Statistics & Decisions 1, 1– 4 (2003)

On the construction of efficient estimators in semiparametric models

J. Forrester, W. Hooper, H. Peng, A. Schick

Summary: This paper deals with the construction of efficient estimators in semiparametric models without the sample splitting technique. Schick (1987) gave sufficient conditions using the leave-one-out technique for a construction without sample splitting. His conditions are stronger and more cumbersome to verify than the necessary and sufficient conditions for the existence of efficient estimators which suffice for the construction based on sample splitting. In this paper we use a conditioning argument to weaken Schick's conditions. We shall then show that in a large class of semiparametric models and for properly chosen estimators of the score function the resulting weaker conditions reduce to the minimal conditions for the construction with sample splitting. In other words, in these models efficient estimators can be constructed without sample splitting under the same conditions as those used for the construction with sample splitting. We demonstrate our results by constructing an efficient estimator using these ideas in a semiparametric additive regression model.

1 Introduction

Bickel [2] used sample splitting techniques to give a general procedure for constructing adaptive estimators in semiparametric models. His construction is essentially an existence result as only a small part of the sample was used to construct the influence function. The moderate sample behavior of his construction is not expected to be good. The sample splitting idea was further developed by Schick [11]. He used a symmetrization argument to give a procedure for the construction of efficient estimators in semiparametric models using two estimators of the efficient score function each based on about half the sample. This construction works under minimal assumptions as shown by Klaassen [7] who demonstrated that Schick's sufficient conditions are also necessary. These conditions require the estimate of the efficient score function to be consistent in the L_2 norm and its "mean" to converge to zero fast enough. Recently, Schick [21] has generalized the sample splitting approach to semiparametric Markov chain models.

Schick [12] gave sufficient conditions for the construction of efficient estimates without sample splitting. His conditions strengthen the consistency condition and impose addi-

AMS 2000 subject classification. 62G05, 62G20

Key words and phrases: sample splitting, leave-one-out estimator, efficient score function, linear regression, partly linear regression model, linear smoothers, under-smoothing.

tional conditions that measure the influence of the individual observations on the estimator of the efficient score function. The latter conditions which require dropping observations from the estimator of the score function can be quite cumbersome to verify. One expects procedures that avoid sample splitting to perform better in moderate sample sizes. This has been substantiated by simulations. However, sample splitting has remained useful due to the simpler conditions.

In this paper we shall use conditioning arguments to relax Schick's [12] sufficient conditions. Conditioning was already utilized by Schick [13, 14] to construct efficient estimators in semiparametric regression models. He conditioned on the covariates to simplify the conditions. Here we push this idea further and condition on transformations of the data which are not necessarily observable anymore. We shall see that if the efficient score function is of a certain type and its estimator is carefully selected, then the additional conditions associated with dropping observations are automatically satisfied. We thus arrive at conditions which are essentially those used in the constructions using sample splitting. Van der Vaart [22] has shown that for a class of semiparametric models with a special structure efficient estimates can be constructed without sample splitting under almost minimal conditions. Our results improve on those of van der Vaart [22] and are applicable to a considerably larger class of models.

We shall formulate our results more generally for the construction of estimators with a desired influence function. If the desired influence function is the efficient influence function, then our construction yields an efficient estimator. However, in some cases one might be interested in influence functions other than the efficient one, say for robustness reasons, and then our construction results in a robust estimator.

Let us now illustrate the idea behind our approach. At the heart of constructing efficient estimators is the following problem. Given independent and identically distributed q-dimensional random vectors $X_{n1}, \dots, X_{nn}$, provide conditions on the function h_n from $\mathbb{R}^{(n+1)q}$ into $\mathbb{R}^m$ that imply

$$H_n := \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_n(X_{nj}, X_{n1}, \dots, X_{nn}) = o_p(1). \tag{1.1}$$

In applications, $h_n(\cdot, X_{n1}, \dots, X_{nn})$ is the difference between the estimated and actual score functions. Now suppose that we can write

$$h_n(X_{nj}, X_{n1}, \dots, X_{nn}) = \sum_{k=1}^{K} g_{knj}(Y_{knj}, Z_{kn1}, \dots, Z_{knn}), \quad j = 1, \dots, n,$$

where, for each k, $(Y_{kn1}, Z_{kn1}), \dots, (Y_{knn}, Z_{knn})$ are independent random vectors of dimension $q_k + p_k$ and $g_{kn1}, \dots, g_{knn}$ are measurable functions from $\mathbb{R}^{q_k} \times \mathbb{R}^{np_k}$ into $\mathbb{R}^m$. Then we have

$$H_n = G_{n1} + \cdots + G_{nK} \tag{1.2}$$

and can use the following basic lemma to treat the terms

$$G_{nk} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} g_{knj}(Y_{knj}, Z_{kn1}, \dots, Z_{knn}).$$

Basic Lemma. *Let $(Y_{n1}, Z_{n1}), \ldots, (Y_{nn}, Z_{nn})$ be independent random vectors of dimension $q+p$ and $g_{n1}, \ldots, g_{nn}$ be measurable functions from $\mathbb{R}^q \times \mathbb{R}^{np}$ into $\mathbb{R}^m$ such that*

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int g_{nj}(y, Z_{n1}, \ldots, Z_{nn}) F_{nj}(dy) = o_p(1) \tag{1.3}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} \int \|g_{nj}(y, Z_{n1}, \ldots, Z_{nn})\|^2 F_{nj}(dy) = o_p(1), \tag{1.4}$$

where $F_{nj}(dy) = F_n(dy \mid Z_{nj})$ is the conditional distribution of Y_{nj} given Z_{nj}. Then

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} g_{nj}(Y_{nj}, Z_{n1}, \ldots, Z_{nn}) = o_p(1). \tag{1.5}$$

Proof: Let $\mathbf{Z}_n = (Z_{n1}, \ldots, Z_{nn})$ and $D_{nj} = g_{nj}(Y_{nj}, \mathbf{Z}_n) - \int g_{nj}(y, \mathbf{Z}_n) F_{nj}(dy)$. In view of (1.3), it suffices to show that $\frac{1}{\sqrt{n}} \sum_{j=1}^{n} D_{nj} = o_p(1)$. By construction, $E(D_{ni}^T D_{nj} \mid \mathbf{Z}_n) = 0$ for $i \neq j$, so that

$$\begin{aligned} E(\|\frac{1}{\sqrt{n}} \sum_{j=1}^{n} D_{nj}\|^2 \mid \mathbf{Z}_n) &= \frac{1}{n} \sum_{j=1}^{n} E(\|D_{nj}\|^2 \mid \mathbf{Z}_n) \\ &\leq \frac{1}{n} \sum_{j=1}^{n} \int \|g_{nj}(y, \mathbf{Z}_n)\|^2 F_{nj}(dy) = o_p(1). \end{aligned}$$

This gives the desired result. □

In some applications we even have

$$\int g_{nj}(y, Z_{n1}, \ldots, Z_{nn}) F_{nj}(dy) = 0, \quad j = 1, \ldots, n,$$

which, of course, implies (1.3). To obtain the representation (1.2) one needs a certain structure for the score function and has to choose an appropriate estimator of the score function.

Our paper is organized as follows. In Section 2 we give an overview of the various constructions used in the literature and present a new result which, with the aid of conditioning arguments, relaxes the conditions given by Schick [12]. This result simplifies to the above Basic Lemma under appropriate structural assumptions and properly chosen estimates of the score function. In this case one no longer has to verify the conditions related to dropping observations. This is addressed in Section 3, where we discuss two structures for the score function for which this is possible. These structures contain the examples of Bickel [2] and the class of models considered by van der Vaart [22]. In Section 4 we generalize these results to a more complicated but frequently occurring type of score function.

There we need the full power of our new approach. In Section 5 we shall use these results and construct an efficient estimator for a semiparametric additive regression model, the so called partly linear model. The construction given there improves on various earlier constructions, by avoiding sample splitting and by working under minimal assumptions on the error density. We only require this density to have finite Fisher information for location. A preliminary estimator for the parameter of interest for this model is constructed in Section 6. There we generalize results of Zhao [23] on bandwidth-matched M-estimation for such models. Section 7 gives a proof of Theorem 2.6, while Section 8 collects some technical details for consistent estimation of the score function of the location model.

2 An overview of construction methods

Let $(\Omega, \mathfrak{A})$ and $(\mathfrak{X}, \mathfrak{B})$ be two measurable spaces and $\xi_1, \dots, \xi_n$ be measurable functions from Ω into $\mathfrak{X}$. Furthermore, let Θ be an open subset of $\mathbb{R}^k$ and Γ be an arbitrary set. For each $(\vartheta, \gamma) \in \Theta \times \Gamma$, let $P_{\vartheta,\gamma}$ be a probability measure on $\mathfrak{A}$ for which $\xi_1, \dots, \xi_n$ are independent and identically distributed with common distribution $F_{\vartheta,\gamma}$, let $L_{\vartheta,\gamma}$ be a measurable function from $\mathfrak{X}$ into $\mathbb{R}^k$ such that

$$\int L_{\vartheta,\gamma}\, dF_{\vartheta,\gamma} = 0 \quad \text{and} \quad \int \|L_{\vartheta,\gamma}\|^2\, dF_{\vartheta,\gamma} < \infty,$$

and let $\Lambda(\vartheta, \gamma)$ be a $k \times k$ matrix. Let (ϑ_0, γ_0) denote a fixed (but unknown) point in $\Theta \times \Gamma$. To simplify notation we suppress dependence on ϑ_0 and γ_0 whenever possible. In particular, we set $P = P_{\vartheta_0,\gamma_0}$, $P_\vartheta = P_{\vartheta,\gamma_0}$, $F_\vartheta = F_{\vartheta,\gamma_0}$, $\Lambda(\vartheta) = \Lambda(\vartheta, \gamma_0)$, $L_\vartheta = L_{\vartheta,\gamma_0}$ and $L(x, \vartheta) = L_\vartheta(x)$. We write E_ϑ for the expectation under P_ϑ. By a local sequence we mean a sequence $\{\vartheta_n\}$ in Θ such that $n^{1/2}(\vartheta_n - \vartheta_0)$ is bounded.
We are interested in constructing functions t_n from $\mathfrak{X}^n$ to $\mathbb{R}^k$ such that the estimator $T_n = t_n(\xi_1, \dots, \xi_n)$ of the Euclidean parameter has influence function $\Lambda(\vartheta_0)L_{\vartheta_0}$ under P_{ϑ_0}, i.e.,

$$T_n = \vartheta_0 + \frac{1}{n}\sum_{j=1}^{n} \Lambda(\vartheta_0)L(\xi_j, \vartheta_0) + o_P(n^{-1/2}). \tag{2.1}$$

We shall do so under the following additional assumptions.

Assumption 2.1 *We have at our disposal a $n^{1/2}$-consistent estimator $\tilde{\vartheta}_n$ of the Euclidean parameter, i.e., $\tilde{\vartheta}_n = \tilde{t}_n(\xi_1, \dots, \xi_n)$ for some measurable function $\tilde{t}_n$ from $\mathfrak{X}^n$ into $\mathbb{R}^k$ such that*

$$n^{1/2}(\tilde{\vartheta}_n - \vartheta_0) = O_P(1).$$

Moreover, this estimator is Θ-valued and discretized, i.e., $\tilde{\vartheta}_n$ takes only values in the grid $\{n^{-1/2}lc : l \in \{\dots, -2, -1, 0, 1, 2, \dots\}^k\} \cap \Theta$ for some positive c.

Assumption 2.2 *The sequences $\{F^n_{\vartheta_n}\}$ and $\{F^n_{\vartheta_0}\}$ of product measures are mutually contiguous for every local sequence $\{\vartheta_n\}$.*

Assumption 2.3 *The maps $\vartheta \mapsto \Lambda(\vartheta)$ and $\vartheta \mapsto \int \|L_\vartheta\|^2 \, dF_\vartheta$ are continuous at ϑ_0.*

Assumption 2.4 *For every local sequence $\{\vartheta_n\}$, we have*

$$\vartheta_n + \frac{1}{n}\sum_{j=1}^{n} \Lambda(\vartheta_n) L(\xi_j, \vartheta_n) = \vartheta_0 + \frac{1}{n}\sum_{j=1}^{n} \Lambda(\vartheta_0) L(\xi_j, \vartheta_0) + o_P(n^{-1/2}).$$

These assumptions are standard in the construction of efficient estimators in semiparametric models, see e.g. Bickel [2] and Schick [11, 12]. In this context, $L_{\vartheta,\gamma}$ is the efficient score function and $\Lambda(\vartheta, \gamma)$ is the inverse of the efficient information matrix

$$J(\vartheta, \gamma) = \int L_{\vartheta,\gamma} L_{\vartheta,\gamma}^\top \, dF_{\vartheta,\gamma}. \tag{2.2}$$

The idea of discretization goes back to Le Cam [8] and has now become a standard technical tool. Discretized $n^{1/2}$-consistent estimates can be treated as if they were non-stochastic sequences in the proof. Combined with contiguity arguments this often leads to considerable simplifications in the proofs.
Under the above assumptions it suffices to construct measurable functions z_n from $\Theta \times \mathfrak{X}^n$ into $\mathbb{R}^k$ and Λ_n from $\Theta \times \mathfrak{X}^n$ into $\mathbb{R}^{k\times k}$, the set of $k \times k$ matrices, such that

$$z_n(\vartheta_n, \xi_1, \ldots, \xi_n) = \frac{1}{n}\sum_{j=1}^{n} L(\xi_j, \vartheta_n) + o_{P_{\vartheta_n}}(n^{-1/2}) \tag{2.3}$$

and

$$\Lambda_n(\vartheta_n, \xi_1, \ldots, \xi_n) = \Lambda(\vartheta_n) + o_{P_{\vartheta_n}}(1) \tag{2.4}$$

for every local sequence $\{\vartheta_n\}$. As $\tilde{\vartheta}_n$ is discrete and $n^{1/2}$-consistent, we obtain from Le Cam's discretization argument that T_n defined by

$$T_n = \tilde{\vartheta}_n + \Lambda_n(\tilde{\vartheta}_n, \xi_1, \ldots, \xi_n) z_n(\tilde{\vartheta}_n, \xi_1, \ldots, \xi_n)$$

satisfies the desired (2.1).
The requirement (2.4) can usually be established via the plug-in principle. Indeed, suppose that Γ is endowed with a metric δ and that $(\vartheta, \gamma) \mapsto \Lambda(\vartheta, \gamma)$ is continuous at (ϑ_0, γ_0), then (2.4) holds with $\Lambda_n(\vartheta_n, \xi_1, \ldots, \xi_n) = \Lambda(\vartheta_n, \hat{\gamma}_n)$ whenever $\hat{\gamma}_n = \gamma_n(\xi_1, \ldots, \xi_n)$ satisfies $\delta(\hat{\gamma}_n, \gamma) = o_{P_{\vartheta_n}}(1)$. In the case of efficient estimation, $\Lambda(\vartheta, \gamma)$ is the inverse of $J(\vartheta, \gamma)$ defined in (2.2). Suppose now that

$$\frac{1}{n}\sum_{j=1}^{n} L(\xi_j, \vartheta_n) L^\top(\xi_j, \vartheta_n) = J(\vartheta_n, \gamma_0) + o_{P_{\vartheta_n}}(1)$$

and that there are functions L_n from $\mathfrak{X} \times \Theta \times \mathfrak{X}^n$ to $\mathbb{R}^k$ such that

$$\frac{1}{n}\sum_{j=1}^{n} \|L_n(\xi_j, \vartheta_n, \xi_1, \ldots, \xi_n) - L(\xi_j, \vartheta_n)\|^2 = o_{P_{\vartheta_n}}(1).$$

Then we have

$$\frac{1}{n}\sum_{j=1}^{n} L_n(\xi_j,\vartheta_n,\xi_1,\dots,\xi_n)L_n^\top(\xi_j,\vartheta_n,\xi_1,\dots,\xi_n) = J(\vartheta_n,\gamma_0)+o_{P_{\vartheta_n}}(1)$$

and obtain (2.4) from the continuity of matrix inversion.
The more difficult part is of course (2.3). Bickel [2] was the first to tackle this problem in generality. Motivated by earlier work of Hájek, he employed a sample splitting scheme in which only a small initial part of the sample was used to estimate the score function $L_{\vartheta,\gamma}$ and the other observations were reserved for averaging this estimator. More precisely, his z_n is of the form

$$z_n(\vartheta,\xi_1,\dots,\xi_n) = \frac{1}{n-m}\sum_{j=m+1}^{n} L_m(\xi_j,\vartheta,\xi_1,\dots,\xi_m)$$

where m increases with n in such a way that $m\to\infty$ and $m/n\to 0$. Let

$$\hat L_n(x,\vartheta) = L_n(x,\vartheta,\xi_1,\dots,\xi_n).$$

Bickel proved (2.3) under the following assumptions:

$$\int \hat L_n(x,\vartheta_n)\,dF_{\vartheta_n}(x) = 0 \tag{B1}$$

$$\int \|\hat L_n(x,\vartheta_n)-L(x,\vartheta_n)\|^2\,dF_{\vartheta_n}(x) = o_{P_{\vartheta_n}}(1). \tag{B2}$$

Schick [11] symmetrized the above construction. His construction calls for two estimates of the score function each based on about half the sample and uses the corresponding other half for averaging. More precisely, his z_n is of the form

$$z_n(\vartheta,\xi_1,\dots,\xi_n) = \frac{1}{n}\left(\sum_{j=1}^{n_1} L_{n_2}(\xi_j,\vartheta,\xi_{n_1+1},\dots,\xi_n) + \sum_{j=n_1+1}^{n} L_{n_1}(\xi_j,\vartheta,\xi_1,\dots,\xi_{n_1})\right)$$

where $n_1+n_2=n$ and $n_1/n\to 1/2$. He proved (2.3) under a weaker set of assumptions, namely under (B2) and

$$\int \hat L_n(x,\vartheta_n)\,dF_{\vartheta_n}(x) = o_{P_{\vartheta_n}}(n^{-1/2}). \tag{S1}$$

It was shown by Klaassen [7] that these two conditions are also necessary.
While sample splitting works from a theoretical point of view, it is undesirable from a practical point of view in moderate samples. Schick [12] has shown that sample splitting can be avoided under additional assumptions. He considered z_n of the form

$$z_n(\vartheta,\xi_1,\dots,\xi_n) = \frac{1}{n}\sum_{j=1}^{n}\hat L_n(\xi_j,\vartheta) \tag{2.5}$$

and proved (2.3) under (S1), under the following strengthening of (B2):

$$E_{\vartheta_n}\Big[\int \|\hat{L}_n(x,\vartheta_n) - L(x,\vartheta_n)\|^2\, dF_{\vartheta_n}(x)\Big] = o(1), \tag{S2}$$

and under the following two additional conditions:

$$\frac{1}{n}\sum_{j=1}^{n}\Big(\hat{L}_n(\xi_j,\vartheta_n) - \hat{L}_{n,-j}(\xi_j,\vartheta_n)\Big) = o_{P_{\vartheta_n}}(n^{-1/2}), \tag{S3}$$

$$\sum_{j=1}^{n} E_{\vartheta_n}\Big[\int \|\hat{L}_n(x,\vartheta_n) - \hat{L}_{n,-j}(x,\vartheta_n)\|^2\, dF_{\vartheta_n}(x)\Big] = o(1), \tag{S4}$$

where

$$\hat{L}_{n,-j}(x,\vartheta_n) = E_{\vartheta_n}(\hat{L}_n(x,\vartheta_n) \mid \xi_1,\dots,\xi_{j-1},\xi_{j+1},\dots,\xi_n).$$

Thus it takes extra effort [(S2)–(S4) rather than (B2)] to avoid sample splitting. For this reason sample splitting has remained useful. See Bhattacharya and Zhao [1], Remark 10, for a comment on this.

Weaker conditions, however, can be derived by conditioning. This was already recognized in Schick [12] and pursued in Schick [13, 14] in the context of regression models by conditioning on the covariates. Let us now formulate a theorem that pushes this idea further. The key is that we can condition not only on observable random variables such as covariates, but generally on random quantities that may even depend on the parameters. This results in weaker versions of (S2)–(S4). We shall see that the weaker versions become automatic in important cases if appropriate estimates are chosen.

Assumption 2.5 *Let $(\mathcal{Y},\mathfrak{C})$ be another measurable space and η be a function from $\mathfrak{X}\times\Theta\times\Gamma$ into $\mathcal{Y}$ measurable in the first argument and such that the conditional distribution of ξ_1 given $\eta(\xi_1,\vartheta,\gamma)$ under $P_{\vartheta,\gamma}$ has a regular version $M_{\vartheta,\gamma}(dx \mid \eta(\xi_1,\vartheta,\gamma))$ for each $(\vartheta,\gamma)\in\Theta\times\Gamma$.*

We then set $\eta_{nj} = \eta(\xi_j,\vartheta_n,\gamma_0)$, abbreviate $M_{\vartheta_n,\gamma_0}(dx \mid \eta_{nj})$ by $M_{nj}(dx)$ and write $E^*_{\vartheta_n}$ for the conditional expectation given $\eta_{n1},\dots,\eta_{nn}$ calculated under P_{ϑ_n}.

Theorem 2.6 *Suppose Assumptions 2.1 to 2.5 hold. Then the following conditions are sufficient for the estimator (2.5) to satisfy (2.3):*

$$\frac{1}{n}\sum_{j=1}^{n}\int(\hat{L}_n(x,\vartheta_n) - L(x,\vartheta_n))M_{nj}(dx) = o_{P_{\vartheta_n}}(n^{-1/2}), \tag{C1}$$

$$\frac{1}{n}\sum_{j=1}^{n} E^*_{\vartheta_n}\Big(\int \|\hat{L}_n(x,\vartheta_n) - L(x,\vartheta_n)\|^2 M_{nj}(dx)\Big) = o_{P_{\vartheta_n}}(1), \tag{C2}$$

$$\frac{1}{n}\sum_{j=1}^{n}\Big(\hat{L}_n(\xi_j,\vartheta_n) - \hat{L}_{nj}(\xi_j,\vartheta_n)\Big) = o_{P_{\vartheta_n}}(n^{-1/2}), \tag{C3}$$

$$\frac{1}{n}\sum_{i\neq j}\sum E^*_{\vartheta_n}\Big(\int \|\hat L_n(x,\vartheta_n)-\hat L_{ni}(x,\vartheta_n)\|^2 M_{nj}(dx)\Big)=o_{P_{\vartheta_n}}(1), \tag{C4}$$

$$\sum_{j=1}^n \int \|\hat L_n(x,\vartheta_n)-\hat L_{nj}(x,\vartheta_n)\|^2 M_{nj}(dx)=o_{P_{\vartheta_n}}(1), \tag{C5}$$

where

$$\hat L_{nj}(x,\vartheta_n)=\int L_n(x,\vartheta_n,\xi_1,\ldots,\xi_{j-1},y,\xi_{j+1},\ldots,\xi_n)M_{nj}(dy).$$

A proof of this theorem is in Section 7. The approach outlined in Theorem 2.6 has been implemented to construct efficient estimates in homoscedastic regression by Schick [13, 15, 19] and in heteroscedastic regression by Schick [14, 17] by conditioning on the covariates. The ideas behind this approach have also proved useful in constructing efficient estimates in some time series models (Schick [16, 18, 20]). Let us now comment on the conditions.

Remark 2.7 Note that the conditions (S1)–(S4) follow from (C1)–(C5) upon taking η to be a constant. For this choice of η, $M_{nj}(dx)$ reduces to $F_{\vartheta_n}(dx)$ and (C1)–(C4) become (S1)–(S4), while (C4) implies (C5).

Remark 2.8 The condition (C5) can be omitted if we slightly change (C1). More precisely, (C1) and (C5) can be replaced by

$$\frac{1}{n}\sum_{j=1}^n \int (\hat L_{nj}(x,\vartheta_n)-L(x,\vartheta_n))M_{nj}(dx)=o_{P_{\vartheta_n}}(n^{-1/2}). \tag{C1'}$$

This follows from the fact that we use (C1) and (C5) in the proof only to conclude (C1'). It is easy to see that (C1) and (C5) yield (C1'). Indeed, it follows from (C5) and the Cauchy-Schwarz inequality that

$$\frac{1}{n}\sum_{j=1}^n \int (\hat L_{n}(x,\vartheta_n)-\hat L_{nj}(x,\vartheta_n))M_{nj}(dx)=o_{P_{\vartheta_n}}(n^{-1/2}).$$

This and (C1) yield (C1').

Remark 2.9 A sufficient condition for (C5) is of course

$$\sum_{j=1}^n E^*_{\vartheta_n}\Big(\int \|\hat L_n(x,\vartheta_n)-\hat L_{nj}(x,\vartheta_n)\|^2 M_{nj}(dx)\Big)=o_{P_{\vartheta_n}}(1). \tag{C5'}$$

Sufficient conditions for (C3)-(C5) are obtained if one replaces in (C3),(C4) and (C5') the quantity $\hat L_{nj}$ by a quantity $\hat L^*_{nj}$ of the form

$$\hat L^*_{nj}(x,\vartheta_n)=L_{nj}(x,\vartheta_n,\xi_1,\ldots,\xi_{j-1},\eta_{nj},\xi_{j+1},\ldots,\xi_n).$$

Refer to these conditions as (C3)*, (C4)* and (C5')*. To see that they are sufficient, note that the left-hand sides of (C4) and (C5') are bounded by the left hand sides of (C4)* and (C5')*, respectively, and that (C3) follows from (C3)* and

$$\frac{1}{n}\sum_{j=1}^{n}\Big(\hat{L}^*_{nj}(\xi_j,\vartheta_n)-\hat{L}_{nj}(\xi_j,\vartheta_n)\Big)=o_{P_{\vartheta_n}}(n^{-1/2}).$$

The latter follows as the conditional expectation of the squared norm of this expression given $\eta_{n1},\ldots,\eta_{nn}$ is bounded by the left-hand side of (C5')*. For this note that

$$E_{\vartheta_n}[\|\hat{L}_{nj}(x,\vartheta_n)-\hat{L}^*_{nj}(x,\vartheta_n)\|^2]\le E_{\vartheta_n}[\|\hat{L}_n(x,\vartheta_n)-\hat{L}^*_{nj}(x,\vartheta_n)\|^2].$$

These sufficient conditions are useful when it is cumbersome to calculate $\hat{L}_{nj}$.

Remark 2.10 A sufficient condition for (C4) and (C5) is

$$\sum_{j=1}^{n}\max_{1\le i\le n}E^*_{\vartheta_n}\Big(\int\|\hat{L}_n(x,\vartheta_n)-\hat{L}_{ni}(x,\vartheta_n)\|^2M_{nj}(dx)\Big)=o_{P_{\vartheta_n}}(1),$$

while a sufficient condition for (C3)–(C5) is

$$\max_{1\le j\le n}E^*_{\vartheta_n}\Big(\sup_{x\in\mathfrak{X}}\|\hat{L}_n(x,\vartheta_n)-\hat{L}_{nj}(x,\vartheta_n)\|\Big)=o_{P_{\vartheta_n}}(n^{-1/2}).$$

In these sufficient conditions we can replace $\hat{L}_{nj}$ by $\hat{L}^*_{nj}$ of the previous remark.

Let us now mention another choice for z_n, namely the leave-one-out estimator

$$z_n(\vartheta,\xi_1,\ldots,\xi_n)=\frac{1}{n}\sum_{j=1}^{n}\tilde{L}_{nj}(\xi_j,\vartheta) \tag{2.6}$$

with

$$\tilde{L}_{nj}(x,\vartheta)=L_{n-1}(x,\vartheta,\xi_1,\ldots,\xi_{j-1},\xi_{j+1},\ldots,\xi_n).$$

Such an estimator was used by van der Vaart [22]. The reason for a leave-one-out estimator is technical. Since ξ_j is not used to construct $\tilde{L}_{nj}(\cdot,\vartheta)$, the estimator $\tilde{L}_{nj}(\cdot,\vartheta)$ of $L(\cdot,\vartheta)$ is independent of ξ_j. For the leave-one-out estimator only analogues of (C1), (C2) and (C4) are needed. More precisely, we have the following result.

Theorem 2.11 *Suppose Assumptions 2.1 to 2.5 hold. Then the following conditions are sufficient for the leave-one-out estimator (2.6) to satisfy (2.3):*

$$\frac{1}{n}\sum_{j=1}^{n}\int(\tilde{L}_{nj}(x,\vartheta_n)-L(x,\vartheta_n))M_{nj}(dx)=o_{P_{\vartheta_n}}(n^{-1/2}), \tag{D1}$$

$$\frac{1}{n}\sum_{j=1}^{n}E^*_{\vartheta_n}\Big(\int\|\tilde{L}_{nj}(x,\vartheta_n)-L(x,\vartheta_n)\|^2M_{nj}(dx)\Big)=o_{P_{\vartheta_n}}(1), \tag{D2}$$

$$\frac{1}{n}\sum_{i\neq j}\sum E^*_{\vartheta_n}\Big(\int \|\tilde L_{nj}(x,\vartheta_n)-\tilde L_{nji}(x,\vartheta_n)\|^2 M_{nj}(dx)\Big)=o_{P_{\vartheta_n}}(1), \tag{D3}$$

where $\tilde L_{nji}(x,\vartheta_n)=E_{\vartheta_n}(\tilde L_{nj}(x,\vartheta_n)\mid \xi_1,\dots,\xi_{i-1},\eta_{ni},\xi_{i+1},\dots,\xi_n)$.

Remark 2.12 Instead of the conditional expectation $\tilde L_{nji}(x,\vartheta_n)$ we could take in (D3) any other estimator $\tilde L_{nji}(x,\vartheta_n)$ based on η_{nj} and the variables ξ_k, $k\neq i,j$. This again is helpful when the conditional expectation is difficult to calculate.

Remark 2.13 Each of the above theorems can be generalized to the case when

$$L(x,\vartheta,\gamma)=A^{[1]}(\vartheta,\gamma)L^{[1]}(x,\vartheta,\gamma)+\cdots+A^{[m]}(\vartheta,\gamma)L^{[m]}(x,\vartheta,\gamma)$$

where $A^{[i]}(\vartheta,\gamma)$ is an $k\times k_i$ matrix and L_i is a function similar to L but into $\mathbb{R}^{k_i}$. Let $A^{[i]}(\vartheta)=A^{[i]}(\vartheta,\gamma_0)$ and $L_i(x,\vartheta)=L_i(x,\vartheta,\gamma_0)$ and assume that $\vartheta\mapsto A^{[i]}(\vartheta)$ and $\vartheta\mapsto\int\|L_i(x,\vartheta)\|^2\,dF_\vartheta(x)$ are continuous at ϑ_0. In this case one takes

$$\hat L_n(x,\vartheta)=\hat A^{[1]}_n(\vartheta)L^{[1]}_n(x,\vartheta)+\cdots+\hat A^{[m]}_n(\vartheta)L^{[m]}_n(x,\vartheta)$$

where $\hat A^{[i]}_n(\vartheta)=A_{ni}(\xi_1,\dots,\xi_n)$ estimates $A^{[i]}(\vartheta)$ and $\hat L^{[i]}_n(x,\vartheta)=L_{ni}(x,\vartheta,\xi_1,\dots,\xi_n)$ estimates $L^{[i]}(x,\vartheta)$ under P_ϑ. To get (2.3) for the full estimate it suffices to show that, for $i=1,\dots,m$,

$$\hat A^{[i]}_n(\vartheta_n)=A^{[i]}(\vartheta_n)+o_{P_{\vartheta_n}}(1), \tag{2.7}$$

and

$$\frac{1}{n}\sum_{j=1}^n\hat L^{[i]}_n(\xi_j,\vartheta_n)=\frac{1}{n}\sum_{j=1}^n L^{[i]}(\xi_j,\vartheta_n)+o_{P_{\vartheta_n}}(n^{-1/2}). \tag{2.8}$$

To obtain (2.8) we can apply Theorem 2.6 with an η which may depend on i.

3 The main idea and a first application

In the previous section we have reviewed methods of constructing estimators with a prescribed influence function and have seen that avoiding sample splitting comes with a price of additional conditions. The conditions (C1) and (C2) for the full estimate (2.5), and (D1) and (D2) for the leave-one-out estimator (2.6), are close to the necessary conditions (S1) and (S2). In most instances they are not more difficult to verify than (S1) and (S2). However, (C3)–(C5) for the full estimator, and (D3) for the leave-one-out estimator, can be cumbersome to verify. We shall now discuss situations in which these latter conditions are automatically satisfied. The basis for this is the following simple, but important observation.

Suppose the estimate $\hat L_n$ can be expressed as

$$\hat L_n(x,\vartheta_n)=\bar L_n(x,\vartheta_n,\eta_{n1},\dots,\eta_{nn}) \tag{3.1}$$

under P_{ϑ_n} for some measurable function $\bar{L}_n$ from $\mathfrak{X} \times \Theta \times \mathcal{Y}^n$ into $\mathbb{R}^k$. Then $\hat{L}_{ni}(\cdot, \vartheta_n) = \hat{L}_n(\cdot, \vartheta_n)$ under P_{ϑ_n} and the conditions (C3)–(C5) are automatically satisfied. Similarly, if the estimate $\tilde{L}_{nj}$ can be expressed as

$$\tilde{L}_{nj}(x, \vartheta_n) = \bar{L}_{n-1}(x, \vartheta_n, \eta_{n1}, \ldots, \eta_{n,j-1}, \eta_{n,j+1}, \ldots, \eta_{nn}), \tag{3.2}$$

under P_{ϑ_n} for some measurable function $\bar{L}_{n-1}$ from $\mathfrak{X} \times \Theta \times \mathcal{Y}^{n-1}$ into $\mathbb{R}^k$, then (D3) is automatically satisfied. Thus we have the following result which can also be viewed as a simple consequence of our Basic Lemma.

Theorem 3.1 *Suppose Assumptions 2.1 to 2.5 hold. If (3.1) holds, then the full estimator (2.5) satisfies (2.3) under (C1) and (C2) alone. If (3.2) holds, then the leave-one-out estimator (2.6) satisfies (2.3) under (D1) and (D2) alone.*

Thus by choosing the estimates of $L_{\vartheta,\gamma}$ carefully and conditioning properly, there is potential for considerable simplifications. In the remainder of this section we shall discuss two simple situations where this idea is easy to implement. Generalizations of this idea to a more complex situation are discussed in the next section.
A first situation where the above idea can be put to good use is when, for every $\vartheta \in \Theta$ and $\gamma \in \Gamma$,

$$L(x, \vartheta, \gamma) = u_\vartheta(x) h_\gamma(v_\vartheta(x)), \quad x \in \mathfrak{X}, \tag{3.3}$$

with h_γ a measurable function from $\mathbb{R}$ to $\mathbb{R}$, v_ϑ a measurable function from $\mathfrak{X}$ into $\mathbb{R}$, u_ϑ a measurable function from $\mathfrak{X}$ into $\mathbb{R}^k$ such that

$$E_{\vartheta,\gamma}(u_\vartheta(\xi_1) \mid v_\vartheta(\xi_1)) = 0. \tag{3.4}$$

Let $U_{nj} = u_{\vartheta_n}(\xi_j)$ and $V_{nj} = v_{\vartheta_n}(\xi_j)$. Under P_{ϑ_n}, we estimate $h = h_{\gamma_0}$ from the observations $V_{n1}, \ldots, V_{nn}$ alone, say by

$$\hat{h}_n(v, \vartheta_n) = \tilde{h}_n(v, V_{n1}, \ldots, V_{nn}), \quad v \in \mathbb{R}. \tag{3.5}$$

The corresponding leave-one-out estimator is

$$\hat{h}_{nj}(v, \vartheta_n) = \tilde{h}_{n-1}(v, V_{n1}, \ldots, V_{n,j-1}, V_{n,j+1}, \ldots, V_{nn}), \quad v \in \mathbb{R}.$$

Given this structure, we can use

$$\hat{L}_n(\xi_j, \vartheta_n) = U_{nj}\, \hat{h}_n(V_{nj}, \vartheta_n),$$

for the full estimator, and

$$\tilde{L}_{nj}(\xi_j, \vartheta_n) = U_{nj}\, \hat{h}_{nj}(V_{nj}, \vartheta_n),$$

for the leave-one-out estimator. The latter estimator was studied by van der Vaart [22] in this context.

We apply Theorem 3.1 with $\eta(x,\vartheta,\gamma) = v_\vartheta(x)$ so that $\eta_{nj} = V_{nj}$. In view of (3.4), (C1) and (D1) are immediate. Consequently, one needs to verify (C2) for the full estimator, which is equivalent to

$$\frac{1}{n}\sum_{j=1}^n \tau_{nj}\big[\hat h_n(V_{nj},\vartheta_n) - h(V_{nj})\big]^2 = o_{P_{\vartheta_n}}(1) \tag{3.6}$$

with $\tau_{nj} = E_{\vartheta_n}(\|U_{nj}\|^2 \mid V_{nj})$. If

$$\max_{1\le j\le n} \tau_{nj} = O_{P_{\vartheta_n}}(1), \tag{3.7}$$

then (3.6) is implied by

$$\frac{1}{n}\sum_{j=1}^n \big[\hat h_n(V_{nj},\vartheta_n) - h(V_{nj})\big]^2 = o_{P_{\vartheta_n}}(1). \tag{3.8}$$

For the leave-one-out estimator one needs to verify (D2), which is

$$\frac{1}{n}\sum_{j=1}^n \tau_{nj}\big[\hat h_{nj}(V_{nj},\vartheta_n) - h(V_{nj})\big]^2 = o_{P_{\vartheta_n}}(1). \tag{3.9}$$

This improves upon the result of van der Vaart [22] who requires that the expected value of the left-hand side of (3.9) tends to zero:

$$E_{\vartheta_n}\Big[\tau_{n1}\big[\hat h_{n1}(V_{n1},\vartheta_n) - h(V_{n1})\big]^2\Big] = o(1). \tag{3.10}$$

We conclude with a closely related situation. It does not quite fit into the setting of Theorem 3.1, but can be treated with a little extra effort. It builds a bridge to what we do in the next section.
Suppose now that for all $\vartheta \in \Theta$ and $\gamma \in \Gamma$,

$$L(x,\vartheta,\gamma) = \Big[u_\vartheta(x) - \int u_\vartheta\, dF_{\vartheta,\gamma}\Big] h_\gamma(v_\vartheta(x)), \quad x \in \mathfrak{X}, \tag{3.11}$$

with h_γ a measurable function from $\mathbb{R}$ to $\mathbb{R}$, v_ϑ a measurable function from $\mathfrak{X}$ into $\mathbb{R}$ and u_ϑ a measurable function from $\mathfrak{X}$ into $\mathbb{R}^k$ such that $u_\vartheta(\xi_1)$ and $v_\vartheta(\xi_1)$ are independent under $P_{\vartheta,\gamma}$ and

$$E_{\vartheta,\gamma}[h_\gamma(v_{\vartheta,\gamma}(\xi_1))] = 0. \tag{3.12}$$

Let h, U_{nj} and V_{nj} be as before. We require that $E_{\vartheta_n}[\|U_{n1}\|^2]$ and $E_{\vartheta_n}[h^2(V_{n1})]$ are bounded. Under P_{ϑ_n} we estimate h as before only from the observations $V_{n1},\dots,V_{nn}$, namely by $\hat h_n(v,\vartheta_n)$ as in (3.5), and take

$$\hat L_n(\xi_j,\vartheta_n) = [U_{nj} - \bar U_n]\,\hat h_n(V_{nj},\vartheta_n) \quad \text{where} \quad \bar U_n = \frac{1}{n}\sum_{i=1}^n U_{ni}.$$

Let us now show that (3.8) implies

$$\frac{1}{n}\sum_{j=1}^{n}\hat{L}_n(\xi_j,\vartheta_n)=\frac{1}{n}\sum_{j=1}^{n}L(\xi_j,\vartheta_n)+o_{P_{\vartheta_n}}(n^{-1/2}).$$

For this write

$$\frac{1}{n}\sum_{j=1}^{n}[\hat{L}_n(\xi_j,\vartheta_n)-L(\xi_j,\vartheta_n)]=T_n-R_n,$$

with

$$T_n=\frac{1}{n}\sum_{j=1}^{n}\big[U_{nj}-E_{\vartheta_n}[U_{nj}]\big]\big[\hat{h}_n(V_{nj},\vartheta_n)-h(V_{nj})\big]$$

and

$$R_n=\frac{1}{n}\sum_{i=1}^{n}\big[U_{ni}-E_{\vartheta_n}[U_{ni}]\big]\,\frac{1}{n}\sum_{j=1}^{n}\hat{h}_n(V_{nj},\vartheta_n).$$

We can apply Theorem 3.1 with $\eta_{nj}=V_{nj}$ to conclude that $T_n=o_{P_{\vartheta_n}}(n^{-1/2})$. Indeed, the left-hand side of (C1) is zero and the left-hand side of (C2) equals $E_{\vartheta_n}[\|U_{n1}\|^2]$ times the left hand side of (3.8). Finally, we have $R_n=o_{P_{\vartheta_n}}(n^{-1/2})$ because

$$\frac{1}{n}\sum_{j=1}^{n}\big[U_{nj}-E_{\vartheta_n}[U_{nj}]\big]=O_{P_{\vartheta_n}}(n^{-1/2})\quad\text{and}\quad\frac{1}{n}\sum_{j=1}^{n}\hat{h}_n(V_{nj},\vartheta_n)=o_{P_{\vartheta_n}}(1),$$

the former since $E_{\vartheta_n}[\|U_{n1}\|^2]$ is bounded, the latter by (3.8) and $E_{\vartheta_n}[h(V_{n1})]=0$ and since $E_{\vartheta_n}[h^2(V_{n1})]$ is bounded.

4 A second application

Let us now generalize the ideas in the previous example. For this we make the following assumption.

Assumption 4.1 *For every (ϑ,γ) in $\Theta\times\Gamma$, $L_{\vartheta,\gamma}$ is of the form*

$$L(x,\vartheta,\gamma)=\Big[u_\vartheta(x)-\mu_{\vartheta,\gamma}(v_\vartheta(x))\Big]h_{\vartheta,\gamma}(v_\vartheta(x),w_{\vartheta,\gamma}(x))\tag{4.1}$$

for a measurable function $h_{\vartheta,\gamma}$ from $\mathbb{R}^m\times\mathbb{R}$ into $\mathbb{R}$ and measurable functions u_ϑ, v_ϑ and $w_{\vartheta,\gamma}$ from $\mathfrak{X}$ to $\mathbb{R}^k$, $\mathbb{R}^m$ and $\mathbb{R}$, respectively, such that $(u_\vartheta(\xi_1),v_\vartheta(\xi_1))$ and $w_{\vartheta,\gamma}(\xi_1)$ are independent under $P_{\vartheta,\gamma}$ and where

$$\mu_{\vartheta,\gamma}(v_\vartheta(\xi_1))=E_{\vartheta,\gamma}(u_\vartheta(\xi_1)\mid v_\vartheta(\xi_1))$$

and

$$E_{\vartheta,\gamma}(h_{\vartheta,\gamma}(v_\vartheta(\xi_1),w_{\vartheta,\gamma}(\xi_1))\mid v_\vartheta(\xi_1))=0.\tag{4.2}$$

This structure arises in various regression models with $w_{\vartheta,\gamma}(\xi_1)$ the error variable and u_ϑ and v_ϑ functions of the covariates only. One such model is the partly linear model which will be treated in detail in the next section. Single index models also have this structure. Let us set $U_{nj} = u_{\vartheta_n}(\xi_j)$, $V_{nj} = v_{\vartheta_n}(\xi_j)$ and $W_{nj} = w_{\vartheta_n}(\xi_j)$. Also, set $h(v,w,\vartheta) = h_{\vartheta,\gamma_0}(v,w)$ and $\mu(v,\vartheta) = \mu_{\vartheta,\gamma_0}(v)$. Given the form (4.1), we should strive to use Theorem 3.1 with $\eta_{nj} = (V_{nj}, W_{nj})$. Consequently, we should estimate $h_{\vartheta_n}(v,w)$ by an expression $\hat{h}_n(v,w,\vartheta_n)$ which can be written as a function of the variables (V_{nj}, W_{nj}), $j = 1,\dots,n$, under P_{ϑ_n}. It is natural to estimate $\mu(v,\vartheta_n)$ by a linear smoother, namely

$$\hat{\mu}_n(v,\vartheta_n) = \sum_{i=1}^n s_{ni}(v, V_{n1},\dots,V_{nn})U_{ni}, \quad v \in \mathbb{R}^m,$$

where $s_{n1},\dots,s_{nn}$ are measurable functions from $(\mathbb{R}^m)^{n+1}$ into $\mathbb{R}^k$. This then leads us to the estimator

$$\hat{L}_n(\xi_j,\vartheta_n) = \Big[U_{nj} - \hat{\mu}_n(V_{nj},\vartheta_n)\Big]\hat{h}_n(V_{nj},W_{nj},\vartheta_n),$$

where $\hat{h}_n$ is expressible as

$$\hat{h}_n(v,w,\vartheta_n) = h_n(v,w,\vartheta_n,V_{n1},W_{n1},\dots,V_{n1},W_{n1})$$

under P_{ϑ_n}, although it needs to be constructed without the knowledge of γ_0. Here Theorem 3.1 is not directly applicable. Instead we proceed as follows. We write

$$\frac{1}{n}\sum_{j=1}^n \hat{L}_n(\xi_j,\vartheta_n) - \frac{1}{n}\sum_{j=1}^n L(\xi_j,\vartheta_n) = T_{n0} - T_{n1} - T_{n2} - T_{n3},$$

where

$$T_{n0} = \frac{1}{n}\sum_{j=1}^n \Big[U_{nj} - \mu(V_{nj},\vartheta_n)\Big]\Delta_{nj},$$

$$T_{n1} = \frac{1}{n}\sum_{j=1}^n \Big[\hat{\mu}_n(V_{nj},\vartheta_n) - \bar{\mu}_n(V_{nj},\vartheta_n)\Big]\Delta_{nj},$$

$$T_{n2} = \frac{1}{n}\sum_{j=1}^n \Big[\bar{\mu}_n(V_{nj},\vartheta_n) - \mu(V_{nj},\vartheta_n)\Big]\Delta_{nj},$$

$$T_{n3} = \frac{1}{n}\sum_{j=1}^n \Big[\hat{\mu}_n(V_{nj},\vartheta_n) - \mu(V_{nj},\vartheta_n)\Big]h(V_{nj},W_{nj},\vartheta_n),$$

with

$$\bar{\mu}_n(v,\vartheta_n) = \sum_{i=1}^n s_{ni}(v,V_{n1},\dots,V_{nn})\mu(V_{ni},\vartheta_n), \quad v \in \mathbb{R}^m,$$

and

$$\Delta_{nj} = \hat{h}_n(V_{nj},W_{nj},\vartheta_n) - h(V_{nj},W_{nj},\vartheta_n).$$

To get the desired

$$\frac{1}{n}\sum_{j=1}^{n}\hat{L}_n(\xi_j,\vartheta_n)=\frac{1}{n}\sum_{j=1}^{n}L(\xi_j,\vartheta_n)+o_{P_{\vartheta_n}}(n^{-1/2}), \tag{4.3}$$

we need to show that, for $i=0,1,2,3$,

$$T_{ni}=o_{P_{\vartheta_n}}(n^{-1/2}). \tag{4.4}$$

We can use Theorem 3.1 or the Basic Lemma to give sufficient conditions for the cases $i=0,1,3$.
A sufficient condition for (4.4) with $i=0$ is

$$\frac{1}{n}\sum_{j=1}^{n}\sigma_{nj}^2\Delta_{nj}^2=o_{P_{\vartheta_n}}(1), \tag{4.5}$$

where

$$\sigma_{nj}^2=E_{\vartheta_n}(\|U_{nj}-\mu(V_{nj},\vartheta_n)\|^2\mid V_{nj}).$$

This follows from Theorem 3.1 applied with $\eta_{nj}=(V_{nj},W_{nj})$. Indeed, in this case, the left-hand side of (C1) equals 0 and (C2) is equivalent to (4.5).
A sufficient condition for (4.4) with $i=3$ is

$$\frac{1}{n}\sum_{j=1}^{n}\|\hat{\mu}_n(V_{nj},\vartheta_n)-\mu(V_{nj},\vartheta_n)\|^2E_{\vartheta_n}(h^2(V_{nj},W_{nj},\vartheta_n)|V_{nj})=o_{P_{\vartheta_n}}(1). \tag{4.6}$$

To see this apply Theorem 3.1 with $\eta_{nj}=(U_{nj},V_{nj})$ and use (4.2). To get a sufficient condition for (4.4) with $i=1$, write T_{n1} as a double sum and change the order of summation to arrive at

$$T_{n1}=\frac{1}{n}\sum_{i=1}^{n}\Big[U_{ni}-\mu(V_{ni})\Big]\bar{\Delta}_{ni}$$

with

$$\bar{\Delta}_{ni}=\sum_{j=1}^{n}s_{ni}(V_{nj},V_{n1},\ldots,V_{nn})\Delta_{nj},$$

a function of $(V_{n1},W_{n1},\ldots,V_{nn},W_{nn})$. By conditioning on these variables, we obtain from Theorem 3.1 or the Basic Lemma that (4.4) with $i=1$ is implied by

$$\frac{1}{n}\sum_{i=1}^{n}\sigma_{ni}^2\bar{\Delta}_{ni}^2=o_{P_{\vartheta_n}}(1). \tag{4.7}$$

Indeed, for this application the left-hand side of (C1) is 0 and (C2) is equivalent to (4.7). In most applications the smoothing matrix S_n with (i,j)-entry

$$S_{nij}=s_{ni}(V_{nj},V_{n1},\ldots,V_{nn})$$

will have a stochastically bounded operator norm:

$$\sup_{x\in\mathbb{R}^n:\|x\|=1} \|S_n x\| = O_{P_{\vartheta_n}}(1). \tag{4.8}$$

Then (4.7) is implied by (4.5). Let us now summarize our findings.

Theorem 4.2 *Suppose Assumption 4.1 holds and the smoothing matrix satisfies (4.8). Then (4.3) is implied by (4.5), (4.6) and*

$$\frac{1}{n}\sum_{j=1}^n \Big[\bar\mu_n(V_{nj},\vartheta_n) - \mu(V_{nj},\vartheta_n)\Big]\Delta_{nj} = o_{P_{\vartheta_n}}(n^{-1/2}). \tag{4.9}$$

Remark 4.3 Note that the squared norm of the left hand side of (4.9) can be bounded via the Cauchy-Schwarz inequality by

$$\frac{1}{n}\sum_{j=1}^n \Delta_{nj}^2 \,\frac{1}{n}\sum_{j=1}^n \|\bar\mu_n(V_{nj},\vartheta_n) - \mu(V_{nj},\vartheta_n)\|^2.$$

From this bound we can derive sufficient conditions for (4.9). Of course, we can also use Theorem 2.6 to verify (4.9) directly.

Simple sufficient conditions can be given if the quantities $E_{\vartheta_n}(h^2(V_{nj},W_{nj},\vartheta_n)|V_{nj})$ and σ_{nj}^2 are bounded.

Theorem 4.4 *Suppose Assumption 4.1 holds, the smoothing matrix satisfies (4.8),*

$$\max_{1\le j\le n}\sigma_{nj}^2 = O_{P_{\vartheta_n}}(1) \quad \textit{and} \quad \max_{1\le j\le n} E_{\vartheta_n}(h^2(V_{nj},W_{nj},\vartheta_n)|V_{nj}) = O_{P_{\vartheta_n}}(1).$$

Then (4.3) follows from

$$\frac{1}{n}\sum_{j=1}^n \Big[\hat h_n(V_{nj},W_{nj},\vartheta_n) - h(V_{nj},W_{nj},\vartheta_n)\Big]^2 = o_{P_{\vartheta_n}}(1), \tag{4.10}$$

$$\frac{1}{n}\sum_{j=1}^n \Big[\hat\mu_n(V_{nj},\vartheta_n) - \mu(V_{nj},\vartheta_n)\Big]^2 = o_{P_{\vartheta_n}}(1) \tag{4.11}$$

and

$$\frac{1}{n}\sum_{j=1}^n \|\bar\mu_n(V_{nj},\vartheta_n) - \mu(V_{nj},\vartheta_n)\|^2 = O_{P_{\vartheta_n}}(n^{-1}). \tag{4.12}$$

Remark 4.5 Of course, we could replace (4.12) by (4.9). We have chosen the stronger (4.12) as it is easier to interpret. Moreover, it can typically be satisfied via under-smoothing as we shall show in the next section.

5 Efficient estimation in a partly linear regression model

In this section we shall apply the results of the previous section to construct efficient estimates for a partly linear regression model. In this model $\xi_j = (Y_j, U_j, V_j)$ is assumed to take values in $\mathbb{R} \times \mathbb{R}^k \times [0,1]$ and satisfies under P_{ϑ_0} the structural relation

$$Y_j = \vartheta_0^\top U_j + \rho(V_j) + \varepsilon_j, \quad j = 1, \ldots, n,$$

where ϑ_0 is an unknown vector in $\mathbb{R}^k$, ρ is an unknown Lipschitz-continuous function on $[0,1]$, the unobserved error variable ε_j has unknown density f with finite Fisher information J and is independent of the covariate (U_j, V_j) which has distribution Q. The nuisance parameter γ_0 is the triple (ρ, f, Q). We impose the following additional assumptions on the covariate distribution.

Assumption 5.1 *If (U,V) has distribution Q, then $E[\|U\|^4] < \infty$, there is a Lipschitz continuous function μ such that $\mu(V) = E(U|V)$, the matrix $W = E[(U - \mu(V))(U - \mu(V))^\top]$ is positive definite, and the marginal distribution G of V has a density g that is bounded and bounded away from 0 on the interval $[0,1]$.*

An estimator $\hat{\vartheta}_n$ is efficient in this model if

$$\hat{\vartheta}_n = \vartheta_0 + \frac{1}{n}\sum_{j=1}^{n} (JW)^{-1}(U_j - \mu(V_j))\ell(Y_j - \vartheta_0^\top U_j - \rho(V_j)) + o_P(n^{-1/2}),$$

where $\ell = -f'/f$ is the score function for location. Such estimators have been constructed by Cuzick [3] and Schick [13] under the additional assumption that the error density f has mean zero and finite variance. Cuzick used Bickel's [2] sample splitting scheme, while Schick avoided sample splitting by conditioning on the covariates. Bhattacharya and Zhao [1] constructed efficient estimates without these moment assumptions, but required the error density f to be symmetric with a bounded derivative and positive in a neighborhood of 0. Their construction utilized the sample splitting scheme of Schick [11]. In their Remark 10, they explain that *intractable calculations* associated with the verification of the leave-one-out type of conditions for their M-type regression estimate forced them to use sample splitting. Because this model satisfies Assumption 4.1 as we shall see below, the approach described in the previous section will avoid the verification of these conditions.
We shall show that efficient estimates exist without the moment or symmetry assumptions of these papers. We shall, however, require the identifiability condition

$$\int \psi(x) f(x)\, dx = 0$$

for some bounded odd function ψ with a positive and bounded derivative. A possible choice is

$$\psi(x) = \arctan(x), \quad x \in \mathbb{R}.$$

The efficient score function for this case is given by

$$L(\xi_j, \vartheta) = (U_j - \mu(V_j))\ell(\varepsilon_j(\vartheta)), \quad \text{where} \quad \varepsilon_j(\vartheta) = Y_j - \vartheta^\top U_j - \rho(V_j).$$

Thus Assumption 4.1 holds with

$$u_\vartheta(\xi_j) = U_j, \quad v_\vartheta(\xi_j) = V_j \quad \text{and} \quad w_{\vartheta,\gamma_0}(\xi_j) = \varepsilon_j(\vartheta) = Y_j - \vartheta^\top U_j - \rho(V_j).$$

Note that (4.2) follows from $\int \ell(x) f(x)\, dx = 0$.
Let now ϕ be a symmetric continuously differentiable density with support $[-1,1]$ and set

$$\phi_c(x) = \frac{1}{c}\,\phi\Big(\frac{x}{c}\Big), \quad x \in \mathbb{R}, c > 0.$$

Modifying the arguments of Zhao [23] slightly we can show that a $n^{1/2}$-consistent estimator of ϑ_0 is given as a minimizer of

$$D_n(t) := \frac{2}{n(n-1)} \sum\sum_{1\le i<j\le n} |Y_i - Y_j - t^\top(U_i - U_j)|\phi_{\tilde c_n}(V_i - V_j),$$

provided the bandwidth $\tilde c_n$ is chosen such that $n\tilde c_n \to \infty$ and $n^{1/2}\tilde c_n^2 \to 0$. Details are given in Section 6. Thus Assumption 2.1 holds with $\tilde\vartheta_n$ a discretized version of this estimator. Assumptions 2.2 to 2.4 are verified in Schick [13].
For $v \in [0,1]$, we estimate $\rho(v)$ by $\hat\rho_n(v,\tilde\vartheta_n)$ which is the solution t to the equation $\Psi_n(t,v,\tilde\vartheta_n) = 0$, where

$$\Psi_n(t,v,\tilde\vartheta_n) = \frac{1}{n}\sum_{j=1}^n \phi_{c_n}(v - V_j)\psi(Y_j - \tilde\vartheta_n^\top U_j - t),$$

and $c_n \sim n^{-1/3}(\log n)^{1/3}$. One can show that this estimator satisfies

$$\sup_{0\le v\le 1} |\hat\rho_n(v,\vartheta_n) - \rho(v)| = O_{P_{\vartheta_n}}(n^{-1/3}(\log n)^{1/3}). \tag{5.1}$$

This result is in the spirit of Härdle and Luckhaus [5] and Härdle et al. [4]. It does not immediately follow from their results, but needs some modifications. Here are the details. Since $s \mapsto \Psi_n(s,v,\vartheta_n)$ is increasing, we see that $|\hat\rho_n(v,\vartheta_n) - \rho(v)| > c_n(L+t)$ if either $\Psi_n(\rho(v) + c_n(L+t), v, \vartheta_n) < 0$ or $\Psi_n(\rho(v) - c_n(L+t), v, \vartheta_n) > 0$ for all positive t and L the Lipschitz constant of ρ. Furthermore, since ϕ has support $[-1,1]$, we have the inequalities

$$\Psi_n(\rho(v) + c_n(L+t), v, \vartheta_n) \ge \tilde\Psi_n(c_n t, v)$$

and

$$\Psi_n(\rho(v) - c_n(L+t), v, \vartheta_n) \le \tilde\Psi_n(-c_n t, v),$$

where

$$\tilde\Psi_n(t,v) = \frac{1}{n}\sum_{j=1}^n \phi_{c_n}(v - V_j)\psi(Y_j - \vartheta_n^\top U_j - \rho(V_j) + t).$$

This shows that for $t > 0$,

$$\begin{aligned}\pi_n(L+t) &= P_{\vartheta_n}\Big(\sup_{0\le v\le 1} |\hat\rho_n(v,\vartheta_n) - \rho(v)| > c_n(L+t)\Big)\\ &\le P_{\vartheta_n}\Big(\inf_{0\le v\le 1} \tilde\Psi_n(c_n t, v) < 0\Big) + P_{\vartheta_n}\Big(\sup_{0\le v\le 1} \tilde\Psi_n(-c_n t, v) > 0\Big).\end{aligned}$$

A standard argument involving the Bernstein inequality yields now that

$$\Psi_n^* = \sup_{0\le v\le 1}\sup_{|t|<1} |\tilde\Psi_n(t,v) - E_{\vartheta_n}[\tilde\Psi_n(t,v)]| = O_{P_{\vartheta_n}}(c_n). \tag{5.2}$$

It is easy to check that

$$E_{\vartheta_n}[\tilde\Psi_n(t,v)] = \int \psi(y+t)f(y)\,dy \int g(v-c_nw)\phi(w)\,dw.$$

Thus, for $t>0$,

$$\inf_{0\le v\le 1} c_n^{-1}E_{\vartheta_n}[\tilde\Psi_n(c_nt,v)] \ge \frac{1}{2}\inf_{0\le v\le 1} g(v)c_n^{-1}\int \psi(y+c_nt)f(y)\,dy$$

and

$$\sup_{0\le v\le 1} c_n^{-1}E_{\vartheta_n}[\tilde\Psi_n(-c_nt,v)] \le \sup_{0\le v\le 1} g(v)c_n^{-1}\int \psi(y-c_nt)f(y)\,dy.$$

By the assumption on ψ, the map $\tau(s)=\int\psi(y+s)f(y)\,dy$ satisfies $\tau(0)=0$ and has a continuous positive derivative. Combining the above we see that there are positive constants d_1 and d_2 such that, for each $t>0$,

$$\pi_n(L+t) \le P_{\vartheta_n}(d_1t - c_n^{-1}\Psi_n^* < 0) + P_{\vartheta_n}(-d_2t + c_n^{-1}\Psi_n^* > 0).$$

From this and (5.2), the desired (5.1) is immediate.
We use the residuals

$$\hat\varepsilon_j(\tilde\vartheta_n) = Y_j - \tilde\vartheta_n^\top U_j - \hat\rho_n(V_j,\tilde\vartheta_n), \quad j=1,\dots,n,$$

to estimate ℓ by

$$\hat\ell_n(x,\tilde\vartheta_n) = \frac{\frac{1}{n}\sum_{j=1}^n K_n'(x-\hat\varepsilon_j(\tilde\vartheta_n))}{b_n + \frac{1}{n}\sum_{j=1}^n K_n(x-\hat\varepsilon_j(\tilde\vartheta_n))}, \quad x\in\mathbb{R}.$$

Here a_n and b_n are sequences of positive numbers that converge to zero at a rate to be determined later and $K_n(x)=K(x/a_n)/a_n$ for a positive, bounded symmetric density K that is twice continuously differentiable with $|K'|/K$ and $|K''|/K$ bounded. At this moment, we only assume that $\hat\mu_n$ is a linear smoother

$$\hat\mu_n(v) = \sum_{j=1}^n s_{nj}(v,V_1,\dots,V_n)U_j$$

with a smoothing matrix S_n that satisfies (4.8). This allows us to discuss various choices later. Finally, our candidate for an efficient estimator of ϑ_0 is

$$\tilde\vartheta_n + \frac{1}{n}\sum_{j=1}^n (\hat W_n\hat J_n(\tilde\vartheta_n))^{-1}(U_j-\hat\mu_n(V_j))\hat\ell_n(\hat\varepsilon_j(\tilde\vartheta_n),\tilde\vartheta_n),$$

where

$$\hat{W}_n = \frac{1}{n}\sum_{j=1}^{n}(U_j - \hat{\mu}_n(V_j))(U_j - \hat{\mu}_n(V_j))^\top \quad \text{and} \quad \hat{J}_n(\tilde{\vartheta}_n) = \frac{1}{n}\sum_{j=1}^{n}\hat{\ell}_n^2(\hat{\varepsilon}_j(\tilde{\vartheta}_n), \tilde{\vartheta}_n).$$

In view of Theorem 4.4, this estimator is efficient if the following three conditions are true.

$$\frac{1}{n}\sum_{j=1}^{n}[\hat{\ell}_n(\hat{\varepsilon}_j(\vartheta_n), \vartheta_n) - \ell(\varepsilon_j(\vartheta_n))]^2 = o_{P_{\vartheta_n}}(1), \tag{5.3}$$

$$\frac{1}{n}\sum_{j=1}^{n}\|\hat{\mu}_n(V_j) - \mu(V_j)\|^2 = o_{P_{\vartheta_n}}(1), \tag{5.4}$$

$$\frac{1}{n}\sum_{j=1}^{n}\|\bar{\mu}_n(V_j) - \mu(V_j)\|^2 = O_{P_{\vartheta_n}}(n^{-1}), \tag{5.5}$$

where

$$\bar{\mu}_n(v) = \sum_{j=1}^{n} s_{nj}(v, V_1, \ldots, V_n)\mu(V_j).$$

It follows from Lemma 8.1 that (5.3) holds if $na_n^3 b_n \to \infty$ and $n^{2/3}a_n^4/(\log n)^{2/3} \to \infty$, e.g., $a_n \sim n^{-1/7}$ and $b_n \sim n^{-1/2}$ work. Finally, (5.4) and (5.5) are satisfied by under-smoothed kernel estimates. Indeed, for a kernel estimate with bandwidth d_n and kernel ϕ as above, the left hand side of (5.4) is of order $O_{P_{\vartheta_n}}(d_n^2 + n^{-1}d_n^{-1})$, while the left hand side of (5.5) is of order $O(d_n^2)$. Keep in mind that we assumed μ to be Lipschitz. Thus if we take $d_n \sim n^{-1/2}$, we obtain both (5.4) and (5.5). Of course, the corresponding smoothing matrix satisfies (4.8).
Instead of kernel estimators we could have also used locally linear smoothers. Larger bandwidths are possible under additional smoothness assumptions.

6 Bandwidth-matched M-estimation in partly linear models

In this section we utilize the fact that our error density has finite Fisher information to relax some of the conditions used by Zhao [23] to obtain the $n^{1/2}$-consistency and asymptotic normality of his bandwidth-matched M-estimator for partly linear regression models. In particular, we allow the random variable appearing in the linear part to be unbounded and relax the smoothness assumptions on his loss function.
We consider again the partly linear model in which the observations $\xi_j = (Y_j, U_j, V_j)$ take values in $\mathbb{R} \times \mathbb{R}^k \times [0,1]$ and satisfy the structural relation

$$Y_j = \vartheta_0^\top U_j + \rho(V_j) + \varepsilon_j, \quad j = 1, \ldots, n,$$

for some ϑ_0 in $\mathbb{R}^k$ and some Lipschitz-continuous function ρ on $[0,1]$. The error variable ε_1 has density f with finite Fisher information and is independent of the covariate (U_1, V_1) whose distribution Q fulfills Assumption 5.1.
Let τ be an even, convex and Lipschitz continuous function from $\mathbb{R}$ to $[0,\infty)$. The corresponding bandwidth-matched M-estimator of ϑ_0 is then a minimizer of the convex U-process

$$D_n(t) = \frac{2}{n(n-1)} \sum\sum_{1\le i<j\le n} \tau(Y_i - Y_j - t^\top (U_i - U_j))\phi_{c_n}(V_i - V_j)$$

where ϕ_c is as in Section 5 with a bounded symmetric density ϕ with support $[-1,1]$. Zhao's [23] version also includes the weight factors $H(V_i)H(V_j)$. To keep our notation simple, we have not included these here.
Since τ is Lipschitz, it is absolutely continuous with a bounded almost everywhere derivative ψ, say $|\psi| \le B$. Since τ is convex and even, this derivative can be taken to be non-decreasing and odd. Now define functions ψ_1 and ψ_2 by

$$\psi_1(t) = E[\psi(t - \varepsilon_1)] = \int \psi(t-y) f(y)\, dy, \quad t \in \mathbb{R},$$

and

$$\psi_2(t) = E[\psi(t + \varepsilon_1 - \varepsilon_2)] = \iint \psi(t+x-y) f(x)\, dx f(y)\, dy, \quad t \in \mathbb{R}.$$

Since ψ is bounded and f has finite Fisher information, the function ψ_1 is differentiable with bounded and uniformly continuous derivative

$$\psi_1'(t) = \int \psi(t-y) f'(y)\, dy, \quad t \in \mathbb{R}.$$

Indeed, the almost everywhere derivative f' of f is integrable with

$$\|f'\|_1 = \int |\ell(y)| f(y)\, dy \ \le J^{1/2}.$$

Thus ψ_1' is bounded by $B\|f'\|_1$ and uniformly continuous as

$$\begin{aligned} |\psi_1'(s+t) - \psi_1'(s)| &= \Big| \int \psi(s-y)(f'(y-t) - f'(y))\, dy \Big| \\ &\le B \int |f'(y-t) - f'(y)|\, dy \end{aligned}$$

and $\int |f'(y-t) - f'(y)|\, dy \to 0$ as $t \to 0$ by the translation continuity in L_1, see Theorem 9.5 in Rudin [10]. That ψ_1' is the derivative of ψ_1 is easy to check. Similarly, one can show that ψ_2 has a bounded and uniformly continuous first and second order derivatives ψ_2' and ψ_2'', namely $\psi_2'(t) = \iint \psi(t+x-y) f(x) f'(y)\, dx\, dy$ and $\psi_2''(t) = -\iint \psi(t+x-y) f'(x) f'(y)\, dx\, dy$.

Theorem 6.1 *Suppose f has finite Fisher information, Q satisfies Assumption 5.1, τ is an even, convex and Lipschitz continuous function from $\mathbb{R}$ to $\mathbb{R}$, and*

$$\psi_2'(0) = \iint \psi(x-y)f(x)f'(y)\,dx\,dy > 0. \tag{6.1}$$

Suppose also that $nc_n \to \infty$ and $n^{1/2}c_n^2 \to 0$. Let $\bar\vartheta_n$ be a minimizer of $D_n(t)$. Then

$$\bar\vartheta_n = \vartheta_0 + \frac{1}{n}\sum_{i=1}^n A^{-1}(U_i - \mu(V_i))\psi_1(\varepsilon_i)g(V_i) + o_P(n^{-1/2})$$

where $A = \psi_2'(0)E[g(V_1)(U_1-\mu(V_1))(U_1-\mu(V_1))^\top]$.

The above theorem relaxes some of the conditions of Zhao [23]. We do not need the variable $\|U_1\|$ to be bounded, but require it instead to have finite fourth moment. We remove Zhao's smoothness assumption (A2)(a) on the joint law Q and weaken Zhao's assumption (A4)(b).
If $\tau(x) = |x|$, then $\psi(x) = \operatorname{sign}(x)$, $\psi_1(t) = 2F(t)-1$, $\psi_2(t) = \int(2F(t+x)-1)f(x)\,dx$, and $\psi_2'(0) = 2\int f^2(x)\,dx > 0$.

Proof: We follow the proof of Zhao [23] who utilizes Pollard's [9] convexity lemma; see also Hjort and Pollard [6]. Let

$$S_n(t) = \frac{2}{n(n-1)}\sum\sum_{1\le i<j\le n}(U_j-U_i)\psi(\varepsilon_i-\varepsilon_j+\rho(V_i)-\rho(V_j)+n^{-1/2}t^\top(U_j-U_i))\phi_{c_n}(V_i-V_j).$$

We shall prove that, for each fixed $t \in \mathbb{R}^k$,

$$n\Big[D_n(\vartheta_0 + n^{-1/2}t) - D_n(\vartheta_0) - n^{-1/2}t^\top S_n(0)\Big] = t^\top At + o_p(1) \tag{6.2}$$

and that

$$S_n(0) = -\frac{1}{n}\sum_{i=1}^n 2(U_i - \mu(V_i))\psi_1(\varepsilon_i)g(V_i) + o_p(n^{-1/2}). \tag{6.3}$$

It then follows from the convexity lemma that $n^{1/2}(\bar\vartheta_n - \vartheta_0) = -\frac{1}{2}A^{-1}n^{1/2}S_n(0) + o_p(1)$ from which the desired result follows.
To verify (6.2) and (6.3) we need the following simple fact about U-statistics. Let

$$H_n = \frac{2}{n(n-1)}\sum\sum_{1\le i<j\le n}h_n(\xi_i,\xi_j)$$

be a U-statistic with a symmetric square-integrable kernel h_n. Let $\bar h_n(x) = E[h_n(x,\xi_1)]$. Suppose that $E[\|h_n(\xi_1,\xi_2)\|^2] = o(n)$, that $E[\|\bar h_n(\xi_1) - \bar h(\xi_1)\|^2] \to 0$ for some function $\bar h$ with $E[\|\bar h(\xi_1)\|^2] < \infty$, and that $\sqrt{n}E[H_n] \to a$, then

$$n^{1/2}(H_n - \frac{1}{n}\sum_{i=1}^n 2[\bar h(\xi_i) - E[h(\xi_i)]]) = a + o_p(1).$$

This is an easy consequence of the Hoeffding decomposition

$$H_n = E[H_n] + \frac{1}{n}\sum_{i=1}^{n} 2(\bar{h}_n(\xi_i) - E[\bar{h}_n(\xi_i)]) + R_n,$$

where $n(n-1)E[\|R_n\|^2] \le 2E[\|h_n(\xi_1,\xi_2)\|^2]$.
To prove (6.3) we use this with

$$h_n(\xi_1,\xi_2) = (U_2 - U_1)\psi(\varepsilon_1 - \varepsilon_2 + \rho(V_1) - \rho(V_2))\phi_{c_n}(V_1 - V_2)$$

and

$$\bar{h}(\xi_1) = -(U_1 - \mu(V_1))\psi_1(\varepsilon_1)g(V_1).$$

Since

$$g_{c_n}(v) = E[\phi_{c_n}(v - V_2)] = \int g(v - c_n s)\phi(s)\,ds$$

is bounded, we immediately obtain that

$$\begin{aligned} E[\|h_n(\xi_1,\xi_2)\|^2] &\le 2B^2 E[(\|U_1\|^2 + \|U_2\|^2)\phi^2_{c_n}(V_1 - V_2)] \\ &\le 4B^2\|\phi_{c_n}\|_\infty E[\|U_1\|^2\phi_{c_n}(V_1 - V_2)] = O(c_n^{-1}) = o(n). \end{aligned}$$

It is well known that $\int |g_{c_n}(v) - g(v)|\,dv \to 0$. Since g is bounded, this gives $\int |g_{c_n}(v) - g(v)|g(v)\,dv \to 0$. From the latter we derive that $g_{c_n}(V_1) \to g(V_1)$ in probability. Note that

$$\bar{h}_n(\xi_1) = -\int (U_1 - \mu(V_1 + c_n v))\psi_1(\varepsilon_1 + \rho(V_1) - \rho(V_1 + c_n v))g(V_1 + c_n v)\phi(v)\,dv.$$

Since ψ_1, ρ and μ are continuous and ψ_1 and g are bounded, one finds that $E[\|\bar{h}_n(\xi_1) - \bar{h}(\xi_1)\|^2] \to 0$. Since ψ is odd and $\varepsilon_1 - \varepsilon_2$ has an even density, we see that $\psi_2(0) = 0$. Thus $|\psi_2(t)| \le B\|f'\|_1|t|$. Since ρ and μ are Lipschitz and g is bounded, we get that the expectation $E[h_n(\xi_1,\xi_2)]$, which equals

$$E[\int (\mu(V_1 + c_n v) - \mu(V_1))\psi_2(\rho(V_1) - \rho(V_1 + c_n v))g(V_1 + c_n v)\phi(v)\,dv],$$

is of order $o(c_n^2) = o(n^{-1/2})$. This shows that (6.3) holds.
To prove (6.2) we first note that its left hand side can be expressed as $n^{1/2}H_n$ with

$$H_n = \int_0^1 t^\top (S_n(ut) - S_n(0))\,du,$$

a U-statistic with kernel

$$h_n(\xi_1,\xi_2) = t^\top(U_2 - U_1)\phi_{c_n}(V_1 - V_2)\int_0^1 [\psi(\eta + n^{-1/2}ut^\top(U_2 - U_1)) - \psi(\eta)]\,du$$

with $\eta = \varepsilon_1 - \varepsilon_2 + \rho(V_1) - \rho(V_2)$. We have as above that

$$E[h_n^2(\xi_1,\xi_2)] \le 4B\|t\|^2 E[\|U_2 - U_1\|^2\phi^2_{c_n}(V_1 - V_2)] = O(c_n^{-1}) = o(n).$$

Since ψ_1 is Lipschitz with Lipschitz constant $B\|f'\|_1$, we get

$$E[\bar{h}_n^2(\xi_1)] \le B^2\|f'\|_1^2\|t\|^4 n^{-1}\|\|E[\|U_2-U_1\|^4\phi_{c_n}^2(V_1-V_2)] = O(c_n^{-1}n^{-1}) = o(1).$$

Finally, since ψ_2' is Lipschitz and ρ and μ are continuous, we have, with $\Delta = \rho(V_1)-\rho(V_2)$ and $\Upsilon = (U_2-U_1)\phi_{c_n}(V_1-V_2)$ and

$$\begin{aligned}
n^{1/2}E[H_n] &= n^{1/2}E\Big[t^\top\Upsilon\int_0^1\Big(\psi_2(\Delta+n^{-1/2}ut^\top(U_2-U_1))-\psi_2(\Delta)\Big)\,du\Big]\\
&= \frac{1}{2}\,E\Big[|t^\top(U_2-U_1)|^2\phi_{c_n}(V_1-V_2)\psi_2'(\Delta)\Big]+o(1)\\
&= \frac{1}{2}\,\psi_2'(0)E\Big[|t^\top(U_2-U_1)|^2\phi_{c_n}(V_1-V_2)\Big]+o(1)\\
&= \frac{1}{2}\,\psi_2'(0)E\Big[\big[|t^\top(U_2-\mu(V_2))|^2+|t^\top(U_1-\mu_Q(V_1))|^2\big]\phi_{c_n}(V_1-V_2)\Big]+o(1)\\
&= \psi_2'(0)E\Big[\sigma^2(V_1)g_{c_n}(V_1)\Big]+o(1),
\end{aligned}$$

where $\sigma^2(V_1) = E(|t^\top(U_1-\mu(V_1))|^2 \mid V_1)$. Since $g_{c_n}(V_1)\to g(V_1)$ in probability, we obtain that

$$n^{1/2}E[H_n]\to\psi_2'(0)E[\sigma^2(V_1)g(V_1)] = t^\top At.$$

This completes the proof of (6.2). □

7 Proof of Theorem 2.6

Let us now give the proof of Theorem 2.6. We have already seen in Remark 2.8 that (C1) and (C5) imply (C1'). Thus we only have to show that (C1') and (C2)–(C4) imply (2.3). To simplify the notation, we abbreviate the conditional expectation $E^*_{\vartheta_n}$ appearing in the lemma by $\mathbb{E}_n$, the conditional expectation given $(\xi_1,\dots,\xi_{j-1},\eta_{nj},\xi_{j+1},\dots,\xi_n)$ under P_{ϑ_n} by $\mathbb{E}_{nj}$. We set

$$\begin{aligned}
T_{nj} &= \hat{L}_{nj}(\xi_j,\vartheta_n)-L(\xi_j,\vartheta_n)\\
&= \int L_n(\xi_j,\vartheta_n,\xi_1,\dots,\xi_{j-1},y,\xi_{j+1},\dots,\xi_n)M_{nj}(dy)-L(\xi_j,\vartheta_n)
\end{aligned}$$

and

$$T^*_{nj} = \mathbb{E}_{nj}(T_{nj}) = \int(\hat{L}_{nj}(x,\vartheta_n)-L(x,\vartheta_n))M_{nj}(dx).$$

Then (C1') can be written as $\frac{1}{n}\sum_{j=1}^n T^*_{nj} = o_{P_{\vartheta_n}}(n^{-1/2})$. Let $D_{nj} = T_{nj}-T^*_{nj}$. In view of (C1') and (C3), it suffices to show that $\bar{D}_n = \frac{1}{n}\sum_{j=1}^n D_{nj} = o_{P_{\vartheta_n}}(n^{-1/2})$. But this follows if we show that $\mathbb{E}_n(n\|\bar{D}_n\|^2) = o_{P_{\vartheta_n}}(1)$. Let us now establish this.
We have $\mathbb{E}_{nj}(\|D_{nj}\|^2)\le\mathbb{E}_{nj}(\|T_{nj}\|^2)$ and

$$\begin{aligned}
\mathbb{E}_{nj}(\|T_{nj}\|^2) &= \int\|\mathbb{E}_{nj}(\hat{L}_n(x,\vartheta_n))-L(x,\vartheta_n,\gamma)\|^2M_{nj}(dx)\\
&\le \int\mathbb{E}_{nj}(\|\hat{L}_n(x,\vartheta_n)-L(x,\vartheta_n,\gamma)\|^2)M_{nj}(dx),
\end{aligned}$$

so that (C2) implies

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_n(\|D_{nj}\|^2) = o_{P_n}(1). \tag{7.1}$$

For $i \neq j$, we have $\mathbb{E}_{ni}(T^*_{nj}) = \mathbb{E}_{nj}(\mathbb{E}_{ni}(T_{nj}))$ and thus

$$\begin{aligned}\mathbb{E}_{nj}(\|D_{nj} - \mathbb{E}_{ni}(D_{nj})\|^2) &= \mathbb{E}_{nj}(\|T_{nj} - \mathbb{E}_{ni}(T_{nj}) - \mathbb{E}_{nj}(T_{nj} - \mathbb{E}_{ni}(T_{nj}))\|^2)\\ &\leq \mathbb{E}_{nj}(\|T_{nj} - \mathbb{E}_{ni}(T_{nj})\|^2),\end{aligned}$$

and furthermore

$$\begin{aligned}\mathbb{E}_{nj}(\|T_{nj} - \mathbb{E}_{ni}(T_{nj})\|^2) &= \int \|\mathbb{E}_{nj}(\hat{L}_n(x,\vartheta_n)) - \mathbb{E}_{nj}(\mathbb{E}_{ni}(\hat{L}_n(x,\vartheta_n)))\|^2 M_{nj}(dx)\\ &\leq \int \mathbb{E}_{nj}(\|\hat{L}_n(x,\vartheta_n) - \mathbb{E}_{ni}(\hat{L}_n(x,\vartheta_n))\|^2) M_{nj}(dx).\end{aligned}$$

Thus (C4) implies

$$\frac{1}{n}\sum\sum_{i\neq j}\mathbb{E}_n(\|D_{nj} - \mathbb{E}_{ni}(D_{nj})\|^2) = o_{P_n}(1). \tag{7.2}$$

As $\mathbb{E}_{ni}(D_{ni}) = \mathbb{E}_{ni}(T_{ni} - T^*_{ni}) = T^*_{ni} - T^*_{ni} = 0$, we have, for $i \neq j$,

$$\mathbb{E}_n(D_{ni}^\top \mathbb{E}_{ni}(D_{nj})) = \mathbb{E}_n(\mathbb{E}_{ni}(D_{ni}^\top \mathbb{E}_{ni}(D_{nj}))) = \mathbb{E}_n(\mathbb{E}_{ni}(D_{ni}^\top)\mathbb{E}_{ni}(D_{nj})) = 0.$$

Similarly, one obtains $\mathbb{E}_n(\mathbb{E}_{nj}(D_{ni}^\top)D_{nj}) = 0$ and $\mathbb{E}_n(\mathbb{E}_{nj}(D_{ni}^\top)\mathbb{E}_{ni}(D_{nj})) = 0$. This shows that

$$\mathbb{E}_n(D_{ni}^\top D_{nj}) = \mathbb{E}_n((D_{ni} - \mathbb{E}_{nj}(D_{ni}))^\top (D_{nj} - \mathbb{E}_{ni}(D_{nj}))), \quad i \neq j.$$

From this and an application of the Cauchy-Schwarz inequality we obtain that

$$\left|\frac{1}{n}\sum\sum_{i\neq j}\mathbb{E}_n(D_{ni}^\top D_{nj})\right| \leq \frac{1}{n}\sum\sum_{i\neq j}\mathbb{E}_n(\|D_{nj} - \mathbb{E}_{ni}(D_{nj})\|^2) = o_{P_n}(1).$$

From the above we obtain that

$$\mathbb{E}_n(n\|\bar{D}_n\|^2) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_n(\|D_{nj}\|^2) + \frac{1}{n}\sum\sum_{i\neq j}\mathbb{E}_n(D_{ni}^\top D_{nj}) = o_{P_n}(1),$$

which is the desired result.

8 Some technical details

In this section we shall give some properties of the estimator of the score function ℓ of a density f with finite Fisher information. Throughout this section we assume that $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ are independent random variables with this density f and that $\delta_{n1}, \ldots, \delta_{nn}$ are

other random variables. We set $\tilde{\varepsilon}_{nj} = \varepsilon_{nj} + \delta_{nj}$. Let K be a positive, bounded symmetric density that is twice continuously differentiable with $|K'|/K$ and $|K''|/K$ bounded. Let a_n and b_n be sequences of positive numbers that converge to zero at a rate to be determined later. For $x \in \mathbb{R}$, let $K_n(x) = K(x/a_n)/a_n$ and set

$$\tilde{l}_n(x) = \frac{-\tilde{f}_n'(x)}{b_n + \tilde{f}_n(x)}, \qquad \hat{l}_n(x) = \frac{-\hat{f}_n'(x)}{b_n + \hat{f}_n(x)}, \qquad \bar{l}_n(x) = \frac{-\bar{f}_n'(x)}{b_n + \bar{f}_n(x)},$$

where

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{j=1}^n K_n(x - \tilde{\varepsilon}_{nj}), \qquad \hat{f}_n(x) = \frac{1}{n}\sum_{j=1}^n K_n(x - \varepsilon_{nj}),$$

and

$$\bar{f}_n(x) = \int f(x - a_n u)K(u)\,du = E[\hat{f}_n(x)].$$

Note also that $\bar{f}_n'(x) = E[\hat{f}_n'(x)]$. Elaborating on arguments of Bickel [2], Schick [12] showed that

$$\int |\bar{l}_n(x) - \ell(x)|^2 f(x)\,dx \to 0$$

whenever $a_n \to 0$ and $b_n \to 0$. We now improve upon his (3.16). Arguing as there one gets the bound

$$|\hat{l}_n(x) - \bar{l}_n(x)| \le |\hat{l}_n(x)|\frac{|\hat{f}_n(x) - \bar{f}_n(x)|}{b_n + \bar{f}_n(x)} + \frac{|\hat{f}_n'(x) - \bar{f}_n'(x)|}{b_n + \bar{f}_n(x)}.$$

Using the bounds $na_n Var(\hat{f}_n(x)) \le c\bar{f}_n(x)$, $na_n^3 Var(\hat{f}_n'(x)) \le c\bar{f}_n(x)$ and $|\hat{l}_n(x)| \le c/a_n$ used in Schick [12] one gets

$$\sup_{x\in\mathbb{R}} E[|\hat{l}_n(x) - \bar{l}_n(x)|^2] \le C/(na_n^3 b_n)$$

for a constant C. This shows that

$$E\Big[\int (\hat{l}_n(x) - \ell(x))^2 f(x)\,dx\Big] = o(1) + O((na_n^3 b_n)^{-1}).$$

Define now functions l_n from $\mathbb{R} \times \mathbb{R}^n \times (0,\infty) \times (0,\infty)$ by

$$l_n(x, y, a, b) = -\frac{\sum_{i=1}^n a^{-2}K'(a^{-1}(x - y_i))}{nb + \sum_{i=1}^n a^{-1}K(a^{-1}(x - y_i))},$$

for $x \in \mathbb{R}$, $a, b > 0$ and $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Then we can express

$$\tilde{l}_n(x) = l_n(x, (\tilde{\varepsilon}_{n1}, \ldots, \tilde{\varepsilon}_{nn}), a_n, b_n)$$

and

$$\hat{l}_n(x) = l_n(x, (\varepsilon_{n1}, \ldots, \varepsilon_{nn}), a_n, b_n), \quad x \in \mathbb{R}.$$

Since $K'(0) = 0$, it is easy to see that

$$l_n(y_j, y, a, b) = l_{n-1}(y_j, (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n), a, (nb + K(0)/a)/(n-1)),$$

From this and the above we now immediately obtain

$$E\Big(\frac{1}{n}\sum_{j=1}^{n}(\hat{l}_n(\varepsilon_{nj}) - \ell(\varepsilon_{nj}))^2\Big) = o(1) + O((na_n^3 b_n)^{-1}).$$

For $y, z \in \mathbb{R}^n$ and $\nu = 1, 2$, we have

$$\sum_{j=1}^{n} |z_j K_n^{(\nu)}(y_j)| \le c a_n^{-\nu} \max_{1\le i\le n} |z_i| \sum_{j=1}^{n} K_n(y_j)$$

for some constant c. So we can bound the derivative of the map $g_j(t) = l_n(y_j + tz_j, y + tz, a_n, b_n)$ by a constant times $a_n^{-2} \max_{1\le i\le n} |z_i|$. This shows that

$$\frac{1}{n}\sum_{j=1}^{n} |\tilde{l}_n(\tilde{\varepsilon}_{nj}) - \hat{l}_n(\varepsilon_{nj})|^2 = O_p(a_n^{-4} \max_{1\le j\le n} \delta_{nj}^2).$$

Let us now summarize our findings.

Lemma 8.1 *Suppose* $na_n^3 b_n \to \infty$ *and* $\max_{1\le j\le n} \delta_{nj}^2 = o_p(a_n^4)$. *Then*

$$\frac{1}{n}\sum_{j=1}^{n} |\tilde{l}_n(\tilde{\varepsilon}_{nj}) - \ell(\varepsilon_{nj})|^2 = o_p(1).$$

Moreover,

$$\frac{1}{n}\sum_{j=1}^{n} |\tilde{l}_n(\tilde{\varepsilon}_{nj}) - \bar{l}_n(\varepsilon_{nj})|^2 = O_p(a_n^{-4} \max_{1\le j\le n} \delta_{nj}^2 + (na_n^3 b_n)^{-1}).$$

Acknowledgements
A. Schick was supported in part by NSF Grant DMS 0072174. We are thankful to two anonymous referees for their comments which helped improve the presentation.

References

[1] P.K. Bhattacharya and P. Zhao. Semiparametric inference in a partly linear model. *Ann. Statist.*, **25**, 244–262, 1997.

[2] P.J. Bickel. On adaptive estimation. *Ann. Statist.*, **10**, 647–671, 1982.

[3] J. Cuzick. Efficient estimates in semiparametric additive regression models with unknown error distributions. *Ann. Statist.*, **20**, 1129–1136, 1992.

[4] W. Härdle, P. Janssen and R. Serfling. Strong uniform consistency rates of estimators of conditional functionals. *Ann. Statist.*, **16**, 1428–1449, 1988.

[5] W. Härdle and S. Luckhaus. Uniform consistency of a class of regression function estimators. *Ann. Statist.*, **12**, 612–623, 1984.

[6] N.L. Hjort and D. Pollard. Asymptotics for minimisers of convex processes. Unpublished manuscript. 1993.

[7] C.A.J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, **15**, 1548–1563, 1987.

[8] L. Le Cam. Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.*, **3**, 37–98, 1960.

[9] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186–199, 1991.

[10] W. Rudin. *Real and Complex Analysis. 2nd ed.* New York, 1974, McGraw Hill.

[11] A. Schick. On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, **14**, 1139–1151, 1986.

[12] A. Schick. A note on the construction of asymptotically linear estimators. *J. Statist. Planning Inference*, **16**, 89–105, 1987. Correction, **22**, 269–270, 1989.

[13] A. Schick. On efficient estimation in regression models. *Ann. Statist.*, **21**, 1486–1521, 1993. Correction and addendum, **23**, 1862–1863, 1995.

[14] A. Schick. Efficient estimation in regression models with unknown scale functions. *Math. Meth. Statist.*, **3**, 171–212.

[15] A. Schick. Root-n consistent and efficient estimation in semiparametric additive regression models. *Statist. Probab. Lett.*, **30**, 45–51, 1996.

[16] A. Schick. Efficient estimation in a semiparametric additive regression model with autoregressive errors. *Stoch. Process. Appl.*, **61**, 339–361, 1996.

[17] A. Schick. Efficient estimates in linear and nonlinear regression with heteroscedastic errors. *J. Statist. Inference Plann.*, **58**, 371–387, 1997.

[18] A. Schick. An adaptive estimator of the autocorrelation coefficient in regression models with autoregressive errors. *J. Time Series*, **19** , 575–589, 1998.

[19] A. Schick, A. (1999a) Efficient estimation of a shift in nonparametric regression. *Statist. Probab. Lett.*, **41**, 287–301.

[20] A. Schick. Efficient estimation in a semiparametric additive regression model with ARMA errors. In *Asymptotics, Nonparametrics, and Time Series*, S. Ghosh ed., 395–425. New York, 1999, Marcel Dekker.

[21] A. Schick. Sample splitting with Markov chains. *Bernoulli*, **7**, 243–266, 2001.

[22] A.W. van der Vaart. Estimating a real parameter in a class of semiparametric models. *Ann. Statist.*, **16**, 1450–1474, 1988.

[23] P-L. Zhao. Bandwidth-matched M-estimation in a partial linear model. Unpublished manuscript. 1995

Jeffrey S. Forrester
Department of Pharmacology
406 Preston Research Building
Vanderbilt University Medical Center
Nashville, TN 37232, USA
jeffrey.s.forrester@Vanderbilt.edu

William J. Hooper
Department of Mathematics and Computer Science
Clarkson University
Potsdam, NY 13699, USA
hooperw@clarkson.edu

Hanxiang Peng
Department of Mathematics
University of Mississippi
University, MS 38677-1848, USA
mmpeng@oldmiss.edu

Anton Schick
Department of Mathematical Sciences
Binghamton University
Binghamton, NY 13902-6000, USA
anton@math.binghamton.edu