Submitted to the Annals of Statistics
arXiv: math.PR/0000000

MINIMAX ESTIMATION OF A FUNCTIONAL ON A STRUCTURED HIGH-DIMENSIONAL MODEL

By James M. Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen and Aad van der Vaart

Harvard T.H. Chan School of Public Health and Sanofi Genzyme and Stanford University and Universiteit Leiden

We introduce a new method of estimation of parameters in semiparametric and nonparametric models. The method is based U-statistics that are based on higher order influence functions that extend ordinary linear influence functions of the parameter of interest, and represent higher derivatives of this parameter. For parameters for which the representation cannot be perfect the method often leads to a bias-variance trade-off, and results in estimators that converge at a slower than $\sqrt{n}$-rate. In a number of examples the resulting rate can be shown to be optimal. We are particularly interested in estimating parameters in models with a nuisance parameter of high dimension or low regularity, where the parameter of interest cannot be estimated at $\sqrt{n}$-rate, but we also consider efficient $\sqrt{n}$-estimation using novel nonlinear estimators. The general approach is applied in detail to the example of estimating a mean response when the response is not always observed.

1. Introduction. Let $X_1, X_2, \ldots, X_n$ be a random sample from a density p relative to a measure μ on a sample space $(\mathcal{X}, \mathcal{A})$. It is known that p belongs to a collection $\mathcal{P}$ of densities, and the problem is to estimate the value $\chi(p)$ of a functional $\chi\colon \mathcal{P} \to \mathbb{R}$. Our main interest is in the situation of a semiparametric or nonparametric model, where $\mathcal{P}$ is infinite dimensional, and especially in the case when the model is described through parameters of low regularity. In this case the parameter $\chi(p)$ may not be estimable at the "usual" $\sqrt{n}$-rate.

In low-dimensional semiparametric models estimating equations have been found a good strategy for constructing estimators [2, 30, 36]. In our present setting it will be more convenient to consider one-step versions of

*The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637 and from the National Institutes of Health grants AI112339, AI32475, AI113251 and ES020337.

AMS 2000 subject classifications: Primary 62G05, 62G20, 62G20, 62F25

Keywords and phrases: Nonlinear functional, nonparametric estimation, U-statistic, influence function, tangent space

such estimators, which take the form

$$\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{P}_n \chi_{\hat{p}_n}, \tag{1.1}$$

for $\hat{p}_n$ an initial estimator for p and $x \mapsto \chi_p(x)$ a given measurable function, for each $p \in \mathcal{P}$, and $\mathbb{P}_n f$ short hand notation for $n^{-1}\sum_{i=1}^n f(X_i)$.

One possible choice in (1.1) is $\chi_p = 0$, leading to the plug-in estimator $\chi(\hat{p}_n)$. However, unless the initial estimator $\hat{p}_n$ possesses special properties, this choice is typically suboptimal. Better functions χ_p can be constructed by consideration of the *tangent space* of the model. To see this, we write (with $P\chi_{\hat{p}}$ shorthand for $\int \chi_{\hat{p}}(x)\, dP(x)$)

$$\hat{\chi}_n - \chi(p) = \big[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\big] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}. \tag{1.2}$$

Because it is properly centered, we may expect the sequence $\sqrt{n}(\mathbb{P}_n - P)\chi_{\hat{p}_n}$ to tend in distribution to a mean-zero normal distribution. The term between square brackets on the right of (1.2), which we shall refer to as the *bias term*, depends on the initial estimator $\hat{p}_n$, and it would be natural to construct the function χ_p such that this term does not contribute to the limit distribution, or at least is not dominating the expression. Thus we would like to choose this function such that the "bias term" is no bigger than of the order $O_P(n^{-1/2})$. A good choice is to ensure that the term $P\chi_{\hat{p}_n}$ acts as minus the first derivative of the functional χ in the "direction" $\hat{p}_n - p$. Functions $x \mapsto \chi_p(x)$ with this property are known as *influence functions* in semiparametric theory [13, 19, 31, 5, 2], go back to the *von Mises calculus* due to [29], and play an important role in robust statistics [11, 9], or [36], Chapter 20.

For an influence function we may expect that the "bias term" is quadratic in the error $d(\hat{p}_n, p)$, for an appropriate distance d. In that case it is certainly negligible as soon as this error is of order $o_P(n^{-1/4})$. Such a "no-bias" condition is well known in semiparametric theory (e.g. condition (25.52) in [36] or (11) in [17]). However, typically it requires that the model $\mathcal{P}$ be "not too big". For instance, a regression or density function on d-dimensional space can be estimated at rate $n^{-1/4}$ if it is a-priori known to have at least $d/2$ derivatives (indeed $\alpha/(2\alpha + d) \geq 1/4$ if $\alpha \geq d/2$). The purpose of this paper is to develop estimation procedures for the case that no estimators exist that attain a $O_P(n^{-1/4})$ rate of convergence. The estimator (1.1) is then suboptimal, because it fails to make a proper trade-off between "bias" and "variance": the two terms in (1.2) have different magnitudes. Our strategy is to replace the linear term $\mathbb{P}_n\chi_p$ by a general U-statistic $\mathbb{U}_n\chi_p$, for an appropriate m-dimensional *influence function* $(x_1, \ldots, x_m) \mapsto \chi_p(x_1, \ldots, x_m)$, chosen using a type of von Mises expansion of $p \mapsto \chi(p)$. Here the order

m is adapted to the size of the model $\mathcal{P}$ and the type of functional to be estimated.

Unfortunately, "exact" higher-order influence functions turn out to exist only for special functionals χ. To treat general functionals χ we approximate these by simpler functionals, or use approximate influence functions. The rate of the resulting estimator is then determined by a trade-off between bias and variance terms. It may still be of order $1/\sqrt{n}$, but it is typically slower. In the former case, surprisingly, one may obtain semiparametric efficiency by estimators whose variance is determined by the linear term, but whose bias is corrected using higher order influence functions. The latter case will be of more interest.

The conclusion that the "bias term" in (1.2) is quadratic in the estimation error $d(\hat{p}_n, p)$ is based on a worst case analysis. First, there exist a large number of models and functionals of interest that permit a first order influence function that is unbiased in the nuisance parameter. (E.g. adaptive models as considered in [1], models allowing a sufficient statistic for the nuisance parameter as in [34, 35], mixture models as considered in [16, 21, 32], and convex-linear models in survival analysis.) In such models there is no need for higher-order influence functions. Second, the analysis does not take special, structural properties of the initial estimators $\hat{p}_n$ into account. An alternative approach would be to study the bias of a particular estimator in detail, and adapt the influence function to this special estimator. The strategy in this paper is not to use such special properties and focus on influence functions that work with general initial estimators $\hat{p}_n$.

The motivation for our new estimators stems from studies in epidemiology and econometrics that include covariates whose influence on an outcome of interest cannot be reliably modelled by a simple model. These covariates may themselves not be of interest, but are included in the analysis to adjust the analysis for possible bias. For instance, the mechanism that describes why certain data is missing is in terms of conditional probabilities given several covariates, but the functional form of this dependence is unknown. Or, to permit a causal interpretation in an observational study one conditions on a set of covariates to control for confounding, but the form of the dependence on the confounding variables is unknown. One may hypothesize in such situations that the functional dependence on a set of (continuous) covariates is smooth (e.g. $d/2$ times differentiable in the case of d covariates), or even linear. Then the usual estimators will be accurate (at order $O_P(n^{-1/2})$) if the hypothesis is true, but they will be badly biased in the other case. In particular, the usual normal-theory based confidence intervals may be totally misleading: they will be both too narrow and wrongly located. The

methods in this paper yield estimators with (typically) wider corresponding confidence intervals, but they are correct under weaker assumptions.

The mathematical contributions of the paper are to provide a heuristic for constructing minimax estimators in semiparametric models, and to apply this to a concrete model, which is a template for a number of other models (see [23, 33]). The methods connect to earlier work [10, 18] on the estimation of functionals on nonparametric models, but differ by our focus on functionals that are defined in terms of the structure of a semiparametric model. This requires an analysis of the inverse map from the density of the observations to the parameters, in terms of the semiparametric tangent spaces of the models. Our second order estimators are related to work on quadratic functionals, or functionals that are well approximated by quadratic functionals, as in [8, 12, 3, 4, 14, 15, 6, 7]. While we place the construction of minimax estimators for these special functionals in a wider framework, our focus differs by going beyond quadratic estimators and to consider semiparametric models.

Our mathematical results are in part conditional on a scale of regularity parameters (through the dimension given in (9.9) and a partition of this dimension that depends on two of these parameters). We hope to discuss adaptation to these parameters in future work.

General heuristics of our construction are given in Section 4. Sections 5–9 are devoted to constructing new estimators for the mean response effect in missing data problems. The latter are introduced in Section 3, so that they can serve as illustration to the general heuristics in Section 4. In Section S11 (in the supplement [22]) we briefly discuss other problems, including estimating a density at a point, where already first order influence functions do not exist and our heuristics naturally lead to projection estimators, and estimating a quadratic functional, where our approach produces standard estimators from the literature in a natural way. Section 10 (partly in the supplement [22]) collects technical proofs. Sections S12, S13 and S14 (in the supplement [22]) discuss three key concepts of the paper: influence functions, projections and U-statistics. Numbers referring to the supplement are preceded by the symbol "S".

2. Notation. Let $\mathbb{U}_n$ denote the *empirical U-statistic* measure, viewed as an operator on functions. For given $m \leq n$ and a function $f: \mathcal{X}^m \to \mathbb{R}$ on the sample space this is defined by

$$\mathbb{U}_n f = \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{1 \leq i_1 \neq i_2 \neq \cdots \neq i_m \leq n} \sum \cdots \sum f(X_{i_1}, X_{i_2}, \cdots, X_{i_m}).$$

We do not let the order m show up in the notation $\mathbb{U}_n f$. This is unnecessary, as the notation is consistent in the following sense: if a function $f\colon \mathcal{X}^l \to \mathbb{R}$ of $l < m$ arguments is considered a function of m arguments that is constant in its last $m-l$ arguments, then the right side of the preceding display is well defined and is exactly the corresponding U-statistic of order l. In particular, $\mathbb{U}_n f$ is the *empirical distribution* $\mathbb{P}_n$ applied to f if $f\colon \mathcal{X} \to \mathbb{R}$ depends on only one argument.

We write $P^n \mathbb{U}_n f = P^m f$ for the expectation of $\mathbb{U}_n f$ if $X_1, \ldots, X_n$ are distributed according to the probability measure P, and for the expectation of f under the product measure P^m of m copies of P. We also use this operator notation for the expectations of statistics in general. If the distribution of the observations is given by a density p, then we use P as the measure corresponding to p, and use the preceding notations likewise. Finally $\mathbb{U}_n - P^m$ denotes the centered U-statistic empirical measure, defined by $(\mathbb{U}_n - P^m)f = \mathbb{U}_n f - P^m f$, for any integrable function f.

We call f *degenerate* relative to P if $\int f(x_1, \ldots, x_m)\, dP(x_i) = 0$ for every i and every $(x_j\colon j \neq i)$, and we call f *symmetric* if $f(x_1, \ldots, x_m)$ is invariant under permutation of the arguments $x_1, \ldots, x_m$. Given an arbitrary measurable function $f\colon \mathcal{X}^m \to \mathbb{R}$ we can form a function that is degenerate relative to P by subtracting the orthogonal projection in $L_2(P^m)$ onto the functions of at most $m-1$ variables. This degenerate function can be written in the form (e.g. [36], Lemma 11.11)

$$(D_P f)(X_1, \ldots, X_m) = \sum_{A \subset \{1,\ldots,m\}} (-1)^{m-|A|} \mathrm{E}_P \Big[f(X_1, \ldots, X_m) \big| \, X_i\colon i \in A \Big], \tag{2.1}$$

where the sum is over all subsets A of $\{1, \ldots, m\}$, including the empty set. Here the conditional expectation $\mathrm{E}\big[f(X_1, \ldots, X_m) \big| \, X_i\colon i \in \emptyset\big]$ is understood to be the unconditional expectation $\mathrm{E} f(X_1, \ldots, X_m) = P^m f$. If the function f is symmetric, then so is the function $D_P f$.

Given two functions $g, h\colon \mathcal{X} \to \mathbb{R}$ we write $g \times h$ for the function $(x, y) \mapsto g(x)h(y)$. More generally, given m functions $g_1, \ldots, g_m$ we write $g_1 \times \cdots \times g_m$ for the tensor product of these functions. Such product functions are degenerate iff all functions in the product have mean zero.

A *kernel operator* $K\colon L_r(\mathcal{X}, \mathcal{A}, \mu) \to L_r(\mathcal{X}, \mathcal{A}, \mu)$ takes the form $(Kf)(x) = \int \bar{K}(x, y) f(y)\, d\mu(y)$ for some measurable function $\bar{K}\colon \mathcal{X}^2 \to \mathbb{R}$. We shall abuse notation in denoting the operator K and the *kernel* $\bar{K}$ with the same symbol: $K = \bar{K}$. A (weighted) projection on a finite-dimensional space is a kernel operator. We discuss such projections in Section S13.

The set of measurable functions whose rth absolute power is μ-integrable is denoted $L_r(\mu)$, with norm $\|\cdot\|_{r,\mu}$, or $\|\cdot\|_r$ if the measure is clear; or also

as $L_r(w)$ with norm $\|\cdot\|_{r,w}$ if w is a density relative to a given dominating measure. For $r=\infty$ the notation $\|\cdot\|_\infty$ refers to the uniform norm.

3. Estimating the mean response in missing data models. In this section we introduce our main example, which will be used as a running example in the next section. We also summarize the results obtained for this example in the remainder of the paper.

Suppose that a typical observation is distributed as $X=(YA,A,Z)$, for Y and A taking values in the two-point set $\{0,1\}$ and conditionally independent given Z.

This model is standard in biostatistical applications, with Y an "outcome" or "response variable", which is observed only if the indicator A takes the value 1. The covariate Z is chosen such that it contains all information on the dependence between the response and the missingness indicator A, thus making the response *missing at random*. Alternatively, we think of Y as a "counterfactual" outcome if a treatment were given ($A=1$) and estimate (half) the treatment effect under the assumption of *no unmeasured confounders*. (The results also apply without the "missing-at-random" assumption, but with a different interpretation; see Remark 3.1.)

The model can be parameterized by the marginal density f of Z (relative to some dominating measure ν) and the probabilities $b(z)=\mathrm{P}(Y=1|\,Z=z)$ and $a(z)^{-1}=\mathrm{P}(A=1|\,Z=z)$. (Using a for the inverse probability simplifies later formulas.) Alternatively, the model can be parameterized by the pair (a,b) and the function $g=f/a$, which is the conditional density of Z given $A=1$, up to the norming factor $\mathrm{P}(A=1)$. Thus the density p of an observation X is described by the triplet (a,b,f), or equivalently the triplet (a,b,g). For simplicity of notation we write p instead of $p_{a,b,f}$ or $p_{a,b,g}$, with the implicit understanding that a generic p corresponds one-to-one to a generic (a,b,f) or (a,b,g).

We wish to estimate the *mean response* $\mathrm{E}Y=\mathrm{E}b(Z)$, i.e. the functional

$$\chi(p)=\int bf\,d\nu=\int abg\,d\nu.$$

Estimators that are $\sqrt{n}$-consistent and asymptotically efficient in the semiparametric sense have been constructed using a variety of methods (e.g. [26, 27], or see Section 5), but only if a or b, or both, parameters are restricted to sufficiently small regularity classes. For instance, if the covariate Z ranges over a compact, convex subset $\mathcal{Z}$ of $\mathbb{R}^d$, then the mentioned papers provide $\sqrt{n}$-consistent estimators under the assumption that a and b belong

to Hölder classes $C^\alpha(\mathcal{Z})$ and $C^\beta(\mathcal{Z})$ with α and β large enough that

$$\frac{\alpha}{2\alpha+d}+\frac{\beta}{2\beta+d}\geq\frac{1}{2}. \tag{3.1}$$

(See e.g. Section 2.7.1 in [37] for the definition of Hölder classes.) For moderate to large dimensions d this is a restrictive requirement. In the sequel we consider estimation for arbitrarily small α and β.

3.1. *Summary of results.* Throughout we assume that the parameters a, b and g are contained in Hölder spaces $C^\alpha(\mathcal{Z})$, $C^\beta(\mathcal{Z})$ and $C^\gamma(\mathcal{Z})$ of functions on a compact, convex domain in $\mathbb{R}^d$. We derive two types of results:

(a) In Section 8 we show that a $\sqrt{n}$-rate is attainable by using a higher order influence function (of order determined by γ) as long as

$$\frac{\alpha+\beta}{2}\geq\frac{d}{4}. \tag{3.2}$$

This condition is strictly weaker than the condition (3.1) under which the linear estimator attains a $\sqrt{n}$-rate. Thus even in the $\sqrt{n}$-situation higher order estimating equations may yield estimators that are applicable in a wider range of models. For instance, in the case that $\alpha=\beta$ the cut-off (3.1) arises for $\alpha=\beta\geq d/2$, whereas (3.2) reduces to $\alpha=\beta\geq d/4$.

(b) We consider minimax estimation in the case $(\alpha+\beta)/2<d/4$, when the rate becomes slower than $1/\sqrt{n}$. It is shown in [25] that even if $g=f/a$ were known, then the minimax rate for a and b ranging over balls in the Hölder classes $C^\alpha(\mathcal{Z})$ and $C^\beta(\mathcal{Z})$ cannot be faster than $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$. In Section 9 we show that this rate is attainable if g is known, and also if g is unknown, but is a-priori known to belong to a Hölder class $C^\gamma(\mathcal{Z})$ for sufficiently large γ, as given by (9.11). (Heuristic arguments, not discussed in this paper, appear to indicate that for smaller γ the minimax rate is slower than $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.)

We start by discussing the first and second order estimators in Sections 5 and 6, where the first is merely a summary of well known facts, but the second already contains some key elements of the new approach of the present paper. The preceding results (a) and (b) are next obtained in Sections 8 ($\sqrt{n}$-rate if $(\alpha+\beta)/2\geq d/4$) and 9 (slower rate if $(\alpha+\beta)/2<d/4$), using the higher-order influence functions of an approximate functional, which is defined in the intermediate Section 7. In the next section we discuss the general heuristics of our approach.

ASSUMPTION 3.1. *We assume throughout that the functions* $1/a, b, g$ *and their preliminary estimators* $1/\hat{a}, \hat{b}, \hat{g}$ *are bounded away from their extremes: 0 and 1 for the first two, and 0 and* ∞ *for the third.*

REMARK 3.1. *The assumption that the responses are "missing at random (MAR)" is used to identify the mean response functional. Without this assumption the results of the paper are still valid, but concern the functional* $\int b_1(z)f(z)\,dz$, *in which* $b_1(z) = \mathrm{E}(Y|\,A = 1, Z = z)$ *has taken the place of* $b(z) = \mathrm{E}(Y|\,Z = z)$, *two functions that are identical under MAR. This follows from the fact that the likelihoods of* $X = (YA, A, Z)$ *without or with assuming MAR take exactly the same form, as given in (4.5), but with* b *replaced by* b_1. *After this replacement all results go through. However, the functional* $\int b_1(z)f(z)\,dz$ *has the interpretation of the mean response only when MAR holds.*

4. General heuristics. Our basic estimator has the form (1.1) except that we replace the linear term by a general U-statistic. Given measurable functions $\chi_p\colon \mathcal{X}^m \to \mathbb{R}$, for a fixed *order* m, we consider estimators $\hat{\chi}_n$ of $\chi(p)$ of the type

$$\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n\chi_{\hat{p}_n}. \tag{4.1}$$

The initial estimators $\hat{p}_n$ are thought to have a certain (optimal) convergence rate $d(\hat{p}_n, p) \to 0$, but need not possess (further) special properties. Throughout we shall treat these estimators as being based on an independent sample of observations, so that $\hat{p}_n$ and $\mathbb{U}_n$ in (4.1) are independent. This takes away technical complications, and allows us to focus on rates of estimation in full generality. (A simple way to avoid the resulting asymmetry would be to swap the two samples, calculate the estimator a second time and take the average.)

4.1. *Influence functions.* The key is to find suitable "influence functions" χ_p. A decomposition of type (1.2) for the estimator (4.1) yields

$$\hat{\chi}_n - \chi(p) = \big[\chi(\hat{p}_n) - \chi(p) + P^m\chi_{\hat{p}_n}\big] + (\mathbb{U}_n - P^m)\chi_{\hat{p}_n}. \tag{4.2}$$

This suggests to construct the influence functions such that $-P^m\chi_{\hat{p}_n}$ represents the first m terms of the Taylor expansion of $\chi(\hat{p}_n) - \chi(p)$. We shall translate this requirement into a manageable form, and next work it out in detail for the missing data problem.

First the requirement implies that the influence function used in (4.1) must be unbiased:

$$P^m\chi_p = 0. \tag{4.3}$$

Next, to operationalize a "Taylor expansion" on the (infinite-dimensional) "manifold" $\mathcal{P}$ we employ "smooth" submodels $t \mapsto p_t$. These are defined as maps from a neighbourhood of $0 \in \mathbb{R}$ to $\mathcal{P}$ that pass through p at $t = 0$ (i.e. $p_0 = p$) such that the derivatives in the following exist. For a large model there will be many such submodels, approaching p from various "directions". Given a collection of submodels we determine χ_p such that, for each submodel $t \mapsto p_t$,

$$\frac{d^j}{dt^j}_{|t=0} \chi(p_t) = -\frac{d^j}{dt^j}_{|t=0} P^m \chi_{p_t}, \qquad j = 1, \ldots, m.$$

The subscript $|t = 0$ on the differential quotients means "derivative evaluated at $t = 0$", i.e. at $p = p_0$. A slight strengthening is to impose this condition "everywhere" on the path, i.e. the jth derivative of $t \mapsto \chi(p_t)$ at t is the jth derivative of $h \mapsto -P_t^m \chi_{p_{t+h}}$ at $h = 0$, for every t. (Here P_t is the measure corresponding to the density p_t and $P_t^m f$ the expectation of a function f under the m-fold product of these measures.) If the map $(s, t) \mapsto P_s^m \chi_{p_t}$ is smooth, then the latter implies (cf. Lemma S12.1 applied with $\chi = f$ and $g(s, t) = -P_t^m \chi_{p_s}$)

$$\frac{d^j}{dt^j}_{|t=0} \chi(p_t) = \frac{d^j}{dt^j}_{|t=0} P_t^m \chi_p, \qquad j = 1, \ldots, m. \tag{4.4}$$

Relative to the previous formula the subscript t on the right hand side has changed places, and the negative sign has disappeared. This is similar to the "Bartlett equalities" familiar from manipulating expectations of scores and their higher derivatives. We take this equation together with unbiasedness as the defining property. Thus a measurable function $\chi_p \colon \mathcal{X}^m \to \mathbb{R}$ is said to be an mth order *influence function* at p of the functional $p \mapsto \chi(p)$ relative to a given collection of one-dimensional submodels $t \mapsto p_t$ (with $p_0 = p$) if it satisfies (4.3) and (4.4), for every submodel under consideration.

Equation (4.4) implies a Taylor expansion of $t \mapsto \chi(p_t)$ at $t = 0$ of order m, but in addition requires that the derivatives of this map can be represented as *expectations* involving a function χ_p. The latter is made operational by requiring the derivatives to be identical to those of the map $t \mapsto P_t^m \chi_p$, which automatically have the desired representation. The representation as an expectation is essential for the construction of estimators. For exploiting derivatives up to the mth order, groups of m observations can be used to match the expectation P^m; this leads to U-statistics of order m.

It is also essential that the expectation is relative to the law of the observations $X_1, \ldots, X_n$. In a structured model, such as the missing data problem, the law P_η of the observations depends on a parameter η and the functional

of interest is a quantity $\psi(\eta)$ defined in term of η. Then the representation requires to represent the derivative of the map $\eta \mapsto \psi(\eta)$ as an expectation relative to P_η. An expansion of just $\eta \mapsto \psi(\eta)$ without reference to the data distribution is not sufficient. Expressing the derivates in P_η implicitly utilises the inverse map $P_\eta \mapsto \eta$, but by directly defining the influence function by (4.4) we sidestep an expansion of $\eta \mapsto \psi(\eta)$ and explicit inversion of the latter map.

We allow that there may be more than one influence function. In particular, we do not require χ_p in (4.4) to be symmetric in its arguments, although a given influence function can always be symmetrized without loss of generality. Furthermore, as the collection of paths $t \mapsto p_t$ is restricted by the model, which may be smaller than the set of all possible densities on the sample space, certain projections of an influence function may also be influence functions.

EXAMPLE 4.1 (Classical U-statistic). The mean functional $\chi(p) = \mathrm{E}_p \mathbb{U}_n f = P^k f$ of a kth order U-statistic has mth order influence function given by $\chi_p(x_1, \ldots, x_m) = f(x_1, \ldots, x_k) - P^k f$, for every $m \geq k$. Alternatively, the symmetrized version $\mathbb{U}_m f - P^k f$ of this function is also an influence function. This example connects to classical U-statistic theory, and may serve to gain some insight in the definition, but our interest in influence functions will go in a different direction.

In the preceding claim we did not specify the set of paths $t \mapsto p_t$. In fact the claim is true for the nonparametric model and all reasonable paths. The claim follows trivially from the fact that $t \mapsto \chi(p_t) = P_t^k f$ has the same derivatives as $t \mapsto P_t^m \chi_p = P_t^m f - P^k f = P_t^k f - P^k f$, where in the last equality we use that $m \geq k$. (The jth derivative for $j > k$ vanishes.)

For $1 \leq m < k$ one can verify, with more effort, that the orthogonal projection in $L_2(P^k)$ of f on the subspace of functions of m variables is an influence function.

EXAMPLE 4.2 (Missing data, paths). The missing data model introduced in Section 3 is parameterized by the parameter triplet (a, b, f). The likelihood of a typical observation $X = (YA, A, Z)$ can be seen to take the form

$$p_{a,b,f}(X) = f(Z)\Big(\frac{1}{a(Z)} b(Z)^Y \big(1 - b(Z)\big)^{1-Y}\Big)^A \Big(1 - \frac{1}{a(Z)}\Big)^{1-A}. \tag{4.5}$$

Submodels are naturally constructed as $t \mapsto p_{a_t, b_t, f_t}$, for given curves $t \mapsto a_t$, $t \mapsto b_t$ and $t \mapsto f_t$ in the respective parameter spaces.

In view of Assumption 3.1 paths of the form $a_t = a + t\underline{\mathfrak{a}}$ and $b_t = b + t\underline{\mathfrak{b}}$, for given bounded, measurable functions $\underline{\mathfrak{a}}, \underline{\mathfrak{b}} \colon \mathcal{Z} \to \mathbb{R}$ are valid curves in the

parameter space, at least for t in a neighbourhood of 0. We may restrict the perturbations $\underline{\mathfrak{a}}$ and $\underline{\mathfrak{b}}$ to be sufficiently smooth to ensure that these paths also belong to the appropriate Hölder spaces.

It is convenient to define the perturbation of the marginal density slightly differently in the form $f_t = f(1+t\underline{\mathfrak{f}})$. For a given bounded function $\underline{\mathfrak{f}}\colon \mathcal{Z} \to \mathbb{R}$ with $\int \underline{\mathfrak{f}} f \, d\nu = 0$, and sufficiently small $|t|$, each f_t is indeed a probability density. The advantage of defining the perturbation by $f\underline{\mathfrak{f}}$ instead of $\underline{\mathfrak{f}}$ is simply that in the present form $\underline{\mathfrak{f}} = d/dt_{|t=0} \log f_t$ can be interpreted as the score function of the model $t \mapsto f_t$.

These paths are usually enough to identify influence functions. By slightly changing the definitions one might also allow non-bounded functions as "directions" of the perturbations.

4.2. *Relation to semiparametric theory and tangent spaces.* In semiparametric theory (e.g. [2, 19, 35, 31]) influence functions are described through inner products with score functions. We do not follow this route here, but make the connection in this section. Scores give a way of rewriting (4.4), which will be useful mainly for first order influence functions.

For a sufficiently regular submodel $t \mapsto p_t$ equation (4.4) for $m = 1$ can be written in the form

$$\frac{d}{dt}_{|t=0} \chi(p_t) = \frac{d}{dt}_{|t=0} P_t \chi_p = P(\chi_p g), \tag{4.6}$$

where $g = (d/dt)_{t=0} p_t/p$ is the *score function* of the model $t \mapsto p_t$ at $t = 0$. A function χ_p satisfying (4.6) is exactly what is called an *influence function* in semiparametric theory. The linear span of all scores attached some submodel $t \mapsto p_t$ is called the *tangent space* of the model at p and an influence function is an element of $L_2(p)$ whose inner products with the elements of the tangent space represent the derivative of the functional in the sense of (4.6) ([36], page 363, or [2, 19, 35, 31]).

EXAMPLE 4.3 (Missing data, score functions). To obtain the score functions at $t = 0$ of the one-dimensional submodels $t \mapsto p_t := p_{a_t, b_t, f_t}$ induced by paths of the form $a_t = a + t\underline{\mathfrak{a}}$, $b_t = b + t\underline{\mathfrak{b}}$, and $f_t = f(1 + t\underline{\mathfrak{f}})$, for given measurable functions $\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}}\colon \mathcal{Z} \to \mathbb{R}$ (where $\int \underline{\mathfrak{f}} f \, d\nu = 0$), we substitute these paths in the right side of equation (4.5) for the likelihood, take the logarithm, and differentiate at $t = 0$. If we insert the perturbations for the three parameters separately, keeping the other parameters fixed, we obtain what

could be called "partial score functions" given by

$$B_p^a \underline{\mathfrak{a}}(X) = -\frac{Aa(Z)-1}{a(Z)(a-1)(Z)}\underline{\mathfrak{a}}(Z), \qquad a-\text{score},$$
$$B_p^b \underline{\mathfrak{b}}(X) = \frac{A\big(Y-b(Z)\big)}{b(Z)(1-b)(Z)}\underline{\mathfrak{b}}(Z), \qquad b-\text{score},$$
$$B_p^f \underline{\mathfrak{f}}(X) = \underline{\mathfrak{f}}(Z), \qquad f-\text{score}.$$

The scores are deliberately written in a form suggesting operators B_p^a, B_p^b, B_p^f working on the three directions $\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}}$. These are called *score operators* in semiparametric theory, and their direct sum is the overall score operator, which we write as B_p. Thus $B_p(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})(X)$ is defined as the sum of the three left sides of the preceding equation.

We claim that the first-order influence function of the functional $\chi\colon p_{a,b,f} \mapsto \int bf\, d\nu$ is given by

$$\chi_p^{(1)}(X) = Aa(Z)\big(Y-b(Z)\big) + b(Z) - \chi(p). \tag{4.7}$$

To prove this well-known fact, it suffices to verify that this function satisfies, for every path $t \mapsto p_t$ as described previously,

$$\frac{d}{dt}_{|t=0}\chi(p_t) = \mathrm{E}_p\big[\chi_p^{(1)}(X)\, B_p(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})(X)\big].$$

This follows by straightforward calculations, where it suffices to verify the equation for each of the three perturbations separately. For instance, for a perturbation of only the parameter a, the left side of the display is clearly zero, as the functional does not depend on a. The right side with $\underline{\mathfrak{b}} = \underline{\mathfrak{f}} = 0$ reduces to $\mathrm{E}_p\big[\chi_p^{(1)}(X) B_p^a \underline{\mathfrak{a}}(X)\big]$, which can be seen to be zero from the fact that $Aa(Z)-1$ and $Y-b(Z)$ are uncorrelated given Z. The validity of the display for the two other types of scores can be verified similarly.

The advantage of choosing a an inverse probability is clear from the form of the (random part of the) influence function (4.7), which is bilinear in (a, b).

Computing (approximate) higher order influence functions for this model is a main achievement of this paper. Expressions are given later on.

For $m > 1$ equation (4.4) can be expanded similarly in terms of inner products of the influence function with score functions, but "higher-order score functions" arise next to ordinary score functions. Here we do not follow this route, but have defined an higher order influence function through

(4.4), and leave the alternative route to other papers. Suitable higher-order tangent spaces are discussed in [23] (also see [33]), using score functions as defined in [39]. A discussion of *second order* scores and tangent spaces can be found in [24]. Second order tangent spaces are also discussed in [20], from a different point of view of and with the purpose of defining *higher order efficiency* of estimators. Higher-order efficient estimators attain the first order efficiency bound (the "asymptotic Cramér-Rao bound") and also optimize certain lower order terms in their distribution or risk. In the present paper we are interested in *first order efficiency*, measured mostly by the convergence rate, which in the most interesting cases is slower than $\sqrt{n}$, and not in refinements of the first order behaviour.

4.3. *Computing the influence function.* Equation (4.4) involves multiple derivatives and many paths and is not easy to solve for χ_p. For actual computation of an influence function it is usually easier to derive higher order influence functions as influence functions of lower order ones.

To describe this operation, we need to decompose the influence function χ_p, or rather its symmetrized version in degenerate functions. Any mth order, zero-mean U-statistic can be decomposed as the sum of m degenerate U-statistics of orders $1, 2, \ldots, m$, by way of its Hoeffding decomposition. In the present situation we can write

$$\mathbb{U}_n\chi_p = \mathbb{U}_n\chi_p^{(1)} + \tfrac{1}{2}\mathbb{U}_n\chi_p^{(2)} + \cdots + \frac{1}{m!}\mathbb{U}_n\chi_p^{(m)},$$

where $\chi_p^{(j)}\colon \mathcal{X}^j \to \mathbb{R}$ is a degenerate kernel of j arguments, defined uniquely as a projection of χ_p (cf. [38] and (2.1)). Since χ_p is a function of m arguments, for $m = n$ the left side evaluates to the symmetrization of the function χ_p, and it is equal to χ_p if χ_p is already permutation symmetric in its arguments. The functions on the right side are similarly symmetric, and the equation can be read as a decomposition of the symmetrized version of χ_p into symmetrizations of certain degenerate functions $\chi_p^{(j)}$. Suitable (symmetric) functions $\chi_p^{(j)}$ in this decomposition can be found by the following algorithm:

[1] Let $x_1 \mapsto \bar\chi_p^{(1)}(x_1)$ be a first order influence function of the functional $p \mapsto \chi(p)$.
[2] Let $x_j \mapsto \bar\chi_p^{(j)}(x_1, \ldots, x_j)$ be a first order influence function of the functional $p \mapsto \bar\chi_p^{(j-1)}(x_1, \ldots, x_{j-1})$, for each $x_1, \ldots, x_{j-1}$, and $j = 2, \ldots, m$.
[3] Let $\chi_p^{(j)} = D_P\bar\chi_p^{(j)}$ be the degenerate part of $\bar\chi_p^{(j)}$ relative to P, as defined in (2.1).

See Lemma S12.2 for a proof. Thus higher order influence functions are constructed as first order influence functions of influence functions. Somewhat abusing language we shall refer to the function $\chi_p^{(j)}$ also as a "jth order influence function". The overall order m will be fixed at a suitable value; for simplicity we do not let this show up in the notation χ_p.

The starting influence function $\bar{\chi}_p^{(1)}$ in step [1] may be any first order influence function (thus satisfying (4.4) for $m = 1$, or alternatively a function χ_p that satisfies (4.6) for every score g); it does not have to possess mean zero, or be an element of the first order tangent space. A similar remark applies to the (first order) influence functions found in step [2]. It is only in step [3] that we make the influence functions degenerate.

EXAMPLE 4.4 (Missing data, higher order scores). The second order score function for the missing data problem is computed as a derivative of the first order score function (4.7) in Section 6. As will be explained momentarily the result (6.3) is actually only a partial second order score function.

Higher order score functions are computed in Sections 8 and 9.

4.4. *Bias-variance trade-off.* Because it is centered, the "variance part" in (4.2), the variable $(\mathbb{U}_n - P^m)\chi_{\hat{p}_n}$, should not change noticeably if we replace $\hat{p}_n$ by p, and be of the same order as $(\mathbb{U}_n - P^m)\chi_p$. For a fixed square-integrable function χ_p the latter centered U-statistic is well known to be of order $O_P(n^{-1/2})$, and asymptotically normal if suitably scaled. A completely successful representation of the "bias" $R_n = \chi(\hat{p}_n) - \chi(p) + P^m\chi_{\hat{p}_n}$ in (4.2) would lead to an error $R_n = O_P\big(d(\hat{p}_n, p)^{m+1}\big)$, which becomes smaller with increasing order m. Were this achievable for any m, then a $\sqrt{n}$-estimator would exist no matter how slow the convergence rate $d(\hat{p}_n, p)$ of the initial estimator. Not surprisingly, in many cases of interest this ideal situation is not real. This is due to the non-existence of influence functions that can exactly represent the Taylor expansion of $\chi(\hat{p}_n) - \chi(p)$.

In general, we have to content ourselves with a partial representation. Next to a first bias in the form of the remainder term R_n of order $O_P\big(d(\hat{p}_n, p)^{m+1}\big)$, we then also incur a "representation bias". The latter bias can be made arbitrarily small by choice of the influence function, but only at the cost of increasing its variance. We thus obtain a trade-off between a variance and two biases. This typically results in a variance that is larger than $1/n$, and a rate of convergence that is slower than $1/\sqrt{n}$, although sometimes a nontrivial bias correction is possible without increasing the variance.

EXAMPLE 4.5 (Missing data, variance and bias terms). The missing data problem is parameterized by the triple (a, b, g) and hence the preliminary estimator $\hat{p}$ is constructed from estimates $\hat{a}$ and $\hat{b}$ and $\hat{g}$ of these parameters.

The remainder bias R_n of the estimator for $m = 1$ is given in (5.1). It is bounded by $\|\hat{a} - a\|_2 \, \|\hat{b} - b\|_2$ and hence is quadratic in the preliminary estimator, as expected. There is no representation bias at this order. The variance of the linear estimator is of order $1/n$. If the preliminary estimators can be constructed so that the product $\|\hat{a} - a\|_2 \, \|\hat{b} - b\|_2$ is of lower or equal order than $1/n$, then the estimator is rate-optimal. Otherwise a higher order estimator is preferable.

The bias and variance terms of the estimator for $m = 2$ are given in Theorem 6.1. The remainder bias R_n is of the order $\|\hat{a}-a\|_r \, \|\hat{b}-b\|_r \, \|\hat{g}-g\|_r$, cubic in the preliminary estimator, while the representation bias is of the order the product of the remainders after projecting $\hat{a} - a$ and $\hat{b} - b$ onto a linear space chosen by the statistician. The dimension k of this space determines the variance of the estimator, adding a contribution of the order k/n^2. Following the statement of the theorem it is shown how the variance can be traded off versus the two biases. It turns concluded that in case the remainder bias of order $\|\hat{a} - a\|_r \, \|\hat{b} - b\|_r \, \|\hat{g} - g\|_r$ actively determines the outcome of this trade-off, then an estimator of higher order is preferable.

For higher orders $m > 2$ the remainder bias decreases to $\|\hat{a} - a\|_r \, \|\hat{b} - b\|_r \, \|\hat{g} - g\|_r^{m-1}$, but the representation bias becomes increasingly complex. A discussion is deferred to Sections 8 and 9.

4.5. *Approximate functionals.* An attractive method to find approximating influence functions is to compute exact influence functions for an approximate functional. Because smooth functionals on finite-dimensional models typically possess influence functions to any order, projections on finite-dimensional models may deliver such approximations.

A simple approximation would be $\chi(\widetilde{p})$ for a given map $p \mapsto \widetilde{p}$ mapping the model $\mathcal{P}$ onto a suitable "smaller" model $\widetilde{\mathcal{P}}$ (typically a submodel $\widetilde{\mathcal{P}} \subset \mathcal{P}$). A closer approximation can be obtained by also including a derivative term. Consider the functional $\widetilde{\chi}\colon \mathcal{P} \to \mathbb{R}$ defined by, for a given map $p \mapsto \widetilde{p}$,

$$\widetilde{\chi}(p) = \chi(\widetilde{p}) + P\chi_{\widetilde{p}}^{(1)}. \tag{4.8}$$

(A complete notation would be $\widetilde{p}(p)$; the right hand side depends on p at three places.) By the definition of an influence function the term $-P\chi_{\widetilde{p}}^{(1)}$ acts as the first order Taylor expansion of $\chi(\widetilde{p}) - \chi(p)$. Consequently, we may expect that

$$\big|\widetilde{\chi}(p) - \chi(p)\big| = O\big(d(\widetilde{p}, p)^2\big). \tag{4.9}$$

This ought to be true for any "projection" $p \mapsto \widetilde{p}$. If we choose the projection such that, for any path $t \mapsto p_t$,

$$\frac{d}{dt}_{|t=0}\Big(\chi(\widetilde{p}_t) + P_0\chi^{(1)}_{\widetilde{p}_t}\Big) = 0, \tag{4.10}$$

then the functional $p \mapsto \widetilde{\chi}(p)$ will be locally (around p_0) equivalent to the functional $p \mapsto \chi\big(\widetilde{p}_0\big) + P\chi^{(1)}_{\widetilde{p}_0}$ (which depends on p in only one place, p_0 being fixed) in the sense that the first order influence functions are the same. The first order influence function of the latter (linear) functional at p_0 is equal to $\chi^{(1)}_{\widetilde{p}_0}$, and hence for a projection satisfying (4.10) the first order influence function of the functional $p \mapsto \widetilde{\chi}(p)$ will be

$$\widetilde{\chi}^{(1)}_p = \chi^{(1)}_{\widetilde{p}}. \tag{4.11}$$

In words, this means that the influence function of the approximating functional $\widetilde{\chi}$ satisfying (4.8) and (4.10) at p is obtained by substituting $\widetilde{p}$ for p in the influence function of the original functional.

This is relevant when obtaining higher order influence functions. As these are recursive derivatives of the first order influence function (see [1]–[3] in Section 4.1), the preceding display shows that we must compute influence functions of

$$p \mapsto \chi^{(1)}_{\widetilde{p}}(x),$$

i.e. we "differentiate on the model $\widetilde{\mathcal{P}}$". If the latter model is sufficiently simple, for instance finite-dimensional, then exact higher order influence functions of the functional $p \mapsto \widetilde{\chi}(p)$ ought to exist. We can use these as approximate influence functions of $p \mapsto \chi(p)$.

EXAMPLE 4.6 (Missing data, approximate functional). In the missing data problem the density p corresponds one-to-one to a triplet of parameters (a, b, g) and hence the projection $p \mapsto \widetilde{p}$ can be described as projections of the parameters. We leave g invariant, and map a and b onto a finite-dimensional affine space, as follows.

We fix a given finite-dimensional subspace L of $L_2(\nu)$ that has good approximation properties for our model classes, the Hölder spaces $C^\alpha(\mathcal{Z})$ and $C^\beta(\mathcal{Z})$, for instance constructed from a wavelet basis. For fixed functions $\hat{a}, \underline{a}, \hat{b}, \underline{b} \colon \mathcal{Z} \to \mathbb{R}^+$ we now let $\widetilde{a}$ and $\widetilde{b}$ be the functions such that $(\widetilde{a} - \hat{a})/\underline{a}$ and $(\widetilde{b} - \hat{b})/\underline{b}$ are the orthogonal projections of the functions $(a - \hat{a})/\underline{a}$ and $(b - \hat{b})/\underline{b}$ onto L in $L_2(\underline{a}\underline{b}g)$. Finally we define the map $p \mapsto \widetilde{p}$ by correspondence to $(a, b, g) \mapsto (\widetilde{a}, \widetilde{b}, g)$.

In Section 7 we shall see that the orthogonal projections follow (4.10), while the concrete form of (4.9) is valid in that

$$\Big|\int abg\,d\nu - \int \widetilde{a}\widetilde{b}g\,d\nu\Big|^2 \le \int \Big|\frac{a-\widetilde{a}}{\underline{a}}\Big|^2 \underline{a}bg\,d\nu \int \Big|\frac{b-\widetilde{b}}{\underline{b}}\Big|^2 \underline{a}\underline{b}g\,d\nu.$$

This approximation error can be made arbitrarily small by making the space L large enough. In that case the approximate functional $p \mapsto \int \widetilde{a}\widetilde{b}g\,d\nu$ is close to the parameter of interest, and we may focus instead on estimating this functional. The advantage is that by construction this depends only on finitely many unknowns, e.g. the coefficients of $(\widetilde{a}-\hat{a})/\underline{a}$ and $(\widetilde{b}-\hat{b})/\underline{b}$ in a basis of L. Higher order influence functions exist to any order.

The bias-variance trade-off of Section 4.4 arises as the approximation error must be traded off against the "variance of estimating the coefficients" as well as against the remainder of using an mth order estimator.

5. First order estimator. The first order estimator (1.1) is well studied for the missing data problem. The first order influence function is given in (4.7), where $\chi_p = \chi_p^{(1)}$. As it depends on the parameter (a, b, f) only through a and b, preliminary estimators $\hat{a}$ and $\hat{b}$ suffice.

The "first order bias" of this estimator, the first term in (1.2), can explicitly be computed as

$$\begin{aligned}\chi(\hat{p}) - \chi(p) + P\chi_{\hat{p}}^{(1)} &= \mathrm{E}_p\big[(A\hat{a}(Z)-1)(Y-\hat{b}(Z)) + \hat{b}(Z)\big] - \int bf\,d\nu \\ &= -\int (\hat{a}-a)(\hat{b}-b)\,g\,d\nu. \qquad (5.1)\end{aligned}$$

In agreement with the heuristics given in Sections 1 and 4 this bias is quadratic in the errors of the initial estimator.

Actually, the form of the bias term is special in that square estimation errors $(\hat{a}-a)^2$ and $(\hat{b}-b)^2$ of the two initial estimators $\hat{a}$ and $\hat{b}$ do not arise, but only the product $(\hat{a}-a)(\hat{b}-b)$ of their errors. This property, termed "double robustness" in [30], makes that for first order inference it suffices that one of the two parameters be estimated well. A prior assumption that the parameters a and b are α and β regular, respectively, would allow estimation errors of the orders $n^{-\alpha/(2\alpha+d)}$ and $n^{-\beta/(2\beta+d)}$. If the product of these rates is $O(n^{-1/2})$, then the bias term matches the variance. This leads to the (unnecessarily restrictive) condition (3.1).

If the preliminary estimators $\hat{a}$ and $\hat{b}$ are solely selected for having small errors $\|\hat{a}-a\|$ and $\|\hat{b}-b\|$ (e.g. minimax in the L_2-norm), then it is hard to see why (5.1) would be small unless the product $\|\hat{a}-a\|\|\hat{b}-b\|$ of the

errors is small. Special estimators might exploit that the bias is an integral, in which cancellation of errors could occur. As we do not wish to use special estimators, our approach will be to replace the linear estimating equation by a higher order one, leading to an analogue of (5.1) that is a cubic or higher order polynomial of the estimation errors.

As noted the marginal density f (or g) does not enter into the first order influence function (4.7). Even though the functional depends on f (or g), a rate on the initial estimator of this function is not needed for the construction of the first order estimator. This will be different at higher orders.

6. Second order estimator. In this section we derive a second order influence function for the missing data problem, and analyze the risk of the corresponding estimator. This estimator is minimax if $(\alpha+\beta)/2 \geq d/4$ and

$$\frac{\gamma}{2\gamma+d} \geq \frac{1}{2} \wedge \frac{2\alpha+2\beta}{d+2\alpha+2\beta} - \frac{\alpha}{2\alpha+d} - \frac{\beta}{2\beta+d}. \tag{6.1}$$

In the other case, higher order estimators have smaller risk, as shown in Sections 8-9. However, it is worth while to treat the second order estimator separately, as its construction exemplifies essential elements, without involving technicalities attached to the higher order estimators.

To find a second order influence function, we follow the strategy [1]–[3] of Section 4.1, and try and find a function $\chi_p^{(2)}\colon \mathcal{X}^2 \to \mathbb{R}$ such that, for every $x_1 = (y_1 a_1, a_1, z_1)$, and all directions $\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}}$,

$$\frac{d}{dt}_{|t=0}\Big[\chi_{p_t}^{(1)}(x_1) + \chi(p_t)\Big] = \mathrm{E}_p \chi_p^{(2)}(x_1, X_2)\, B_p(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})(X_2).$$

Here the expectation E_p on the right side is relative to the variable X_2 only, with x_1 fixed. This equation expresses that $x_2 \mapsto \chi_p^{(2)}(x_1, x_2)$ is a first order influence function of $p \mapsto \chi_p^{(1)}(x_1) + \chi(p)$, for fixed x_1. On the left side we added the "constant" $\chi(p_t)$ to the first order influence function (giving another first order influence function) to facilitate the computations. This is justified as the strategy [1]–[3] works with any influence function. In view of (4.7) and the definitions of the paths $t \mapsto a + t\underline{\mathfrak{a}}$, $t \mapsto b + t\underline{\mathfrak{b}}$ and $t \mapsto f(1+t\underline{\mathfrak{f}})$, this leads to the equation

$$\begin{aligned} a_1\big(y_1 - b(z_1)\big)\underline{\mathfrak{a}}(z_1) - \big(a_1 a(z_1) - 1\big)\underline{\mathfrak{b}}(z_1) & \\ = \mathrm{E}_p \chi_p^{(2)}(x_1, X_2)\, B_p(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})(X_2). & \end{aligned} \tag{6.2}$$

Unfortunately, no function $\chi_p^{(2)}$ that solves this equation for every $(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})$ exists. To see this note that for the special triplets with $\underline{\mathfrak{b}} = \underline{\mathfrak{f}} = 0$ the

requirement can be written in the form

$$\underline{\mathfrak{a}}(z_1) = \mathrm{E}_p\left[\frac{\chi_p^{(2)}(x_1, X_2)}{a_1\big(y_1 - b(z_1)\big)}\frac{1 - A_2 a(Z_2)}{a(Z_2)(a-1)(Z_2)}\right]\underline{\mathfrak{a}}(Z_2).$$

The right side of the equation can be written as $\int K(z_1, z_2)\underline{\mathfrak{a}}(z_2)\,dF(z_2)$, for $K(z_1, Z_2)$ the conditional expectation of the function in square brackets given Z_2. Thus it is the image of $\underline{\mathfrak{a}}$ under the kernel operator with kernel K. If the equation were true for any $\underline{\mathfrak{a}}$, then this kernel operator would work as the identity operator. However, on infinite-dimensional domains the identity operator is not given by a kernel. (Its kernel would be a "Dirac function on the diagonal".)

Therefore, we have to be satisfied with an influence function that gives a partial representation only. In particular, a projection onto a finite-dimensional linear space possesses a kernel, and acts as the identity on this linear space. A "large" linear space gives representation in "many" directions. By reducing the expectation in (6.2) to an integral relative to the marginal distribution of Z_2, we can use an orthogonal projection $\Pi_p\colon L_2(g) \to L_2(g)$ onto a subspace L of $L_2(g)$. Writing also Π_p for its kernel, and letting $S_2 h$ denote the symmetrization $\big(h(X_1, X_2) + h(X_2, X_1)\big)/2$ of a function $h\colon \mathcal{X}^2 \to \mathbb{R}$, we define

$$\chi_p^{(2)}(X_1, X_2) = -2S_2\Big[A_1\big(Y_1 - b(Z_1)\big)\Pi_p(Z_1, Z_2)\big(A_2 a(Z_2) - 1\big)\Big]. \tag{6.3}$$

LEMMA 6.1. *For $\chi_p^{(2)}$ defined by (6.3) with Π_p the kernel of an orthogonal projection $\Pi_p\colon L_2(g) \to L_2(g)$ onto a subspace $L \subset L_2(g)$, equation (6.2) is satisfied for every path $t \mapsto p_t$ corresponding to directions $(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})$ such that $\underline{\mathfrak{a}} \in L$ and $\underline{\mathfrak{b}} \in L$.*

PROOF. By definition $\mathrm{E}(A|\,Z) = (1/a)(Z)$ and $\mathrm{E}(Y|\,Z) = b(Z)$. Also $\mathrm{var}\big(Aa(Z)|\,Z\big) = a(Z) - 1$ and $\mathrm{var}(Y|\,Z) = b(Z)(1-b)(Z)$. By direct computation using these identities, we find that for the influence function (6.3) the right side of (6.2) reduces to

$$a_1\big(y_1 - b(z_1)\big)\Pi_p\underline{\mathfrak{a}}(z_1) - \big(a_1 a(z_1) - 1\big)\Pi_p\beta(z_1).$$

Thus (6.2) holds for every $(\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}})$ such that $\Pi_p\underline{\mathfrak{a}} = \underline{\mathfrak{a}}$ and $\Pi_p\underline{\mathfrak{b}} = \underline{\mathfrak{b}}$. ∎

Together with the first order influence function (4.7) the influence function (6.3) defines the (approximate) influence function $\chi_p = \chi_p^{(1)} + \frac{1}{2}\chi_p^{(2)}$. For an

initial estimator $\hat{p}$ based on independent observations we now construct the estimator (4.1), i.e.

$$\hat{\chi}_n = \chi(\hat{p}) + \mathbb{P}_n \chi^{(1)}_{\hat{p}} + \tfrac{1}{2}\mathbb{U}_n \chi^{(2)}_{\hat{p}}. \tag{6.4}$$

Unlike the first order influence function, the second order influence function does depend on the density f of the covariates, or rather the function $g = f/a$ (through the kernel Π_p, which is defined relative to $L_2(g)$), and hence the estimator (6.4) involves a preliminary estimator of g. As a consequence, the quality of the estimator of the functional χ depends on the precision by which g (as part of the plug-in $\hat{p} = (\hat{a}, \hat{b}, \hat{g})$) can be estimated. The intuitive reason is that the bias (5.1) depends on g, and it can only be made smaller by estimating it.

Let $\hat{\mathrm{E}}_p$ and $\hat{\mathrm{var}}_p$ denote conditional expectations given the observations used to construct $\hat{p}$, let $\|\cdot\|_r$ be the norm of $L_r(g)$, and let $\|\Pi\|_r$ denote the norm of an operator $\Pi\colon L_r(g) \to L_r(g)$.

THEOREM 6.1. *The estimator $\hat{\chi}_n$ given in (6.4) with influence functions $\chi^{(1)}_p$ and $\chi^{(2)}_p$ defined by (4.7) and (6.3), for Π_p the kernel of an orthogonal projection in $L_2(g)$ onto a k-dimensional linear subspace, satisfies, for $r \geq 2$ (with $r/(r-2) = \infty$ if $r = 2$),*

$$\begin{aligned}\hat{\mathrm{E}}_p\hat{\chi}_n - \chi(p) &= O_P\Big(\|\Pi_p\|_r\|\Pi_{\hat{p}}\|_r\|\hat{a}-a\|_r\|\hat{b}-b\|_r\|\hat{g}-g\|_{r/(r-2)}\Big) \\ &\qquad + O_P\Big(\big\|(I-\Pi_p)(a-\hat{a})\big\|_2\big\|(I-\Pi_p)(b-\hat{b})\big\|_2\Big), \\ \hat{\mathrm{var}}_p\hat{\chi}_n &= O_P\Big(\frac{1}{n} + \frac{k}{n^2}\Big).\end{aligned}$$

The two terms in the bias result from having to estimate p in the second order influence function (giving "third order bias") and using an approximate influence function (leaving the remainders $I - \Pi_p$ after projection), respectively. The terms $1/n$ and k/n^2 in the variance appear as the variances of $\mathbb{U}_n\chi^{(1)}_p$ and $\mathbb{U}_n\chi^{(2)}_p$, the second being a degenerate second order U-statistic (giving $1/n^2$, see (S14.1)) with a kernel of variance k.

The proof of the theorem is deferred to Section 10.1.

Assume now that the range space of the projections Π_p can be chosen such that, for some constant C,

$$\|a - \Pi_p a\|_2 \leq C\Big(\frac{1}{k}\Big)^{\alpha/d}, \qquad \|b - \Pi_p b\|_2 \leq C\Big(\frac{1}{k}\Big)^{\beta/d}. \tag{6.5}$$

Furthermore, assume that there exist estimators $\hat{a}$ and $\hat{b}$ and $\hat{g}$ that achieve convergence rates $n^{-\alpha/(2\alpha+d)}$, $n^{-\beta/(2\beta+d)}$ and $n^{-\gamma/(2\gamma+d)}$, respectively, in

$L_r(g)$ and $L_{r/(r-2)}(g)$, uniformly over these a-priori models and a model for g (e.g. for $r = 3$), and that the preceding displays also hold for $\hat{a}$ and $\hat{b}$. These assumptions are satisfied if the unknown functions a and b are "regular" of orders α and β on a compact subset of $\mathbb{R}^d$ (see e.g. [28]). Then the estimator $\hat{\chi}_n$ of Theorem 6.1 attains the square rate of convergence

$$\left(\frac{1}{n}\right)^{2\alpha/(2\alpha+d)+2\beta/(2\beta+d)+2\gamma/(2\gamma+d)} \vee \left(\frac{1}{k}\right)^{(2\alpha+2\beta)/d} \vee \frac{1}{n} \vee \frac{k}{n^2}. \tag{6.6}$$

We shall see in the next section that the first of the four terms in this maximum can be made smaller by choosing an influence function of order higher than 2, while the other three terms arise at any order. This motivates to determine a "second order 'optimal" value of k by balancing the second, third and fourth terms. We next would use the second order estimator if γ is large enough so that the first term is negligible relative to the other terms.

For $(\alpha+\beta)/2 \geq d/4$ we can choose $k = n$ and the resulting rate (the square root of (6.6)) is $n^{-1/2}$ provided that (6.1) holds. The latter condition is certainly satisfied under the sufficient condition (3.1) for the linear estimator to yield rate $n^{-1/2}$.

More interestingly, for $(\alpha + \beta)/2 < d/4$ we choose $k \sim n^{2d/(d+2\alpha+2\beta)}$ and obtain the rate, provided that (6.1) holds,

$$n^{-(2\alpha+2\beta)/(d+2\alpha+2\beta)}.$$

This rate is slower than $n^{-1/2}$, but better than the rate $n^{-\alpha/(2\alpha+d)-\beta/(2\beta+d)}$ obtained by the linear estimator. In [25] this rate is shown to be the fastest possible in the minimax sense, for the model in which a and b range over balls in $C^\alpha(\mathcal{Z})$ and $C^\beta(\mathcal{Z})$, and g being known.

In both cases the second order estimator is better than the linear estimator, but minimax only for sufficiently large γ. This motivates to consider higher order estimators.

7. Approximate functional. Even though the functional of interest does not possess an exact second-order influence function, we might proceed to higher orders by differentiating the approximate second-order influence function $\chi_p^{(2)}$ given in (6.3), and balancing the various terms obtained. However, the formulas are much more transparent if we compute *exact* higer-order influence functions of an approximating functional instead. In this section we first define a suitable functional and next compute its influence functions.

Following the heuristics of Section 4.5, we define an approximate functional by equation (4.8), using a particular projection $p \mapsto \widetilde{p}$ of the param-

eters. We choose this projection to map the parameters a and b onto finite-dimensional models and leave the parameter g unaltered: p is mapped into an element $\widetilde{p}$ of the approximating model, or equivalently a triplet (a, b, g) into a triplet $(\widetilde{a}, \widetilde{b}, g)$ in the approximating model for the three parameters (where g is unaltered). (Even though this is not evident in the notation, the projection is joint in the three parameters: the induced maps $(a, b, g) \mapsto \widetilde{a}$ and $(a, b, g) \mapsto \widetilde{b}$ do not reduce to maps $a \mapsto \widetilde{a}$ and $b \mapsto \widetilde{b}$, but $\widetilde{a}$ and $\widetilde{b}$ depend on the full triplet (a, b, g).)

As "model" for (a, b) we consider the product of two affine linear spaces

$$\big(\hat{a} + \underline{a}L\big) \times \big(\hat{b} + \underline{b}L\big), \tag{7.1}$$

for a given finite-dimensional subspace L of $L_2(\nu)$ and fixed functions $\hat{a}, \underline{a}, \hat{b}, \underline{b} \colon \mathcal{Z} \to \mathbb{R}$ that are bounded away from zero and infinity. (Later the functions $\hat{a}$ and $\hat{b}$ are taken equal to the preliminary estimators; one choice for the other functions is $\underline{a} = \underline{b} = 1$.) The pair $(\widetilde{a}, \widetilde{b})$ of projections are defined as elements of the model (7.1) satisfying equation (4.10). In view of (5.1), for any path $\widetilde{p}_t \leftrightarrow (\widetilde{a}_t, \widetilde{b}_t, g) = (\widetilde{a} + t\underline{a}\, l, \widetilde{b} + t\underline{b}\, l', g)$, for given $l, l' \in L$,

$$\chi(\widetilde{p}_t) + P\chi^{(1)}_{\widetilde{p}_t} = \chi(p) - \int \big(\widetilde{a} + t\underline{a}\, l - a\big)\big(\widetilde{b} + t\underline{b}\, l' - b\big)\, g\, d\nu. \tag{7.2}$$

Equation (4.10) requires that the derivative of this expression with respect to t at $t = 0$ vanishes. Thus the functions $\widetilde{a}$ and $\widetilde{b}$ must be chosen to satisfy the set of stationary equations, for every $l, l' \in L$,

$$0 = \int (\widetilde{a} - a)\underline{b}\, l'\, g\, d\nu = \int \Big(\frac{\widetilde{a} - \hat{a}}{\underline{a}} - \frac{a - \hat{a}}{\underline{a}}\Big)\, l'\, \underline{a}\underline{b} g\, d\nu, \qquad l' \in L, \tag{7.3}$$

$$0 = \int \underline{a}\, l(\widetilde{b} - b)\, g\, d\nu = \int \Big(\frac{\widetilde{b} - \hat{b}}{\underline{b}} - \frac{b - \hat{b}}{\underline{b}}\Big)\, l\, \underline{a}\underline{b} g\, d\nu, \qquad l \in L. \tag{7.4}$$

Because the functions $(\widetilde{a} - \hat{a})/\underline{a}$ and $(\widetilde{b} - \hat{b})/\underline{b}$ are required to be in L, the second way of writing these equations shows that the latter two functions are the orthogonal projections of the functions $(a - \hat{a})/\underline{a}$ and $(b - \hat{b})/\underline{b}$ onto L in $L_2(\underline{a}\underline{b}g)$.

As explained in Section 4.5, as it satisfies (4.10) the projection $(a, b, g) \mapsto (\widetilde{a}, \widetilde{b}, g)$ renders the first order influence function of the approximate functional $\widetilde{\chi}$ equal to the first order influence function of χ evaluated at the projection. Furthermore, the difference between χ and $\widetilde{\chi}$ is quadratic in the distance between $\widetilde{p}$ and p (see (4.9)). The following theorem summarizes the preceding and verifies these properties in the present concrete situation.

THEOREM 7.1. *For given measurable functions $\hat{a}, \underline{a}, \hat{b}, \underline{b}\colon \mathcal{Z} \to \mathbb{R}$ with $\underline{a}$ and $\underline{b}$ bounded away from zero and infinity, define a map $(a, b, g) \mapsto (\widetilde{a}, \widetilde{b}, g)$ by letting $(\widetilde{a} - \hat{a})/\underline{a}$ and $(\widetilde{b} - \hat{b})/\underline{b}$ be the orthogonal projections of $(a - \hat{a})/\underline{a}$ and $(b - \hat{b})/\underline{b}$ in $L_2(\underline{a}\underline{b}g)$ onto a closed subspace L. Let $\widetilde{p}$ correspond to $(\widetilde{a}, \widetilde{b}, g)$ and define $\widetilde{\chi}(p) = \chi(\widetilde{p}) + P\chi^{(1)}_{\widetilde{p}}$. Then $\widetilde{\chi}$ has influence function*

$$\widetilde{\chi}^{(1)}_p(X) = A\widetilde{a}(Z)\big(Y - \widetilde{b}(Z)\big) + \widetilde{b}(Z) - \chi\big(\widetilde{p}\big). \tag{7.5}$$

Furthermore, for $\underline{g} = \underline{a}\underline{b}g$,

$$\big|\widetilde{\chi}(p) - \chi(p)\big| \le \Big\|(I - \Pi_p)\frac{\hat{a} - a}{\underline{a}}\Big\|_{2,\underline{g}}\Big\|(I - \Pi_p)\frac{\hat{b} - b}{\underline{b}}\Big\|_{2,\underline{g}}.$$

PROOF. The formula for the influence function agrees with the combination of equations (4.11) and (4.7), and can also be verified directly. In view of (4.8) and (5.1),

$$\widetilde{\chi}(p) - \chi(p) = -\int \big(\widetilde{a} - a\big)\big(\widetilde{b} - b\big)\, g\, d\nu.$$

We rewrite the right side as an integral relative to $\underline{g}\, d\nu$, and next apply the Cauchy-Schwarz inequality. Finally we note that $(\widetilde{a} - a)/\underline{a} = (\widetilde{a} - \hat{a})/\underline{a} - (a - \hat{a})/\underline{a} = (I - \Pi_p)\big((\hat{a} - a)/\underline{a}\big)$, and similarly for b. ∎

The approximation error $\widetilde{\chi}(p) - \chi(p)$ can be rendered arbitrarily small by choosing the space L large enough. Of course, we choose L to be appropriate relative to a-priori assumptions on the functions a and b. If these functions are known to belong to Hölder classes, then L can for instance be chosen as the linear span of the first k basis elements of a suitable orthonormal wavelet basis of $L_2(\nu)$.

To compute higher order influence functions of $\widetilde{\chi}$ we recursively determine influence functions of influence functions, according to the algorithm [1]–[3] in Section 4.3, starting with the influence function of $p \mapsto \widetilde{\chi}^{(1)}_p(x_1) + \chi(\widetilde{p})$, for a fixed x_1. We defer the details of this derivation to Section S10.5, and summarize the result in the following theorem.

To simplify notation, define

$$\begin{aligned} \widetilde{Y} &= A\big(Y - \widetilde{b}(Z)\big)\underline{a}(Z), \\ \widetilde{A} &= \big(A\widetilde{a}(Z) - 1\big)\underline{b}(Z), \\ \underline{A} &= A\underline{a}(Z)\underline{b}(Z). \end{aligned} \tag{7.6}$$

These are the generic variables; indexed versions $\widetilde{Y}_i, \widetilde{A}_i, \underline{A}_i, \ldots$ are defined by adding an index to every variable in the equalities. With this notation and with $\underline{a} = \underline{b} = 1$ the second order influence function (6.3) at $p = \widetilde{p}$ can be written as the symmetrization of $-2\widetilde{Y}_1\Pi_p(Z_1, Z_2)\widetilde{A}_2$. This function was derived in an ad-hoc manner as an approximate or partial influence function of χ, but it is also the exact influence function of $\widetilde{\chi}$. The higher order influence functions of $\widetilde{\chi}$ possess an equally attractive form.

THEOREM 7.2. *An mth order influence function $\widetilde{\chi}_p^{(m)}$ evaluated at $(X_1, \ldots, X_m)$ of the functional $\widetilde{\chi}$ defined in Theorem 7.1 is the degenerate (in $L_2(p)$) part of the variable*

$$(-1)^{j-1} j!\, \widetilde{A}_1 \Pi_{1,2} \underline{A}_2 \Pi_{2,3} \underline{A}_3 \Pi_{3,4} \underline{A}_4 \times \cdots \times \underline{A}_{m-1} \Pi_{m-1,m} \widetilde{Y}_m.$$

Here $\Pi_{i,j}$ is the kernel of the orthogonal projection in $L_2(\underline{abg})$ onto L, evaluated at (Z_i, Z_j).

To obtain the degenerate part of the variable in the preceding lemma, we apply the general formula (2.1) together with Lemma S10.2. Assertions (i) and (ii) of the latter lemma show that the variable is already degenerate relative to X_1 and X_m, while assertion (iii) shows that integrating out the variable X_i for $1 < i < m$ simply collapses $\Pi_{i-1,i}\underline{A}_i\Pi_{i,i+1}$ into $\Pi_{i-1,i+1}$. For instance, with S_m denoting symmetrization of a function of m variables,

$$\begin{aligned}
\widetilde{\chi}_p^{(2)}(X_1, X_2) &= -2S_2[\widetilde{A}_1 \Pi_{1,2} \widetilde{Y}_2], \\
\widetilde{\chi}_p^{(3)}(X_1, X_2, X_3) &= 6S_3\Big[\widetilde{A}_1 \Pi_{1,2} \underline{A}_2 \Pi_{2,3} \widetilde{Y}_3 - \widetilde{A}_1 \Pi_{1,3} \widetilde{Y}_3\Big], \\
\widetilde{\chi}_p^{(4)}(X_1, X_2, X_3, X_4) &= -24S_4\Big[\widetilde{A}_1 \Pi_{1,2} \underline{A}_2 \Pi_{2,3} \underline{A}_3 \Pi_{3,4} \widetilde{Y}_4 \\
&\qquad - \widetilde{A}_1 \Pi_{1,3} \underline{A}_3 \Pi_{3,4} \widetilde{Y}_4 - \widetilde{A}_1 \Pi_{1,2} \underline{A}_2 \Pi_{2,4} \widetilde{Y}_4 + \widetilde{A}_1 \Pi_{1,4} \widetilde{Y}_4\Big].
\end{aligned} \tag{7.7}$$

As shown on the left, but not on the right of the equations, these quantities depend on the unknown parameter $p = (a, b, g)$. In the right sides, the variables $\widetilde{Y}_i$ and $\widetilde{A}_i$ depend on p through $\widetilde{b}$ and $\widetilde{a}$, and hence are not observable variables. Furthermore, the kernels $\Pi_{i,j}$ depend on g as they are orthogonal projections in $L_2(\underline{abg})$.

8. Parametric rate ($(\alpha + \beta)/2 \geq d/4$). In this section we show that the parameter $\chi(p)$ is estimable at $1/\sqrt{n}$-rate provided the average smoothness $(\alpha + \beta)/2$ is at least $d/4$. We achieve this using the estimator

$$\hat{\chi}_n = \chi(\hat{p}) + \mathbb{U}_n\Big(\widetilde{\chi}_{\hat{p}}^{(1)} + \tfrac{1}{2}\widetilde{\chi}_{\hat{p}}^{(2)} + \cdots + \frac{1}{m!}\widetilde{\chi}_{\hat{p}}^{(m)}\Big), \tag{8.1}$$

with the influence functions $\widetilde{\chi}_p^{(j)}$ those of the approximate functional $\widetilde{\chi}$ in Section 7: they are given in Theorems 7.1 and 7.2 for $j = 1$, and $j = 2, \ldots, m$, respectively. (Because the map $p \mapsto \widetilde{p}$ maps $\hat{p}$ into itself, the influence function for $j = 1$ in the display is also the first order influence function (7.5) of of χ, when evaluated at $p = \hat{p}$.)

We assume that the projections Π_p and $\Pi_{\hat{p}}$ map $L_s\big(\underline{a}bg\big)$ to $L_s\big(\underline{a}bg\big)$, for every $s \in \big[r/(r-1), r\big]$, with uniformly bounded norms. (For $r = 2$ this entails only $s = 2$; in this case we define $r/(r-2) = \infty$.)

THEOREM 8.1. *The estimator (8.1), with Π_p a kernel of an orthogonal projection in $L_2\big(\underline{a}bg\big)$ satisfying (S13.1) with $\sup_x \Pi_p(x,x) \lesssim k$, satisfies, for a constant c that depends on $\|p/\hat{p}\|_\infty$ only, and $r \geq 2$,*

$$\hat{\mathrm{E}}_p\hat{\chi}_n - \chi(p) = O\Big(\|\hat{a}-a\|_r\|\hat{b}-b\|_r\|\hat{g}-g\|_{(m-1)r/(r-2)}^{m-1}\Big) + O\Big(\Big\|(I-\Pi_p)\frac{\hat{a}-a}{\underline{a}}\Big\|_2\Big\|(I-\Pi_p)\frac{\hat{b}-b}{\underline{b}}\Big\|_2\Big),$$

$$\hat{\mathrm{var}}_p\hat{\chi}_n \leq \sum_{j=1}^m \frac{1}{\binom{n}{j}} c^j k^{j-1}.$$

The first term in the bias is of the order $1 + 1 + (m-1) = m+1$, as to be expected for an estimator based on an mth order influence function; the second term is due to estimating $\widetilde{\chi}$ rather than χ; it is independent of m, and the same as in Theorem 6.1 if $\underline{a} = \underline{b} = 1$. The bound on the variance can roughly be understood in that each of the degenerate U-statistics $\mathbb{U}_n\widetilde{\chi}_{\hat{p}}^{(j)}$ in (8.1) contributes a term of order k^{j-1}/n^j.

For α-, β- and γ-regular parameters a, b, g on a d-dimensional domain the range space of the projections Π_p can be chosen so that (6.5) holds and such that there exist estimators $\hat{a}, \hat{b}, \hat{g}$ of a, b, g, with the first two taking values in this range space, with convergence rates $n^{-\alpha/(2\alpha+d)}$, $n^{-\beta/(2\beta+d)}$ and $n^{-\gamma/(2\gamma+d)}$. Then the second term in the bias (with $\underline{a} = \underline{b} = 1$) is of order $(1/k)^{\alpha/d+\beta/d}$. If $(\alpha+\beta)/2 \geq d/4$ and we choose $k = n$, then this is of order $1/\sqrt{n}$. For $k = n$ the standard deviation of the resulting estimator is also of the order $1/\sqrt{n}$, while the first term in the bias can be made arbitrarily small by choosing a sufficiently large order m. Specifically, the estimator $\hat{\chi}_n$ attains a $\sqrt{n}$-rate of convergence as soon as

$$m - 1 \geq \Big(\frac{1}{2} - \frac{\alpha}{2\alpha+d} - \frac{\beta}{2\beta+d}\Big)\Big(\frac{2\gamma+d}{\gamma}\Big). \tag{8.2}$$

For any $\gamma > 0$ there exists an order m that satisfies this, and hence the parameter is $\sqrt{n}$-estimable as soon as $(\alpha+\beta)/2 \geq d/4$.

More ambitiously, we may aim at attaining the parametric rate for every $\gamma > 0$, without a-priori knowledge of γ. This can be achieved if $(\alpha+\beta)/2 > d/4$ by using orders $m = m_n$ that increase to infinity with the sample size. In this case the estimator can also be shown to be asymptotically efficient in the semiparametric sense.

THEOREM 8.2. *If $(\alpha+\beta)/2 > d/4$, then the estimator (8.1), with $m = \log n$ and Π_p a kernel of an orthogonal projection in $L_2(\underline{abg})$ on a $k = n/(\log n)^2$-dimensional space satisfying (6.5) and (S13.1) with $\sup_x \Pi_p(x,x) \lesssim k$, based on preliminary estimators $\hat{a}, \hat{b}, \hat{g}$ that attain rates $(\log n/n)^{-\delta/(2\delta+d)}$ relative to the uniform norm, satisfies*

$$\sqrt{n}\big(\hat{\chi}_n - \chi(p) - \mathbb{P}_n\widetilde{\chi}_p^{(1)}\big) \xrightarrow{P} 0.$$

An estimator that is asymptotically linear in the first order efficient influence function, as in the theorem, is asymptotically optimal in terms of the local asymptotic minimax and convolution theorems (see e.g. [36], Chapter 25). The present estimator $\hat{\chi}_n$ actually looses its efficiency by splitting the sample in a part used to construct the preliminary estimators and a part to form $\mathbb{P}_n$. This can be easily remedied by crossing over the two parts of the split, and taking the average of the two estimators so obtained. By the theorem these are both asymptotically linear in their sample, and hence their average is asymptotically linear in the full sample and asymptotically efficient.

The proofs of the theorems are deferred to Section 10.2.

9. Minimax rate at lower smoothness ($(\alpha+\beta)/2 < d/4$). If the average a-priori smoothness $(\alpha+\beta)/2$ of the functions a and b falls below $d/4$, then the functional χ cannot be estimated any more at the parametric rate ([25]). The estimator (8.1) of Theorem 8.1 can still be used and, with its bias and variance as given in the theorem properly balanced, attains a certain rate of convergence, faster than the current state-of-the-art linear estimators. However, in this section we present an estimator that is always better, and attains the minimax rate of convergence $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ provided that the parameter g is sufficiently regular.

This estimator takes the same general form

$$\hat{\chi}_n = \chi(\hat{p}) + \mathbb{U}_n\Big(\widetilde{\chi}_{\hat{p}}^{(1)} + \tfrac{1}{2}\widetilde{\chi}_{\hat{p}}^{(2)} + \cdots + \frac{1}{m!}\chi_{\hat{p}}^{(m)}\Big), \tag{9.1}$$

as the estimator (8.1), but the influence functions $\chi_p^{(j)}$ for $j \geq 3$ will be different. The idea is to "cut out" certain terms from the influence functions

in (8.1) in order to decrease the variance, but without increasing the bias. For clarity we first consider the third order estimator, and next extend to the general mth order. To attain the minimax rate the order m must be fixed to a large enough value so that the first term in the bias given in Theorem 8.1 is no larger than $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$. (Apart from added complexity there is no loss in choosing m *larger* than needed.)

The third order kernel $\widetilde{\chi}_p^{(3)}$ in (7.7) is the symmetrization of the variable

$$6\widetilde{A}_1\Big(\Pi_p(Z_1,Z_2)\underline{A}_2\Pi_p(Z_2,Z_3)-\Pi_p(Z_1,Z_3)\Big)\widetilde{Y}_3.$$

Here Π_p is the kernel of an orthogonal projection in $L_2(\underline{abg})$ onto a k-dimensional linear space, which we may view as the sum of k projections on one-dimensional spaces. The quantity k^2 in the order $O(k^2/n^3)$ of the variance in Theorem 8.1 for $m=3$ arises as the number of terms in the product $\Pi_p(Z_1,Z_2)\underline{A}_2\Pi_p(Z_2,Z_3)$ of the two k-dimensional projection kernels. It turns out that this order can be reduced without increasing the bias by cutting out "products of projections on higher base elements".

To make this precise, we partition the projection space in blocks, and decompose the two projections in the influence function over the blocks:

$$\Pi_p=\sum_{r=0}^{R}\Pi_p^{(k_{r-1},k_r]},\qquad \Pi_p=\sum_{s=0}^{S}\Pi_p^{(l_{s-1},l_s]}. \tag{9.2}$$

Here $\Pi_p^{(m,n]}$ is the projection on the subspace spanned by base elements with index in intervals $(m,n]$, and $1=k_{-1}<k_0<k_1<\cdots<k_R=k$ and $1=l_{-1}<l_0<l_1<\cdots<l_S=k$ are suitable partitions of the set $\{1,\ldots,k\}$. ("Full" partitions in singleton sets would make the construction conceptual simpler, but a small number of blocks will be needed in our proofs.) The product of the two kernels now becomes a double sum, from which we retain only terms with small values of (r,s). The improved third order influence function is, with as before S_3 denoting symmetrization,

$$\begin{aligned}\chi_p^{(3)}(X_1,X_2,X_3)=6S_3\Bigg[\sum_{\substack{(r,s):r+s\le D\\ \vee r=0\vee s=0}}\sum\widetilde{A}_1\Big(&\Pi_p^{(k_{r-1},k_r]}(Z_1,Z_2)\underline{A}_2\Pi_p^{(l_{s-1},l_s]}(Z_2,Z_3)\\ &-\Pi_p^{(k_{r-1}\vee l_{s-1},k_r\wedge l_s]}(Z_1,Z_3)\Big)\widetilde{Y}_3\Bigg].\end{aligned} \tag{9.3}$$

The negative term in the display is the conditional expectation given Z_1,Z_3 of the leading term, and maintains the degeneracy of the kernel.

For the decomposition (9.2) to be valid, the subspaces corresponding to the blocks must be orthogonal in $L_2(\underline{abg})$. We may achieve this by starting

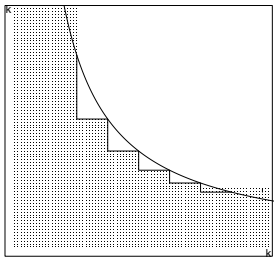

FIG 1. *Both axis carry the indices of the basis functions spanning the projection space L, and point in the plane refers to a product of two projections. Products of projections on pairs of basis functions in the shaded area are included in the third order influence function. The step function refers to the partitions of the indices as in (9.2).*

with a standard basis $e_1, \varepsilon_2, \ldots$, with good approximation properties for a target model, and next replacing this by an orthonormal basis in $L_2(\underline{abg})$ by the Gram-Schmidt procedure. For a bounded g the approximation properties will be preserved.

The grids are defined by

$$k_{-1} = 1, \qquad k_r \sim n2^{r/\alpha}, \qquad r = 0, \ldots, R, \tag{9.4}$$

$$l_{-1} = 1, \qquad l_s \sim n2^{s/\beta}, \qquad s = 0, \ldots, S, \tag{9.5}$$

where R and S are chosen such that $k_R \sim l_S \sim k$ (note that $k_0 = l_0 = n$). In these definitions the notation $\sim$ means "equal up to a fixed multiple" (needed to allow that k_r and l_s are (dyadic) integers). For ease of notation let $l_s = l_{-1}$ for $s \leq -1$, and $l_s = l_S$ for $s \geq S$.

The grids $k_0 < k_1 < \cdots < k_R$ and $l_0 < l_1 < \cdots < l_S$ partition the integers $n, n+1, \ldots, k$ in R and S groups. As $k_r^\alpha l_s^\beta = 2^{r+s} n^{\alpha+\beta}$, for every $r, s \geq 0$, the cut-off $r+s \leq D$ in (9.3) is delimited by the "hyperbola" $i^\alpha j^\beta \sim 2^D n^{\alpha+\beta}$ in the space of indices $(i, j) \in \{1, \ldots, k\}^2$ of base elements used in the two kernels, with only the pairs below the hyperbola retained (see Figure 1). The intuition behind this hyperbolic cut-off is the product form of the bias (5.1): a higher order correction on the estimator of a may combine with a lower order correction on b, and vice versa, to give an overall correction of the desired order. The overall bias is smaller if the cut-off D is chosen larger, but then more terms are included in the estimator and the variance will be bigger.

Before deriving an optimal value of D, we introduce the mth order estimator for general $m \geq 3$. Again we take the estimator of Theorem 8.1 as starting point, but modify the higher order influence functions $\widetilde{\chi}_p^{(j)}$, for

$j = 4, \ldots, m$, similar and in addition to the modification of the third order influence function. For given j the former influence function is given in Theorem 7.2 (with m of the theorem taken equal to j), and is based on a product of $j-1$ projection kernels. We modify this in two steps. For each of the $j-2$ contiguous pairs of kernels $((1st, 2nd), (2nd, 3rd), \ldots, ((j-2)th, (j-1)th))$ we form a new kernel by truncating the pair at the hyperbola as described previously for the third order kernel, and truncating all other kernels at n. Next the modified jth order kernel is the sum of the resulting $j-2$ kernels. More formally, the modified jth order kernel is equal to

$$\chi_p^{(j)}(X_1, \ldots, X_j) = \sum_{i=1}^{j-2} \chi_p^{(j,i)}(X_1, \ldots, X_j), \tag{9.6}$$

where $\chi_p^{(j,i)}(X_1, \ldots, X_j)$ is the symmetrized, degenerate (relative to $L_2(p)$) part of the variable, for $i = 1, \ldots, j-2$, written in the notation of Theorem 8.1,

$$j!(-1)^{j-1}\, \widetilde{Y}_1 \Pi_{1,2}^{(0,n]} \underline{A}_2 \times \cdots \times \underline{A}_{i-1} \Pi_{i-1,i}^{(0,n]} \underline{A}_i \times$$
$$\times \Big[\sum_{\substack{(r,s): r+s \le D \\ \vee r=0 \vee s=0}} \sum \Pi_{i,i+1}^{(k_{r-1},k_r]} \underline{A}_{i+1} \Pi_{i+1,i+2}^{(l_{s-1},l_s]} \Big] \underline{A}_{i+2} \Pi_{i+2,i+3}^{(0,n]} \times \cdots \times \underline{A}_{j-1} \Pi_{j-1,j}^{(0,n]} \widetilde{A}_j.$$

For $j = 3$ there is only one pair of kernels, and the construction reduces to the modification (9.3) as discussed previously.

We assume that the projections $\Pi_p^{(0,l]}$ and $\Pi_{\hat{p}}^{(0,l]}$ map $L_s(\underline{abg})$ to $L_s(\underline{abg})$, for every $s \in \big[r/(r-1), r\big]$, with uniformly bounded norms.

THEOREM 9.1. *The estimator (9.1) for $m \ge 3$ with the influence functions $\widetilde{\chi}_p^{(j)}$ and $\chi_p^{(j)}$ given in (7.5) and (7.7) for $j = 1, 2$, respectively, and in (9.6) for $j \ge 3$, and with $\Pi_p^{(0,l]}$ kernels of orthogonal projections in $L_2(\underline{abg})$ satisfying (S13.1) with $\sup_x \Pi_{\hat{p}}^{(0,l]}(x,x) \lesssim l$, satisfies, for $r \ge 2$*

(and $r/(r-2) = \infty$ if $r = 2$),

$$\hat{\mathrm{E}}_p\hat{\chi}_n - \chi(p) = O\Big(\|\hat{a}-a\|_r\|\hat{b}-b\|_r\|\hat{g}-g\|^{m-1}_{\frac{mr}{r-2}}\Big)$$
$$+ O\Big(\Big\|(I-\Pi_p^{(0,k]})\frac{\hat{a}-a}{\underline{a}}\Big\|_2\Big\|(I-\Pi_p^{(0,k]})\frac{\hat{b}-b}{\underline{b}}\Big\|_2\Big),$$
$$+ O\Big(\sum_{r=1}^{R}\Big\|(I-\Pi_{\hat{p}}^{(0,k_{r-1}]})\Big(\frac{\hat{a}-a}{\underline{a}}\Big)\Big\|_r\Big\|(I-\Pi_{\hat{p}}^{(0,l_{D-r}]})\Big(\frac{\hat{b}-b}{\underline{b}}\Big)\Big\|_r\|\hat{g}-g\|_{\frac{r}{r-2}}\Big)$$
$$+ O\Big(R\Big\|(I-\Pi_{\hat{p}}^{(0,n]})\frac{\hat{a}-a}{\underline{a}}\Big\|_r\Big\|(I-\Pi_{\hat{p}}^{(0,n]})\frac{\hat{b}-b}{\underline{b}}\Big\|_r\|\hat{g}-g\|^2_{\frac{mr}{r-2}}\Big),$$
$$\hat{\mathrm{var}}_p\hat{\chi}_n \lesssim \frac{1}{n} + \frac{k}{n^2} + \frac{D2^{(\frac{1}{\alpha}\vee\frac{1}{\beta})D}}{n}.$$

A proof of the theorem is presented in Sections 10.3 and S10.4.

The first two terms in the bias are the same as in Theorem 8.1; the third and fourth terms are the price paid for cutting out terms from the influence function. The benefit is a reduced variance. We shall show that the boundary parameter D can be chosen such that the third term in the variance (resulting from the third and higher order parts of the influence function) is not bigger than the second term, while the increase in bias is negligible.

Assume that the functions a and b and their estimates are known to belong to models that are well approximated by the base functions $e_1, e_2, \ldots$ in the sense that, for $p \in \{p, \hat{p}\}$, and every value l in one of the two grids (9.4)-(9.5),

$$\Big\|(I-\Pi_p^{(0,l]})\Big(\frac{\hat{a}-a}{\underline{a}}\Big)\Big\|_r \lesssim \Big(\frac{1}{l}\Big)^{\alpha/d}, \tag{9.7}$$

$$\Big\|(I-\Pi_p^{(0,l]})\Big(\frac{\hat{b}-b}{\underline{b}}\Big)\Big\|_r \lesssim \Big(\frac{1}{l}\Big)^{\beta/d}. \tag{9.8}$$

Then the second term in the bias is of the order $(1/k)^{\alpha/d+\beta/d}$, as in Theorem 8.1, which is smaller than the minimax rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ for

$$k \sim n^{2d/(2\alpha+2\beta+d)}. \tag{9.9}$$

With this choice of k, the upper bound on the variance is of the square minimax rate $n^{-(4\alpha+4\beta)/(2\alpha+2\beta+d)}$ if D is chosen to satisfy

$$2^{(\frac{1}{\alpha}\vee\frac{1}{\beta})D} \sim \frac{1}{\log n}n^{(d-2\alpha-2\beta)/(d+2\alpha+2\beta)}. \tag{9.10}$$

Furthermore, under (9.9) the numbers R, S of grid points are of the order $\log n$.

In the third term of the bias we apply assumptions (9.7)-(9.8) and the identity $k_{r-1}^{\alpha} l_{D-r}^{\beta} \sim n^{\alpha+\beta} 2^D$, which results from (9.4)-(9.5), to see that the third term of the bias is of order

$$\sum_{r=1}^{R} \Big(\frac{1}{k_{r-1}}\Big)^{\alpha/d} \Big(\frac{1}{l_{D-r}}\Big)^{\beta/d} \|\hat{g} - g\|_{r/(r-2)} \le R \Big(\frac{1}{n^{\alpha+\beta} 2^D}\Big)^{1/d} \|\hat{g} - g\|_{r/(r-2)}.$$

If the convergence rate of $\hat{g}$ is $n^{-\gamma/(2\gamma+d)}$, then, for the choice of D given in (9.10), this can (by a calculation) seen to be of smaller order than the minimax rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ if γ is large enough that

$$\frac{\gamma}{2\gamma + d} > \Big(\frac{\alpha \vee \beta}{d}\Big)\Big(\frac{d - 2\alpha - 2\beta}{d + 2\alpha + 2\beta}\Big). \tag{9.11}$$

The fourth term in the bias can by a similar analysis be seen to be of the order

$$R \Big(\frac{1}{n}\Big)^{\alpha/d} \Big(\frac{1}{n}\Big)^{\beta/d} \|\hat{g} - g\|^2_{(m-2)r/(r-2)}.$$

Again this is smaller than the minimax rate if γ satisfies assumption (9.11).

Finally, if the convergence rates of $\hat{a}$ and $\hat{b}$ are $n^{-\alpha/(2\alpha+d)}$ and $n^{-\beta/(2\beta+d)}$, then the first term in the upper bound of the bias is of the order

$$\Big(\frac{1}{n}\Big)^{\alpha/(2\alpha+d)+\beta/(2\beta+d)+(m-1)\gamma/(2\gamma+d)}.$$

We choose m large enough so that this is of smaller order than the preceding terms. In particular, we can choose it so that this is smaller than the minimax rate.

We summarize this in the following corollary, which is the most advanced result of the paper.

COROLLARY 9.1. *If (9.7)–(9.11) hold, and $\Pi_p^{(0,l]}$ are kernels of orthogonal projections in $L_2(\underline{abg})$ satisfying (S13.1) with $\sup_x \Pi_p^{(0,l]}(x,x) \le l$, then the mth order estimator with the kernels (9.6) for $j \ge 3$ and sufficiently large m and suitable initial estimators, attains the rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ for estimating $\chi(p)$.*

10. Proofs.

10.1. *Proof of Theorem 6.1.* Write $\hat{\Pi}$ and Π for $\Pi_{\hat{p}}$ and Π_p, respectively, for both the kernels and the corresponding projection operators, and drop p also in $\hat{\mathrm{E}}_p$ and $\hat{\mathrm{var}}_p$. From (5.1) and (6.3) we have

$$\begin{aligned}
&\hat{\mathrm{E}}\hat{\chi}_n - \chi(p) \\
&\quad = -\int (\hat{a}-a)(\hat{b}-b)\, g\, d\nu - \hat{\mathrm{E}} A_1\big(Y_1 - \hat{b}(Z_1)\big)\big(A_2\hat{a}(Z_2) - 1\big)\hat{\Pi}(Z_1, Z_2) \\
&\quad = -\int (\hat{a}-a)(\hat{b}-b)\, g\, d\nu + \iint \big[(\hat{a}-a)\times(\hat{b}-b)\big]\,\big(g\times g\big)\,\hat{\Pi}\, d\nu\times\nu.
\end{aligned}$$

The double integral on the far right with $\hat{\Pi}$ replaced by Π can be written as the single integral $\int(\hat{a}-a)\Pi(\hat{b}-b)\, g\, d\nu$, for $\Pi(\hat{b}-b)$ the image of $\hat{b}-b$ under the projection Π. Added to the first integral on the right this gives $-\int(\hat{a}-a)(I-\Pi)(\hat{b}-b)\, g\, d\nu$, which is bounded in absolute value by the second term in the upper bound for the bias.

Replacement of $\hat{\Pi}$ by Π in the double integral gives a difference

$$\begin{aligned}
&\iint \big[(\hat{a}-a)\times(\hat{b}-b)\big]\, g\times g\,(\hat{\Pi}-\Pi)\, d\nu\times\nu \\
&\qquad = \int (\hat{a}-a)\left(\hat{\Pi}\Big((\hat{b}-b)\frac{g}{\hat{g}}\Big) - \Pi(\hat{b}-b)\right) g\, d\nu \\
&\qquad \le \|\hat{a}-a\|_s \left\|\hat{\Pi}\Big((\hat{b}-b)\frac{g}{\hat{g}}\Big) - \Pi(\hat{b}-b)\right\|_{r,\hat{g}} \|g/\hat{g}\|_\infty^{1/r},
\end{aligned}$$

by Hölder's inequality, for a conjugate pair (r, s). Considering $\hat{\Pi}$ as the projection in $L_2(\hat{g})$ with weight 1, and Π as the weighted projection in $L_2(\hat{g})$ with weight function $\hat{w} = g/\hat{g}$, we can apply Lemma S13.7(i) (with $q = s/r$ and $rp = s/(s-2)$) to see that this is bounded in absolute value by

$$\|\hat{a}-a\|_s \|\hat{\Pi}\|_{s,\hat{g}} \|\Pi\|_{s,\hat{g}} \|\hat{b}-b\|_{s,\hat{g}} \|\hat{w}-1\|_{s/(s-2),\hat{g}} \|w\|_\infty^{1/r}.$$

Because $\hat{w}$ is assumed bounded away from 0 and infinity, this is of the same order as the first term in the upper bound on the bias (if r replaces s).

Because the function $\chi^{(1)}_{\hat{p}}$ is uniformly bounded, the (conditional) variance of $\mathbb{U}_n\chi^{(1)}_{\hat{p}}$ is of the order $O(1/n)$. Thus for the variance bound it suffices to consider the (conditional) variance of $\mathbb{U}_n\chi^{(2)}_{\hat{p}}$. In view of Lemma S14.1 and (S14.1) this is bounded above by a multiple of

$$\Big(1 + \Big\|\frac{p}{\hat{p}}\Big\|_\infty\Big)^2 \hat{P}^n(\mathbb{U}_n\chi^{(2)}_{\hat{p}})^2 = \Big(1 + \Big\|\frac{p}{\hat{p}}\Big\|_\infty\Big)^2 \binom{n}{2}^{-2} \hat{P}^2(\chi^{(2)}_{\hat{p}})^2.$$

The variables $A\big(Y-\hat{b}(Z)\big)$ and $\big(A\hat{a}(Z)-1\big)$ are uniformly bounded. Hence the last term on the right is bounded above by a multiple of $n^{-2}\int\hat{\Pi}^2\,(\hat{g}\times\hat{g})\,d\nu\times\nu$, which is equal to k/n^2, by Lemma S13.3.

10.2. *Proof of Theorems 8.1 and 8.2.* Let $\hat{A}$ and $\hat{Y}$ be $\widetilde{A}$ and $\widetilde{Y}$ as in (7.6) with a and b in their definitions replaced by $\hat{a}$ and $\hat{b}$. Because $\hat{a}$ and $\hat{b}$ are projected onto themselves under the map $(a,b,g)\mapsto(\widetilde{a},\widetilde{b},g)$ (see Theorem 7.1), we actually obtain the same variables by replacing $\widetilde{a}$ and $\widetilde{b}$ by $\hat{a}$ and $\hat{b}$, respectively: $\hat{A}=\big(A\hat{a}(Z)-1\big)\underline{b}(Z)$ and $\hat{Y}=A\big(Y-\hat{b}(Z)\big)\underline{a}(Z)$. Furthermore, let Π and $\hat{\Pi}$ denote the operators Π_p and $\Pi_{\hat{p}}$, respectively, and $\Pi_{i,j}$ and $\hat{\Pi}_{i,j}$ their kernels evaluated at (Z_i,Z_j).

By explicit calculations,

$$\chi(\hat{p})+\hat{\mathrm{E}}_p\widetilde{\chi}^{(1)}_{\hat{p}}-\chi(p)=-\int(\hat{a}-a)(\hat{b}-b)\,g\,d\nu=\hat{\mathrm{E}}\hat{A}_1\Pi_{1,2}\hat{Y}_2-\hat{R},\tag{10.1}$$

for $\hat{R}$ defined by

$$\hat{R}=\int\Big(\frac{\hat{a}-a}{\underline{a}}\Big)(I-\Pi)\Big(\frac{\hat{b}-b}{\underline{b}}\Big)\underline{a}\underline{b}g\,d\nu.$$

The variable $\hat{R}$ is bounded by the second term in the expression for $\hat{\mathrm{E}}_p\hat{\chi}_n-\chi(p)$ in the statement of the theorem. We next show by induction on m that

$$\begin{aligned}&\hat{R}+\chi(\hat{p})+\hat{\mathrm{E}}\widetilde{\chi}^{(1)}_{\hat{p}}+\cdots+\frac{1}{m!}\hat{\mathrm{E}}\widetilde{\chi}^{(m)}_{\hat{p}}-\chi(p)\\&=(-1)^{m-1}\hat{\mathrm{E}}\hat{A}_1(\hat{\Pi}-\Pi)_{1,2}\underline{A}_2(\hat{\Pi}-\Pi)_{2,3}\times\cdots\times\underline{A}_{m-1}(\hat{\Pi}-\Pi)_{m-1,m}\hat{Y}_m.\end{aligned}\tag{10.2}$$

The analysis of the bias can then be concluded by showing that the right side of (10.2) is of the order as the first term given in the theorem.

Equation (10.1) and the definition of $\widetilde{\chi}^{(2)}_p$ readily show that identity (10.2) is true for $m=2$. We proceed to general m by induction. Relative to its value for m the left side receives for $(m+1)$ the extra term $\hat{\mathrm{E}}\chi^{(m+1)}_{\hat{p}}/(m+1)!$, which is equal to $(-1)^m$ times $\hat{\mathrm{E}}\hat{A}_1\hat{\Pi}_{1,2}\underline{A}_2\hat{\Pi}_{2,3}\times\cdots\times\underline{A}_m\hat{\Pi}_{m,m+1}\hat{Y}_{m+1}$ minus a sum of terms resulting from projections of this leading term. This extra term without the factor $(-1)^m$ (but including the projections) can be written (cf. (7.7) and (2.1))

$$\sum_{i=0}^{m-1}\binom{m-1}{i}\hat{\mathrm{E}}\hat{A}_1\hat{\Pi}_{1,2}\underline{A}_2\hat{\Pi}_{2,3}\times\cdots\times\underline{A}_{m-i}\hat{\Pi}_{m-i,m-i+1}\hat{Y}_{m-i+1}(-1)^i.\tag{10.3}$$

To prove the induction hypothesis for $m+1$ it suffices to show that this is equal to

$$\begin{aligned}&\hat{\mathrm{E}}\hat{A}_1(\hat{\Pi}-\Pi)_{1,2}\underline{A}_2(\hat{\Pi}-\Pi)_{2,3}\times\cdots\times\underline{A}_{m-1}(\hat{\Pi}-\Pi)_{m-1,m}\hat{Y}_m\\ &\quad+\hat{\mathrm{E}}\hat{A}_1(\hat{\Pi}-\Pi)_{1,2}\underline{A}_2(\hat{\Pi}-\Pi)_{2,3}\times\cdots\times\underline{A}_m(\hat{\Pi}-\Pi)_{m,m+1}\hat{Y}_{m+1}.\end{aligned}\tag{10.4}$$

To achieve this we expand the two terms of the preceding display into sums of expressions of the form, with each $K^{(j)}_{j,j+1}$ equal to $\hat{\Pi}_{j,j+1}$ or $\Pi_{j,j+1}$ and l the number of j for which the first alternative is true,

$$B_l := (-1)^{m-1-l}\,\hat{\mathrm{E}}\hat{A}_1K^{(1)}_{1,2}\underline{A}_2K^{(2)}_{2,3}\times\cdots\times\underline{A}_{m-1}K^{(m-1)}_{m-1,m}\hat{Y}_m,\tag{10.5}$$

and of the same form with $m+1$ replacing m for the second term of (10.4). As the notation suggests the expression in (10.5) depends on l (and m, but this is fixed), but not on which K are equal to $\hat{\Pi}$ or Π. To see this we use that Π is a projection onto L in $L_2(\underline{abg})$, so that $\int \Pi_{1,2}\,\gamma(z_2)\,(\underline{abg})(z_2)\,d\nu(z_2)=\gamma(z_1)$ for every $\gamma\in L$; and $\hat{\Pi}$ is also a projection onto L, so that as a function of one argument $\hat{\Pi}_{1,2}$ is contained in L. This observation yields the identities, for K equal to $\hat{\Pi}$ or Π,

$$\hat{\mathrm{E}}_{Z_j}\Pi_{j-1,j}\underline{A}_jK_{j,j+1}=K_{j-1,j+1}=\hat{\mathrm{E}}_{Z_j}K_{j-1,j}\underline{A}_j\Pi_{j,j+1}.$$

This allows to reduce (10.5) to

$$\begin{aligned}B_l&=(-1)^{m-1-l}\,\hat{\mathrm{E}}\hat{A}_1\hat{\Pi}_{1,2}\underline{A}_2\hat{\Pi}_{2,3}\times\cdots\times\underline{A}_l\hat{\Pi}_{l,l+1}\hat{Y}_{l+1},\qquad l\geq 1,\\ B_0&=(-1)^{m-1}\hat{\mathrm{E}}\hat{A}_1\Pi_{1,2}\hat{Y}_2.\end{aligned}$$

Thus after expanding the two terms of (10.4) in the quantities B_l, and simplifying these quantities, we can write their sum (10.4)

$$(B_0-B_0)+\sum_{l=1}^{m-1}\left(\binom{m}{l}-\binom{m-1}{l}\right)B_l(-1)^{m-l}+B_m.$$

The difference of the binomial coefficients is $\binom{m-1}{l-1}$. The expression is equal to (10.3), as claimed. This completes the proof of (10.2).

Next we bound the right side of (10.2), by taking the expectation in turn with respect to $X_m, X_{m-1},\ldots,X_1$. For $M_{\hat{w}}$ multiplication by the function $\hat{w}=g/\hat{g}$,

$$\hat{\mathrm{E}}_{X_m}(\hat{\Pi}-\Pi)_{m-1,m}\hat{Y}_m=(\hat{\Pi}M_{\hat{w}}-\Pi)\Big(\frac{\hat{b}-b}{\underline{b}}\Big)(Z_{m-1}).$$

Next, for any function h and $i = m-1, m-2, \ldots, 2$,

$$\hat{\mathrm{E}}_{X_i}(\hat{\Pi} - \Pi)_{i-1,i}\underline{A}_i h(Z_i) = (\hat{\Pi} M_{\hat{w}} - \Pi)h(Z_{i-1}).$$

Combining these equations, we can write the right side of (10.2) in the form

$$(-1)^{m-1}\int \Big(\frac{a-\hat{a}}{\underline{a}}\Big)\Big[\big(\hat{\Pi}M_{\hat{w}} - \Pi\big)^{m-1}\Big(\frac{\hat{b}-b}{\underline{b}}\Big)\Big]\,\underline{ab}g\,d\nu.$$

We bound this by first applying Hölder's inequality, with conjugate pair (τ, t) with τ equal to r as in the statement of the theorem, and next Lemma S13.7(iii), with $\hat{\Pi}$ and Π viewed as weighted orthogonal projections in $L_2(\underline{ab}\hat{g})$ with weights 1 and $\hat{w}$, respectively, and $r = \tau(m-1)/(m+\tau-3)$, $p = (m+\tau-3)/(\tau-2)$ and $q = (m+\tau-3)/(m-1)$, so that $rp = (m-1)\tau/(\tau-2)$ and $rq = \tau$ (and m of the lemma taken equal to the present m minus 1).

To bound the (conditional) variance of $\hat{\chi}_n$ we use Lemma S14.1 to see that

$$P^n(\mathbb{U}_n\widetilde{\chi}^{(j)}_{\hat{p}})^2 \le 2j\Big[1 + \Big\|\frac{p}{\hat{p}}\Big\|_\infty\Big]^{2j}\hat{P}^n(\mathbb{U}_n\widetilde{\chi}^{(j)}_{\hat{p}})^2 \lesssim \Big[1 + \Big\|\frac{p}{\hat{p}}\Big\|_\infty\Big]^{2j}\frac{2j}{\binom{n}{j}}\hat{P}^j(\widetilde{\chi}^{(j)}_{\hat{p}})^2,$$

because $\widetilde{\chi}^{(j)}_{\hat{p}}$ is degenerate under $\hat{P}$. The variable $\widetilde{\chi}^{(j)}_{\hat{p}}(X_1, \ldots, X_j)/j!$ is the symmetrization of the projection of $\hat{A}_1\hat{\Pi}_{1,2}\underline{A}_2\cdots\hat{\Pi}_{j-1,j}\hat{Y}_j$ onto the degenerate variables. Because the second moment of a mean of (arbitrary) random variables is bounded above by the maximum of the second moments of the terms, we can ignore the symmetrization, while the projection decreases the second moment. This shows that

$$\frac{1}{(j!)^2}\hat{P}^j(\widetilde{\chi}^{(j)}_{\hat{p}})^2 \le \hat{P}^j(\hat{A}_1\hat{\Pi}_{1,2}\underline{A}_2\cdots\hat{\Pi}_{j-1,j}\hat{Y}_j)^2 \lesssim k^{j-1},$$

by Lemma S13.4 and the assumption that the kernels are bounded by k on the diagonal.

We complete the proof of Theorem 8.1 by bounding the square of $\hat{\chi}_n - \chi(\hat{p})$ by $\sum_{j=1}^m 2^j(\mathbb{U}_n\widetilde{\chi}^{(j)}_{\hat{p}}/j!)^2\sum_j 2^{-j}$. The extra factor 2^j can be incorporated in the constant c in the theorem.

For the proof of Theorem 8.2 it clearly suffices to show that

$$\hat{\mathrm{E}}_p\sqrt{n}\big(\hat{\chi}_n - \chi(p) - \mathbb{P}_n\widetilde{\chi}^{(1)}_p\big) \xrightarrow{P} 0,$$
$$\hat{\mathrm{var}}_p\sqrt{n}\big(\hat{\chi}_n - \chi(p) - \mathbb{P}_n\widetilde{\chi}^{(1)}_p\big) \xrightarrow{P} 0.$$

Because an influence function is centered at mean zero, the first is simply $\sqrt{n}$ times the bias of $\hat{\chi}_n$. By Theorem 8.1 the bias is of the order

$$\Big(\frac{\log n}{n}\Big)^{\alpha/(2\alpha+d)+\beta/(2\beta+d)+\gamma(m-1)/(2\gamma+d)} + \Big(\frac{1}{k}\Big)^{(\alpha+\beta)/d}.$$

The first term is trivially $o(n^{-1/2})$, as $m_n \to \infty$. In the second we write $(\alpha+\beta)/d = r/2$, where $r > 1$ by assumption, and see that it is $o(n^{-1/2})$, since $kn^{-1/r} \to \infty$.

To handle the variance we split the estimator $\hat{\chi}_n$ in its linear and higher order terms. The sum of the variances of the U-statistics of orders 2 to m in $\hat{\chi}_n$ is bounded by the sum of the terms $j \geq 2$ in Theorem 8.1, i.e.

$$\sum_{j=2}^{m} \frac{c^j k^{j-1}}{\binom{n}{j}} \leq \frac{1}{n} \sum_{j=2}^{m} \Big(\frac{2ckj}{n}\Big)^{j-1} \frac{c2j\sqrt{j}}{e^j},$$

by the inequalities $\binom{n}{j} \geq (n/2)^j/j!$, for $j < n/2$, and $j! \lesssim (j/e)^j\sqrt{j}$, by Stirling's approximation with bound. The expression in brackets is bounded by $2ckm/n \lesssim 1/\log n$, for $m \sim \log n$ and $k \sim n/(\log n)^2$. Thus the sum tends to zero by dominated convergence. Finally the linear term in $\hat{\chi}_n$ gives the contribution

$$\hat{\mathrm{var}}_p \sqrt{n}\big(\mathbb{P}_n \widetilde{\chi}^{(1)}_{\hat{p}} - \chi(p) - \mathbb{P}_n \widetilde{\chi}^{(1)}_p\big) = \hat{\mathrm{var}}(\widetilde{\chi}^{(1)}_{\hat{p}} - \widetilde{\chi}^{(1)}_p).$$

From the explicit expression (4.7) for the first order influence function (or (7.5) in the case of $\hat{p}$, which gives an identical function), this is seen to tend to zero by the dominated convergence theorem.

10.3. *Proof of Theorem 9.1 for $m = 3$.* The theorem asserts that the bias of the estimator $\hat{\chi}_n$ is equal to the sum of four terms, the first two of which also arise in the bias of the estimator considered in Theorem 8.1. Therefore, we can prove the assertion on the bias by showing that the expected values of the current estimator $\hat{\chi}_n$ (for $m = 3$) and the estimator in Theorem 8.1 differ by less than the additional bias terms in Theorem 9.1.

The two estimators differ only in their third order influence functions, where the present estimator retains only the terms in the double sum (9.3) with $r = 0$, $s = 0$, or $r + s \leq D$. Thus the difference of the expectations of the two estimators is equal to

$$\sum_{\substack{r+s>D \\ r,s\geq 1}} \sum \hat{\mathrm{E}}_p \hat{A}_1 \Big[\hat{\Pi}^{(k_{r-1},k_r]}(Z_1, Z_2) \underline{A}_2 \hat{\Pi}^{(l_{s-1},l_s]}(Z_2, Z_3) - \hat{\Pi}^{(k_{r-1}\vee l_{s-1},k_r\wedge l_s]}(Z_1, Z_3)\Big] \hat{Y}_3.$$

The expectation $\hat{\mathrm{E}}_p$ refers to the variable (X_1, X_2, X_3) for fixed values of the preliminary samples, which are indicated in the "hat" symbols on $\hat{A}_1, \hat{Y}_3$ and the kernels, and hence is an integral relative to the density $(x_1, x_2, x_3) \mapsto p(x_1)p(x_2)p(x_3)$. If we replace $p(x_2)$ in this density by $\hat{p}(x_2)$, then the integral will be zero, as the kernel is degenerate under $\hat{P}$. Thus we may integrate against $(x_1, x_2, x_3) \mapsto p(x_1)(p-\hat{p})(x_2)p(x_3)$. In that case the projection term $\hat{A}_1\hat{\Pi}^{(k_{r-1}\vee l_{s-1}, k_r\wedge l_s]}(Z_1, Z_3)\hat{Y}_3$ integrates to zero, as it does not depend on X_2 and $\int (p-\hat{p})(x_2)\, d\mu(x_2) = 0$, and hence can be dropped. Next we condition $\hat{A}_1$ and $\hat{Y}_3$ on Z_1, Z_2, Z_3 and write the preceding display in the form

$$\sum_{\substack{r+s>D\\ r,s\geq 1}} \sum \iiint \frac{\hat{a}-a}{\underline{a}}(z_1)\hat{\Pi}^{(k_{r-1},k_r]}(z_1,z_2)\hat{\Pi}^{(l_{s-1},l_s]}(z_2,z_3)\frac{\hat{b}-b}{\underline{b}}(z_3) \\ \times\, d\rho(z_1)\, d(\rho-\hat{\rho})(z_2)\, d\rho(z_3).$$

for ρ and $\hat{\rho}$ the measures defined by $d\rho = \underline{abg}\, d\nu$ and $d\hat{\rho} = \underline{ab}\hat{\underline{g}}\, d\nu$. The double sum can be rewritten as the sum over r running from 1 to R and over s from $D-r+1$ to S, which gives the equivalent representation, with the $\times$ referring to "tensor products" as explained in Section 2,

$$\sum_{r=1}^{R} \int \Big(\frac{\hat{a}-a}{\underline{a}} \times 1 \times \frac{\hat{b}-b}{\underline{b}}\Big)\, \Big(\hat{\Pi}^{(k_{r-1},k_r]} \times \hat{\Pi}^{(l_{D-r},k]}\Big)\, d\big(\rho \times (\rho-\hat{\rho}) \times \rho\big).$$

We write $\hat{\Pi}^{(k_{r-1},k_r]} = \hat{\Pi}^{(k_{r-1},k]} - \hat{\Pi}^{(k_r,k]}$, and next arrive at the difference of two expressions of the type, with $k_r' = k_{r-1}$ and $k_r' = k_r$, respectively,

$$\sum_{r=1}^{R} \int \Big(\frac{\hat{a}-a}{\underline{a}} \times 1 \times \frac{\hat{b}-b}{\underline{b}}\Big)\, \Big(\hat{\Pi}^{(k_r',k]} \times \hat{\Pi}^{(l_{D-r},k]}\Big)\, d\big(\rho \times (\rho-\hat{\rho}) \times \rho\big).$$

If the measure of integration were $\hat{\rho} \times (\rho-\hat{\rho}) \times \hat{\rho}$ (with $\hat{\rho}$ instead of ρ), then we could perform the integrals on z_1 and z_3 and next apply Hölder's inequality to bound the resulting expression in absolute value by

$$\sum_{r=1}^{R} \Big\|\hat{\Pi}^{(k_r',k]}\Big(\frac{\hat{a}-a}{\underline{a}}\Big)\Big\|_r \Big\|\hat{\Pi}^{(l_{D-r},k]}\Big(\frac{\hat{b}-b}{\underline{b}}\Big)\Big\|_r \Big\|\frac{g}{\hat{g}} - 1\Big\|_{r/(r-2)},$$

where the norms are those of $L_2(\underline{ab}\hat{\underline{g}})$, which are equivalent to those of $L_2(\nu)$, by assumption. We can write $\hat{\Pi}^{(l,k]} = \hat{\Pi}^{(0,k]}(I - \hat{\Pi}^{(0,l]})$ and use the assumed boundedness of $\hat{\Pi}^{(0,l]}$ as an operator on $L_r(\underline{abg})$ to bound this by the third term in the bias.

Replacing $\rho\times(\rho-\hat\rho)\times\rho$ by $\hat\rho\times(\rho-\hat\rho)\times\hat\rho$ can be achieved by writing the first and last occurrence of ρ as $\rho=\hat\rho+(\rho-\hat\rho)$ and expanding the resulting expression on the $+$ signs into four terms. One of these has the measure $\hat\rho\times(\rho-\hat\rho)\times\hat\rho$. The other three terms have two or three occurrences of $\rho-\hat\rho$, and can be bounded by the first term in the bias (with $m=3$). This is argued precisely under (S10.9) below.

Because the first and second order influence functions are equal to those of the estimator considered in Theorem 8.1, the (conditional) variances of $\mathbb{U}_n\widetilde\chi^{(j)}_{\hat p}$ for $j=1,2$ can be seen to be of the orders $O(1/n)$ and $O(k/n^2)$, respectively, by the same proof. By Lemma S14.1 the variance for $j=3$ is bounded by (see (S14.1))

$$6\Big(1+\Big\|\frac{p}{\hat p}\Big\|_\infty\Big)^6\hat P^n(\mathbb{U}_n\chi^{(3)}_{\hat p})^2$$
$$\lesssim\frac{1}{\binom{n}{3}}\hat P^3\Big(\sum_{r=0}^R\hat A_1\hat\Pi^{(k_{r-1},k_r]}(Z_1,Z_2)\underline{A}_2\hat\Pi^{(0,l'_{D-r}]}(Z_2,Z_3)\hat Y_3\Big)^2,$$

where $l'_{D-r}=l_{D-r}\vee n$. After bounding out $\hat A_1^2$ and $\hat Y_3^2$, we write the squared sum as a double sum. From the fact that the projections $\hat\Pi^{(k_{r-1},k_r]}$ are orthogonal for different r, it follows that the off-diagonal terms of the double sum vanish (the expectation with respect to X_1 is zero). Thus the preceding display is bounded above by a multiple of

$$\frac{1}{n^3}\sum_{r=0}^R\hat P^3\big(\hat\Pi^{(k_{r-1},k_r]}(Z_1,Z_2)\underline{A}_2\hat\Pi^{(0,l'_{D-r}]}(Z_2,Z_3)\big)^2.$$

By Lemmas S13.4 and S13.3 and the assumption that $\sup_z\hat\Pi^{(0,l]}(z,z)\lesssim l$ this is bounded by a multiple of

$$\frac{1}{n^3}\sum_{r=0}^R(k_r-k_{r-1})l'_{D-r}\le\frac{1}{n^3}\Big(nk+\sum_{r=1}^R(k_r-k_{r-1})(l_{D-r}+n)\Big).$$

By (9.4) $k_r-k_{r-1}=(1-2^{-\alpha})k_r\lesssim k_r=n2^{r/\alpha}$ for $r\ge1$. On substituting this in the display, and noting that $l_{D-r}=0$ if $r>D$, we see that this is bounded $k/n^2+2^{D/\alpha\vee D/\beta}/n$ if $\alpha\neq\beta$ and bounded by $k/n^2+D2^{D/\alpha}/n$ if $\alpha=\beta$.

SUPPLEMENTARY MATERIAL

Supplement: ***Estimation of a Functional on a Structured Model under Low Regularity***
(doi: COMPLETED BY THE TYPESETTER; .pdf). The remainder of the paper is given in the supplement.

REFERENCES

[1] Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 3, 647–671. MR663424 (84a:62045)

[2] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD. MR1245941 (94m:62007)

[3] Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50**, 3, 381–393. MR1065550 (91e:62079)

[4] Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23**, 1, 11–29. MR1331653 (96c:62065)

[5] Bolthausen, E., Perkins, E., and van der Vaart, A. (2002). *Lectures on probability theory and statistics*. Lecture Notes in Mathematics, Vol. **1781**. Springer-Verlag, Berlin. Lectures from the 29th Summer School on Probability Theory held in Saint-Flour, July 8–24, 1999, Edited by Pierre Bernard. MR1915443 (2003d:60004)

[6] Cai, T. T. and Low, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33**, 6, 2930–2956. http://dx.doi.org/10.1214/009053605000000147. MR2253108 (2007k:62058)

[7] Cai, T. T. and Low, M. G. (2006). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34**, 5, 2298–2325. http://dx.doi.org/10.1214/009053606000000849. MR2291501 (2008m:62054)

[8] Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6**, 3, 290–323. http://dx.doi.org/10.1016/0885-064X(90)90025-9. MR1081043 (91m:65343)

[9] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. The approach based on influence functions. MR829458 (87k:62054)

[10] Has′minskiĭ, R. Z. and Ibragimov, I. A. (1979). On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics (Hradec Králové, 1978)*. North-Holland, Amsterdam-New York, 41–51. MR571174 (81j:62076)

[11] Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*, Second ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ. http://dx.doi.org/10.1002/9780470434697. MR2488795 (2010j:62004)

[12] Kerkyacharian, G. and Picard, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *Ann. Statist.* **24**, 2, 485–507. MR1394973 (97e:62062)

[13] Koševnik, J. A. and Levit, B. J. (1976). On a nonparametric analogue of the information matrix. *Teor. Verojatnost. i Primenen.* **21**, 4, 759–774. MR0428578 (55 #1599)

[14] Laurent, B. (1997). Estimation of integral functionals of a density and its derivatives. *Bernoulli* **3**, 2, 181–211. MR1466306 (99c:62144)

[15] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, 5, 1302–1338. MR1805785 (2002c:62052)

[16] Lindsay, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11**, 2, 486–497. MR696061 (84h:62050)

[17] Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95**, 450, 449–485. With comments and a rejoinder by the authors. MR1803168 (2002a:62143)

[18] Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*. Lecture Notes in Math., Vol. **1738**. Springer, Berlin, 85–277. MR1775640 (2001h:62074)

[19] Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory.* Lecture Notes in Statistics, Vol. **13**. Springer-Verlag, New York. With the assistance of W. Wefelmeyer. MR675954 (84i:62036)

[20] Pfanzagl, J. (1985). *Asymptotic expansions for general statistical models.* Lecture Notes in Statistics, Vol. **31**. Springer-Verlag, Berlin. With the assistance of W. Wefelmeyer. MR810004 (87i:62004)

[21] Pfanzagl, J. (1990). *Estimation in semiparametric models.* Lecture Notes in Statistics, Vol. **63**. Springer-Verlag, New York. Some recent developments. MR1048589 (91f:62074)

[22] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. Supplement to "higher order estimating equations for high-dimensional models".

[23] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman.* Inst. Math. Stat. Collect., Vol. **2**. Inst. Math. Statist., Beachwood, OH, 335–421. http://dx.doi.org/10.1214/193940307000000527. MR2459958 (2010b:62115)

[24] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2009a). Quadratic semiparametric von mises calculus. *Metrika 69*, 227–247.

[25] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2009b). Semiparametric minimax rates. *Electron. J. Stat. 3*, 1305–1321.

[26] Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90**, 429, 122–129. MR1325119 (96d:62084)

[27] Rotnitzky, A. and Robins, J. M. (1995). Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand. J. Statist.* **22**, 3, 323–333. MR1363216 (96j:62090)

[28] Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique.* Mathématiques & Applications (Berlin) [Mathematics & Applications], Vol. **41**. Springer-Verlag, Berlin. MR2013911 (2005a:62007)

[29] v. Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statistics 18*, 309–348. MR0022330 (9,194h)

[30] van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality.* Springer Series in Statistics. Springer-Verlag, New York. MR1958123 (2003m:62003)

[31] van der Vaart, A. (1991). On differentiable functionals. *Ann. Statist.* **19**, 1, 178–204. MR1091845 (92i:62100)

[32] Van der Vaart, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.* **24**, 2, 862–878. http://dx.doi.org/10.1214/aos/1032894470. MR1394993 (97d:62096)

[33] van der Vaart, A. (2014). Higher Order Tangent Spaces and Influence Functions. *Statist. Sci.* **29**, 4, 679–686. http://dx.doi.org/10.1214/14-STS478. MR3300365

[34] VAN DER VAART, A. W. (1988a). Estimating a real parameter in a class of semiparametric models. *Ann. Statist.* **16**, 4, 1450–1474. http://dx.doi.org/10.1214/aos/1176351048. MR964933 (89m:62032)

[35] VAN DER VAART, A. W. (1988b). *Statistical estimation in large parameter spaces*. CWI Tract, Vol. **44**. Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica, Amsterdam. MR927725 (89e:62049)

[36] VAN DER VAART, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. **3**. Cambridge University Press, Cambridge. MR1652247 (2000c:62003)

[37] VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. MR1385671 (97g:60035)

[38] VAN ZWET, W. R. (1984). A Berry-Esseen bound for symmetric statistics. *Z. Wahrsch. Verw. Gebiete* **66**, 3, 425–440. http://dx.doi.org/10.1007/BF00533707. MR751580 (86h:60063)

[39] WATERMAN, R. P. AND LINDSAY, B. G. (1996). Projected score methods for approximating conditional scores. *Biometrika* **83**, 1, 1–13. http://dx.doi.org/10.1093/biomet/83.1.1. MR1399151 (98g:62044)

JAMES M. ROBINS
ERIC TCHETGEN TCHETGEN
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH
E-MAIL: robins@hsph.harvard.edu
etchetge@hsph.harvard.edu

RAJARSHI MUKHERJEE
STANFORD UNIVERSITY
E-MAIL: rajmrt23@gmail.com

LINGLING LI
SANOFI GENZYME
E-MAIL: lingling07.li@gmail.com

AAD VAN DER VAART
MATHEMATICAL INSTITUTE
LEIDEN UNIVERSITY
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS
E-MAIL: avdvaart@math.leidenuniv.nl