Submitted to the Statistical Science
arXiv: math.PR/0000000

Higher Order Tangent Spaces and Influence Functions

Aad van der Vaart

Abstract. We review higher order tangent spaces and influence functions and their use to construct minimax efficient estimators for parameters in high-dimensional semiparametric models.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Key words and phrases: Semiparametric model, U-statistic, minimax rate of convergence.

1. MAIN DISCUSSION

The concept of *influence function* of an estimator was originally coined in the theory of robust statistics, and as *asymptotic influence function* played a role in the development of semiparametric statistics ([2],[3]). If an estimator T_n of a quantity μ based on a random sample of observations $X_1, X_2, \ldots, X_n$ possesses an asymptotic expansion of the form

$$T_n = \mu + \frac{1}{n}\sum_{i=1}^{n}\psi(X_i) + o_P(n^{-1/2}), \tag{1.1}$$

then the function ψ is its *asymptotic influence function.* The name derives from the fact that if an observation X_i is replaced by a value x, then the change in the estimator is $n^{-1}\big(\psi(x) - \psi(X_i)\big)$, at least if the remainder term $o_P(n^{-1/2})$ is neglected. The estimator is 'asymptotically robust' if this change is bounded in x, i.e. if the influence function ψ is bounded.

Semiparametric theory as developed in the 1980s/90s was not concerned with robustness, but with efficient estimation. Provided that the variables $\psi(X_i)$ have zero mean and finite variance, the expansion (1.1) implies that the sequence $\sqrt{n}(T_n - \mu)$ is asymptotically normally distributed with mean zero. Among different asymptotically unbiased estimators the ones with small asymptotic variance are preferred. Semiparametric lower bound theory showed that under so-called 'asymptotic regularity' estimators with an expansion (1.1) with ψ the *efficient influence function* attain the smallest variance. Furthermore, it showed how to compute the latter function from the *tangent space* of the underlying semiparametric model ([4], [7], [1], and [17]).

Higher order tangent spaces and influence functions are generalizations of these concepts, but were developed by Robins et al. [9] from the perspective of constructing estimators rather than asymptotic efficiency. Thus it will be fruitful to

Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, Netherlands (e-mail: avdvaart@math.leidenuniv.nl)

gives the definitions of influence functions and tangent spaces also from the point of view of constructing estimators.

Assume that the observations $X_1, \ldots, X_n$ are a random sample from a distribution P_η with density p_η relative to a measure μ on a sample space $(\mathcal{X}, \mathcal{A})$. The parameter η is known to belong to a subset $\mathcal{H}$ of a normed space, and it is desired to estimate the value $\chi(\eta)$ of a functional $\chi\colon \mathcal{H} \to \mathbb{R}$. Interest is in the situation of a semiparametric or nonparametric model, where $\mathcal{H}$ is infinite-dimensional, and the dependence $\eta \mapsto p_\eta$ is assumed smooth (as in [16]).

Given a 'consistent' initial estimator $\hat\eta$ of η, the 'plug-in estimator' $\chi(\hat\eta)$ is typically consistent for the parameter of interest $\chi(\eta)$, but it may not be a good estimator. In particular, if $\hat\eta$ is a general purpose estimator, not specially constructed to yield a good plug-in, then $\chi(\hat\eta)$ will often have a suboptimal precision. To gain insight in this situation assume that the parameter permits a Taylor expansion of the form

$$\chi(\eta) = \chi(\hat\eta) + \chi'_{\hat\eta}(\eta - \hat\eta) + O\big(\|\eta - \hat\eta\|^2\big). \tag{1.2}$$

Such an expansion suggests that the plug-in estimator will have an error of the order $O_P\big(\|\eta - \hat\eta\|\big)$, unless the linear term $\chi'_{\hat\eta}(\eta - \hat\eta)$ in the expansion vanishes and the error has the square of this order. For a large parameter set the latter estimation error will typically be large.

The expansion (1.2) also suggests that better estimators can be obtained by 'estimating' the linear term. To achieve this assume a 'generalized von-Mises representation' of the derivative of the form

$$\chi'_{\hat\eta}(\eta - \hat\eta) = \int \dot\chi^1_{\hat\eta}\, d(P_\eta - P_{\hat\eta}) = P_\eta \dot\chi^1_{\hat\eta} + O\big(\|\eta - \hat\eta\|^2\big), \tag{1.3}$$

for some measurable function $\dot\chi^1_{\hat\eta}\colon \mathcal{X} \to \mathbb{R}$. Here Pf is short for the integral $\int f\, dP$, and it is assumed that $P_\eta \dot\chi^1_\eta = 0$ for every η (which can always be arranged by a recentering, as $\int 1\, d(P_\eta - P_{\hat\eta}) = 0$). The von-Mises representation (1.3) and (1.2) suggest the 'corrected plug-in estimator'

$$T_n = \chi(\hat\eta) + \mathbb{P}_n \dot\chi^1_{\hat\eta}, \tag{1.4}$$

where $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$ is the expectation $n^{-1}\sum_{i=1}^n f(X_i)$ of a function f under the empirical measure $\mathbb{P}_n$. It is reasonable to assume that $(\mathbb{P}_n - P_\eta)\dot\chi^1_{\hat\eta}$ is asymptotically equivalent to $(\mathbb{P}_n - P_\eta)\dot\chi^1_\eta$ up to the order $o_P(n^{-1/2})$, as the difference $(\mathbb{P}_n - P_\eta)\dot\chi^1_{\hat\eta}$ is 'centered' and ought to have 'variance' of the order $O(1/n)$. (We put 'centered' and 'variance' in quotes, because the randomness in the initial estimator $\hat\eta$ prevents a simple calculation of mean and variance.) Thus under reasonable regularity conditions the corrected plug-in estimator (1.4) will satisfy

$$\begin{aligned} T_n - \chi(\eta) &= \chi(\hat\eta) - \chi(\eta) + P_\eta \dot\chi^1_{\hat\eta} + (\mathbb{P}_n - P_\eta)\dot\chi^1_{\hat\eta} \\ &= O\big(\|\hat\eta - \eta\|^2\big) + (\mathbb{P}_n - P_\eta)\dot\chi^1_\eta + o_P(n^{-1/2}). \end{aligned} \tag{1.5}$$

If the first term on the right is sufficiently small, specifically $\|\hat\eta - \eta\| = o_P(n^{-1/4})$, then T_n satisfies (1.1) with $\dot\chi^1_\eta$ as the influence function.

The improvement of the estimator (1.4) over the ordinary plug-in estimator is that the estimation error $\|\hat{\eta} - \eta\|$ need have order $O_P(n^{-1/4})$ rather than $O_P(n^{-1/2})$ for the estimator to have error $O_P(n^{-1/2})$. For small 'parametric' models this is not very relevant, but for semi- or nonparametric models the gain can be substantial. For instance, if $\hat{\eta}$ involves an ordinary smoothing estimator of a regression function on a d-dimensional domain, then a typical rate of estimation is $n^{-\alpha/(2\alpha+d)}$, for α the number of derivatives of the true regression function. This is never $O_P(n^{-1/2})$, but $O_P(n^{-1/4})$ for $\alpha \geq d/2$.

The function $\dot{\chi}_\eta^1$ in the von Mises representation (1.3) is exactly an 'influence function' as in the theory of semiparametric models (see [4], [7], [17], [2]), and can be related to the 'tangent set'. Informally, a *tangent set* (at P_η) of a model $(P_\eta : \eta \in \mathcal{H})$ is the set of all *score functions* at $t = 0$

$$\dot{g}_\eta := \frac{\partial}{\partial t}_{|t=0} \log p_{\eta_t} = \frac{\frac{\partial}{\partial t}_{|t=0} p_{\eta_t}}{p_\eta}, \tag{1.6}$$

of (smooth) one-dimensional submodels $(P_{\eta_t} : t \geq 0)$ with $\eta_0 = \eta$. (Here $t \mapsto \eta_t$ is a map from a neighbourhood of $0 \in \mathbb{R}$ to $\mathcal{H}$ such that the derivative (1.6) exists.) An *influence function* (of the real parameter $\chi(\eta)$ at P_η) is defined as a measurable map $x \mapsto \dot{\chi}_\eta^1(x)$ such that, for all paths $t \mapsto \eta_t$ considered,

$$\frac{d}{dt}_{|t=0} \chi(\eta_t) = P_\eta \dot{\chi}_\eta^1 \dot{g}_\eta. \tag{1.7}$$

Combining (1.2)-(1.3) (with η_t in the role of η and η in the role of $\hat{\eta}$) we see that $\chi(\eta_t)$ is to first order given by $\chi(\eta) + P_{\eta_t} \dot{\chi}_\eta^1$. Since, according to (1.6), $\dot{g}_\eta \, dP_\eta$ is the derivative at $t = 0$ of dP_{η_t}, we next conclude that the function $\dot{\chi}_\eta^1$ in the von-Mises expansion (1.3) is an influence function also in the sense of (1.7).

An influence function is not necessarily unique, as only its inner products with elements $\dot{g}_\eta$ of the tangent set matter. An influence function that is contained in the closed linear span of the tangent set is called the *efficient influence function*. It minimizes the variance $\operatorname{var}_\eta \mathbb{P}_n \dot{\chi}_\eta^1$ over all influence functions, and is the influence function of asymptotically efficient estimators.

The theory developed by Robins et al. in [9] extends the preceding from linear to higher order approximations. The motivation is that the parameter η may be so high-dimensional that no estimator $\hat{\eta}$ attains the rate $O_P(n^{-1/4})$. The preceding suggests that then the corrected plug-in estimator will be suboptimal, as in the expansion (1.5) the 'bias' $\chi(\hat{\eta}) - \chi(\eta) + P_\eta \dot{\chi}_{\hat{\eta}}^1$ dominates the 'variance' $(\mathbb{P}_n - P_\eta)\dot{\chi}_{\hat{\eta}}^1$. For this situation Robins et al. [9] introduced higher-order expansions and influence functions, as follows.

A *tangent set of order* m (at P_η) are all derivatives of the type, for given one-dimensional submodels $(P_{\eta_t} : t \geq 0)$,

$$\dot{g}_\eta(x_1, \ldots, x_m) = \frac{\frac{\partial^j}{\partial t^j}_{|t=0} \prod_{i=1}^m p_{\eta_t}(x_i)}{\prod_{i=1}^m p_\eta(x_i)}, \qquad j = 1, 2, \ldots, m. \tag{1.8}$$

The functions on the right side are *higher order score functions* ([14], [6]). These are defined relative to the joint density $(x_1, \ldots, x_m) \mapsto \prod_{i=1}^m p_\eta(x_i)$ of m observations, not as higher-order derivatives of a single density, because higher order

derivatives of the log likelihood of n observations do not reduce to sums over single observations, as do first order derivatives. The relationship between expansions on a single observation and the joint likelihood can be seen from:

$$\begin{aligned}\prod_{i=1}^n \frac{p_{\eta_t}}{p_\eta}(x_i) &= \prod_{i=1}^n \big(1 + t\dot g_\eta(x_i) + \tfrac12 t^2 \ddot g_\eta(x_i) + \cdots\big)\\ &= 1 + t\sum_{i=1}^n \dot g_\eta(x_i) + t^2\Big(\tfrac12\sum_{i=1}^n \ddot g_\eta(x_i) + \sum\sum_{1\le i<j\le n} \dot g_\eta(x_i)\dot g_\eta(x_j)\Big) + \cdots,\end{aligned}$$

Inspection of this expansion shows that the coefficient of t^j is a U-statistic of degree j (cf. equation (1.11) below). The kernels of these U-statistics up to order m can also be obtained as higher order derivatives of products of m densities, as in (1.8). Furthermore, they are *degenerate* in the sense that the integral of a kernel with respect to a single coordinate relative to the true density p_η is zero, generalizing the property that a score function has mean zero; equivalently this property can be described as orthogonality of higher order score functions relative to lower order score functions.

Correspondingly an *influence function of order* m (of the map $\eta \mapsto \chi(\eta)$ at P_η) is a measurable map $(x_1, \ldots, x_m) \mapsto \dot\chi_\eta(x_1, \ldots, x_m)$ such that, for every given one-dimensional submodel $(P_{\eta_t}\colon t \ge 0)$,

$$\frac{\partial^j}{\partial t^j}_{|t=0} \chi(p_{\eta_t}) = P_\eta^m \dot\chi_\eta \dot g_\eta, \qquad j = 1, 2, \ldots, m. \tag{1.9}$$

This influence function is determined only up to its inner products with the tangent set and hence is not unique. A minimal version could be defined as one such that the variance of the U-statistic with kernel $\dot\chi_\eta$ is minimal.

For computation in examples the defining equations (1.9) of a higher order influence function can be tedious. It is usually easier to apply the rule that a higher order derivative is the derivative of the previous order derivative (as shown for second order influence functions in [8], 4.3.11). One computes the first order influence function $x_1 \mapsto \dot\chi_\eta^1(x_1)$ of the functional $\eta \mapsto \chi(\eta)$ as usual. Next one recursively for $j = 2, 3, \ldots, m$ determines influence functions, written, $x_j \mapsto \dot\chi_\eta^j(x_1, \ldots, x_j)$ as influence functions of the functionals $\eta \mapsto \dot\chi_\eta^{j-1}(x_1, \ldots, x_{j-1})$, for fixed $(x_1, \ldots, x_{j-1})$. The function $\dot\chi_\eta^j$ can be made degenerate (in the sense defined previously) by subtracting its projection on the linear span of all functions of one argument less. Then

$$\dot\chi_\eta(x_1, \ldots, x_m) = \sum_{j=1}^m \tfrac{1}{j!}\, \dot\chi_\eta^j(x_1, \ldots, x_j)$$

is an mth order influence function. As we consider only a single value of m at a time, we do not let m show up in the notation on the left. As a consequence the formulas in the following will look as in the linear case.

Given an influence function of order m we may now generalize the definition of the improved plug-in estimator (1.4) to

$$T_n = \chi(\hat\eta) + \mathbb{U}_n \dot\chi_{\hat\eta}, \tag{1.10}$$

for $\mathbb{U}_n f$ denoting a U-statistic of order m with kernel f:

$$\mathbb{U}_n f = \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{1\le i_1\ne i_2\ne\cdots\ne i_m\le n} f(X_{i_1},\ldots,X_{i_m}). \tag{1.11}$$

The term $\mathbb{U}_n\dot\chi_{\hat\eta}$ should correct the plug-in estimator $\chi(\hat\eta)$ up to order m and hence an argument similar to (1.5) should give the expansion

$$T_n - \chi(\eta) = O\big(\|\hat\eta-\eta\|^{m+1}\big) + (\mathbb{U}_n - P_\eta^m)\dot\chi_\eta + o_P(n^{-1/2}). \tag{1.12}$$

The bias of the plug-in estimator $\chi(\hat\eta)$ would be corrected to the order $O\big(\|\hat\eta-\eta\|^{m+1}\big)$, and good estimators for $\chi(\eta)$ exist even in situations where η is estimable only with low precision. The only cost would be a slightly larger variance in the U-statistic relative to the empirical measure.

Unfortunately, there is no such free lunch: one cannot seriously correct bias without seriously increasing the variance. Although (1.12) and the preceding heuristics are correct, they do not apply, as higher order influence functions typically do not exist. Besides by a lack of invertibility of the map $\eta\to p_\eta$, this is caused by failure of a higher order von-Mises type representation. Whereas a continuous, linear map $B\colon L_2(P_\eta)\to\mathbb{R}$, such as arises from the first derivative χ_η' in (1.2), is always representable as an inner product $B(g) = P_\eta\dot\chi_\eta^1 g$ for some function $\dot\chi_\eta^1$, a continuous, multilinear map $B\colon L_2(P_\eta)^j\to\mathbb{R}$ is not necessarily representable as a repeated integral of the type

$$B(g_1,\ldots,g_j) = \int\cdots\int g_1(x_1)\cdots g_j(x_j)\,\dot\chi_\eta(x_1,\cdots,x_j)\,dP_\eta(x_1)\cdots dP_\eta(x_j).$$

The definition (1.10) uses such a 'von-Mises representation' in order to estimate the higher derivatives using the data, by a U-statistic.

We must therefore set a more modest aim: correcting the bias in certain directions only. A key observation is that a multilinear map on a *finite-dimensional* subspace $L\times\cdots\times L\subset L_2(P_\eta)^m$ is always representable by a kernel. If the invertibility $\eta\mapsto p_\eta$ can be resolved, we can therefore always 'represent' and estimate the mth order derivative at differences $\eta-\hat\eta$ *within* a given finite-dimensional linear space. The bias in non-represented directions then remains, and the challenge is to determine the directions that balance three terms:

- the bias in the non-represented directions, *representation bias*,
- the estimation error $O_P\big(\|\hat\eta-\eta\|^{m+1}\big)$, the *estimation bias*,
- the variance of the resulting U-statistic.

Regarding the third component we note that although the variance of a U-statistic with a fixed kernel is dominated by its linear term and is of order $O(1/n)$, the need to represent the functionals in more and more directions given larger sample size n results in kernels that become more and more complex with n. The resulting variance of $\mathbb{U}_n\dot\chi_{\hat\eta}$ is therefore typically larger than $O(1/n)$. A new balance should be found with the squared biases, which will also be larger than parametric.

The preceding heuristic scheme is general, but its implementation requires finding the appropriate influence functions that create the correct bias-variance trade-off. Robins et al. [9] achieved this for estimating a functional in a class of high-dimensional semiparametric models that includes some popular models for

missing data or causal inference. The high dimensions arise by the inclusion of a multivariate 'control covariate'. The models have a technical characterization, through a certain form of the first order influence function. They are *structured semiparametric models* in that their natural parameterization is in terms of three or more parameters, which vary independently. Thus the full parameter takes the form $\eta = (a, b, c, f)$, that is partitioned in three subparameters a, b, c and f. The parameter f is the marginal density of an observable covariate Z. The technical characterization is that the first order influence function of the parameter of interest $\eta \mapsto \chi(\eta)$ can be written in the form

$$\dot{\chi}^1_\eta(x) = a(z)b(z)S_1(x) + a(z)S_2(x) + b(z)S_3(x) + S_4(x) - \chi(\eta), \tag{1.13}$$

for known functions $S_i(x)$ of the data (i.e. $S = (S_1, S_2, S_3, S_4)$ is a given statistic). The covariate Z is assumed to range over a compact d-dimensional domain and the parameters a, b, f are unknown functions on this domain, restricted only nonparametrically by smoothness assumptions. The parameter c is an additional parameter to complete the identification of the distribution of X, but it does not appear in (1.13).

As the higher order corrections are based on von Mises representations of higher order influence functions, which are derivatives of the first order influence function, it is not unnatural to base a theory on the form of the first order influence function. However, by itself (1.13) appears not insightful. The following examples illustrate the class of models.

Example 1.1 (Missing data). *In a version of the* missing data problem *we observe the triple $X = (YA, A, Z)$, where Y and A are random variables that take values in the two-point set $\{0, 1\}$ that are conditionally independent given the variable Z. We can think of Y as a response, which is observed only if the indicator A takes the value 1. To ensure independence of the response and missingness, the covariate Z would be chosen such that it contains all information on the dependence between Y and A ('missing at random'). Alternatively, we can 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'. Both applications may require that Z is high-dimensional (e.g. of dimension 10), where there is typically insufficient a-priori information to model the form of the dependence of A and Y on Z. The three parameters are the marginal density f of Z and the (inverse) probabilities $b(z) = \mathrm{P}(Y = 1|\, Z = z)$ and $a(z)^{-1} = \mathrm{P}(A = 1|\, Z = z)$. The functional of interest is the mean response $\mathrm{E}Y$, i.e.*

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

The representation (1.13) can be shown to be valid with $S_1 = -A$, $S_2 = AY$, $S_3 = 1$ and $S_4 = 0$ (see e.g. [10]). The parameters a and b are (transformed) regression functions and are nonparametrically estimable at the rates $n^{-\alpha/(2\alpha+d)}$ and $n^{-\beta/(2\beta+d)}$ if they are a-priori known to be α- and β-smooth, where d is the dimension of Z. The parameter f is a density and can be estimated from the covariates. Closer inspection (see (1.14) below) shows that a more crucial parameter is the quotient f/a, which is proportional to the conditional density of Z given $A = 1$ and can be estimated directly from the observed covariates and

treatment indicators, at a rate $n^{-\gamma/(2\gamma+d)}$ *if this function is known to be* γ*-smooth. The purpose of constructing higher order influence functions is to ensure that standard nonparametric regression or density estimators can replace the unknown parameters in theoretical expressions with optimal estimators as a result.*

EXAMPLE 1.2 (Covariance model). *Let a typical observation be a triple* $X = (Y, A, Z)$*, where* Y *and* A *are binary variables with values in* $\{0, 1\}$*. We are interested in estimating the expected conditional product moment* $\mathrm{E}\big[\mathrm{E}(Y|\, Z)\mathrm{E}(A|\, Z)\big]$*. In terms of the parameters* $a(Z) = \mathrm{E}(A|\, Z)$ *and* $b(Z) = \mathrm{E}(Y|\, Z)$*, and* $\eta = (a, b, f, c)$*, for* f *the marginal density of* Z *and* c *an additional parameter, this target can be written as*

$$\chi(\eta) = \int abf\, d\nu.$$

Representation (1.13) can be seen to hold with $S_1 = -1$*,* $S_2 = A$*,* $S_3 = Y$ *and* $S_4 = 0$*. The parameters* a *and* b *are regression functions of* Y *and* A *on* Z *and hence can be estimated at the rates* $n^{-\alpha/(2\alpha+d)}$ *and* $n^{-\beta/(2\beta+d)}$ *if they are a-priori known to be* α*- and* β*-smooth. The marginal density* f *can be similarly estimated nonparametrically from the observed covariates.*

The triple (a, b, f) *does not fully parameterize the joint distribution of an observation, but the remaining part* c *of the parameter does not seem to play a role when estimating* $\chi(\eta)$*. A full parameterization is obtained by adding the* treatment effect function $c(Z) = \mathrm{E}(Y|\, A = 1, Z) - \mathrm{E}(Y|\, A = 0, Z)$*. The conditional distribution of* Y *given* A *can then be expressed in* (a, b, c, f) *through* $\mathrm{P}(Y = 1|\, A, Z) = c(Z)(A - a(Z)) + b(Z)$*.*

Estimating $\chi(\eta)$ *is relevant to the biostatistical setup through a detour, which relates* $\chi(\eta)$ *to the treatment effect function* c*. First, in terms of statistical difficulty the functional* $\chi(\eta)$ *is equivalent to the functional* $\mathrm{E}\,\mathrm{cov}(Y, A|\, Z) = \mathrm{E}(YA) - \chi(\eta)$*, as* $\mathrm{E}(YA)$ *can be estimated at the rate* $n^{-1/2}$ *by a simple sample average. Second, the problem of estimating* $\mathrm{E}\,\mathrm{cov}(Y, A|\, Z)$ *is a template for estimating* $\psi(t) := \mathrm{E}\,\mathrm{cov}(Y - tA, A|\, Z)$*, for every given* t*, which can next be inverted to give an estimate for the value* τ *that satisfies* $\psi(\tau) = 0$*. The latter value can be shown to be equal to the* variance weighted average treatment effect

$$\tau = \frac{\mathrm{E}\,\mathrm{var}(A|\, Z)c(Z)}{\mathrm{E}\,\mathrm{var}(A|\, Z)}.$$

(See [12], Section 4 for details.) Under the assumption of non-confounding this parameter is nonzero if and only if the treatment A *has a nonzero causal effect, and it may be the ultimate purpose to ascertain this.*

EXAMPLE 1.3 (Average treatment effect). *Suppose a clinical trial with two possible treatments, indicated by* $A \in \{0, 1\}$*, has two binary outcome variables* Y_1 *and* Y_2*, and let* $a_j(Z) = \mathrm{E}(Y_j|\, A = 1, Z) - \mathrm{E}(Y_j|\, A = 0, Z)$ *be the treatment effects at level* Z *of an observed covariate, for* $j = 1, 2$*. We observe a random sample of the variables* (Y_1, Y_2, A, Z)*, and are interested in estimating the* average treatment effect

$$\chi(\eta) = \int a_1 a_2\, f\, d\nu.$$

Here η *parameterizes the distribution of* (Y_1, Y_2, A, Z)*, and* f *is the density of the covariate* Z*, relative to some measure* ν*, for instance Lebesgue measure on a*

compact subset of $\mathbb{R}^d$*. The parameter* η *includes the triplet* (a_1, a_2, f)*, and possibly other unknown aspects of the distribution of an observation. In a clinical trial the probability* $\pi(Z) = \mathrm{P}(A = 1|\, Z)$ *that an individual with covariate* Z *is treated will be a known function of the covariate.*

As the tangent space is a true subspace of the full tangent space, there are multiple influence functions for χ*. It can be shown that any influence function of* χ *can be represented in the form (1.13) with, for some measurable function* C*,*

$$S_1 = 1 - \frac{2A(A-\pi(Z))}{\pi(Z)(1-\pi)(Z)}, \qquad S_2 = Y_2 \frac{A-\pi(Z)}{\pi(Z)(1-\pi)(Z)},$$
$$S_3 = Y_1 \frac{A-\pi(Z)}{\pi(Z)(1-\pi)(Z)}, \qquad S_4 = C(Z) \frac{A-\pi(Z)}{\pi(Z)(1-\pi)(Z)}.$$

Perhaps the special case that $Y_1 = Y_2$ *is of most interest. The parameter* (a_1, a_2, f) *then reduces to a pair* (a, f)*, and* $S_2 = S_3$*, but the general setup remains the same.*

In models with first order influence function of the form (1.13) the error of the first order von-Mises representation (1.2) -(1.3) can be computed to be, for a given initial estimator $\hat\eta = (\hat a, \hat b, \hat f)$,

$$\chi(\hat\eta) - \chi(\eta) + P_\eta \dot\chi^1_{\hat\eta} = \int (\hat a - a)(\hat b - b)\, \tilde s_{\eta,1} f\, d\nu, \tag{1.14}$$

for $\tilde s_{\eta,i}(z) = \mathrm{E}_\eta(S_i|\, Z = z)$. (From the fact that a, b and f are only nonparametrically restricted and that (1.13) gives the influence function it can be shown that necessarily $\tilde s_{\eta,1} b + \tilde s_{\eta,2} = 0 = \tilde s_{\eta,1} a + s_{\eta,3}$, after which identity (1.14) follows by algebra.) This is quadratic in the errors $\hat a - a$ and $\hat b - b$ of the initial estimators, but is special in that the squares of the estimation errors $|\hat a - a|$ and $|\hat b - b|$ of the two initial estimators $\hat a$ and $\hat b$ do no arise, but only their product. This property, termed 'double robustness' in [11], [13], makes that in first order inference it suffices that one of the two parameters is estimated well. If initial estimators of a and b attain estimation rates $n^{-\alpha/(2\alpha+d)}$ and $n^{-\beta/(2\beta+d)}$, respectively, then the order of the remainder term in the expansion is the product of these rates. This shows that the linear estimator (1.4) attains a rate $O_P(n^{-1/2})$ if

$$\frac{\alpha}{2\alpha+d} + \frac{\beta}{2\beta+d} \geq \frac{1}{2}. \tag{1.15}$$

If this condition fails, then the 'bias' (1.14) is greater than $O_P(n^{-1/2})$. The linear estimator (1.4) then does not balance bias and variance and is suboptimal.

For moderate to large dimensions d inequality (1.15) is a restrictive requirement, whose validity is questionable for many applications. Higher order influence functions allow to construct better estimators than the linear estimator (1.4). As shown in [9], [10], [15], [12], [5] there are two cases:

- $(\alpha+\beta)/2 \geq d/4$. In this case estimation at rate $n^{-1/2}$ is possible by using a higher order estimator (1.10) of sufficiently large order m. If the inequality is strict, then this estimator is also semiparametrically regular and efficient, even though (1.15) need not be satisfied.
- $(\alpha+\beta)/2 < d/4$. In this case the minimax rate of estimation is slower than $n^{-1/2}$. If the function $\tilde s_{\eta,1} f$ has a regularity γ bigger than a certain cut-off

(that depends on (α, β)), then the minimax rate is $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+1)}$ and is attainable by a higher order estimator (1.10) with a carefully constructed approximate influence function $\dot{\chi}_\eta$.

In both cases it is necessary to estimate the marginal density f, or rather the function $\tilde{s}_{\eta,1}f$, notwithstanding the fact that it does not enter the first order influence function (1.13). Robins et al. [9] construct minimax estimators under the assumption that this function has a minimal smoothness. A completely general solution is apparently still more complicated.

The details of the constructions are beyond the scope of the present paper. The approximations are based on expanding the parameters a and b on bases that express their regularity (e.g. suitable wavelets), and representing the higher order derivatives of the functional χ on the subspaces obtained by truncating these bases. The truncation point is chosen relative to the functional to be estimated (and not necessarily the usual one used to estimate the functions themselves). For orders three and up, it is in addition necessary to remove pairs of basis functions (resulting from the pair (a, b)) whose combined index is 'large', in order to cut variance without increasing bias. For an introduction to constructing truncated second order influence functions we refer to [10].

2. CONCLUDING REMARKS

One may look at the work of Robins et al. [9] and its sequel from two perspectives. The mathematical statistical point of view is the simplest: higher order estimating functions are a means to construct estimators that are theoretically minimax in complex semiparametric models, where the interest is not simply in a mean of the observations, but in a parameter defined through the structure of the model. As always in high-dimensional models minimaxity is about the bias-variance trade-off. Inspection of higher order tangent spaces reveals in what form the bias arises, and the connected von Mises calculus allows to correct for it. So far no completely general method exists for trading this against variance (other than the abstract idea to use 'finite-dimensional approximations'), and in fact beyond the application to models characterized by (1.13) nothing much is known.

The second perspective is practically oriented. The models dealt with in this paper are relevant in studies in epidemiology, econometrics, and the social sciences. The parameter of interest is defined through the substantial application, for instance measuring a response to treatment or the consequence of an intervention. High dimensions arise to identify this parameter of interest from data. Observational studies, where covariates must be included in the statistical analysis to control for possible confounding are a typical case. One has a choice to adopt a relatively simple statistical model for this complex reality, maybe even a classical parametric model or a one-dimensional propensity score, or to let the data 'speak for itself', as much as possible. Without any model restriction one runs into the 'curse of dimensionality' and no conclusions are possible. Semiparametric models as developed in the 1980s and 1990s are between these extremes, but from the present perspective relatively close to finite-dimensional models. In fact, they focus on functionals in situations where a bias-variance trade-off is unnecessary, as the bias is negligible. The main purpose of methods based on high-dimensional influence functions is to fill the huge gap between 'classical semi-

parametric models' and the model in which nothing is assumed. In a situation with fewer or less stringent a-priori assumptions on the model, statistical bias starts playing a role and must be traded versus variance. Estimators with bigger standard errors result, but bias due to model misspecification decreases. The choice between model bias with smaller variance and larger estimation variance is not easy to make with current statistical methodology. However, larger and larger data bases certainly make the methodology of higher influence functions feasible.

Thus these methods are potentially useful to answer a wide range of questions. We close with some remarks about further research that needs to be done to make the methods fully operational.

The improved estimators based on higher order influence functions combine good preliminary estimators for deviations of the parameter of interest $\chi(\eta)$ in some directions with a-priori assumptions that the deviations in other 'nonestimable' directions are small. The latter a-priori assumptions are always questionable. It is an open problem to develop estimation procedures that can 'adapt' to 'scales of a-priori conditions', for instance by implicitly estimating unknown smoothness levels from the data.

For practical application estimation without error indications are insufficient. Although there is some preliminary work on confidence intervals related to the higher order estimators, these procedures remain to be explored.

The models (1.13) considered in [9] are structured semiparametric models (with a partitioned parameter (a, b, c, f) and the functional of interest defined naturally in terms of the partition), but typically nonparametric in the sense that any law on the sample space is realized by some choice of the parameters (a, b, c, f). Genuinely semiparametric problems, such as partially linear regression, pose a further challenge. For such models the first order influence function is non-unique, and as the estimation error is bigger than the first order variance, the efficient first order influence function may not play a special role, thus increasing the degrees of freedom in constructing suitable higher order influence functions.

REFERENCES

[1] Begun, J. M., Hall, W. J., Huang, W.-M., and Wellner, J. A. Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist. 11*, 2 (1983), 432–452.

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

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

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

[5] Li, L., Tchetgen Tchetgen, E., van der Vaart, A., and Robins, J. M. Higher order inference on a treatment effect under low regularity conditions. *Statist. Probab. Lett. 81*, 7 (2011), 821–828.

[6] Lindsay, B. G. Efficiency of the conditional score in a mixture setting. *Ann. Statist. 11*, 2 (1983), 486–497.

[7] Pfanzagl, J. *Contributions to a general asymptotic statistical theory*, vol. 13 of *Lecture Notes in Statistics.* Springer-Verlag, New York, 1982. With the assistance of W. Wefelmeyer.

[8] Pfanzagl, J. *Asymptotic expansions for general statistical models*, vol. 31 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1985. With the assistance of W. Wefelmeyer.

[9] Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, vol. 2 of *Inst. Math. Stat. Collect.* Inst. Math. Statist., Beachwood, OH, 2008, pp. 335–421.

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

[11] Robins, J., and Rotnitzky, A. Comment on the bickel and kwon article, "inference for semiparametric models: Some questions and an answer". *Statistica Sinica 11(4)* (2001), 920–936.

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

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

[14] Small, C. G., and McLeish, D. L. *Hilbert space methods in probability and statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1994. A Wiley-Interscience Publication.

[15] Tchetgen, E., Li, L., Robins, J., and van der Vaart, A. Minimax estimation of the integral of a power of a density. *Statist. Probab. Lett. 78*, 18 (2008), 3307–3311.

[16] van der Vaart, A. On differentiable functionals. *Ann. Statist. 19*, 1 (1991), 178–204.

[17] van der Vaart, A. W. *Statistical estimation in large parameter spaces*, vol. 44 of *CWI Tract*. Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica, Amsterdam, 1988.