE(xx^t)\beta & c^2E(xx^t), \end{pmatrix},$$

which is of course of rank k. To identify the model we ensure the constraint that $\|\beta\|^2 = 1$. For simplicity, we also assume that at least one of the coefficient is non zero and more precisely $\beta_1 \neq 0$ then the matrix $\dot{G}_\theta = \frac{dG_\theta}{d\theta^t}$ is given by

$$\dot{G}_\theta = 2(\beta^t, 0).$$

A matrix of dimension $(k+1, k)$ orthogonal to $\dot{G}_\theta$ is given by

$$\tilde{U}_1 = \begin{pmatrix} -\frac{\beta_2}{\beta_1} & \cdots & -\frac{\beta_k}{\beta_1} & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1 \end{pmatrix},$$

which can be straightforwardly orthogonalized by a Gram-Schmidt orthogonalization technique (see below the examples in part 4).

Example 3 One-way factor analysis.

Consider the effect model $Y_{i,j} = \mu + a_i + \epsilon_{i,j}$ for $i = 1, \ldots, I$ the experimental units and $j = 1, \ldots, n_i$. Here n_i is the number of experimental units in the treatment

group i and $n = \sum_{i=1}^{I} n_i$. Assume for simplicity that $\epsilon_{i,j}$ are i.i.d. $N(0, \sigma^2)$ and that the design is balanced that is $n_i = n/I$ for all i. Several proposals have been considered to identify the effect a_i. One may choose a reference treatment and put $a_i = 0$. Another solution is to consider the treatment effects to satisfy $\sum_{i=1}^{I} a_i = 0$, which is easier to interpret. We will consider this last solution.

Let's define the parameter of the model $\theta = (\mu, a_1, \ldots, a_I, \sigma)$. The log-likelihood is given by

$$l_\theta = \frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i,j}(y_{i,j} - \mu - a_i)^2$$

and the score function by

$$\dot{l}_\theta = \begin{pmatrix} -\frac{1}{\sigma^2}\sum_{i,j}(y_{i,j} - \mu - a_i) \\ -\frac{1}{\sigma^2}\sum_{j=1}^{n_1}(y_{1,j} - \mu - a_1) \\ \vdots \\ -\frac{1}{\sigma^2}\sum_{j=1}^{n_I}(y_{I,j} - \mu - a_I) \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i,j}(y_{i,j} - \mu - a_i)^2 \end{pmatrix}.$$

That yields the Fisher information matrix

$$I(\theta) = \frac{1}{\sigma^2}\begin{pmatrix} 1 & I^{-1} & \cdots & \cdots & I^{-1} & 0 \\ I^{-1} & I^{-1} & 0 & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 & \vdots \\ I^{-1} & 0 & \cdots & 0 & I^{-1} & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 2 \end{pmatrix}.$$

The matrix $I(\theta)$ is clearly not full of rank.

Under the identifiability constraint $G(\theta) = \sum_{i=1}^{I} a_i = 0$, we choose the gradient matrix $\dot{G}_\theta = (0, 1, \ldots, 1, 0)$ of length $I + 2$. A matrix of size $(I + 2, I + 1)$ orthogonal to $\dot{G}_\theta$ is given by

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -1 & -1 & \cdots & -1 & 0 \\ \vdots & 1 & 0 & \cdots & 0 & \vdots \\ \vdots & 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 1 \end{pmatrix}.$$

Its Gram-Schmidt orthogonalization U_θ can be easily computed and is given by

$$U_\theta = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}\sqrt{3}} & \cdots & -\frac{1}{\sqrt{I(I+1)}} & 0 \\ \vdots & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}\sqrt{3}} & \cdots & -\frac{1}{\sqrt{I(I+1)}} & \vdots \\ \vdots & 0 & \sqrt{\frac{2}{3}} & \cdots & -\frac{1}{\sqrt{I(I+1)}} & \vdots \\ \vdots & \vdots & 0 & \cdots & -\frac{1}{\sqrt{I(I+1)}} & 0 \\ 0 & 0 & \cdots & 0 & \sqrt{\frac{I}{I+1}} & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}.$$

Notice that it does not depend on θ. Some lengthy but easy calculations yield the efficiency bound

$$\mathcal{B}_\theta = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & I-1 & -1 & \cdots & \cdots & -1 & 0 \\ \vdots & -1 & I-1 & -1 & \cdots & -1 & \vdots \\ \vdots & -1 & -1 & I-1 & & -1 & \vdots \\ \vdots & -1 & -1 & \cdots & \ddots & -1 & 0 \\ 0 & -1 & \cdots & -1 & -1 & I-1 & 0 \\ 0 & 0 & \cdots & 0 & & 0 & \frac{1}{2} \end{pmatrix}.$$

Notice that the usual estimator of the individual effect $\overline{Y}_{i,\cdot} - \overline{Y}_n$ has precisely its variance equal to $(I-1)\sigma^2/n$ and is thus efficient. A straightforward computation shows that the complete estimator $\overline{Y}_n, \overline{Y}_{i,\cdot} - \overline{Y}_n$ has the correct (degenerate) efficiency bound (for simplicity we do not include the standard deviation which is orthogonal to the other components and is only asymptotically efficient).

3 Semiparametric Efficiency bounds under identifiability constraints

3.1 Notations

We now consider a semiparametric model (dominated by a measure μ) indexed by two parameters $\theta \in \Theta$ the parameter of interest and $\eta \in H$, the nuisance parameter

$$\mathbb{P} = \left\{ P_{\theta,\eta},\ \theta \in \Theta \subset \mathbb{R}^k;\ \eta \in H \right\}. \tag{2}$$

We define $L_2(P_{\theta,\eta})$ the Hilbert space of any real function with finite variance with respect to $P_{\theta,\eta}$ and $\overset{\circ}{L}_2(P_{\theta,\eta})$ the subset of $L_2(P_{\theta,\eta})$ of functions with null expectation. In the following $\xrightarrow{P_{\theta,\eta}}$ will used for convergence in probability according to the law $P_{\theta,\eta}$ of the observations and similarly $\xrightarrow{L_{P_{\theta,\eta}}}$ is convergence in distribution. Rather than using the heavy notation $o_{P_{\theta,\eta}}$ (resp. $O_{P_{\theta,\eta}}$) we will use o_P (resp. O_P) when there is no possible confusion.

We denote the Radon-Nikodym density

$$p_{\theta,\eta} = \frac{dP_{\theta,\eta}}{d\mu}$$

and the corresponding score in the quadratic sense in θ by $\dot{l}_{\theta,\eta}$. When the model is differentiable in the usual sense we have

$$\dot{l}_{\theta,\eta} = \frac{d\log(p_{\theta,\eta})}{d\theta} = \frac{\dot{p}_{\theta,\eta}}{p_{\theta,\eta}}.$$

In the following, we assume that $\mathbb{P}_{\Theta,H}$ is regular in the sense that any parametric submodel in $\mathbb{P}_{\Theta,H}$ is regular in a parametric sense, that is, any submodel is quadratically differentiable and the Fisher information exists and is positive semidefinite. We refer to Bickel et al. (1998) for a more detailed presentation of the following concept and recall only a few useful definitions.

3.2 Tangent spaces

We refer to Kosorok (2008) in particular part III for a nice introduction to tangent space and their use in semiparametric models. The tangent set at a given point $P_0 = P_{\theta,\eta}$ is

$$T\left[P_{\theta,\eta},\mathbb{P}\right] = \left\{ \begin{array}{c} g \in L_2\left(P_{\theta,\eta}\right), \text{ such that there exists a regular path} \\ \left(p_t\right)_{t\in\mathbb{R}} \subset \mathbb{P};\ p_0 = p_{\theta,\eta} \text{ with score } g \text{ in } t=0 \end{array} \right\}.$$

In other words, this tangent space is the set of all the scores (with finite variance) of all the regular parametric models indexed by some real parameter t, passing through $p_{\theta,\eta}$ which can be constructed in the semiparametric model.

We then define the tangent space at $P_0 = P_{\theta,\eta}$ as the closure of the span of $T\left[P_{\theta,\eta},\mathbb{P}\right]$

$$T_L\left[P_{\theta,\eta},\mathbb{P}\right] = \overline{Lin}\left[T\left[P_{\theta,\eta},\mathbb{P}\right]\right].$$

We similarly define the tangent space according to parameter θ, as the tangent space of any parametric model $\mathbb{P}_1\left(\eta_0\right) = \left\{P_{\theta,\eta_0};\quad \theta\in\mathbb{R}^k\right\}$, with a fixed nuisance parameter say

$$\dot{\mathbb{P}}_1 = T_L\left[P_{\theta,\eta_0},\mathbb{P}_1\left(\eta_0\right)\right] = \left\{c^t\dot{l}_{\theta,\eta_0};\ c\in\mathbb{R}^k\right\}.$$

Similarly we consider for a fixed $\theta=\theta_0$ the tangent space with respect to the nuisance parameter in the model

$$\mathbb{P}_2\left(\theta_0\right) = \left\{P_{\theta_0,\eta};\ \eta\in H\right\}$$

say

$$\begin{aligned}\dot{\mathbb{P}}_2 &= T_L\left[P_{\theta_0,\eta}, \mathbb{P}_2\left(\theta_0\right)\right] \\ &= \overline{Lin}\left\{ g \in \overset{\circ}{L}_2\left(P_{\theta_0,\eta}\right) / \exists \text{ a parametric regular model } \left(p_t\right) \subset \mathbb{P}_2\left(\theta_0\right); \right. \\ &\left. \text{such that } p_0 = p_{\theta_0,\eta} \text{ with score } g \text{ at } t=0\right\} \subset \overset{\circ}{L}_2\left(P_{\theta_0,\eta}\right)\Big\}.\end{aligned}$$

As done by Bickel et al. (1998) we will assume that at a given point $P_{\theta,\eta}$ we have

$$T\left[P_{\theta,\eta}, \mathbb{P}\right] = \dot{\mathbb{P}}_1 + \dot{\mathbb{P}}_2.$$

3.3 Definition of the efficiency bound and efficient scores under identification constraints

We define $\Pi_{\dot{\mathbb{P}}_2^{\perp}}$ as the projection onto the orthogonal complement of the nuisance tangent space (recall that it is a closed linear space so that the projection exists). Define the orthogonalized score in the unconstrained model by

$$s_{\theta,\eta} = \Pi_{\dot{\mathbb{P}}_2^{\perp}} \dot{l}_{\theta,\eta}.$$

This can be interpreted as the least favorable score among all parametric models passing through the original model $p_{\theta,\eta}$.

Recall that U_θ is the matrix orthogonal to the gradients of the constraint, depending only on the parameter θ. Then define as in the parametric case the corresponding bound under the given identifiability constraints

$$\mathcal{B}_{\theta,\eta} = U_\theta (U_\theta^t V(s_{\theta,\eta}) U_\theta)^{-1} U_\theta^t.$$

This in turn allows us to define the efficient score by

$$s^c_{\theta,\eta} = U_\theta U_\theta^t s_{\theta,\eta},$$

which can be seen as the projection onto the columns U_θ of the orthogonalized score. The corresponding efficient influence function is then given by

$$\begin{aligned}\psi_{\theta,\eta} &= \mathcal{B}_{\theta,\eta} s^c_{\theta,\eta} \\ &= U_\theta (U_\theta^t V(s_{\theta,\eta}) U_\theta)^{-1} U_\theta^t U_\theta U_\theta^t s_{\theta,\eta} \\ &= U_\theta (U_\theta^t V(s_{\theta,\eta}) U_\theta)^{-1} U_\theta^t s_{\theta,\eta} \\ &= \mathcal{B}_{\theta,\eta} s_{\theta,\eta}.\end{aligned}$$

Notice that $V(\psi_{\theta,\eta}) = \mathcal{B}_{\theta,\eta}$.

3.4 A convolution theorem for non-identifiable parameters.

The following theorem establishes a convolution theorem in semiparametric models which justifies that $\mathcal{B}_{\theta,\eta}$ is the efficiency bound of the parameter θ in the constrained model (notice that this bound is not full rank since the parameter is linked by the constraints). Recall that a locally regular estimator in a semiparametric model $\mathbb{P}$ is an estimator that is locally regular for any sub-parametric model in $\mathbb{P}$

Theorem 3 *Hajek-Le Cam convolution theorem under identifiability constraints in semiparametric models.*

Assume that
(i) the statistics T_n estimating θ is locally regular
(ii) $s_{\theta,\eta}$ belong to the closure of the tangent set
(iii) $\mathcal{B}_\theta$ is well defined. Then we have

$$\begin{pmatrix} \sqrt{n}\left(T_n-\theta\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\eta}s_{\theta,\eta}\left(X_i\right) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\eta}s_{\theta,\eta}\left(X_i\right) \end{pmatrix} \xrightarrow[n\to\infty]{L_{P_\theta}} \begin{pmatrix} \Delta_{\theta,\eta} \\ Z_{\theta,\eta} \end{pmatrix},$$

where $Z_{\theta,\eta}$ is distributed a centered normal distribution $\mathcal{N}\left(0,\mathcal{B}_{\theta,\eta}\right)$ with variance $\mathcal{B}_{\theta,\eta}$ and it is independent from $\Delta_{\theta,\eta}$. It follows that the minimal variance attainable by any regular statistics is given by

$$\mathcal{B}_{\theta,\eta}=U_\theta(U_\theta^tV(s_{\theta,\eta})U_\theta)^{-1}U_\theta^t.$$

Proof The arguments are similar to the ones in Bickel et al. (1998) (Th. 2 p 63) or van der Vaart (1989). Since $s_{\theta,\eta}$ is in the closure of the tangent set, it can be approximated by a sequence of parametric models in $T\left[P_{\theta,\eta},\mathbb{P}\right]$. In each of these submodels, T_n is still locally regular. Thus on these sub-models, the convolution theorem of the first part holds, and taking the limit eventually of subsequences (which is possible because we work in the closure of the tangent set), we get the results (take $V_n=V_n$ and replace U_n by R_n and I^{-1} by $\mathcal{B}_{\theta,\eta}$). □

3.5 Efficient semiparametric estimation under approximate constraints

A natural question is if we can estimate efficiently the parameter under the identifiability constraints (at least asymptotically). A key assumption in the semiparametric estimation is first to find, for a fixed value θ, an estimator of the nuisance parameter η (maybe depending on θ). Assume, for any fixed θ, that there exists a consistent "estimator" of the nuisance parameter η say $\widehat{\eta}_{\theta,n}$ such that

$$\widehat{\eta}_{\theta,n}-\eta\overset{P_{\theta,\eta}}{\to}0 \text{ as } n\to\infty.$$

Define the profile log-likelihood function with estimated nuisance parameter as in Severini and Wong (1992)

$$L_n(\theta) = \sum_{i=1}^{n} \log(p_{\theta,\widehat{\eta}_{\theta,n}}).$$

Then the maximum likelihood estimator with estimated nuisance parameter under the identifiability constraint is defined as

$$\widetilde{\theta}_n = \arg \max_{\theta;\, G(\theta)=0} L_n(\theta).$$

We will later assume that this estimator is defined uniquely. It should be noticed that, in that case, the Lagrangian may be written

$$L = L_n(\theta) - \lambda^t G(\theta), \text{ with multipliers } \lambda \in \mathbb{R}^l,$$

so that the first-order condition becomes

$$\dot{L}_n(\theta) - \dot{G}_\theta^t \lambda = 0,$$

which in turn by multiplying by U_θ^t may be rewritten

$$U_\theta^t \dot{L}_n(\theta) = 0,$$

a fact that may simplify the search for an adequate candidate. If we can check that, for some estimator $\widehat{\theta}_n$ (may be the maximum likelihood estimator $\widetilde{\theta}_n$), $\mathcal{B}_{\widehat{\theta}_n,\widehat{\eta}_{\theta,n}} s_{\widehat{\theta}_n,\widehat{\eta}_{\theta,n}}(x) - \mathcal{B}_{\theta,\eta} s_{\theta,\eta}(x) = o_P(1)$ uniformly in x then, by assuming that the set of function in x $\{\mathcal{B}_{\theta,\eta} s_{\theta,\eta}(x),\, \theta \in \Theta, \eta \in H\}$ is a Donsker class of function and that the bias $E_{P_{\theta,\eta}} \mathcal{B}_{\widehat{\theta}_n,\widehat{\eta}_{\theta,n}} s_{\widehat{\theta}_n,\widehat{\eta}_{\theta,n}}(x)$ is of order $o(n^{-1/2})$ then it is easy by following the same arguments as Cheng (2013) or Cheng and Kosorok (2009) that the constrained profile MLE will be asymptotically efficient (simply replace their condition M1-M4 by the same assumptions on $\mathcal{B}_{\theta,\eta} s_{\theta,\eta}$ rather than on the efficient score function). However, finding this m.l.e may be difficult in practice and establishing its asymptotic validity may require specific structure for the nuisance parameter (Severini & Wong, 1992).

For this reason, we will follow the iterated Newton-Raphson steps approach of Cam (1960) (see also Cheng (2013) for extensions). Indeed a key procedure in the semiparametric estimation literature is the one (or iterated) step procedure based on an estimator of the efficient score, provided that an initial estimator of the parameter converging at rate $\sqrt{n}$ is available. We can proceed similarly here. Consider an initial estimator $\widehat{\theta}_n$ satisfying the constraints $G(\widehat{\theta}_n) = G(\theta) = 0$. This hypothesis may be actually relaxed to

$G(\widehat{\theta}_n) = o_P(n^{-1/2})$. Since we assume that asymptotically $\widehat{\theta}_n \overset{P_{\theta,\eta}}{\underset{n\to\infty}{\to}} \theta$, we will clearly have $G(\theta) = 0$. An initial estimator may be obtained as in Cheng (2013) by using a grid of size proportional to $n^{-1/2}$ following the procedure proposed in Le Cam (1956).

Define the one-step estimator by

$$\widehat{\theta}_n^{(1)} = \widehat{\theta}_n + \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}} s_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i)\,.$$

Notice that, intuitively, with this iteration procedure $\widehat{\theta}_n^{(1)}$ will automatically satisfy the constraints (at least up to $o_P(n^{-1/2})$) since we will have by a straightforward expansion

$$\begin{aligned} G(\widehat{\theta}_n^{(1)}) &= G\left(\widehat{\theta}_n + \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}} s_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i)\right) \\ &= G(\widehat{\theta}_n) + \dot{G}_\theta \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}_{\theta,\eta} s_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i) + o_P(n^{-1/2}) \\ &= o_P(n^{-1/2}) + 0 + o_P(n^{-1/2}), \end{aligned}$$

since we have $\dot{G}_\theta \mathcal{B}_{\theta,\eta} = \dot{G}_\theta U_\theta (U_\theta^t V(s_{\theta,\eta}) U_\theta)^{-1} U_\theta^t = 0$ by construction of $\dot{G}_\theta$ and U_θ.

We will make the following simplifying assumptions (other kinds of hypotheses may be proposed according to the context).

H_0 : The initial estimator is such that $\widehat{\theta}_n - \theta = O_P(n^{-1/2})$ and satisfies $G(\widehat{\theta}_n) = o_P(n^{-1/2})$.

H_1 : For any fixed θ in the neighborhood of the true parameter value, $\widehat{\eta}_{\theta,n}$ is an estimator of η which is symmetric in the observations and bounded. Moreover, we assume that we have the rate of convergence for some $r \in]1/4, 1/2]$

$$\widehat{\eta}_{\theta,n} - \eta = O_P(n^{-r}).$$

H_2 : The quantity $\mathcal{B}_{\theta,\eta} s_{\theta,\eta}(.)$ is continuous in both components θ and η for H endowed with some metric inducing weak convergence.

H_3 : Let $\widehat{\theta}_n$ be the initial parameter estimator then we assume that the "bias" of the estimated efficient score function is such that

$$E_{P_{\theta,\eta}}\left(\mathcal{B}_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}} s_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i)\right) = o(n^{-1/2}).$$

H_4 : The set of functions (in x) $\{s_{\theta,\eta}(x)\,,\ \theta \in \Theta, \eta \in H\}$ is a Donsker class of functions.

Then we have the following result ensuring that the iterated procedure yields an efficient estimator satisfying asymptotically the constraints.

Theorem 4 *Under the assumption $H_1 - H_4$, we have that*

$$G(\widehat{\theta}_n^{(1)}) = o_P(n^{-1/2})$$

and

$$\sqrt{n}(\widehat{\theta}_n^{(1)} - \theta) \overset{L_{P_{\theta,\eta}}}{\underset{n\to\infty}{\longrightarrow}} N(0, \mathcal{B}_{\theta,\eta}).$$

Thus the one-step estimator is efficient.

To prove the theorem, we will make use of the following lemma. It is taken from Bertail (2006). This result is very useful in semiparametric applications in which some permutation invariance property is satisfied by the estimator of the nuisance parameter.

Lemma 5 *Assume $X_1, X_2, \ldots, X_n$ are i.i.d. random variables taking their values in $\mathcal{X}$ and for each n, let $\eta_n \in H$ be a symmetric function of the observations (invariant by permutation of the observations). Let $\omega(x,t)$ $x \in \mathcal{X}$ $t \in H$ be a function taking its value in a separable Banach space, endowed with a norm $\|\ \|$. Assume that there exits a function $\Omega(x)$ such that for all $t \in H$,*
(i) $\|\omega(x,t)\| \leq \Omega(x)$ with $E\Omega(X) < \infty$ and
(ii) $\omega(x,t)$ is continuous in t. Then $\eta_n \overset{a.s.}{\to} \eta$ implies that

$$S_n^\omega = \frac{1}{n}\sum_{i=1}^n \omega(X_i, \eta_n) \overset{a.s.}{\to} E(\omega(X_i, \eta)).$$

Moreover under the same conditions, if $E(\omega(X_i, \eta) = 0$, we also have

$$U_n^\omega = \frac{1}{n}\sum_{i\neq j}^n \omega(X_i, \eta_n)\omega(X_j, \eta_n) \overset{a.s.}{\to} 0.$$

Proof : It is sufficient to write

$$S_n^\omega = E\left(\omega(X_1, G_n) \mid \mathcal{S}^n\right),$$

where $\mathcal{S}^n$ is the symmetric field containing all the symmetric functions of $X_1, X_2, \ldots, X_n$. By the extended backward martingale convergence of Blackwell and Dubins (1962), under (i) and (ii), S_n^ω converges with probability one to $E\left(\omega(X_1, \eta) \mid \mathcal{S}^\infty\right)$. But by the Hewitt-Savage zero-one law, $\mathcal{S}^\infty$ is non trivial and therefore $E\left(\omega(X_1, \eta) \mid \mathcal{S}^\infty\right)$ is constant, equal to $E\left(\omega(X_1, \eta)\right)$.

Now apply similarly the same idea $\omega(X_i, \eta_n)\omega(X_j, \eta_n)$ which is bounded by $H(X_i)H(X_j)$ to get the convergence of U-statistics to $E(\omega(X_i, \eta)\omega(X_j, \eta))$ and use the independence.

Proof of Theorem 4: Define

$$I = \sqrt{n}(\widehat{\theta}_n - \theta) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}}\, s_{\widehat{\theta}_n,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i)$$
$$- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}\; s_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i)$$
$$II = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}\; s_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\eta}s_{\theta,\eta}(X_i)\,.$$

From the convolution theorem, it is sufficient to show that

$$\sqrt{n}\left(\widehat{\theta}_n^{(1)} - \theta\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\eta}s_{\theta,\eta}(X_i) = o_P(1).$$

By construction, this is equivalent to showing that $I + II = o_P(1)$. Now under H_3 we can recenter the last term II to get

$$II = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathcal{B}_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}\; s_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i) - E_{P_\theta,\eta}\mathcal{B}_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}\; s_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i)) \tag{3}$$
$$- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\eta}s_{\theta,\eta}(X_i) + o(1). \tag{4}$$

Notice that by exchangeability of the variables inside the sum, we have using lemma 5,

$$E_{P_{\theta,\eta}}\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}\; s_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}(X_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathcal{B}_{\theta,\eta}s_{\theta,\eta}(X_i)\,\|^2$$
$$\leq E\|\mathcal{B}_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}s_{\theta,\widehat{\eta}_{\widehat{\theta}_n,n}}(X) - \mathcal{B}_{\theta,\eta}s_{\theta,\eta}X)\|^2 + o(1) = o(1).$$

The last equality is a consequence of the hypothesis H_2. It follows that $II = o_P(1)$.

Now let's consider the term I. Since $G(\widehat{\theta}_n) = o_P(n^{1/2})$ and $G(\theta) = 0$ we have $\dot{G}_\theta(\widehat{\theta}_n - \theta) = o_P(n^{1/2})$. Since we have the orthogonal decomposition $P_{U_\theta} + P_{\dot{G}_\theta} = I$, it follows that we have the identity

$$P_U(\widehat{\theta}_n - \theta) = (\widehat{\theta}_n - \theta) + o_P(n^{-1/2}).$$

Following the same steps as Le Cam, we use the *Local Asymptotic Normality property* as well as the asymptotic differentiability property

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\widehat{\theta}_n,\widehat{\eta}_n} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\theta,\widehat{\eta}_n} &= -I(\theta,\eta)\sqrt{n}(\widehat{\theta}_n-\theta)+o_P(1)\\ &= -I(\theta,\eta)P_U\sqrt{n}(\widehat{\theta}_n-\theta)+o_P(1),\end{aligned}$$

where $I(\theta,\eta)=V(s_{\theta,\eta})$. Using again hypothesis H_2, we get

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathcal{B}_{\theta,\widehat{\eta}_n}\sum_{i=1}^{n} s_{\widehat{\theta}_n,\widehat{\eta}_n} - \frac{1}{\sqrt{n}}\mathcal{B}_{\theta,\widehat{\eta}_n}\sum_{i=1}^{n} s_{\theta,\widehat{\eta}_n} &= -\mathcal{B}_{\theta,\widehat{\eta}_n} I(\theta,\eta)P_U\sqrt{n}(\widehat{\theta}_n-\theta)+o_P(1)\\ &= -\mathcal{B}_{\theta,\eta} I(\theta,\eta)P_U\sqrt{n}(\widehat{\theta}_n-\theta)+o_P(1)\\ &= -P_U\sqrt{n}(\widehat{\theta}_n-\theta)+o_P(1)\\ &= -\sqrt{n}(\widehat{\theta}_n-\theta)+o_p(1).\end{aligned}$$

Notice that we use the fact that

$$\mathcal{B}_{\theta,\eta}I(\theta,\eta)P_U = U_\theta(U_\theta^t I(\theta,\eta)U_\theta)^{-1}U_\theta^t I(\theta,\eta)U_\theta U_\theta^t = U_\theta U_\theta^t = P_U.$$

Now remark that the term I is given by

$$\begin{aligned}I &= \sqrt{n}(\widehat{\theta}_n-\theta) + (B_{\widehat{\theta}_n,\widehat{\eta}_n}-\mathcal{B}_{\theta,\widehat{\eta}_n})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\widehat{\theta}_n,\widehat{\eta}_n}\\ &+ \frac{1}{\sqrt{n}}\mathcal{B}_{\theta,\widehat{\eta}_n}\sum_{i=1}^{n} s_{\widehat{\theta}_n,\widehat{\eta}_n} - \frac{1}{\sqrt{n}}\mathcal{B}_{\theta,\widehat{\eta}_n}\sum_{i=1}^{n} s_{\theta,\widehat{\eta}_n}\\ &= (B_{\widehat{\theta}_n,\widehat{\eta}_n}-\mathcal{B}_{\theta,\widehat{\eta}_n})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\widehat{\theta}_n,\widehat{\eta}_n} + o_P(1).\end{aligned}$$

Finally using the same continuity arguments as for II and the fact that the class $\{s_{\theta,\eta},\theta\in\Theta,\eta\in H\}$ is Donsker (H_4), this last quantity is $o_P(1)$.

Remark : In large dimensions, it may be difficult to choose a grid of size $n^{-1/2}$. An initial estimator may be obtained as in Cheng (2013) by using a grid of size proportional to $n^{-\psi}$,for some $1/4\le\psi\le 1/2$. In that case the rate of convergence of the initial estimator may be sub-optimal of order $n^{-\psi}$ for some $1/4\le\psi\le 1/2$. If we relax the hypothesis concerning $\widehat{\theta}_n$ to $G(\widehat{\theta}_n)=o_P(n^{-\psi})$, we can use the same iterated construction as Cheng (2013) and define

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} + \frac{1}{n}\sum_{i=1}^{n}\mathcal{B}_{\widehat{\theta}_n^{(i)},\widehat{\eta}_{\widehat{\theta}_n^{(i)},n}} s_{\widehat{\theta}_n^{(i)},\widehat{\eta}_{\widehat{\theta}_n^{(i)},n}}(X_i),\ i=1,\dots,k^*,$$

up to some well chosen iteration number k^* (see Cheng, 2013). We believe that the same procedure as in Cheng (2013) allows to construct efficient estimators of the parameter after a finite number of iterations. Since the calculations are much more involved, we do not pursue this approach here.

4 Examples

We develop several examples of interest in the semiparametric literature to show how the efficiency bound and the efficient score under a given identifiability constraint can be straightforwardly calculated with our method.

Example 4 (continued, nonparametric latent variable models (probit)

Let $Y \sim B(1, F(X^t\theta))$. This corresponds to a latent model in which one considers

$$Y = \begin{cases} 1 \text{ if } U &= X^t\theta - \varepsilon \geq 0 \\ 0 \text{ if } U &= X^t\theta - \varepsilon < 0, \end{cases}$$

where U is the unobserved latent variable and ε is a random variable with distribution F. In this model there are several possibilities to identify the parameter θ (for instance by fixing the expectation and the variance of the distribution F). To exploit what we did before we will rather consider that $E_F\varepsilon = 0$ (or alternatively that there is no constant among the X) and that $\|\theta\|_2 = 1$. Here the nuisance parameter is either $\eta = F$ or f.

The log-likelihood is given by

$$\dot{l}_{\theta,F}(y) = \log(p_{\theta,F}(y)) = y \log F(X^t\theta) + (1-y)\log(1 - F(X^t\theta)).$$

The gradient or score function is given by

$$\begin{aligned} l_{\theta,F} &= X^t \left(y \frac{f(X^t\theta)}{F(X^t\theta)} - (1-y)\frac{f(X^t\theta)}{1 - F(X^t\theta)} \right) \\ &= X^t \frac{f(X^t\theta)(y - F(X^t\theta))}{F(X^t\theta)(1 - F(X^t\theta))}, \end{aligned}$$

yielding

$$I(\theta) = V\left(\dot{l}_{\theta,F}(X,Y)\right).$$

Notice that $I(\theta)$ is of rank $k-1$. Now the tangent space is given by scores associated with some likelihood of the form

$$l_{\theta,F_t}(y) = y \log F_t(X^t\theta) + (1-y)\log(1 - F_t(X^t\theta))$$

that is the tangent space $\dot{\mathbb{P}}_2$ is generated by the gradients

$$\left.\frac{\partial l_{\theta,F_t}}{\partial t}\right|_{t=0}(y) = y\frac{f(X^t\theta)}{F(X^t\theta)} - (1-y)\frac{f(X^t\theta)}{1-F(X^t\theta)}$$
$$= \frac{f(X^t\theta)}{F(X^t\theta)(1-F(X^t\theta)}(y - F(X^t\theta)).$$

The projection onto the orthogonal of the nuisance tangent space is thus given by

$$s_{\theta,F}(y) = \Pi_{\dot{\mathbb{P}}_2^{\perp}}\dot{l}_{\theta,F} = (X - E(X \mid X^t\theta))^t\frac{f(X^t\theta)(y - F(X^t\theta))}{F(X^t\theta)(1-F(X^t))}.$$

Then we can use the same transformation U_θ as before. Indeed consider the Gram-Schmidt orthonormalization (denoted by orth(.))

$$U_\theta = orth\left(\begin{pmatrix} -\frac{\beta_2}{\beta_1} & \cdots & -\frac{\beta_k}{\beta_1} \\ 1 & & \\ & Id & \\ & & 1 \end{pmatrix}\right),$$

then we get the projected efficient score under the constraints given by

$$s^c_{\theta,\eta}(Y,X) = U_\theta U_\theta^t(X - E(X \mid X^t\theta))^t\frac{f(X^t\theta)(Y - F(X^t\theta))}{F(X^t\theta)(1-F(X^t))}.$$

As a consequence, the semiparametric efficiency bound is given by

$$\mathcal{B}_{\theta,\eta} = U_\theta\left(U_\theta^t Var\left((X - E(X \mid X^t\theta))^t\frac{f(X^t\theta)(y - F(X^t\theta))}{F(X^t\theta)(1-F(X^t\theta))}\right)U_\theta\right)^{-1}U_\theta^t$$

and the efficient influence function is given by

$$\mathcal{B}_{\theta,\eta}s^c_{\theta,\eta}(Y,X) = \mathcal{B}_{\theta,\eta}(X - E(X \mid X^t\theta))^t\frac{f(X^t\theta)(Y - F(X^t\theta))}{F(X^t\theta)(1-F(X^t))}.$$

There are several nuisance parameters in this expression that can be estimated non-parametrically.

Example 5 Continue, One-way factor analysis with general residual distributions

If we do not assume the residuals to be Gaussian in example 3 but with general distribution with a density f, we can still obtain the semiparametric

efficient bound as in the Gaussian case. We denote by f' its derivative. Indeed in a regression model, the tangent space of the nuisance parameter (here the distribution of the residuals ϵ) and the tangent space corresponding to the parameters of the model are orthogonal (see Bickel et al., 1998). It follows that the score with respect to $(\mu, a_1, \ldots, a_I)$ is simply given by

$$\dot{l}_\theta = \begin{pmatrix} -\sum_{i,j} \frac{f'}{f}(y_{i,j} - \mu - a_i) \\ -\sum_{j=1}^{n_1} \frac{f'}{f}(y_{1,j} - \mu - a_1) \\ \vdots \\ -\sum_{j=1}^{n_I} \frac{f'}{f}(y_{I,j} - \mu - a_I) \end{pmatrix}.$$

As a consequence, as in the linear model (see Bickel et al., 1998)), the Fisher Information is exactly the same as in example 3 but with $1/\sigma^2$ replaced by $E\left(\frac{f'(\epsilon)}{f(\epsilon)}\right)$ provided that this quantity exists. The calculations are then exactly the same and we conclude that again in a semiparametric model, the means over each experimental unit recentered by the overall mean (satisfying the constraints) are clearly efficient in the considered semiparametric sense.

Example 6 Mixture models (Bickel et al. (1998) p125-141)

Consider an exponential model of the form

$$f(x, \theta, \xi) = \exp(\xi^t T(x, \theta) - b(x, \theta) + c(\theta, \xi)),$$

for $\theta \in R^k$ and $\xi \in \mathbb{R}^l$ and for some functions $T(.,.) : \mathbb{R}^{l+1} \times \mathbb{R}^k \to \mathbb{R}^l$, $b(.,.)$: $\mathbb{R}^{l+1} \times \mathbb{R}^k \to \mathbb{R}$ continuously differentiable in θ, with derivative with respect to θ, $\dot{T}_\theta$ and $\dot{b_\theta}$ and with $c(.,.) : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$.

The observations $X = (Y, Z) \in \mathbb{R} \times \mathbb{R}^l$ are taken from the mixture model

$$\int f(x, \theta, \xi)\eta(\xi)d\xi,$$

where $\eta(\xi)$ is a mixing density that is unknown and which plays the role of the nuisance parameter. Such models cover a lot of models including error in variable models (see below). Most of the time some components in θ are not identifiable (see for instance Example 2 in Bickel et al., 1998). As a consequence, one may ask what are the corresponding efficient scores under identifiability constraints. The efficient score without identification has

been obtained in Bickel et al. (1998), corollary 1, p. 131 under the following additional hypotheses :

(i) the exponential model is full rank, meaning that $V_{f(.,\theta,\xi)}(T(X,\theta))$ is full rank (for any θ and ξ).

(ii) The quantity $\int \|\xi\|^2 g(\xi)d\xi < \infty$ and the maps

$$\int \mid \dot{T_\theta}(,.) \mid^2 f(x,\theta,\xi)\eta(\xi)d\xi < \infty$$

$$\int \mid \dot{b_\theta}(,.) \mid^2 f(x,\theta,\xi)\eta(\xi)d\xi < \infty$$

are continuous in θ.

(iii) The density $\eta(\xi)$ is lower bounded on a open set. This rules out the possibility to have a discrete measure for the mixture probability.

In that case the efficient gradient in $\mathbb{R}^k$ is given by

$$\begin{aligned} s_{\theta,\eta}(X) = &(\dot{T_\theta}(X,\theta) - E(\dot{T_\theta}(X,\theta) \mid T(X,\theta))E(\xi \mid T(X,\theta)) \\ &+ ((\dot{b_\theta}(X,\theta) - E(\dot{b_\theta}(X,\theta) \mid T(X,\theta)) \end{aligned}$$

Here to give a precise meaning to $E(\xi \mid T(X,\theta))$, notice that the joint distribution of (ξ, X) is given by $\eta(\xi)f(x,\theta,\xi)$ so that the conditional distribution of U knowing X is in fact $\eta(\xi)f(x,\theta,\xi)/\int \eta(\xi)f(x,\theta,\xi)d\xi$ so that

$$E(\xi \mid T(X,\theta) = v) = \frac{\int \xi \exp(\xi^t v - b(x,\theta) + c(\theta,\xi))\eta(\xi)d\xi}{\int \exp(\xi^t v - b(x,\theta) + c(\theta,\xi))\eta(\xi)d\xi}.$$

Now, according to the considered model, several identifiability constraint $G : \mathbb{R}^k \to \mathbb{R}^l, G$ with $G(\theta) = 0$ can be introduced. As before if $\dot{G}_\theta = \frac{dG_\theta}{d\theta^t}$ is full rank l and U_θ in $\mathcal{M}_{k,\,k-l}(\mathbb{R})$ is such that $\dot{G}_\theta U_\theta = 0$ and $U_\theta^t U_\theta = I_{k-l}$ then the constrained efficient gradient is given by the projected score

$$U_\theta U_\theta^t s_{\theta,\eta}(X).$$

A particular case of this model is the restricted error in variable models (see p. 127 of Bickel et al., 1998). In the following, we keep the same notations as in this book. Consider the model

$$Y = \alpha + \beta Z^\star + \varepsilon_2,$$

but we observe Z given by

$$Z = Z^\star + \varepsilon_1$$

In the simplest case, it is assumed that the residuals are Gaussian $\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix} \sim N\left(0, \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right)$ and $Z^\star$ has density $\frac{1}{\sigma}\eta(\frac{\cdot}{\sigma})$ with $Var_\eta(H) = 1$. Thus, for $x = (y, z)$ and $\theta = (\alpha, \beta, \sigma_1^2, \sigma_2^2, \sigma^2)$, the likelihood is given by

$$f(x, \theta, \xi) = \frac{1}{2\pi\sigma_1\sigma_2} \exp(-\frac{1}{2\sigma_1^2}(z - \sigma\xi)^2 - \frac{1}{2\sigma_2^2}(y - \alpha - \sigma\beta\xi)^2),$$

yielding $T(x, \theta) = \frac{\sigma z}{\sigma_1^2} + \frac{\sigma\beta(y-\alpha)}{\sigma_2^2}$ and $b(x, \theta) = \frac{z^2}{2\sigma_1^2} + \frac{(y-\alpha)^2}{2\sigma_2^2}$.The nuisance parameter is η, the distribution of the $Z^\star$which is unobserved.

Identifiability (see discussions in (Kendall & Stuart, 1979), chap. 29) can be obtained using the following constraints.

(i) The variance σ_1, σ_2 are proportionals that is $\sigma_1^2{=}c_0\sigma_2^2$, where c_0 is known.

(ii) One of the variance σ_1^2 or σ_2^2 is known.

(iii) The reliability ratio given by $\frac{\sigma^2}{\sigma^2+\sigma_1^2}$ is known. This can also be written $\sigma^2 = c_1\sigma_1^2$, for some known constant c_1 .

In this simple case, it is always possible to plug the identifiability constraints in the likelihood and to make the calculation for a parameter of dimension 4 instead of the original dimension. However, our approach is more direct and only requires calculating the derivatives of T and b according to θ which is immediate. For each of these restrictions the calculation of the matrix U_θ are respectively given by

(i)

$$U_\theta = \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & \frac{c_0}{\sqrt{1+c_0^2}} & 0\\ 0 & 0 & \frac{1}{\sqrt{1+c_0^2}} & 0\\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(ii)

$$U_\theta = \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(iii)

$$U_\theta = \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & \frac{1}{\sqrt{1+c_1^2}} & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & \frac{c_1}{\sqrt{1+c_1^2}} & 0 \end{pmatrix}.$$

The efficiency bound and the efficient influence function can be easily computed explicitly in all these cases.

Example 7 Single index model

Consider the model

$$Y = g(X^t\beta) + \varepsilon,$$

where, as in example 2, X is the set of explanatory variable and ε some residual independent of X, with density f.

Recall that a single index model (or pursuit regression model) is in between the pure linear model and a nonparametric model (which would suffer from the curse of dimension when X is of large dimension). We look at the direction that is linked to the response variable Y through a nonlinear function g. But since g is a real function, we expect to estimate it at a reasonable non-parametric rate. We assume that g has derivative g'.

Notice that, as in example 2, g is not identifiable up to a constant term. In general, it is assumed that the first component of β is positive and that $\|\beta\|_2 = 1$. The parameter of interest is $\theta = \beta$ and the nuisance parameter is given by $\eta = (f, g)$.

The log-likelihood of one observation is given by

$$l_{\theta,\eta}(y, x) = \log f(y - g(x^t\beta)).$$

It follows that the gradient with respect to β is given by

$$\dot{l}_{\theta,\eta}(y, x) = xg'(x^t\beta)\frac{f'(y - g(x^t\beta))}{f(y - g(x^t\beta))} = xg'(x^t\beta)\frac{f'(\varepsilon)}{f(\varepsilon)}.$$

Similarly considering a path $\eta_t = (f_t, g_t)$ in H the tangent space to the nuisance parameters $\dot{P}_2$ is given by functions of the form

$$\begin{aligned}\frac{d}{dt}\log f_t(y - g_t(x^t\beta))\mid_{t=0} &= g_0^{(1)}(x^t\beta)\frac{f'(\varepsilon)}{f(\varepsilon)} \\ &= b(x^t\beta)a(\varepsilon),\end{aligned}$$

where a is a (centered) score function at ε.

It follows that we have the decomposition

$$\dot{l}_{\theta,\eta}(y, x) = E(x \mid x^t\beta)^t g'(x^t\beta)\frac{f'(\varepsilon)}{f(\varepsilon)} + (x - E(x \mid x^t\beta))g'(x^t\beta)\frac{f'(\varepsilon)}{f(\varepsilon)},$$

where the first term belong to the tangent space $\dot{P}_2$ and the second one is in the orthogonal space . It follows that the efficient score without constraints is

given by

$$s_{\theta,\eta} = (x - E(x \mid x^t\beta))g'(x^t\beta)\frac{f'(\varepsilon)}{f(\varepsilon)}.$$

Notice that $\beta^t s_{\theta,\eta} = 0$ since $(x - E(x \mid x^t\beta))^t\beta = x^t\beta - E(x^t\beta \mid x^t\beta) = 0$. It follows that the Fisher information is degenerate (due to the non-identifiability of the parameter β). Following what was done before, consider the Gram-Schmidt orthonormalization (which can be computed explicitly with R or Mathematica) matrix given by

$$U_\theta = orth\begin{pmatrix}\begin{pmatrix} -\frac{\beta_2}{\beta_1} & \cdots & -\frac{\beta_k}{\beta_1} \\ 1 & & \\ & Id & \\ & & 1 \end{pmatrix}\end{pmatrix}.$$

Then the constrained efficient score is given by

$$s^c_{\theta,\eta} = U_\theta U^t_\theta (x - E(x \mid x^t\beta))g'(x^t\beta)\frac{f'(\varepsilon)}{f(\varepsilon)}$$

and the efficiency bound is given by

$$\mathcal{B}_{\theta,\eta} = U_\theta (U^t_\theta V((x - E(x \mid x^t\beta))g'(x^t\beta))U_\theta)^{-1}U^t_\theta / I_f.$$

where

$$I_f = V\left(\frac{f'(\varepsilon)}{f(\varepsilon)}\right).$$

The corresponding efficient influence function is then given by

$$\psi_{\theta,\eta} = \mathcal{B}_{\theta,\eta}s^c_{\theta,\eta}.$$

Example 8 **Partially linear single-index models**

In this model, we have a structure of the form

$$Y = x^t\alpha + g(x^t\beta) + \varepsilon$$

where now the parameter of interest is $\theta = (\alpha, \beta)^t,\ \alpha \in \mathbb{R}^p,\ \beta \in \mathbb{R}^p$.

As in example 4, g is not identifiable up to a constant term. Similarly, it is assumed that the first (non-null) component of β is positive and that $\|\beta\|_2 = 1$. Moreover we need additional orthogonality constraints between α and β to ensure that part of the linear term can not be "swallowed" by the function g. We thus assume that $\alpha^t\beta = 0$

If ε has density f then the log-likelihood of one observation is given by

$$l_{\theta,\eta}(y,x) = \log f(y - x^t\alpha - g(x^t\beta)).$$

It follows that the gradient with respect to β is given by

$$\dot{l}_{\theta,\eta}(y,x) = \begin{pmatrix} x\frac{f'(y-x^t\alpha-g(x^t\beta))}{f(y-x^t\alpha-g(x^t\beta))} \\ xg'(x^t\beta)\frac{f'(y-x^t\alpha-g(x^t\beta))}{f(y-x^t\alpha-g(x^t\beta))} \end{pmatrix} = \begin{pmatrix} x \\ xg'(x^t\beta) \end{pmatrix} \frac{f'(\varepsilon)}{f(\varepsilon)}.$$

Similarly considering a path $\eta_t = (f_t, g_t)$ in H the tangent space to the nuisance parameters $\dot{P}_2$ is given by functions of the form

$$\begin{aligned} \frac{d}{dt}\log f_t(y - x^t\alpha + g_t(x^t\beta)) \mid_{t=0} &= g'(x^t\beta)\left.\frac{f_t'(\varepsilon)}{f_t(\varepsilon)}\right|_{t=0} \\ &= b(x^t\beta)a(\varepsilon) \end{aligned}$$

where a is a (centered) score function at ε. It follows that we have the decomposition

$$\dot{l}_{\theta,\eta}(y,x) = \begin{pmatrix} E(x \mid x^t\beta) \\ E(x \mid x^t\beta)g'(x^t\beta) \end{pmatrix} \frac{f'(\varepsilon)}{f(\varepsilon)} + \begin{pmatrix} x - E(x \mid x^t\beta) \\ (x - E(x \mid x^t\beta))g'(x^t\beta) \end{pmatrix} \frac{f'(\varepsilon)}{f(\varepsilon)}$$

where the first term belong to the tangent space $\dot{P}_2$ and the second one is in the orthogonal space . Thus now the efficient score without constraints is given by

$$s_{\theta,\eta} = \begin{pmatrix} x - E(x \mid x^t\beta) \\ (x - E(x \mid x^t\beta))g'(x^t\beta) \end{pmatrix} \frac{f'(\varepsilon)}{f(\varepsilon)}.$$

Notice that $\beta^t s_{\theta,\eta} = 0$ since $(x - E(x \mid x^t\beta))^t\beta = x^t\beta - E(x^t\beta \mid x^t\beta) = 0$. Similarly if α is not orthogonal to β then consider the orthogonal decomposition $\alpha = P_\beta\alpha + P_{\beta^\perp}\alpha$ and $P_\beta = \frac{\beta^t\alpha}{\alpha^t\alpha}\beta$ and we have that $(P_\beta\alpha)^t s_{\theta,\eta} = 0$ (this shows that we need $\beta^t\alpha = 0$ to fully identify α). Now the identifiability constraints are given by the function $G(\theta) = \begin{pmatrix} \alpha^t\alpha - 1 \\ \beta^t\beta - 1 \\ \alpha^t\beta, \end{pmatrix} = 0$ yielding

$$\dot{G}_\theta = \begin{pmatrix} 2\alpha^t & 0 \\ 0 & 2\beta^t \\ \beta^t & \alpha^t \end{pmatrix}$$

$$\dot{G}_\theta U_\theta = 0.$$

First notice that the ratio $\frac{\beta_i}{\alpha_i}$ can not always be the same (else it would contradict the fact that the α and β are orthogonal). So up to some renumbering assume that $\alpha_1 \neq 0$, $\beta_1 \neq 0$ and that $\frac{\beta_1}{\alpha_1}\alpha_2 \neq \beta_2$ then we should

find a space $\left\{ \begin{pmatrix} x \\ y \end{pmatrix}, \alpha^t x = 0, \beta^t y = 0, \beta^t x + \alpha^t y = 0, x \in \mathbb{R}^p, y \in \mathbb{R}^p \right\}$. This amounts to finding the kernel of the projection on the columns of the matrix

$$A = \begin{pmatrix} \alpha & 0 & \beta \\ 0 & \beta & \alpha \end{pmatrix},$$

which is given by $\Pi = A(A^tA)^{-1}A$. Some straightforward calculations show that we have

$$\Pi = \begin{pmatrix} \alpha\alpha^t + \beta\beta^t/2 & \beta\alpha^t/2 \\ \alpha\beta^t/2 & \beta\beta^t + \alpha\alpha^t/2 \end{pmatrix}.$$

The projection to the orthogonal complement of the column of A is given by

$$\Pi^\perp = I_{2p} - \Pi.$$

We essentially need a base of the kernel of the matrix Π. Notice that $\alpha\alpha^t + \beta\beta^t/2$ and $\beta\beta^t + \alpha\alpha^t/2$ are of rank 2 (since α and β are orthogonal). So if we take the Gram-Schmidt orthogonalization of $\Pi^\perp$ and only take the non-null $2p-3$ components, then, we get an orthonormal basis of the kernel of Π and thus the matrix U_θ.

The constrained efficient score is now given by

$$s^c_{\theta,\eta} = U_\theta U_\theta^t (x - E(x \mid x^t\beta)) g'(x^t\beta) \frac{f'(\varepsilon)}{f(\varepsilon)}.$$

Moreover, the efficiency bound is given by

$$\mathcal{B}_{\theta,\eta} = U_\theta (U_\theta^t V((x - E(x \mid x^t\beta))g'(x^t\beta))U_\theta)^{-1} U_\theta^t / I_f,$$

where

$$I_f = V\left(\frac{f'(\varepsilon)}{f(\varepsilon)}\right).$$

The corresponding efficient influence function is then given by

$$\psi_{\theta,\eta} = \mathcal{B}_{\theta,\eta} s^c_{\theta,\eta}.$$

5 Conclusion

In this paper, we consider some regular parametric and semiparametric models, which may not be strictly identifiable. Such models appear naturally in many applications, including probit, mixture, single-index models, or even ANOVA-type models. We define a notion of degenerate efficiency bound for parameters that may be not identifiable. For this, we require the existence of

an identifiability constraint. We obtain new convolution theorems for locally regular statistics estimating these parameters and show how these bounds may be calculated explicitly in many usual models. Since the exact computation of the bound requires the computation of the Gram-Schmidt orthogonalization of a well-chosen matrix, we give the explicit form of this matrix for many interesting identifying constraints. These calculations may be adapted to other interesting settings. In turn, the computation of the efficient score and of the efficiency bounds allows building one-step estimators (satisfying approximately the constraint) that attain the efficiency bound.

References

Bahadur, R.R. (1960). On the Asymptotic Efficiency of Tests and Estimates. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *22*(3/4), 229–252. Retrieved 2023-01-11, from http://www.jstor.org/stable/25048458 (Publisher: Springer)

Bertail, P. (2006). Empirical likelihood in some semiparametric models. *Bernoulli*, *12*(2), 299–331. Retrieved 2023-02-14, from http://www.jstor.org/stable/25464804

Bickel, P., Klaassen, C., Ritov, Y., Wellner, J. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer New York. Retrieved from https://books.google.fr/books?id=lSnTm6SC_SMC

Cam, L. (1960). *Locally asymptotically normal families of distributions. certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses.* Berkeley & Los Angeles.

Cheng, G. (2013, September). How Many Iterations are Sufficient for Efficient Semiparametric Estimation?: k-step semiparametric estimation. *Scandinavian Journal of Statistics*, *40*(3), 592–618. Retrieved 2023-01-11, from http://doi.wiley.com/10.1002/sjos.12005

10.1002/sjos.12005

Cheng, G., & Kosorok, M.R. (2009, March). The penalized profile sampler. *Journal of Multivariate Analysis*, *100*(3), 345–362. Retrieved 2023-01-11, from https://linkinghub.elsevier.com/retrieve/pii/S0047259X08001322

10.1016/j.jmva.2008.05.001

Fukumizu, K. (1996). A Regularity Condition of the Information Matrix of a Multilayer Perception. *Neural networks*, *9*(3), 1871-1879.

Fukumizu, K. (2003, June). Likelihood ratio of unidentifiable models and multilayer neural networks. *The Annals of Statistics*, *31*(3). Retrieved 2023-01-11, from https://projecteuclid.org/journals/annals-of-statistics/volume-31/issue-3/Likelihood-ratio-of-unidentifiable-models-and-multilayer-neural-networks/10.1214/aos/1056562464.full

10.1214/aos/1056562464

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (4th ed., Vol. 2: Inference and Relationship). London: Charles Griffin.

Klaassen, C.A., & Susyanto, N. (2019). Semiparametrically efficient estimation of euclidean parameters under equality constraints. *Journal of Statistical Planning and Inference*, *201*, 120-132. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378375818303598

https://doi.org/10.1016/j.jspi.2018.12.005

Kosorok, M.R. (2008). *Introduction to empirical processes and semiparametric inference.* New York: Springer.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proceedings of the third berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics* (Vol. 3, pp. 129–157).

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* New York, NY: Springer New York. Retrieved 2023-01-11, from http://link.springer.com/10.1007/978-1-4612-4946-7 10.1007/978-1-4612-4946-7

Ma, Y., & Zhu, L. (2013, February). Efficient estimation in sufficient dimension reduction. *The Annals of Statistics*, *41*(1). Retrieved 2023-01-11, from https://projecteuclid.org/journals/annals-of-statistics/volume-41/issue-1/Efficient-estimation-in-sufficient-dimension-reduction/10.1214/12-AOS1072.full

10.1214/12-AOS1072

Rothenberg, T.J. (1971, May). Identification in Parametric Models. *Econometrica*, *39*(3), 577. Retrieved 2023-01-12, from https://www.jstor.org/stable/1913267?origin=crossref

10.2307/1913267

Severini, T.A., & Wong, W.H. (1992). Profile likelihood and conditionally parametric models. *The Annals of Statistics*, *20*(4), 1768–1802. Retrieved 2023-01-19, from http://www.jstor.org/stable/2242367

Shen, X. (2001). On bahadur efficiency and maximum likelihood estimation in general parameter spaces. *Statistica Sinica*, *11*(2), 479–498. Retrieved 2023-01-12, from http://www.jstor.org/stable/24306873

Stoica, P., & Ng, B.C. (1998, July). On the Cramer-Rao bound under parametric constraints. *IEEE Signal Processing Letters*, *5*(7), 177–179. Retrieved 2023-01-11, from http://ieeexplore.ieee.org/document/700921/

10.1109/97.700921

Susyanto, N., & Klaassen, C.A.J. (2017, January). Semiparametrically efficient estimation of constrained Euclidean parameters. *Electronic Journal of Statistics*, *11*(2). Retrieved 2023-01-11, from https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue-2/Semiparametrically-efficient-estimation-of-constrained-Euclidean-parameters/10.1214/17-EJS1308.full

10.1214/17-EJS1308

van der Vaart, A. (1989). On the asymptotic information bound. *The Annals of Statistics*, *17*(4), 1487–1500.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press. Retrieved from https://books.google.fr/books?id=UEuQEM5RjWgC