Estimation of a non-parametric variable importance measure of a continuous exposure

Antoine Chambaz, Pierre Neuvial, Mark Van Der Laan

Estimation of a non-parametric variable importance measure of a continuous exposure

Antoine Chambaz[1], Pierre Neuvial[2], Mark J. van der Laan[3]

[1] MAP5, Université Paris Descartes and CNRS
[2] Laboratoire Statistique et Génome, Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA
[3] Division of Biostatistics, UC Berkeley

September 30, 2011

Abstract

We define a new measure of variable importance of an exposure on a continuous outcome, accounting for potential confounders. The exposure features a reference level x_0 with positive mass and a *continuum* of other levels. For the purpose of estimating it, we fully develop the semi-parametric estimation methodology called targeted minimum loss estimation methodology (TMLE) [23, 22]. We cover the whole spectrum of its theoretical study (convergence of the iterative procedure which is at the core of the TMLE methodology; consistency and asymptotic normality of the estimator), practical implementation, simulation study and application to a genomic example that originally motivated this article. In the latter, the exposure X and response Y are, respectively, the DNA copy number and expression level of a given gene in a cancer cell. Here, the reference level is $x_0 = 2$, that is the expected DNA copy number in a normal cell. The confounder is a measure of the methylation of the gene. The fact that there is no clear biological indication that X and Y can be interpreted as an exposure and a response, respectively, is not problematic.

1 Introduction

Consider the following statistical problem: One observes the data structure $O = (W, X, Y)$ on an experimental unit of interest, where $W \in \mathcal{W}$ stands for a vector of baseline covariates, and $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ respectively quantify an exposure and a response; the exposure features a reference level x_0 with positive mass (there is a positive probability that $X = x_0$) and a *continuum* of other levels (a first source of difficulty); one wishes to investigate the relationship between X on Y, accounting for W (a second source of difficulty) and making few assumptions on the true data-generating distribution (a third source of difficulty). Taking

W into account is desirable when one knows (or cannot rule out the possibility) that it contains confounding factors, *i.e.*, common factors upon which the exposure X and the response Y may simultaneously depend.

We illustrate our presentation with an example where the experimental unit is a set of cancer cells, the relevant baseline covariate W is a measure of DNA methylation, the exposure X and response Y are, respectively, the DNA copy number and expression level of a given gene. Here, the reference level is $x_0 = 2$, that is the expected copy number in a normal cell. The fact that there is no clear biological indication that X and Y can be interpreted as an exposure and a response, respectively, is not problematic. Associations between DNA copy numbers and expression levels in genes have already been considered in the literature (see *e.g.*, [11, 26, 1, 17, 10]). In contrast to these earlier contributions, we do exploit the fact that X features both a reference level and a continuum of other levels, instead of discretizing it or considering it as a purely continuous exposure.

We focus on the case that there is very little prior knowledge on the true data-generating distribution P_0 of O, although we know/assume that *(i)* O takes its values in the *bounded set* $\mathcal{O}$ (we will denote $\|O\| = \max\{|W|, |X|, |Y|\}$), *(ii)* $P_0(X \neq x_0) > 0$, and finally *(iii)* $P_0(X \neq x_0|W) > 0$ P_0-almost surely. Accordingly, we see P_0 as a specific element of the non-parametric set $\mathcal{M}$ of all possible data-generating distributions of O satisfying the latter constraints. We define the parameter of interest as $\Psi(P_0)$, for the non-parametric variable importance measure $\Psi : \mathcal{M} \to \mathbb{R}$ characterized by

$$\Psi(P) = \underset{\beta \in \mathbb{R}}{\arg\min}\, E_P \left\{ \left(E_P(Y|X,W) - E_P(Y|X = x_0, W) - \beta(X - x_0) \right)^2 \right\} \tag{1}$$

for all $P \in \mathcal{M}$. The methodology presented in this article straightforwardly extends to situations where one would prefer to replace the expression $\beta(X - x_0)$ in (1) by $\beta f(X)$ for any f such that $f(x_0) = 0$ and $E_P\{f(X)^2\} > 0$ for all $P \in \mathcal{M}$. We emphasize that we *do not* assume a semi-parametric model (which would write here as $Y = \beta(X - x_0) + \eta(W) + U$ with unspecified η and U such that $E_P(U|X,W) = 0$), in contrast to [15, 14, 28, 21, 20]. This fact bears important implications. The parameter of interest, $\Psi(P_0)$, is universally defined (therefore justifying the expression "*non-parametric* variable importance measure of a continuous exposure" in the title), no matter what properties the unknown true data-generating distribution P_0 enjoys, or does not enjoy.

Parameter Ψ quantifies the influence of X and Y on a linear scale, using the reference level x_0 as a pivot (note that this expression conveys the notion that the role of X and Y are not symmetric). As its name suggests, Ψ belongs to the family of variable importance measures (a family that includes the excess risk), which was introduced in [21]. However, its case is not covered by the latter article because X is continuous (we will see how Ψ naturally relates to an excess risk when X takes only two distinct values). Our purpose here is to fully develop the semi-parametric estimation methodology called targeted minimum loss estimation (TMLE) methodology [23, 22]. We cover the whole spectrum of its theoretical study, practical

implementation, simulation study, and application to the aforementioned genomic example.

In Section 2, we study the fundamental properties of parameter Ψ. In Section 3 we provide an overview of the TMLE methodology tailored for the purpose of estimating $\Psi(P_0)$. In Section 4, we state and comment on important theoretical properties enjoyed by the TMLE (convergence of the iterative updating procedure at the core of its definition; its consistency and asymptotic normality). The specifics of the TMLE procedure are presented in Section 5. The properties considered in Section 4 are illustrated by a simulation study inspired by the problem of assessing the importance of DNA copy number variations on expression level in genes, accounting for their methylation (the real data application we are ultimately interested in), as described in Section 6. All proofs are postponed to the appendix.

We assume from now on, without loss of generality, that $\boldsymbol{x_0 = 0}$. For any measure λ and measurable function f, $\lambda f = \int f d\lambda$. We set $L_0^2(P) = \{s \in L^2(P) : Ps = 0\}$. Moreover, the following notation are used throughout the article: for all $P \in \mathcal{M}$, $\theta(P)(X,W) = E_P(Y|X,W)$, $\mu(P)(W) = E_P(X|W)$, $g(P)(0|W) = P(X = 0|W)$, and $\sigma^2(P) = E_P\{X^2\}$. In particular, $\Psi(P)$ can also be written as

$$\Psi(P) = \underset{\beta \in \mathbb{R}}{\arg\min}\, E_P\left\{(\theta(P)(X,W) - \theta(P)(0,W) - \beta X)^2\right\}.$$

2 The non-parametric variable importance parameter

It is of paramount importance to study the parameter of interest in order to better estimate it. Parameter Ψ actually enjoys the following properties [see Chapter 25 in 25, for definitions].

Proposition 1. *For all $P \in \mathcal{M}$,*

$$\Psi(P) = \frac{E_P\{X(\theta(P)(X,W) - \theta(P)(0,W))\}}{E_P\{X^2\}}. \tag{2}$$

Parameter Ψ is pathwise differentiable at every $P \in \mathcal{M}$ with respect to the maximal tangent set $L_0^2(P)$. Its efficient influence curve at P is $D^\star(P) = D_1^\star(P) + D_2^\star(P)$, where $D_1^\star(P) = D_1^\star(\sigma^2(P), \theta(P), \Psi(P))$ and $D_2^\star(P) = D_2^\star(\sigma^2(P), \theta(P), \mu(P), g(P))$ are two $L_0^2(P)$-orthogonal components characterized by

$$\begin{aligned} D_1^\star(\sigma^2, \theta, \psi)(O) &= \frac{1}{\sigma^2}(X(\theta(X,W) - \theta(0,W) - X\psi)), \\ D_2^\star(\sigma^2, \theta, \mu, g)(O) &= \frac{1}{\sigma^2}(Y - \theta(X,W))\left(X - \frac{\mu(W)\mathbf{1}\{X=0\}}{g(0|W)}\right). \end{aligned}$$

Furthermore, the efficient influence curve is double-robust: for any $(P, P') \in \mathcal{M}^2$, if either $\theta(P')(0,\cdot) = \theta(P)(0,\cdot)$ or ($\mu(P') = \mu(P)$ and $g(P') = g(P)$) holds, then $PD^\star(P') = 0$ implies $\Psi(P') = \Psi(P)$.

The proof of Proposition 1 is relegated to Section A.2.

Let us emphasize again that *we do not assume* a semi-parametric model $Y = \beta X + \eta(W) + U$ (with unspecified η and U such that $E_P(U|X,W) = 0$). Setting $R(P,\beta)(X,W) = \theta(P)(X,W) - \theta(P)(0,W) - \beta X$ for all $(P,\beta) \in \mathcal{M} \times \mathbb{R}$, the latter semi-parametric model holds for $P \in \mathcal{M}$ if there exists a unique $\beta(P) \in \mathbb{R}$ such that $R(P,\beta(P)) = 0$. Note that β is *always* solution to the equation $\beta E_P\{X^2\} = E_P\{X\,(\theta(P)(X,W) - \theta(P)(0,W) - R(P,\beta)(X,W))\}$. In particular, if the semi-parametric model holds for a certain $P \in \mathcal{M}$, then $\beta(P) = \Psi(P)$ by (2). On the contrary, if the semi-parametric model does not hold for P, then it is not clear what $\beta(P)$ could even *mean* whereas $\Psi(P)$ is still a well-defined parameter worth estimating. We discuss in Section 4.2 what happens if one estimates $\beta(P)$ when assuming wrongly that the semi-parametric holds (the discussion allows to identify the awkward non-parametric extension of parameter $\beta(P)$ that one therefore estimates).

Equality (2) also teaches us that

$$\Psi(P) = \mathcal{F}(P) - \frac{E_P\{\mu(P)(W)\theta(P)(0,W)\}}{\sigma^2(P)} \tag{3}$$

for the functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$ characterized by

$$\mathcal{F}(P) = \arg\min_{\beta \in \mathbb{R}} E_P\left\{(Y - \beta X)^2\right\} \equiv \frac{E_P\{XY\}}{\sigma^2(P)} \tag{4}$$

(all $P \in \mathcal{M}$). In that view, the second term in the right-hand side of (3) is a correction term added to $\mathcal{F}(P)$ in order to take W into account for the purpose of quantifying the influence of X on Y on a linear scale. Whereas the roles of X and Y are symmetric in the numerator of $\mathcal{F}(P)$, they are obviously not in that of the correction term. Less importantly, (2) also makes clear that there is a connexion between Ψ and an excess risk. Indeed, consider $P \in \mathcal{M}$ such that $P(X \in \{0, x_1\}) = 1$ for $x_1 \neq 0$. Then $\Psi(P)$ satisfies

$$\Psi(P) = \frac{E_P\{(\theta(P)(x_1,W) - \theta(P)(0,W))h(P)(W)\}}{\sigma^2(P)}$$

for $h(P)(W) = P(X = x_1|W)$, *i.e.*, $\Psi(P)$ appears as a *weighted* excess risk (the classical excess risk would be here $E_P\{\theta(P)(x_1,W) - \theta(P)(0,W)\}$).

Since Ψ is pathwise differentiable, the theory of semi-parametric estimation applies, providing a notion of asymptotically efficient estimation. Remarkably, the asymptotic variance of a regular estimator of $\Psi(P_0)$ is lower-bounded by the variance $\mathrm{Var}_{P_0} D^\star(P_0)(O)$ under P_0 of the efficient influence curve at P_0 (a consequence of the convolution theorem). The TMLE procedure takes advantage of the properties of Ψ described in Proposition 1 in order to build a consistent and possibly asymptotically efficient substitution estimator of $\Psi(P_0)$. In view of (3), this is a challenging statistical problem because, whereas estimating $\mathcal{F}(P_0)$ is straightforward (the ratio of the empirical means of XY and X^2 is an efficient estimator of $\mathcal{F}(P_0)$), estimating the correction term in (3) is more delicate, notably because this necessarily involves estimating the infinite-dimensional features $\theta(P_0)(0,\cdot)$ and $\mu(P_0)$.

3 Overview of the TMLE procedure tailored to the estimation of the non-parametric variable importance measure

We assume now that we observe n independent copies $O^{(1)} = (W^{(1)}, X^{(1)}, Y^{(1)}), \ldots, O^{(n)} = (W^{(n)}, X^{(n)}, Y^{(n)})$ of the observed data structure $O \sim P_0 \in \mathcal{M}$. The empirical measure is denoted by P_n. The TMLE procedure iteratively updates an initial substitution estimator $\psi_n^0 = \Psi(P_n^0)$ of $\Psi(P_0)$ (based on an initial estimator P_n^0 of the data-generating distribution P_0), building a sequence $\{\psi_n^k = \Psi(P_n^k)\}_{k \geq 0}$ (with P_n^k the kth update of P_n^0) which converges to the targeted minimum loss estimator (TMLE) ψ_n^* as k increases. This iterative scheme is visually illustrated in Figure 1, and we invite the reader to consult its caption now.

We determine what initializing the TMLE procedure boils down to in Section 3.1. A general one-step targeted updating procedure is described in Section 3.2. How to conduct specifically these initialization and update (as well as two alternative tailored two-step updating procedures) is addressed in Section 5.

3.1 Initial estimator

In this subsection, we describe what it takes to construct an initial substitution estimator of $\Psi(P_0)$. Of course, how one derives the substitution estimator $\Psi(P)$ from the description of (certain features of) P is relevant even if P is not literally an initial estimator of P_0.

By (2), building an initial substitution estimator $\Psi(P_n^0)$ of $\Psi(P_0)$ requires the estimation of $\theta(P_0)$, of $\sigma^2(P_0)$, and of the marginal distribution of (W, X) under P_0. Given P_n^0, initial estimator of P_0 with known $\theta(P_n^0)$, $\sigma^2(P_n^0) > 0$ and marginal distribution of (W, X) under P_n^0, $\Psi(P_n^0)$ can indeed be obtained (or, more precisely, evaluated accurately) by the law of large numbers, as discussed below. We emphasize that such an initial estimator may very well be biased. In other words, one would need strong assumptions on the true data-generating distribution P_0 (which we are not willing to make; typically, assuming that P_0 belongs to a given regular parametric model) and adapting the construction of P_n^0 based on those assumptions (typically, relying on maximum likelihood estimation) in order to obtain the consistency of $\Psi(P_n^0)$.

For B a large integer (say $B = 10^5$), evaluating accurately (rather than computing exactly) the initial substitution estimator $\Psi(P_n^0)$ of $\Psi(P_0)$ boils down to simulating B independent copies $(\tilde{W}^{(b)}, \tilde{X}^{(b)})$ of (W, X) under P_n^0, then using the approximation

$$\psi_n^0 = \Psi(P_n^0) = \frac{B^{-1} \sum_{b=1}^{B} \tilde{X}^{(b)} (\theta(P_n^0)(\tilde{X}^{(b)}, \tilde{W}^{(b)}) - \theta(P_n^0)(0, \tilde{W}^{(b)}))}{\sigma^2(P_n^0)} + O(B^{-1/2}). \quad (5)$$

Knowing the marginal distribution of (W, X) under P_n^0 amounts to knowing *(i)* the marginal distribution of W under P_n^0, *(ii)* the conditional distribution of $Z \equiv \mathbf{1}\{X = 0\}$ given W under P_n^0, and *(iii)* the conditional distribution of X given $(W, X \neq 0)$ under P_n^0. Firstly, we advocate for estimating initially the marginal distribution of W under P_0

by its empirical version, or put in terms of likelihood, to build P_n^0 in such a way that $P_n^0(W) = n^{-1}\sum_{i=1}^n \mathbf{1}\{W^{(i)} = W\}$. Secondly, the conditional distribution of Z given W under P_n^0 is the Bernoulli law with parameter $1 - g(P_n^0)(0|W)$, so it is necessary that $g(P_n^0)$ be known too (and such that, P_n^0-almost surely, $g(P_n^0)(0|W) \in (0,1)$). Thirdly, the conditional distribution of X given $(W, X \neq 0)$ under P_n^0 can be *any* (finite variance) distribution, whose conditional mean can be deduced from $\mu(P_n^0)$:

$$E_{P_n^0}(X|X \neq 0, W) = \frac{\mu(P_n^0)(W)}{1 - g(P_n^0)(0|W)}, \tag{6}$$

and whose conditional second order moment $E_{P_n^0}(X^2|X \neq 0, W)$ satisfies

$$E_{P_n^0}\left\{(1 - g(P_n^0)(0|W))E_{P_n^0}(X^2|X \neq 0, W)\right\} = \sigma^2(P_n^0). \tag{7}$$

In particular, it is also necessary that $\mu(P_n^0)$ be known too.

In summary, the only features of P_n^0 we really care for in order to evaluate accurately (rather than compute exactly) $\psi_n^0 = \Psi(P_n^0)$ are $\theta(P_n^0)$, $\mu(P_n^0)$, $g(P_n^0)$, $\sigma^2(P_n^0)$, and the marginal distribution of W under P_n^0, which respectively estimate $\theta(P_0)$, $\mu(P_0)$, $g(P_0)$, $\sigma^2(P_0)$, and the marginal distribution of W under P_0. We could for instance rely on a working model where the conditional distribution of X given $(W, X \neq 0)$ is *chosen* as the Gaussian distribution with conditional mean as in (6) and any conditional second order moment (which is nothing but a measurable function of W) such that (7) holds. Let us emphasize that we do use here expressions from the semantical field of *choice*, and not from that of *assumption*; a working model is just a tool we use in the construction of the initial estimator, and we do not necessarily assume that it is well-specified. Although such a Gaussian working model would be a perfectly correct choice, we advocate for using another one for computational convenience, as presented in Section 5.1.

3.2 A general one-step updating procedure of the initial estimator

The next step consists in iteratively updating $\psi_n^0 = \Psi(P_n^0)$. Assuming that one has already built $(k-1)$ updates $P_n^1, \ldots, P_n^{k-1}$ of P_n^0, resulting in $(k-1)$ updated substitution estimators $\psi_n^1 = \Psi(P_n^1), \ldots, \psi_n^{k-1} = \Psi(P_n^{k-1})$, it is formally sufficient to describe how the kth update P_n^k is derived from its predecessor P_n^{k-1} in order to fully determine the iterative procedure. Note that the value of $\psi_n^1 = \Psi(P_n^1), \ldots, \psi_n^{k-1} = \Psi(P_n^{k-1})$ are derived as $\psi_n^0 = \Psi(P_n^0)$, by following (5) in Section 3.1 with $P_n^1, \ldots, P_n^{k-1}$ substituted for P_n^0.

We present here a general one-step updating procedure (two alternative tailored two-step updating procedures are also presented in Section 5.2). We invite again the reader to refer to Figure 1 for its visual illustration.

Set $\rho \in (0,1)$ a constant close to 1 and consider the path $\{P_n^{k-1}(\varepsilon) : |\varepsilon| \leq \rho \|D^\star(P_n^{k-1})\|_\infty^{-1}\}$ characterized by

$$\frac{dP_n^{k-1}(\varepsilon)}{dP_n^{k-1}}(O) = \left(1 + \varepsilon D^\star(P_n^{k-1})(O)\right), \tag{8}$$

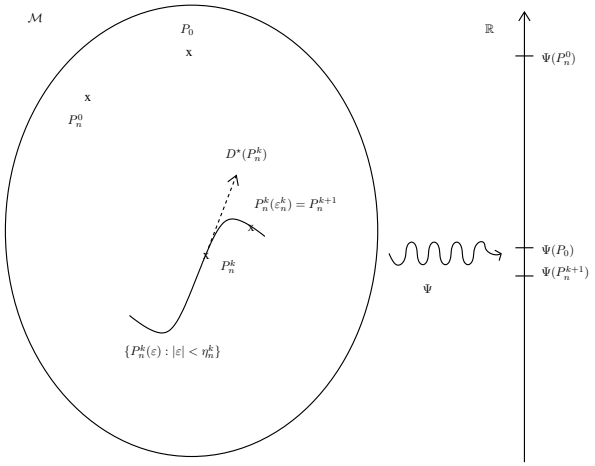

Figure 1: Illustration of the TMLE procedure (with its general one-step updating procedure). We purposedly represent the initial estimator P_n^0 closer to P_0 than its kth and $(k+1)$th updates P_n^k and P_n^{k+1}, heuristically because P_n^0 is as close to P_0 as one can possibly get (given P_n and the specifics of the super-learning procedure) when targeting P_0 itself. However, this obviously does not necessarily imply that $\Psi(P_n^0)$ performs well when targeting $\Psi(P_0)$ (instead of P_0), which is why we also purposedly represent $\Psi(P_n^{k+1})$ closer to $\Psi(P_0)$ than $\Psi(P_n^0)$. Indeed, P_n^{k+1} is obtained by fluctuating its predecessor P_n^k "in the direction of Ψ", *i.e.*, taking into account the fact that we are ultimately interested in estimating $\Psi(P_0)$. More specifically, the fluctuation $\{P_n^k(\varepsilon) : |\varepsilon| < \eta_n^k\}$ of P_n^k is a one-dimensional parametric model (hence its curvy shape in the large model $\mathcal{M}$) such that *(i)* $P_n^k(0) = P_n^k$, and *(b)* its score at $\varepsilon = 0$ equals the efficient influence curve $D^\star(P_n^k)$ at P_n^k (hence the dotted arrow). An optimal stretch ε_n^k is determined (*e.g.* by maximizing the likelihood on the fluctuation), yielding the update $P_n^{k+1} = P_n^k(\varepsilon_n^k)$.

where $D^\star(P_n^{k-1})$ is the current estimator of the efficient influence curve at P_0 obtained as the efficient influence curve at P_n^{k-1}. The path is a one-dimensional parametric model that fluctuates P_n^{k-1} (*i.e.*, $P_n^{k-1}(0) = P_n^{k-1}$) in the direction of $D^\star(P_n^{k-1})$ (*i.e.*, the score of the path at $\varepsilon = 0$ equals $D^\star(P_n^{k-1})$). Here, we choose minus the log-likelihood function as loss function (*i.e.*, we choose $L : \mathcal{M} \times \mathcal{O} \to \mathbb{R}$ characterized by $L(P)(O) = -\log P(O)$). Consequently, the optimal update of P_n^{k-1} is indexed by the maximum likelihood estimator (MLE)

$$\begin{aligned} \varepsilon_n^{k-1} &= \underset{|\varepsilon| \leq \rho \|D^\star(P_n^{k-1})\|_\infty^{-1}}{\arg\max} \sum_{i=1}^n \log P_n^{k-1}(\varepsilon)(O^{(i)}) \\ &= \underset{|\varepsilon| \leq \rho \|D^\star(P_n^{k-1})\|_\infty^{-1}}{\arg\max} \sum_{i=1}^n \log \left(1 + \varepsilon D^\star(P_n^{k-1})(O^{(i)})\right). \end{aligned}$$

The MLE ε_n^{k-1} is uniquely defined (and possibly equal to $\pm\rho\|D^\star(P_n^{k-1})\|_\infty^{-1}$, hence the introduction of the constant ρ in the definition of the path) provided for instance that

$$\max_{i \leq n} |D^\star(P_n^{k-1})(O^{(i)})| > 0$$

(this statement is to be understood *conditionally on* P_n, *i.e.* it is a statement about the sample). Under mild assumptions on P_0, ε_n^{k-1} targets ε_0^{k-1} such that $P_n^{k-1}(\varepsilon_0^{k-1})$ is the Kullback-Leibler projection of P_0 onto the path $\{P_n^{k-1}(\varepsilon) : |\varepsilon| \leq \rho\|D^\star(P_n^{k-1})\|_\infty^{-1}\}$. We now set $P_n^k = P_n^{k-1}(\varepsilon_n^{k-1})$, thus concluding the description of the iterative updating step of the TMLE procedure. Finally, the TMLE ψ_n^* is defined as $\psi_n^* = \lim_{k\to\infty} \psi_n^k$, assuming that the limit exists, or more generally as $\psi_n^{k_n}$ for a conveniently chosen sequence $\{k_n\}_{n\geq 0}$ (see Sections 4.1 and 4.2 regarding this issue).

This is a very general way of dealing with the updating step of the TMLE methodology. The key is that it is possible to determine how the fundamental features of $P_n^k(\varepsilon)$ (*i.e.*, the components of $P_n^k(\varepsilon)$ involved in the definition of $D^\star(P_n^k(\varepsilon))$ and in the definition of Ψ) behave (exactly) as functions of ε relative to their counterparts at $\varepsilon = 0$ (*i.e.*, with respect to (wrt) P_n^k), as shown in the next Lemma (its proof is relegated to Section A.2).

Lemma 1. *Set $s \in L_0^2(P)$ with $\|s\|_\infty < \infty$ and consider the path $\{P_\varepsilon : |\varepsilon| < \|s\|_\infty^{-1}\} \subset \mathcal{M}$ characterized by*

$$\frac{dP_\varepsilon}{dP}(O) = (1 + \varepsilon s(O)). \tag{9}$$

The path has score function s. For all $|\varepsilon| < \|s\|_\infty^{-1}$ and all measurable function f of W,

$$\theta(P_\varepsilon)(X,W) = \frac{\theta(P)(X,W) + \varepsilon E_P(Y s(O)|X,W)}{1 + \varepsilon E_P(s(O)|X,W)}, \tag{10}$$

$$\mu(P_\varepsilon)(W) = \frac{\mu(P)(W) + \varepsilon E_P(X s(O)|W)}{1 + \varepsilon E_P(s(O)|W)}, \tag{11}$$

$$g(P_\varepsilon)(0|W) = \frac{g(P)(0|W) + \varepsilon E_P(\mathbf{1}\{X = 0\} s(O)|W)}{1 + \varepsilon E_P(s(O)|W)}, \tag{12}$$

$$\sigma^2(P_\varepsilon) = \sigma^2(P) + \varepsilon E_P\{X^2 s(O)\}, \quad (13)$$

$$E_{P_\varepsilon}\{f(W)\} = E_P\{f(W)(1 + \varepsilon E_P(s(O)|W))\}. \quad (14)$$

Regarding the computation of $\Psi(P_n^k)$, it is also required to know how to sample independent copies of (W, X) under $P_n^k(\varepsilon)$, see Section 3.1. Finally, we emphasize that by (14), the marginal distribution of W under P_n^k typically deviates from its counterpart under P_n^0 (*i.e.*, from its *empirical* counterpart).

TMLE and one-step estimation methodologies.

By being based on an iterative scheme, the TMLE methodology naturally evokes the *one-step* estimation methodology introduced by Le Cam [8] (see [25, Sections 5.7 and 25.8] for a recent account). The latter estimation methodology draws its inspiration from the method of Newton-Raphson in numerical analysis, and basically consists in updating an initial estimator by relying on a linear approximation to the original estimating equation.

Yet, some differences between the TMLE and one-step estimation methodologies are particularly striking. Most importantly, because the TMLE methodology only involves substitution estimators, how one updates (in the parameter space $\mathbb{R}$) the initial estimator $\psi_n^0 = \Psi(P_n^0)$ of $\Psi(P_0)$ into $\psi_n^1 = \Psi(P_n^1)$ is the consequence of how one updates (in model $\mathcal{M}$) the initial estimator P_n^0 of P_0 into P_n^1. In contrast, the one-step estimator is naturally presented as an update (in the parameter space $\mathbb{R}$) of the initial estimator, for the sake of solving a linear approximation (in $\Psi(P)$) to the estimating equation $P_n D^\star(P) = 0$. The TMLE methodology does not involve such a linear approximation; it nevertheless guarantees by construction $P_n D^\star(P_n^k) \approx 0$ for large k (see Section 4.1 on that issue). Furthermore, on a more technical note, the asymptotic study of the TMLE ψ_n^* does not require that the initial estimator $\psi_n^0 = \Psi(P_n^0)$ be $\sqrt{n}$-consistent (*i.e.*, that $\sqrt{n}(\psi_n^0 - \Psi(P_0))$ be uniformly tight), whereas that of the one-step estimator typically does.

However, there certainly exist interesting relationships between the TMLE and one-step estimation methodologies too. Such relationships are not obvious, and we will investigate them in future work.

4 Convergence and asymptotics

In this section, we state and comment on important theoretical properties enjoyed by the TMLE. In Section 4.1, we study the convergence of the iterative updating procedure which is at the core of the TMLE procedure. In Section 4.2, we derive the consistency and asymptotic normality of the TMLE. By building on the statement of consistency, we also argue why it is more interesting to estimate our non-parametric variable importance measure $\Psi(P_0)$ than its semi-parametric counterpart.

4.1 On the convergence of the updating procedure

Studying the convergence of the updating procedure has several aspects to it. We focus on the general one-step procedure of Section 3.2. All proofs are relegated to Section A.4.

On one hand, the following result (very similar to Result 1 in [23]) trivially holds:

Lemma 2. *Assume (i) that all the paths we consider are included in* $\mathcal{M}' \subset \mathcal{M}$ *such that* $\sup_{P \in \mathcal{M}'} \|D^\star(P)\|_\infty = M < \infty$*, and (ii) that their fluctuation parameters* ε *are restricted to* $[-\rho, \rho]$ *for* $\rho = (2M)^{-1}$*. If* $\lim_{k\to\infty} \varepsilon_n^k = 0$ *then* $\lim_{n\to\infty} P_n D^\star(P_n^k) = 0$*.*

Condition *(i)* is weak, and we refer to Lemma 4 for a set of conditions which guarantee that it holds. Lemma 2 is of primary importance. It teaches us that if the TMLE procedure "converges" (in the sense that $\lim_{k\to\infty} \varepsilon_n^k = 0$) then its "limit" is a solution of the efficient influence curve equation (in the sense that for any arbitrary small deviation from 0, it is possible to guarantee $P_n D^\star(P_n^k) \approx 0$ by choosing k large enough). This is the key to the proofs of consistency and asymptotic linearity, see Section 4.2. Actually, the condition $\lim_{k\to\infty} \varepsilon_n^k = 0$ can be replaced by a more explicit condition on the class of the considered data-generating distributions, as shown in the next lemma.

Lemma 3. *Under the assumptions of Lemma 2, let us suppose additionally that the sample satisfies (iii)* $\inf_{k\geq 0} P_n D^\star(P_n^k)^2 > 0$*, and (iv) that the* log*-likelihood of the data is uniformly bounded on* $\mathcal{M}'$*:* $\sup_{P\in\mathcal{M}'} \sum_{i=1}^n \log P(O^{(i)}) < \infty$*. Then it holds that* $\lim_{k\to\infty} \varepsilon_n^k = 0$ *and* $\lim_{n\to\infty} P_n D^\star(P_n^k) = 0$*.*

On the other hand, it is possible to obtain another result pertaining to the "convergence" of the updating procedure directly put in terms of the convergence of the sequences $\{P_n^k\}_{k\geq 0}$ and $\{\psi_n^k\}_{k\geq 0}$, provided that $\{\varepsilon_n^k\}_{k\geq 0}$ goes to 0 quickly enough. Specifically,

Lemma 4. *Suppose that* $P_n^0(\|O\| \leq C) = 1$ *for some finite* $C > 0$*. Then obviously* $P_n^k(\|O\| \leq C) = P_n^k(|\theta(P_n^k)(X, W)| \leq C) = P_n^k(|\mu(P_n^k)(W)| \leq C) = 1$ *for all* $k \geq 0$*. Suppose moreover that for all* $k \geq 0$*,* $g(P_n^k)(0|W) \geq c > 0$ *and* $\sigma^2(P_n^k) \geq c$ *are bounded away from 0. Then condition (i) of Lemma 2 holds. Assume now that* $\sum_{k\geq 0} |\varepsilon_n^k| < \infty$*. Then the sequence* $\{P_n^k\}_{k\geq 0}$ *converges in total variation (hence in law) to a data-generating distribution* P_n^**. Simultaneously, the sequence* $\{\psi_n^k\}_{k\geq 0}$ *converges to* $\Psi(P_n^*)$*.*

It is necessary to bound $g(P_n^k)$ and $\sigma^2(P_n^k)$ away from 0 because conditions *(i)* and *(ii)* of Lemma 2 only imply that $g(P_n^k)(0|W) \geq g(P_n^0)(0|W)((1-\rho)/(1+\rho))^k$ and $\sigma^2(P_n^k) \geq \sigma^2(P_n^0)(1-\rho)^k$. Now, it makes perfect sense from a computational point of view to resort to lower-thresholding in order to ensure that $g(P_n^k)(0|W)$ and $\sigma^2(P_n^k)$ cannot be smaller than a fixed constant. Assuming that the series $\sum_{k\geq 0} |\varepsilon_n^k|$ converges ensures that $\{P_n^k\}_{k\geq 0}$ converges in total variation rather than weakly only. Interestingly, we do draw advantage from this stronger type of convergence in order to derive the second part of the lemma. In conclusion, note that Newton-Raphson-type algorithms converge at a k^{-2}-rate, which suggests that the condition $\sum_{k\geq 0} |\varepsilon_n^k| < \infty$ is not too demanding.

4.2 Consistency and asymptotic normality

Let us now investigate the statistical properties of the TMLE ψ_n^*. We actually consider a slightly modified version of the TMLE in order to circumvent the issue of the convergence of the sequence $\{\psi_n^k\}_{k\geq 0}$ as k goes to infinity. The modified version is perfectly fine from a practical point of view. All proofs are relegated to Section A.5.

Consistency.

Under mild assumptions, the TMLE is consistent. Specifically:

Proposition 2 (consistency)**.** *We assume (i) that there exist finite $C > c > 0$ such that $\|\theta(P_n^{k_n})\|_\infty \leq C$, $g(P_n^{k_n})(0|W) \geq c$ and $\sigma^2(P_n^{k_n}) \geq c$ for all $n \geq 1$, (ii) that $\theta(P_n^{k_n})$, $\mu(P_n^{k_n})$, $g(P_n^{k_n})$ and $\sigma^2(P_n^{k_n})$ respectively converge to θ_0 such that $\|\theta_0\|_\infty \leq C$, μ_0, g_0 and $\sigma_0^2 \geq c$ in such a way that $P_0(\theta(P_n^{k_n}) - \theta_0)^2 = o_P(1)$, $P_0(\theta(P_n^{k_n})(0,\cdot) - \theta_0(0,\cdot))^2 = o_P(1)$, $P_0(\mu(P_n^{k_n}) - \mu_0)^2 = o_P(1)$, $P_0(g(P_n^{k_n})(0|\cdot) - g_0(0|\cdot))^2 = o_P(1)$ and $\sigma^2(P_n^{k_n}) = \sigma_0^2 + o_P(1)$, and (iii) that $D_1^\star(P_n^{k_n})$ and $D_2^\star(P_n^{k_n})$ belong to a P_0-Donsker class with P_0-probability tending to 1. In addition, we suppose that all assumptions of Lemma 3 are met, and that the (possibly random) integer $k_n \geq 0$ is chosen so that $P_n D^\star(P_n^{k_n}) = o_P(1/\sqrt{n})$.*

Define $\widetilde{\psi}_n^* = \psi_n^{k_n} = \Psi(P_n^{k_n})$. If the limits satisfy either $\theta_0(0,\cdot) = \theta(P_0)(0,\cdot)$ or ($\mu_0 = \mu(P_0)$ and $g_0 = g(P_0)$) then $\widetilde{\psi}_n^*$ consistently estimates $\Psi(P_0)$.

It is remarkable that the consistency of the TMLE $\widetilde{\psi}_n^* = \Psi(P_n^{k_n})$ is granted essentially whenever the estimators $\theta(P_n^{k_n})$, $\mu(P_n^{k_n})$, $g(P_n^{k_n})$, $\sigma^2(P_n^{k_n})$ converge and that *one only* of the limits $\theta_0(0,\cdot)$ of $\theta(P_n^{k_n})(0,\cdot)$ and (μ_0, g_0) of $(\mu(P_n^{k_n}), g(P_n^{k_n}))$ coincides with the corresponding truth $\theta(P_0)(0,\cdot)$ or $(\mu(P_0), g(P_0))$. This property is mostly inherited from the double-robustness of the efficient influence curve $D^\star$ of parameter Ψ (*i.e.*, $PD^\star(P') = 0$ implies $\Psi(P') = \Psi(P)$) and from the fact that the TMLE solves the efficient influence curve equation (*i.e.*, $P_n D^\star(P_n^{k_n}) \approx 0$).

Merit of the non-parametric variable importance measure over its semi-parametric counterpart.

Let us repeat that we do not assume a semi-parametric model $Y = \beta X + \eta(W) + U$ (with unspecified η and U such that $E_P(U|X,W) = 0$). However, if $P \in \mathcal{M}$ is such that $\theta(P)(X,W) = \beta(P)X + \theta(P)(0,W)$ (*i.e.*, if the semi-parametric model holds under P) then $\Psi(P) = \beta(P)$. Let us denote by $\mathcal{M}_{\mathrm{SP}} \subset \mathcal{M}$ the set of all such data-generating distributions. It is known (see for instance [28]) that $\beta : \mathcal{M}_{\mathrm{SP}} \to \mathbb{R}$ is a pathwise differentiable parameter (wrt the corresponding maximal tangent space), and that its efficient influence curve at $P \in \mathcal{M}_{\mathrm{SP}}$ is given by

$$D_{\mathrm{SP}}^\star(P)(O) = \frac{Y - \beta(P)X - \theta(P)(0,W)}{v^2(P)(X,W)} \left(X - \frac{E_P\left(\frac{X}{v^2(P)(X,W)}\middle|W\right)}{E_P\left(\frac{1}{v^2(P)(X,W)}\middle|W\right)} \right),$$

with $v^2(P)(X,W) = E_P((Y-\theta(P)(X,W))^2|X,W)$ is the conditional variance of Y given (X,W) under P. Note that the second factor in the right-hand side expression reduces to $(X-\mu(P)(W))$ whenever $v^2(P)(X,W)$ only depends on W.

For the purpose of emphasizing the merit of the non-parametric variable importance measure over its semi-parametric counterpart, say that one estimates $\beta(P_0)$ assuming (temporarily) that $P_0 \in \mathcal{M}_{\mathrm{SP}}$ (hence $\Psi(P_0) = \beta(P_0)$). Say that one builds $P^*_{n,\mathrm{SP}} \in \mathcal{M}_{\mathrm{SP}}$ such that *(i)* $v^2(P^*_{n,\mathrm{SP}})(X,W)$ does not depend on (X,W), and *(ii)* $P_n D^\star_{\mathrm{SP}}(P^*_{n,\mathrm{SP}}) = 0$. Let us assume that $\beta(P^*_{n,\mathrm{SP}})$, $v^2(P^*_{n,\mathrm{SP}})$, $\mu(P^*_{n,\mathrm{SP}})$ and $\theta(P^*_{n,\mathrm{SP}})$ respectively converge to β_1, $v_1^2 > 0$, μ_1 and θ_1 (such that $\theta_1(X,W) = \beta_1 X + \theta_1(0,W)$), and finally that one solves in the limit the efficient influence curve equation:

$$E_{P_0}\{(Y-\beta_1 X-\theta_1(0,W))(X-\mu_1(W))\} = 0 \tag{15}$$

(this is typically derived from *(ii)* above; see the proof of Proposition 2 for a typical derivation). Then (by double-robustness of $D^\star_{\mathrm{SP}}$), the estimator $\beta(P^*_{n,\mathrm{SP}})$ of $\beta(P_0)$ is consistent (*i.e.*, $\beta_1 = \beta(P_0)$) if either $\theta_1 = \theta(P_0)$ (that is obvious) or $\mu_1 = \mu(P_0)$. For example, let us suppose that $\mu_1 = \mu(P_0)$. In particular, one can deduce from equalities $E_{P_0}\{X(X-\mu(P_0)(W))\} = E_{P_0}\{(X-\mu(P_0)(W))^2\}$ and (15) that

$$\beta_1 = \frac{E_{P_0}\{(\theta(P_0)(X,W)-\theta_1(0,W))(X-\mu(P_0)(W))\}}{E_{P_0}\{(X-\mu(P_0)(W))^2\}}$$

(provided that X does not coincide with $\mu(P_0)(W)$ under P_0). Equivalently, $\beta_1 = b(P_0)$ for the functional $b : \mathcal{M}' = \mathcal{M} \setminus \{P \in \mathcal{M} : X = \mu(P)(W)\} \to \mathbb{R}$ such that, for every $P \in \mathcal{M}'$,

$$b(P) = \underset{\beta\in\mathbb{R}}{\arg\min}\, E_P\left\{[\theta(P)(X,W)-\theta_1(0,W)-\beta(X-\mu(P)(W))]^2\right\}.$$

Note that one can interpret parameter b as a non-parametric extension of the semi-parametric parameter β (*non-parametric*, because its definition does not involve a semi-parametric model anymore). Now, we want to emphasize that b arguably defines a sensible target if $\theta_1(0,\cdot) = \theta(P)(0,\cdot)$ (in addition to $\mu_1 = \mu(P_0)$), but not otherwise! This illustrates the danger of relying on a semi-parametric model when it is not absolutely certain that it holds, thus underlying the merit of targeting the non-parametric variable importance measure rather than its semi-parametric counterpart.

Asymptotic normality.

In addition to being consistent under mild assumptions, the TMLE is also asymptotically linear, and thus satisfies a central limit theorem. Let us start with a partial result:

Proposition 3. *Suppose that the assumptions of Proposition 2 are met. If* $\sigma^2(P_n^{k_n}) = \sigma_0^2 + O_P(1/\sqrt{n})$ *then it holds that*

$$\widetilde{\psi}_n^* - \Psi(P_0) = (P_n - P_0)D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0)) \\ + P_0 D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0))(1 + o_P(1)) + o_P(1/\sqrt{n}). \quad (16)$$

Expansion (16) sheds some light on the first order properties of the TMLE $\widetilde{\psi}_n^*$. It notably makes clear that the convergence of $\widetilde{\psi}_n^*$ is affected by how fast the estimators $\theta(P_n^{k_n})$, $\mu(P_n^{k_n})$ and $g(P_n^{k_n})$ converge to their limits (see second term). If the rates of convergence are collectively so slow that they only guarantee $P_0 D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0)) = O_P(1/n^r)$ for some $r \in [0, 1/2[$, then expansion (16) becomes

$$\widetilde{\psi}_n^* - \Psi(P_0) = P_0 D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0)) + o_P(1/n^r)$$

and asymptotic linearity fails to hold. On the contrary, we easily deduce from Proposition 3 what happens when $\theta_0(0, \cdot) = \theta(P_0)(0, \cdot)$, $\mu_0 = \mu(P_0)$, $g_0 = g(P_0)$, with fast rates of convergence:

Corollary 1 (asymptotic normality)**.** *Suppose that the assumptions of Proposition 3 are met. If in addition it holds that* $\theta_0(0, \cdot) = \theta(P_0)(0, \cdot)$, $\mu_0 = \mu(P_0)$, $g_0 = g(P_0)$ *and*

$$P_0(\theta(P_n^{k_n})(0, \cdot) - \theta_0(0, \cdot))^2 \times \Big(P_0(\mu(P_n^{k_n}) - \mu_0)^2 + P_0(g(P_n^{k_n})(0|\cdot) - g_0(0|\cdot))^2\Big) = o_P(1/n)$$

then

$$\widetilde{\psi}_n^* - \Psi(P_0) = (P_n - P_0)D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0)) + o_P(1/\sqrt{n})$$

i.e.*, the TMLE* $\widetilde{\psi}_n^*$ *is asymptotically linear with influence function* $D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0))$*. Thus,* $\sqrt{n}(\widetilde{\psi}_n^* - \Psi(P_0))$ *is asymptotically distributed from a centered Gaussian law with variance* $P_0 D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0))^2$*. In particular, if* $\theta_0 = \theta(P_0)$ *then the TMLE* $\widetilde{\psi}_n^*$ *is efficient.*

Corollary 1 covers a simple case in the sense that, by being $o_P(1/\sqrt{n})$, the second right-hand side term in (16) does not significantly contribute to the linear asymptotic expansion *i.e.*, the influence curve actually is $D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0))$. Depending on how $\theta(P_n^0)$, $\mu(P_n^0)$ and $g(P_n^0)$ are obtained (again, we recommend relying on super-learning), the contribution to the linear asymptotic expansion may be significant (but determining this contribution would be a very difficult task to address on a case by case basis when relying on super-learning).

5 Specifics of the TMLE procedure tailored to the estimation of the non-parametric variable importance measure

In this section, we present practical details on how we conduct the initialization and updating steps of the TMLE procedure as described in Section 3. We introduce in Section 5.1 a working model for the conditional distribution of X given $(W, X \neq 0)$ which proves very efficient in computational terms. In Section 5.2, we introduce two alternative two-step updating

procedures which can be substituted to the general one-step updating procedure presented in Section 3.2. Finally, we describe carefully what are all the features of interest of P_0 that must be considered for the purpose of targeting the parameter of ultimate interest, $\Psi(P_0)$, *via* the construction of the TMLE.

5.1 Working model for the conditional distribution of X given $(W, X \neq 0)$

The working model for the conditional distribution of X given $(W, X \neq 0)$ under P_n^0 that we build relies on two ideas:

- we link the conditional second order moment $E_{P_n^0}(X^2|X \neq 0, W)$ to the conditional mean $E_{P_n^0}(X|X \neq 0, W)$ (both under P_n^0) through the equality

$$E_{P_n^0}(X^2|X \neq 0, W) = \varphi_{n,\lambda}\left(E_{P_n^0}(X|X \neq 0, W)\right) \tag{17}$$

  where $\varphi_{n,\lambda}(t) = \lambda t^2 + (1-\lambda)(t(m_n + M_n) - m_n M_n)$ (with $m_n = \min_{i \leq n} X^{(i)}$, $M_n = \max_{i \leq n} X^{(i)}$), and $\lambda \in [0, 1]$ is a fine-tune parameter;

- under P_n^0 and conditionally on $(W, X \neq 0)$, X takes its values in the set $\{X^{(i)} : i \leq n\} \setminus \{0\}$ of the observed X's different from 0.

Since the conditional distribution of X given $(W, X \neq 0)$ under P_n^0 is subject to two constraints, X cannot take fewer than three different values in general. Elegantly, it is possible (under a natural assumption on P_n^0) to fine-tune λ and to select three values in $\{X^{(i)} : i \leq n\} \setminus \{0\}$ in such a way that X only takes the latter values:

Lemma 5. *Assume that P_n^0 guarantees that $\sigma^2(P_n^0) > 0$, $P_n^0(X \neq 0) > 0$, $g(P_n^0)(0|W) \in (0, 1)$ P_n^0-almost surely, and $X \in [m_n + c, M_n - c]$ for some $c > 0$ when $X \neq 0$. It is possible to construct $P_n^{00} \in \mathcal{M}$ in such a way that (i) W has the same marginal distribution under P_n^{00} and P_n^0, $\mu(P_n^{00}) = \mu(P_n^0)$, $g(P_n^{00}) = g(P_n^0)$, $\sigma^2(P_n^{00}) = \sigma^2(P_n^0)$, and (ii) for all $W \in \mathcal{W}$, there exist three different values $x^{(1)}, x^{(2)}, x^{(3)} \in \{X^{(i)} : i \leq n\} \setminus \{0\}$ and three non-negative weights p_1, p_2, p_3 summing up to 1 such that, conditionally on $(W, X \neq 0)$ under P_n^{00}, $X = x^{(k)}$ with conditional probability p_k.*

Hence, we directly construct a P_n^0 of the same form as P_n^{00}. Note that, by (8), because the conditional distribution of X given $(W, X \neq 0)$ under P_n^0 has its support included in $\{X^{(i)} : i \leq n\} \setminus \{0\}$, then so do the conditional distributions of X given $(W, X \neq 0)$ under P_n^k (all $k \geq 1$) obtained by following the general one-step updating procedure of Section 3.2. Similarly, because we initially estimate the marginal distribution of W under P_0 by its empirical counterpart, then the marginal distributions of W under P_n^0 and P_n^k (all $k \geq 1$) have their supports included in $\{W_i : i \leq n\}$.

We discuss in Section 5.4 why it is computationally more interesting to consider such a working model (instead of a Gaussian working model for instance). We emphasize that

assuming $X \in [m_n + c, M_n - c]$ when $X \neq 0$ (for a possibly tiny $c > 0$) is hardly a constraint, and that the latter must be accounted for while estimating $\mu(P_0)$, $g(P_0)$, and $\sigma^2(P_0)$. The proof of the lemma is relegated to Section A.2.

5.2 Two tailored alternative two-step updating procedures

We present in Section 3.2 a general one-step updating procedure. Alternatively, it is also possible to decompose each update into a first update of the conditional distribution of Y given (W, X), followed by a second update of the marginal distribution of (W, X).

First update: fluctuating the conditional distribution of Y given (W, X).

We actually propose two different fluctuations for that purpose: a *Gaussian fluctuation* on one hand and a *logistic fluctuation* on the other hand, depending on what one knows or wants to impose.

Gaussian fluctuation. In this case too, minus the log-likelihood function is used as a loss function. Specifically, we first fluctuate only the conditional distribution of Y given (W, X), by introducing the path $\{P_{n,1}^{k-1}(\varepsilon) : \varepsilon \in \mathbb{R}\}$ such that *(i)* (W, X) has the same distribution under $P_{n,1}^{k-1}(\varepsilon)$ as under P_n^{k-1}, and *(ii)* under $P_{n,1}^{k-1}(\varepsilon)$ and given (W, X), Y is distributed from the Gaussian law with conditional mean $\theta(P_n^{k-1})(X, W) + \varepsilon H(P_n^{k-1})(X, W)$ and conditional variance 1, where the so-called clever covariate $H(P)$ is characterized for any $P \in \mathcal{M}$ by

$$H(P)(X, W) = \frac{1}{\sigma^2(P)} \left(X - \frac{\mu(P)(W)\mathbf{1}\{X = 0\}}{g(P)(0|W)} \right).$$

This definition guarantees that the path fluctuates P_n^{k-1} (*i.e.*, $P_{n,1}^{k-1}(0) = P_n^{k-1}$, provided that Y is conditionally Gaussian given (W, X) under P_n^0) in the direction of $D_2^\star(P_n^{k-1})$ (*i.e.*, the score of the path at $\varepsilon = 0$ equals $D_2^\star(P_n^{k-1})$). Introducing the MLE

$$\begin{aligned} \varepsilon_{n,1}^{k-1} &= \arg\max_{\varepsilon \in \mathbb{R}} \sum_{i=1}^n \log P_{n,1}^{k-1}(\varepsilon)(O^{(i)}) \\ &= \frac{\sum_{i=1}^n (Y^{(i)} - \theta(P_n^{k-1})(X^{(i)}, W^{(i)})) H(P_n^{k-1})(X^{(i)}, W^{(i)})}{\sum_{i=1}^n H(P_n^{k-1})(X^{(i)}, W^{(i)})^2}, \end{aligned}$$

the first intermediate update bends P_n^{k-1} into $P_{n,2}^{k-1} = P_{n,1}^{k-1}(\varepsilon_{n,1}^{k-1})$.

Logistic fluctuation. There is yet another interesting option in the case that $Y \in [a, b]$ is bounded (or in the case that one wishes to impose $Y \in [a, b]$, typically then with $a = \min_{i \leq n} Y^{(i)}$ and $b = \max_{i \leq n} Y^{(i)}$), which allows to incorporate this known fact (or wish) into the procedure. Let us assume that $\theta(P_0)$ takes its values in $]a, b[$ and also that $\theta(P_n^{k-1})$ is constrained in such a way that $\theta(P_n^{k-1})(X, W) \in]a, b[$. Introduce

for clarity the function on the real line characterized by $F_{a,b}(t) = (t-a)/(b-a)$. Here, we choose the loss function characterized by $-L_{a,b}(P)(O) = F_{a,b}(Y)\log F_{a,b} \circ \theta(P)(X,W) + (1-F_{a,b}(Y))\log(1-F_{a,b}\circ\theta(P)(X,W))$, with convention $L_{a,b}(P)(O) = +\infty$ if $\theta(P)(X,W) \in \{a,b\}$. Note that the loss $L_{a,b}(P)$ depends on the conditional distribution of Y given (W,X) under P only through its conditional mean $\theta(P)$. This straightforwardly implies that in order to describe a fluctuation $\{P_{n,1}^{k-1}(\varepsilon) : \varepsilon \in \mathbb{R}\}$ of P_n^{k-1}, it is only necessary to detail the form of the marginal distribution of (W,X) under $P_{n,1}^{k-1}(\varepsilon)$ and how $\theta(P_{n,1}^{k-1}(\varepsilon))$ depends on $\theta(P_n^{k-1})$ and ε. Specifically, we first fluctuate only the conditional distribution of Y given (W,X), by making $P_{n,1}^{k-1}(\varepsilon)$ be such that *(i)* (W,X) has the same distribution under $P_{n,1}^{k-1}(\varepsilon)$ as under P_n^{k-1}, and *(ii)*

$$\theta(P_{n,1}^{k-1}(\varepsilon))(X,W) = F_{a,b}^{-1}\left(\text{expit}\left(\text{logit}F_{a,b}\circ\theta(P_n^{k-1})(X,W) + \varepsilon H(P_n^{k-1})(X,W)\right)\right).$$

Now, introduce the $L_{a,b}$-minimum loss estimator

$$\varepsilon_{n,1}^{k-1} = \arg\min_{\varepsilon\in\mathbb{R}} \sum_{i=1}^{n} L_{a,b}(P_{n,1}^{k-1}(\varepsilon))(O^{(i)}),$$

which finally yields the first intermediate update $P_{n,2}^{k-1} = P_{n,1}^{k-1}(\varepsilon_{n,1}^{k-1})$. The following lemma (whose proof is relegated to Section A.2) justifies our interest in the loss function $L_{a,b}$ and fluctuation $\{P_{n,1}^{k-1}(\varepsilon) : \varepsilon \in \mathbb{R}\}$:

Lemma 6. *Assume that the conditions stated above are met. Then $L_{a,b}$ is a valid loss function for the purpose of estimating $\theta(P_0)$ in the sense that*

$$\theta(P_0) = \arg\min_{P\in\mathcal{M}} P_0 L_{a,b}(P).$$

Moreover, it holds that

$$\tfrac{\partial}{\partial\varepsilon} L_{a,b}(P_{n,1}^{k-1}(\varepsilon))\Big|_{\varepsilon=0}(O) = -D_2^\star(P_n^{k-1})(O).$$

The second inequality is the counterpart of the fact that, when using the Gaussian fluctuation, the score of the path at $\varepsilon = 0$ equals $D_2^\star(P_n^{k-1})$.

Second update: fluctuating the marginal distribution of (W,X).

Next, we preserve the conditional distribution of Y given (W,X) and only fluctuate the marginal distribution of (W,X), by introducing the path $\{P_{n,2}^{k-1}(\varepsilon) : |\varepsilon| \le \rho\|D_1^\star(P_{n,2}^{k-1})\|_\infty^{-1}\}$ such that *(i)* Y has the same conditional distribution given (W,X) under $P_{n,2}^{k-1}(\varepsilon)$ as under $P_{n,2}^{k-1}$, and *(ii)* the marginal distribution of (W,X) under $P_{n,2}^{k-1}(\varepsilon)$ is characterized by

$$\frac{dP_{n,2}^{k-1}(\varepsilon)}{dP_{n,2}^{k-1}}(X,W) = \left(1 + \varepsilon D_1^\star(P_{n,2}^{k-1})(X,W)\right). \tag{18}$$

This second path fluctuates $P_{n,2}^{k-1}$ (*i.e.*, $P_{n,2}^{k-1}(0) = P_{n,2}^{k-1}$) in the direction of $D_1^\star(P_{n,2}^{k-1})$ (*i.e.*, the score of the path at $\varepsilon = 0$ equals $D_1^\star(P_{n,2}^{k-1})$). Consider again minus the log-likelihood as loss function, and introduce the MLE

$$\varepsilon_{n,2}^{k-1} \quad = \quad \underset{|\varepsilon| \leq \rho \|D_1^\star(P_{n,2}^{k-1})\|_\infty^{-1}}{\arg\max} \sum_{i=1}^{n} \log P_{n,2}^{k-1}(\varepsilon)(O^{(i)}) :$$

the second update bends $P_{n,2}^{k-1}$ into $P_n^k = P_{n,2}^{k-1}(\varepsilon_{n,2}^{k-1})$, concluding the description of how we can alternatively build P_n^k based on P_n^{k-1}.

Note that, by (18), because the conditional distribution of X given $(W, X \neq 0)$ under P_n^0 has its support included in $\{X^{(i)} : i \leq n\} \setminus \{0\}$ (a consequence of our choice of working model, see Section 5.1), then so do the conditional distributions of X given $(W, X \neq 0)$ under P_n^k (all $k \geq 1$) obtained by following either one of the tailored two-step updating procedure. Furthermore, it still holds that the marginal distributions of W under P_n^0 and P_n^k (all $k \geq 1$) have their supports included in $\{W_i : i \leq n\}$ (because we initially estimate the marginal distribution of W under P_0 by its empirical counterpart).

5.3 Super-learning of the features of interest

It still remains to specify how we wish to carry out the initial estimation and updating of the features of interest $\theta(P_0)$, $\mu(P_0)$, $g(P_0)$, and $\sigma^2(P_0)$. As for $\sigma^2(P_0) = E_{P_0}\{X^2\}$, we simply estimate it by its empirical counterpart *i.e.*, construct P_n^0 in such a way that $\sigma^2(P_n^0) = n^{-1}\sum_{i=1}^n (X^{(i)})^2$. The three other features $\theta(P_0)$, $\mu(P_0)$ and $g(P_0)$ are estimated by *super-learning*, and P_n^0 is constructed in such a way that $\theta(P_n^0)$, $\mu(P_n^0)$ and $g(P_n^0)$ equal their corresponding estimators. Super-learning is a cross-validation based aggregation method that builds a predictor as a convex combination of base predictors [24, 22] (we briefly describe in Section 6.5 the specifics of the super-learning procedure that we implement for our application to simulated and real data). The weights of the convex combination are chosen so as to minimize the prediction error, which is expressed in terms of the non-negative least squares (NNLS) loss function [7] and estimated by V-fold cross-validation. Heuristically the obtained predictor is by construction at least as good as the best of the base predictors (this statement has a rigorous form implying oracle inequalities, see [24, 22]).

Lemma 1 teaches us what additional features of P_n^{k-1} must be known in order to derive the kth update P_n^k from its predecessor P_n^{k-1}, starting from $k = 1$. Specifically, if we rely on the general one-step updating procedure of Section 3.2 then we need to know:

- $E_{P_n^{k-1}}(YD^\star(P_n^{k-1})(O)|X, W)$ and $E_{P_n^{k-1}}(D^\star(P_n^{k-1})(O)|X, W)$ for the update of $\theta(P_n^{k-1})$ (see (10));

- $E_{P_n^{k-1}}(D^\star(P_n^{k-1})(O)|W)$ for the updates of $\mu(P_n^{k-1})$, $g(P_n^{k-1})$, and the marginal distribution of W under P_n^{k-1} (see the right-hand side denominators in (11), (12), (14));

- $E_{P_n^{k-1}}(XD^\star(P_n^{k-1})(O)|W)$ for the update of $\mu(P_n^{k-1})$ (see the right-hand side numerator in (11));

- $E_{P_n^{k-1}}(\mathbf{1}\{X=0\}D^\star(P_n^{k-1})(O)|W)$ for the update of $g(P_n^{k-1})$ (see the right-hand side numerator in (12));

- $E_{P_n^{k-1}}\{X^2D^\star(P_n^{k-1})(O)\}$ for the update of $\sigma^2(P_n^{k-1})$ (see (13)).

It is noteworthy that if either one of the two-step updating procedures of Section 5.2 is used then the first two conditional expectations do not need to be known, because updating $\theta(P_n^{k-1})$ relies on the clever covariate $H(P_n^{k-1})$, which is entirely characterized by the current estimators $\mu(P_n^{k-1})$, $g(P_n^{k-1})$, and $\sigma^2(P_n^{k-1})$ of the features $\mu(P_0)$, $g(P_0)$, and $\sigma^2(P_0)$, respectively. In the sequel of this sub-section, we focus on the general one-step updating procedure of Section 3.2. How to proceed when relying on either of the two-step updating procedures of Section 5.2 can be easily deduced from that case.

Once $\theta(P_n^0)$, $\mu(P_n^0)$, $g(P_n^0)$, and $\sigma^2(P_n^0)$ are determined (see the first paragraph of this sub-section) hence $D^\star(P_n^0)$ is known, we therefore also estimate by super-learning the conditional expectations $E_{P_0}(YD^\star(P_n^0)(O)|X,W)$, $E_{P_0}(D^\star(P_n^0)(O)|X,W)$, $E_{P_0}(D^\star(P_n^0)(O)|W)$, $E_{P_0}(XD^\star(P_n^0)(O)|W)$, $E_{P_0}(\mathbf{1}\{X=0\}D^\star(P_n^0)(O)|W)$; as for $E_{P_0}\{X^2D^\star(P_n^0)(O)\}$, we simply estimate it by its empirical counterpart. Then we constrain P_n^0 in such a way that the conditional expectations $E_{P_n^0}(YD^\star(P_n^0)(O)|X,W)$, $E_{P_n^0}(D^\star(P_n^0)(O)|X,W)$, $E_{P_n^0}(D^\star(P_n^0)(O)|W)$, $E_{P_n^0}(XD^\star(P_n^0)(O)|W)$, $E_{P_n^0}(\mathbf{1}\{X=0\}D^\star(P_n^0)(O)|W)$, and expectation $E_{P_n^0}\{X^2D^\star(P_n^0)(O)\}$ equal their corresponding estimators. This completes the construction of P_n^0, and suffices for characterizing the features $\theta(P_n^1)$, $\mu(P_n^1)$, $g(P_n^1)$ and $\sigma^2(P_n^1)$ of the first update P_n^1.

Now, if one wished to follow exactly the conceptual road consisting in relying on Lemma 1 in order to derive the second update P_n^2 from its predecessor P_n^1, one would have to describe how each conditional (and unconditional) expectation of the above list behaves, as a function of ε, on the path $\{P_n^1(\varepsilon) : |\varepsilon| \leq \rho\|D^\star(P_n^1)\|_\infty^{-1}\}$. This would in turn enlarge the above list of the features of interest of P_0 that one would have to consider in the initial construction of P_n^0. Note that the length of the list would increase quadratically in the number of updates. Instead, once $D^\star(P_n^{k-1})$ is known, we estimate by super-learning the conditional expectations $E_{P_0}(YD^\star(P_n^{k-1})(O)|X,W)$, $E_{P_0}(D^\star(P_n^{k-1})(O)|X,W)$, $E_{P_0}(D^\star(P_n^{k-1})(O)|W)$, $E_{P_0}(XD^\star(P_n^{k-1})(O)|W)$, $E_{P_0}(\mathbf{1}\{X=0\}D^\star(P_n^{k-1})(O)|W)$; as for $E_{P_0}\{X^2D^\star(P_n^{k-1})(O)\}$, we simply estimate it by its empirical counterpart. Then we proceed *as if* the conditional expectations $E_{P_n^{k-1}}(YD^\star(P_n^{k-1})(O)|X,W)$, $E_{P_n^{k-1}}(D^\star(P_n^{k-1})(O)|X,W)$, $E_{P_n^{k-1}}(D^\star(P_n^{k-1})(O)|W)$, $E_{P_n^{k-1}}(XD^\star(P_n^{k-1})(O)|W)$, $E_{P_n^{k-1}}(\mathbf{1}\{X=0\}D^\star(P_n^{k-1})(O)|W)$, and $E_{P_n^{k-1}}\{X^2D^\star(P_n^{k-1})(O)\}$ were equal to their corresponding estimators. By doing so, the length of the list of the features of interest of P_0 is fixed no matter how many steps of the updating procedure are carried out. Arguably, following this alternative road has little if no effect relative to following exactly the conceptual road consisting in relying on Lemma 1, because only second (or more) order expressions in ε are involved.

5.4 Merit of the working model for the conditional distribution of X given $(W, X \neq 0)$

Let us explain here why *(a)* initially estimating the marginal distribution of W under P_0 by its empirical counterpart and *(b)* relying on the working model for the conditional distribution of X given $(W, X \neq 0)$ that we described in Section 5.1 is computationally very interesting. The key is that, under P_n^0 and its successive updates P_n^k (all $k \geq 1$), the distributions of (W, X) have their supports included in $\{(W^{(i)}, X^{(j)}) : i \leq j \leq n\}$ (we say they are "parsimonious").

Indeed, Lemma 1 and a simple induction yield that, for each $k \geq 1$, a single call to $\theta(P_n^k)$, $\mu(P_n^k)$ or $g(P_n^k)$ involves a number of (nested) calls to the "past" features of interest $\theta(P_n^{k'})$, $\mu(P_n^{k'})$ and $g(P_n^{k'})$ $(0 \leq k' < k)$ which is $O(k)$. Furthermore, the evaluation of $\Psi(P_n^k)$ (following (5) with P_n^k substituted to P_n^0) requires in turn B calls (assuming for simplicity that the functions are not vectorized) to $\theta(P_n^k)$ (in order to evaluate the numerator of the right-hand side term of (5)), $\mu(P_n^k)$ and $g(P_n^k)$ (in order to simulate $\{(\tilde{W}^{(b)}, \tilde{X}^{(b)}) : b \leq B\}$). Overall, at least $O(Bk)$ calls to the set of all features of interest are performed at the kth updating step of the TMLE procedure. In practice (even if functions are vectorized) this leads to a large memory footprint and prohibitive running time of the algorithm, as each of these calls consists in the *prediction* of the corresponding feature, as described in Section 5.3.

By taking advantage of the "parsimony" of the distributions of (W, X) under the successive P_n^k $(k \geq 0)$, we manage to alleviate dramatically the time and memory requirements of our implementation. Indeed, the "parsimony" implies that, at the kth step of the TMLE procedure $(k \geq 0)$, it is only required to compute and store $O(n^2)$ quantities (including, but not limited to, $\theta(P_n^k)(X^{(i)}, W^{(j)})$, $\mu(P_n^k)(W^{(i)})$ and $g(P_n^k)(W^{(j)})$ for all $1 \leq i, j \leq n$) — see Section 5.3). In particular, the evaluation of $\Psi(P_n^k)$ now requires retrieving $O(B)$ values from a handful of vectors instead of performing $O(Bk)$ memory and time-consuming (nested) function calls.

6 Application

We first present the genomic problem that motivated this study, in Section 6.1, and earlier contributions on the same topic, in Section 6.2. Two real datasets are described in Section 6.3. They play a central role in this article. We both *(a)* draw inspiration from one of them and *(b)* use it in order to set up our simulation study, as presented in Section 6.4. We also apply the TMLE methodology directly to the other. The specifics of the TMLE procedures that we undertake both on simulated and real data are given in Section 6.5, and their results are summarized in Section 6.6, for the simulation study, and in Section 6.7, for the real data application.

6.1 Association between DNA copy number and gene expression in cancers

The activity of a gene in a cell is directly related to its *expression level*, that is, the number of messenger RNA (mRNA) fragments corresponding to this gene. Cancer cells are characterized by changes in their gene expression patterns. Such alterations have been shown to be caused directly or indirectly by genetic events, such as *changes in the number of DNA copies*, and epigenetic events, such as *DNA methylation*. Some changes in DNA copy number have been reported to be positively associated with gene expression levels [11]. Conversely, DNA methylation is a chemical transformation of cytosines (one of the four types of DNA nucleotides) which is thought to lead to gene expression silencing [5]. Therefore, DNA methylation levels are generally negatively associated with gene expression levels.

We propose to apply the methodology developed in the previous sections to the search for genes for which there exists an association between DNA copy number variation and gene expression level, accounting for DNA methylation.

6.2 Related works

In the context of cancer studies, various methods have been proposed in order to find associations between DNA copy number and gene expression at the level of genes. Because we cannot cite all of them, we try here to cite one relevant publication for each broad type of method. Most of them can be classified into two groups, depending on whether DNA copy number is viewed as a continuous or a discrete variable. When DNA copy number is viewed as a continuous variable, associations between X and Y are generally quantified using a correlation coefficient [11]. When it is viewed as a discrete variable, associations are typically quantified using a test of differential expression between DNA copy number states [26]. A common limitation to this two types of methods is that they are generally good at identifying genes that were already known, but less so at finding novel candidates. This is not surprising: for correlation-based methods, high correlation between X and Y requires both X and Y to vary substantially, in which case it is likely that these (marginal) variations have already been reported. For methods based on differential expression between copy number states, the latter often correspond to biological or clinical groups which are already known and for which differential expression analyses have already been carried out.

In the present paper, we acknowledge the fact that while DNA copy number is observed as a quantitative variable, the copy neutral state (two copies of DNA) generally has positive mass, in the sense that for a given gene, a positive proportion of samples have two copies of DNA.

Another major difference between our method and the ones cited above is that we explicitly incorporate DNA methylation into the analysis. Several papers where DNA copy number, gene expression and DNA methylation are combined have been published recently, but they typically analyze one dimension of (W, X, Y) at a time, and then use an *ad hoc* rule to

merge or intersect the results [1, 17]. The CNAmet method [10] relies on two scores: a score of differential expression between copy number levels on the one hand, and between DNA methylation levels on the other hand. Then both scores are summed. In the method proposed here, the three dimensions are studied *jointly*.

6.3 Datasets

We exploit glioblastoma multiforme (GBM, the most common type of primary adult brain cancers) and ovarian cancers (OvCa, a cancerous growth arising from the ovary) data from The Cancer Genome Atlas (TCGA) project [2], a collaborative initiative to better understand several types of cancers using existing large-scale whole-genome technologies. TCGA has recently completed a comprehensive genomic characterization of these types of tumor, including DNA copy number (X) , gene expression (Y), and DNA methylation (W) microarray experiments [18, 19].

Probe-level normalized GBM and OvCa data can be downloaded from the TCGA repository at `http://tcga-data.nci.nih.gov/tcga/`. In order to study associations between X, Y and W at the level of genes, these probe-level measurements first need to be aggregated into gene-level summaries. We choose to define X, Y and W as follows for a given gene:

- DNA methylation W is the proportion of "methylated" signal at a CpG locus in the gene's promoter region;
- DNA copy number X is a locally smoothed total copy number relative to a set of reference samples;
- expression Y is the "unified" gene expression level across three microarray platforms, as defined by [27].

After this pre-processing step, each gene is represented by a $3 \times n$ matrix, where 3 is the number of data types and n is the number of samples. Figure 2(a) represents DNA methylation, DNA copy number, and gene expression data for one particular gene, `EGFR`, which is known to be altered in GBM. The association between copy number and expression is non-linear, and high methylation levels are associated with low expression levels.

6.4 Simulation scheme

Because association patterns between copy number, expression and methylation are generally non-linear, setting up a realistic simulation model is a difficult task. We design here a simulation strategy based on perturbations of real observed data structures, which mimics situations such as the one observed in the Figure 2(a) for the `EGFR` gene in GBM. This strategy implements the following constraints:

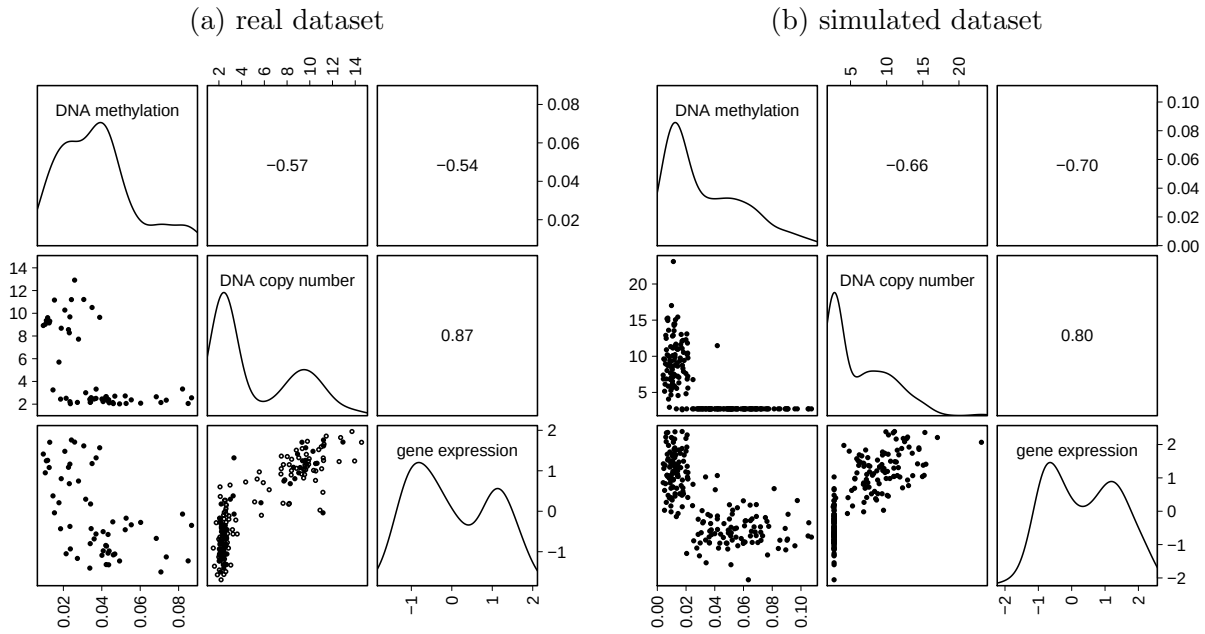

Figure 2: Illustrating DNA methylation, DNA copy number, and gene expression data. In both graphics, we represent kernel density estimates (diagonal panels), pairwise plots (lower panels), and report the pairwise Pearson correlation coefficients (upper panels). **(a).** Real dataset corresponding to the `EGFR` gene in 187 GBM tumor samples. For 130 among the 187 samples, only DNA copy number and gene expression data were available (circles in lower middle plot). **(b).** Simulated dataset consisting of $n = 200$ independent copies of the synthetic observed data structure described in Section 6.6. Note that the constant O_2^X is added to each value of X so that graphics corresponding to real and simulated data can be more easily compared.

- there are generally up to three copy number classes: normal regions, and regions of copy number gains and losses;

- in normal regions, expression is negatively correlated with methylation;

- in regions of copy number alteration, copy number and expression are positively correlated.

Our simulation scheme relies on three real observed data structures $O_1 = (O_1^W, O_1^X, O_1^Y)$, $O_2 = (O_2^W, O_2^X, O_2^Y)$, $O_3 = (O_3^W, O_3^X, O_3^Y)$ corresponding to three samples from different copy number classes: loss (class 1), normal (class 2), and gain (class 3). We simulate a synthetic observed data structure $O = (W, X, Y) \sim P^s$ as follows. Given a vector $p = (p_1, p_2, p_3)$ of proportions such that $p_1 + p_2 + p_3 = 1$, we first draw a class assignment U from the multinomial distribution with parameter $(1, p)$ (in other words, $U = u$ with probability p_u). Conditionally on U, a measure W of DNA methylation is drawn randomly as a perturbation of the DNA methylation in the corresponding real observed data structure O_U: given a vector $\omega = (\omega_1, \omega_2, \omega_3)$ of positive numbers,

$$W = \text{expit}\left(\text{logit}\left(O_U^W\right) + \omega_U Z\right),$$

where Z is a standard normal random variable independent of U. Finally, a couple (X, Y) of DNA copy number and DNA expression is drawn conditionally on (U, W) as a perturbation of the couple (O_U^X, O_U^Y) in the corresponding real observed data structure O_U (with an additional centering applied to X so that the pivot value be equal to 0): Given $\sigma_2 > 0$, two variance-covariance 2×2-matrices Σ_1 and Σ_3 and a *non-increasing* mapping $\lambda_0 : [0, 1] \to [0, 1]$,

- if $U = 2$, then $(X, Y) = (0, O_2^Y + \lambda_0(W) + \sigma_2 Z')$, where Z' is a standard normal random variable independent of (U, W);

- if $U \neq 2$, then (X, Y) is drawn conditionally on (U, W) from the bivariate Gaussian distribution with mean $(O_U^X - O_2^X, O_U^Y)$ and variance-covariance matrix Σ_U.

In particular, the reference/pivot value $x_0 = 0$. Note that λ_0 is chosen non-increasing in order to account for the negative association between DNA expression and methylation. Furthermore, the synthetic observed data structure O drawn from P^s is not bounded.

We easily derive closed-form expressions for the features of interest $\theta(P^s)$, $\mu(P^s)$, $g(P^s)$, and $\sigma^2(P^s)$, which we report in the Appendix (see Lemma 7). Relying on Lemma 7 makes it possible to evaluate the value of $\Psi(P^s)$, by following the procedure described in Section 3.1 (see details in Section 6.6).

Finally we provide in Figure 2(b), for the sake of illustration, a visual summary of a simulation run with $n = 200$ independent copies of the synthetic observed data structure O drawn from P^s and based on real observed data structure from two GBM samples for the

EGFR gene which are described in Table 1. The parameters for this simulation were chosen as follows: $p = (0, 1/2, 1/2)$, $\omega = (0, 3, 3)$, $\lambda_0 : w \mapsto -w$, $\sigma_2 = 1$, $\Sigma_3 = \begin{pmatrix} 9.96 & 1 \\ 1 & 0.43 \end{pmatrix}$.

6.5 Library of algorithms for super-learning

We explain in Section 5.3 that we rely on super-learning [24, 22] in order to estimate some relevant infinite-dimensional features of P_0, including (but not limited to) $\theta(P_0)$, $\mu(P_0)$ and $g(P_0)$. This algorithmic challenge is easily overcome, thanks to the remarkable R-package SuperLearner [12] and the possibility to rely on the library of R-packages [13] built by the statistical community. As for the base predictors, they involve (by alphabetical order):

- Generalized additive models: we use the gam R-package [4], with its default values.
- Generalized linear models: we use the glm R-function with identity link (for learning $\theta(P_0)$ and $\mu(P_0)$) and logit link (for learning $g(P_0)$), and with linear combinations of $(1, X, W)$ or $(1, X, W, XW)$ (for learning $\theta(P_0)$) and linear combinations of $(1, W)$ or $(1, W, W^2)$ (for learning $\mu(P_0)$ and $g(P_0)$).
- Piecewise linear splines: we use polymars R-function from the polspline R-package [6], with its default values.
- Random forests: we use the randomForest R-package [9], with its default values.
- Support vector machines: we use the svm R-function from the e1071 R-package [3], with its default values.

Note that none of the statistical models associated to the above estimation procedures contains P^s (see Lemma 7).

6.6 Simulation study

We conduct *twice* a simulation study where $B' = 10^3$ datasets of $n = 200$ independent observed data structures are (independently) generated under P^s (*i.e.*, under the simulation scheme described in Section 6.4). In each simulation study and for every simulated dataset, we perform the TMLE methodology for the purpose of estimating the target parameter $\Psi(P^s)$. From one simulation study to the other, we only change the set up of the super-learning procedure, by modifying the library of algorithms involved in the super-learning of the features of interest:

- the first time, we proceed exactly as described in Section 6.5 (we say that the *full-SL* is undertaken);
- the second time, we decide to include only algorithms based on generalized linear models (we say that the *light-SL* is undertaken).

We do not use any index to refer to the super-learning set up (full-SL or light-SL) for the sake of alleviating notations.

In each simulation study (*i.e.*, for each set up of the super-learning procedure full-SL and light-SL) and for each $b \leq B'$, we record the values $\psi_{n,b}^k = \Psi(P_{n,b}^k)$ of the initial substitution estimator ($k = 0$) and subsequent updated substitution estimators ($k = 1, 2, 3$) targeting $\Psi(P^s)$, as derived on the bth simulated dataset (whose empirical measure is denoted by $P_{n,b}$). The targeted update steps rely on the *Gaussian fluctuations* presented in Section 5.2 (the results are very similar when one applies either the general one-step updating procedure of Section 3.2 or the second tailored alternative two-step updating procedure of Section 5.2). We do not record the next updates because the *ad hoc* stopping criterion that we devise systematically indicates that this is not necessary (heuristically, the criterion elaborates on the gains in likelihood and the variations in the resulting estimates).

The value of $\Psi(P^s)$ is evaluated by simulations, following (5) in Section 3.1 with P^s substituted for P_n^0 (we rely on $B = 10^5$ simulated observed data structures, whose empirical measure is denoted by P_B; the features $\theta(P^s)$ and $\sigma^2(P^s)$ are explicitly known, see Lemma 7). In order to get a sense of how accurate our evaluation of $\Psi(P^s)$ is, we also use the same large simulated dataset to evaluate $\mathrm{Var}_{P^s} D^\star(P^s)(O)$ (as the empirical variance $\mathrm{Var}_{P_B} D^\star(P^s)(O)$; again, $D^\star(P^s)$ is known explicitly by Lemma 7). Denoting by $\psi_B(P^s)$ and $v_B(P^s)$ the latter evaluations, we interpret the intervals $[\psi_B(P^s) \pm \xi_{1-\alpha/2}\sqrt{v_B(P^s)/n}]$ and $[\psi_B(P^s) \pm \xi_{1-\alpha/2}\sqrt{v_B(P^s)/B}]$ as $(1-\alpha)$-accuracy intervals for the evaluation of $\Psi(P^s)$ based on $n = 200$ and $B = 10^5$ independent observed data structures. The gray intervals in Figure 3 represent these accuracy intervals for $\alpha = 5\%$, $n = 200$ (light gray) and $B = 10^5$ (dark gray). Note that (by the convolution theorem) the length of $[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/n}]$ is the optimal length of a 95%-confidence interval based on an efficient (regular) estimator of $\Psi(P^s)$ relying on n observations (assuming that the asymptotic regime is reached). The numerical values are reported in Table 2.

The results of this joint simulation study are summarized by Figure 3 (which shows kernel density estimates of the empirical distributions of $\{\psi_{n,b}^k : b \leq B'\}$ for $0 \leq k \leq 3$) and Table 3. They illustrate some of the fundamental characteristics of the TMLE estimator and related confidence intervals: convergence of the iterative updating procedure, robustness, asymptotic normality, and coverage.

Convergence of the iterative updating procedure, and robustness. A substantial bias in the initial estimation is revealed by the location of the mode of $\{\psi_{n,b}^0 : b \leq B'\}$ in Figure 3, both for the full-SL and light-SL procedures. We see that the full-SL initial estimator is less biased than its light-SL counterpart. As one can judge visually or by the first rows of Tables 3(a) and 3(b), this initial bias is diminished (if not perfectly corrected) at the first updating step of the TMLE procedure, illustrating the robustness of the targeted estimator. The empirical distributions of $\{\psi_{n,b}^k : b \leq B'\}$ for $k = 1, 2, 3$

sample name	methylation O_i^W	copy number O_i^X	expression O_i^Y
`TCGA-02-0001` ($i = 2$)	0.05	2.72	-0.46
`TCGA-02-0003` ($i = 3$)	0.01	9.36	1.25

Table 1: Real methylation, copy number and expression data used as a baseline for simulating the dataset according to the simulation scheme presented in Section 6.6. A visual of the simulated dataset is provided in Figure 2(b).

$\psi_B(P^s)$	$v_B(P^s)$	$[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/N}]$	
		$N = 200$	$N = 10^5$
0.2345	0.05980232	$[0.2006; 0.2684]$	$[0.2329; 0.2360]$

Table 2: Values of $\psi_B(P^s)$ and $v_B(P^s)$, estimators of $\Psi(P^s)$ and $\mathrm{Var}_{P^s} D^\star(P^s)(O)$, and 95%-accuracy intervals $[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/n}]$, $[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/B}]$ ($n = 200$, $B = 10^5$).

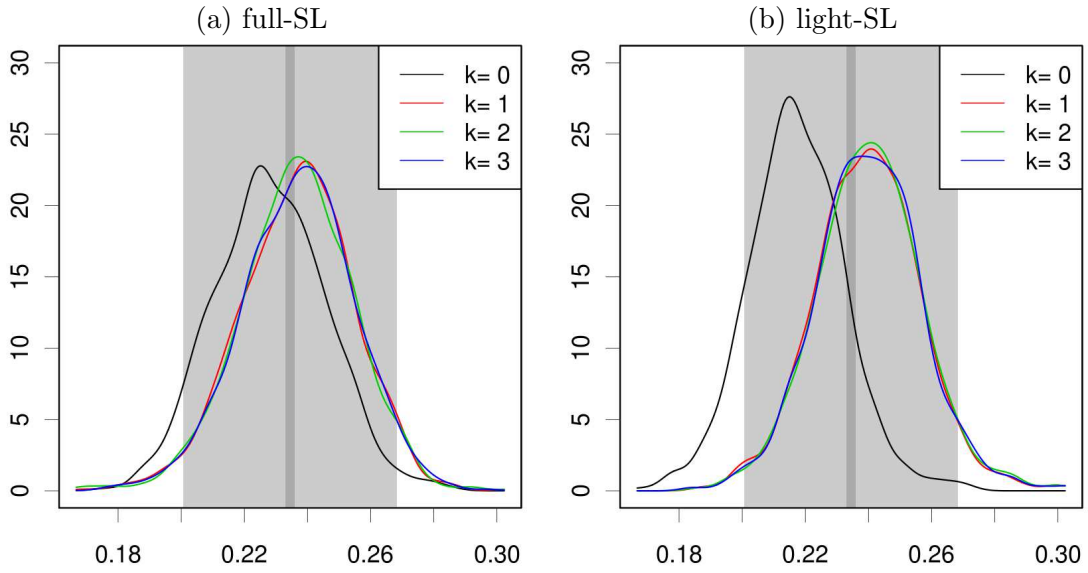

Figure 3: Empirical distribution of $\{\psi_{n,b}^k : b \leq B'\}$ based on $n = 200$ independent observed data structures for $k = 0$ (initial estimator) and k iterations of the updating procedure ($k = 1, 2, 3$), as obtained from $B' = 10^3$ independent replications of the simulation study (using a Gaussian kernel density estimator). **(a).** The super-learning procedure involves all algorithms described in Section 6.5. **(b).** The super-learning procedure only involves algorithms based on generalized linear models. In both graphics, gray rectangles represent 95%-accuracy intervals $[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/n}]$ and $[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/B}]$ for the true parameter $\Psi(P^s)$ based on 200 observed data structures (light gray) and $B = 10^5$ observed data structures (dark gray). The length of $[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/n}]$ is the optimal length of a 95%-confidence interval based on an efficient (regular) estimator of $\Psi(P^s)$ relying on n observations (assuming that the asymptotic regime is reached).

are not (visually) markedly different, an empirical indication that the TMLE procedure converges quickly.

Asymptotic normality. In order to check the asymptotic normality of the TMLE estimator (*e.g.* under the conditions of Corollary 1), we first perform Lilliefors tests of normality based on the empirical distributions of $\{\psi_{n,b}^k : b \leq B'\}$ for $k = 0, 1, 2, 3$ (*i.e.*, we perform Kolmogorov-Smirnov tests of normality *without* specification of the means and variances under the null). We report the values of the test statistics and corresponding p-values in the third and fourth rows of Tables 3(a) and 3(b). If we take into account the multiplicity of tests, there is no clear indication that the limit distributions are not Gaussian.

Second, we test the fit of the empirical distributions of $\{\psi_{n,b}^k : b \leq B'\}$ to a Gaussian distribution with mean and variance given by the estimates $\psi_B(P^s)$ and $v_B(P^s)$ (which are independent of $\{\psi_{n,b}^k : b \leq B'\}$). We report in the fifth rows of Tables 3(a) and 3(b) the obtained values of the KS test statistics. If all p-values are smaller than 10^{-4}, one notices that the test statistics are strikingly smaller for $k \geq 1$ than for $k = 0$. Performing Anderson-Darling tests of normality with only the null mean *or* the null variance specified (*i.e.*, KS tests of normality with specified null mean, equal to $\psi_B(P^s)$, and unspecified null variance *or* specified null variance, equal to $v_B(P^s)$, and unspecified null mean) teaches us that *it is mainly the little remaining bias and not the choice of the variance under the null* that makes the KS tests have so small p-values [values not shown].

Coverage. The theoretical convergence in distribution of the TMLE estimator to a Gaussian limit (*e.g.* under the conditions of Corollary 1) promotes the use of intervals $[\psi_{n,b}^k \pm \xi_{1-\alpha/2} s_{n,b}^k/\sqrt{n}]$ as $(1-\alpha)$-confidence intervals for $\Psi(P^s)$ $(k = 1, 2, 3)$, with $(s_{n,b}^k)^2 = \mathrm{Var}_{P_{n,b}} D^\star(P_{n,b}^k)(O)$. Interestingly, the theoretical result of Corollary 1 *do not* guarantee that it is safe to estimate the limit variance by $(s_{n,b}^k)^2$ (additional assumptions on the construction and convergence of $\theta(P_n^{k_n})$, $\mu(P_n^{k_n})$ and $g(P_n^{k_n})$ would be required to get such a result). We nonetheless check whether the latter intervals provide the wished coverage or not. For this purpose, we compute and report in the sixth and seventh rows of Tables 3(a) and 3(b) the empirical coverages $c_n^k = \frac{1}{B'} \sum_{b=1}^{B'} \mathbf{1}\{\psi_B(P^s) \in [\psi_{n,b}^k \pm \xi_{1-\alpha/2} s_{n,b}^k/\sqrt{n}]\}$ and their *optimistic* counterpart $c_n^{k+} = \frac{1}{B'} \sum_{b=1}^{B'} \mathbf{1}\{[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/B}] \cap [\psi_{n,b}^k \pm \xi_{1-\alpha/2} s_{n,b}^k/\sqrt{n}] \neq \emptyset\}$ (the latter incorporates the remaining uncertainty of the true value of $\Psi(P^s)$). We conclude that the provided coverage is good for the light-SL procedure (with excellent *optimistic* coverage), but disappointing for the full-SL procedure (even for the *optimistic* coverage). The results may have been better if one had relied on the bootstrap in order to estimate the asymptotic variance of the TMLE. We will investigate this issue in future work.

6.7 Real data application

For the real data application, we focus on all 130 genes $g \in \mathcal{G}$ of chromosome 18 in the OvCa dataset. This choice is notably motivated by the associated sample size, approximately equal to 500 (thus much larger than the sample size associated to the GBM dataset). We estimate the non-parametric variable importance measure of X on Y accounting for W for each gene separately (*i.e.*, $\Psi(P_0^g)$ where $P_0^g \in \mathcal{M}$ is the true distribution of $O = (W, X, Y)$ for gene g), following exactly one of the statistical methodologies developed in the simulation study. Specifically, the targeted update step relies on the Gaussian fluctuations presented in Section 5.2, and the super-learning involves the library of algorithms that we report in Section 6.5. In particular, we estimate for each gene g the asymptotic variance of the TMLE $\psi_n^{g,*}$ of $\Psi(P_0^g)$ with the empirical variance $(s_n^{g,*})^2$ of the efficient influence curve at $P_n^{g,*}$. In a future work solely devoted to this real data application, we will use the bootstrap in order to derive a more robust estimator of the asymptotic variance (again, Corollary 1 requires some conditions on P_0^g and $P_n^{g,*}$ in order to guarantee that $(s_n^{g,*})^2$ is a consistent estimator). We will also "extend" W, by adding to the DNA methylation of the gene of interest the DNA methylations, DNA copy numbers and gene expressions of its neighboring genes.

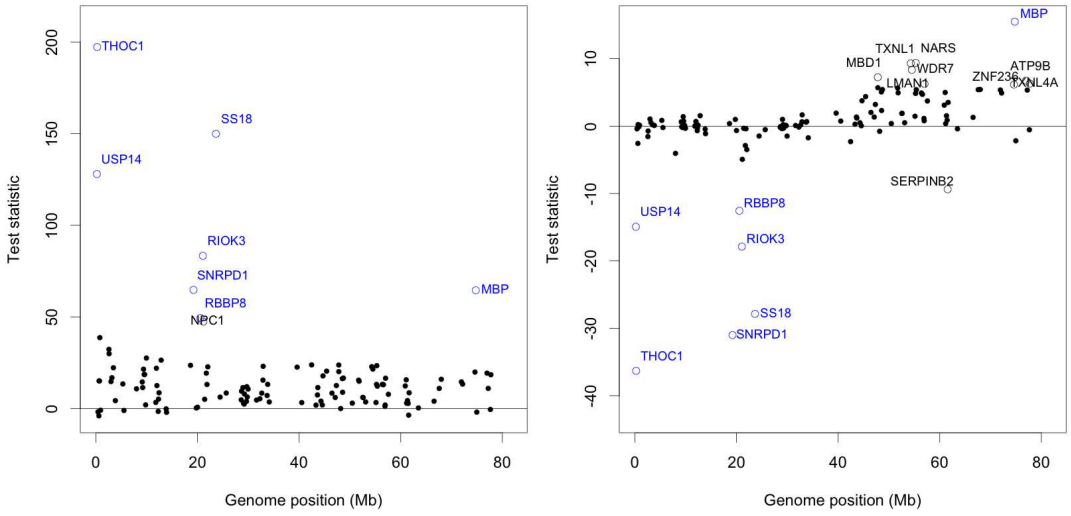

Figure 4: Real data application to the 130 genes of chromosome 18 in the OvCa dataset (ovarian cancers). We represent the tests statistics $\sqrt{n}(\psi_n^{g,3} - \psi_{\text{ref}}^g)/s_n^{g,3}$ for $\psi_{\text{ref}}^g = 0$ (left graphic) and $\psi_{\text{ref}}^g = \mathcal{F}(P_n^g)$ (right graphic) along the position of gene g on the genome. We report the names of the genes such that $\sqrt{n}|\psi_n^{g,3}|/s_n^{g,3} > 45$ (left graphic) and $\sqrt{n}|\psi_n^{g,3} - \mathcal{F}(P_n^g)|/s_n^{g,3} > 6$ (right graphic), the cut-offs being arbitrarily chosen.

We only briefly summarize the results of the real data application. For this purpose, we report in Figure 4 the values of the test statistics $\sqrt{n}(\psi_n^{g,3} - \psi_{\text{ref}}^g)/s_n^{g,3}$ $(g \in \mathcal{G})$ derived from the TMLE after three updates, using two different reference values $\psi_{\text{ref}}^g \in \{0, \mathcal{F}(P_n^g)\}$.

Here, $\mathcal{F}(P_n^g) = \sum_{i=1}^n X^{(i)}Y^{(i)} / \sum_{i=1}^n (X^{(i)})^2$ is the least square (substitution, asymptotically efficient) estimator of parameter $\mathcal{F}(P_0^g)$, see (4), a parameter which overlooks the role potentially played by W while quantifying the influence of X on Y. We are aware that $\mathcal{F}(P_n^g)$ is not independent of $\psi_n^{g,3}$ and $s_n^{g,3}$, and will make sure in a future work solely devoted to this real data application that our estimator of $\mathcal{F}(P_0^g)$ is derived from an independent dataset (or we will undertake a cross-validated procedure). The reference value $\psi_{\text{ref}}^g = 0$ is a natural null value to rely on from a testing perspective. Using $\psi_{\text{ref}}^g = \mathcal{F}(P_n^g)$ as another null value is relevant because that allows us to identify those genes for which the (possibly intricate) role played by W in quantifying the influence of X on Y is especially important and results in a stark deviation of $\Psi(P_0^g)$ from $\mathcal{F}(P_0^g)$.

Looking at the left graphic in Figure 4 teaches us that a majority of the $\Psi(P_0^g)$ ($g \in \mathcal{G}$) are likely positive. Eight genes stand up (by having a test statistic $\sqrt{n}\psi_n^{g,3}/s_n^{g,3} > 45$): two genes at 18p11.32 (`USP14` and `THOC1`), a cluster of five genes at 18q11.2 (`SNRPD1`, `RBBP8`, `RIOK3`, `NPC1`, `SS18`), and gene `MBP` at 18q23. This suggests that the region 18q11.2 (especially 19-24 Mb) is of particular relevance in this set of ovarian cancers. Seven out of the latter eight genes (specifically: all of them but gene `NPC1`) also stand up in the right graphic of Figure 4: six out of the latter seven genes standing up in both graphics (specifically: all of them but gene `MBP`) exhibit a significantly *small* test statistic (by having $\sqrt{n}(\psi_n^{g,3} - \mathcal{F}(P_n^g))/s_n^{g,3} < -6$), as does the additional gene `SERPINB2`, while gene `MBP` exhibits a significantly *large* test statistic (by having $\sqrt{n}(\psi_n^{g,3} - \mathcal{F}(P_n^g))/s_n^{g,3} > 6$), as do eight additional genes (`MBD1`, `TXNL1`, `LMAN1`, `WDR7`, `NARS`, `ZNF236`, `ATP9B`, `TXNL4A`). All genes standing up in the right graphic of Figure 4 are located at 18q2 (41-76 Mb).

Acknowledgments

The topic of this article originates from a presentation [16] by Terry Speed (Department of Statistics, UC Berkeley) in the UC Berkeley Statistics and Genomics Seminar. We would like to thank him for a series of instructive discussions that followed. We also would like to thank The Cancer Genome Atlas project [2] for kindly providing the datasets.

A Appendix

A.1 Miscellanea

Recall that P^s denotes the data-generating distribution of the synthetic observed data structure $O = (W, X, Y)$ described in Section 6.6. We easily derive the following closed-form expressions for the features of interest $\theta(P^s)$, $\mu(P^s)$, $g(P^s)$, and $\sigma^2(P^s)$.

Lemma 7. *Let φ denote the density of the standard normal distribution. The following*

equalities hold:

$$\begin{aligned}
\theta(P^s)(X,W) &= (O_2^Y + \lambda_0(W))P^s(U=2|X,W) \\
&\quad + \sum_{u=1,3}\left(O_u^Y + \frac{\Sigma_u(1,2)}{\Sigma_u(1,1)}(X-(O_u^X-O_2^X))\right)P^s(U=u|X,W), \\
\mu(P^s)(W) &= \sum_{u=1}^{3}(O_u^X - O_2^X)P^s(U=u|W), \\
g(P^s)(0|W) &= P^s(U=2|W), \\
\sigma^2(P^s) &= \sum_{u=1,3} p_u\left(\Sigma_u(1,1) + (O_u^X - O_2^X)^2\right),
\end{aligned}$$

where, for each $u = 1, 2, 3$*,*

$$\begin{aligned}
P^s(U=2|X,W) &\propto \frac{p_2}{\omega_2}\varphi\left(\frac{\text{logit}(W) - \text{logit}(O_2^W)}{\omega_2}\right)\mathbf{1}\{X=0\}, \\
P^s(U=u|X,W) &\propto \frac{p_u}{\omega_u}\varphi\left(\frac{\text{logit}(W) - \text{logit}(O_u^W)}{\omega_u}\right) \times \varphi\left(\frac{X-(O_u^X - O_2^X)}{\sqrt{\Sigma_u(1,1)}}\right), \\
P^s(U=u|W) &\propto \frac{p_u}{\omega_u}\varphi\left(\frac{\text{logit}(W) - \text{logit}(O_u^W)}{\omega_u}\right).
\end{aligned}$$

A.2 Proofs of Lemmas 1, 6 and Proposition 1

Proof of Lemma 1. Let us consider (10). For any non-negative measurable function f of (X, W), it holds that

$$\begin{aligned}
E_{P_\varepsilon}\{Yf(X,W)\} &= E_P\{Yf(X,W)(1+\varepsilon s(O))\} \\
&= E_P\{\theta(P)(X,W)f(X,W)\} + \varepsilon E_P\{Yf(X,W)s(O)\} \\
&= E_P\{(\theta(P)(X,W) + \varepsilon E_P(Ys(O)|X,W))f(X,W)\} \\
&= E_{P_\varepsilon}\{h(X,W)f(X,W)\}
\end{aligned}$$

for $h(X, W)$ equal to the right-hand side expression of (10), since (9) implies

$$\frac{dP_\varepsilon}{dP}(X,W) = (1 + \varepsilon E_P(s(O)|X,W)).$$

The function f being arbitrarily chosen, the latter equalities yield (10). The remaining relationships are easily proven in the same spirit. □

Proof of Lemma 6. Note that

$$P_0 L_{a,b}(P) = E_{P_0}\{\text{KL}(F_{a,b}\circ\theta(P_0)(X,W), F_{a,b}\circ\theta(P)(X,W))\} + c(P_0),$$

where $\text{KL}(p, q)$ is the Kullback-Leibler divergence between the Bernoulli distributions of parameters $p, q \in]0, 1[$ and $c(P_0)$ is a constant depending on P_0 only. Since $\text{KL}(p, q) \geq 0$ with

equality iff $p = q$, we obtain that $\theta(P_0)$ minimizes $P \mapsto P_0 L_{a,b}(P)$ and also that another minimizer must satisfy $\theta(P)(X,W) = \theta(P_0)(X,W)$ P_0-almost surely. The second equality is easily obtained by differentiating. □

Proof of Proposition 1. By expanding the squared sum in (1), we obtain that

$$\Psi(P) = \arg\min_{\beta\in\mathbb{R}} \left\{-2\beta E_P\{X(\theta(P)(X,W) - \theta(P)(0,W))\} + \beta^2 E_P\{X^2\}\right\},$$

which straightforwardly yields (2). It is easily seen that $PD_1^\star(P)D_2^\star(P) = 0$, or in other words that the two components are orthogonal in $L_0^2(P)$.

Regarding the pathwise differentiability, it is sufficient to consider paths of the form (9) for arbitrarily chosen $s \in L_0^2(P)$ with $\|s\|_\infty < \infty$. Set such a s and $|\varepsilon| < \|s\|_\infty^{-1}$, $\varepsilon \neq 0$. Using the telescopic equality $a_1/b_1 - a_0/b_0 = (a_1 - a_0)/b_1 - (a_0/b_0)(b_1 - b_0)/b_1$ yields

$$\varepsilon^{-1}(\Psi(P_\varepsilon) - \Psi(P)) = \frac{T_\varepsilon^1}{\sigma^2(P_\varepsilon)} - \Psi(P)\frac{T_\varepsilon^2}{\sigma^2(P_\varepsilon)}, \tag{19}$$

with

$$\begin{aligned} T_\varepsilon^1 &= \varepsilon^{-1}\Big(E_{P_\varepsilon}\{X(\theta(P_\varepsilon)(X,W) - \theta(P_\varepsilon)(0,W))\} - E_P\{X(\theta(P)(X,W) - \theta(P)(0,W))\}\Big), \\ T_\varepsilon^2 &= \varepsilon^{-1}(\sigma^2(P_\varepsilon) - \sigma^2(P)) = E_P\{s(O)X^2\} \end{aligned} \tag{20}$$

by (13). Now, the same telescopic equality also yields that

$$\begin{aligned} T_\varepsilon^1 = {} & E_P\{X(\theta(P_\varepsilon)(X,W) - \theta(P_\varepsilon)(0,W))s(O)\} \\ & + E_P\left\{X\left(\varepsilon^{-1}(\theta(P_\varepsilon)(X,W) - \theta(P)(X,W)) - \varepsilon^{-1}(\theta(P_\varepsilon)(0,W) - \theta(P)(0,W))\right)\right\}. \end{aligned}$$

By (10) and the dominated convergence theorem (indeed, $\{\|\theta(P_\varepsilon)\|_\infty : |\varepsilon| < \|s\|_\infty^{-1}\}$ is bounded),

$$\begin{aligned} T_\varepsilon^1 = {} & E_P\{X(\theta(P)(X,W) - \theta(P)(0,W))s(O)\} + o(\varepsilon) \\ & + E_P\left\{X\left(\varepsilon^{-1}(\theta(P_\varepsilon)(X,W) - \theta(P)(X,W)) - \varepsilon^{-1}(\theta(P_\varepsilon)(0,W) - \theta(P)(0,W))\right)\right\}. \end{aligned}$$

Furthermore, (10) also yields that

$$\varepsilon^{-1}(\theta(P_\varepsilon)(X,W) - \theta(P)(X,W)) = E_P((Y - \theta(P)(X,W))s(O)|X,W) + o(\varepsilon).$$

Consequently, applying the dominated convergence theorem finally yields (by using the above telescopic equality and (10), one easily checks that $\{\sup_{O\in\mathcal{O}} \varepsilon^{-1}|\theta(P_\varepsilon)(X,W) - \theta(P)(X,W)| : |\varepsilon| < \|s\|_\infty^{-1}\}$ is bounded)

$$\begin{aligned} T_\varepsilon^1 = {} & E_P\{X(\theta(P)(X,W) - \theta(P)(0,W))s(O)\} \\ & + E_P\{E_P(X(Y - \theta(P)(X,W))s(O)|X,W) \end{aligned}$$

$$- XE_P((Y - \theta(P)(X,W))s(O)|X=0,W)\} + o(\varepsilon), \quad (21)$$

where we emphasize that

$$E_P((Y - \theta(P)(X,W))s(O)|X=0,W) = E_P\left(\frac{\mathbf{1}\{X=0\}}{g(P)(0|W)}(Y - \theta(P)(X,W))s(O)\Big|W\right).$$

Combining (19), (20), (21) and (13) teaches us that, for all $s \in L_0^2(P)$ with $\|s\|_\infty < \infty$,

$$\varepsilon^{-1}(\Psi(P_\varepsilon) - \Psi(P)) = E_P\{D^\star(P)(O)s(O)\} + o(\varepsilon),$$

where $D^\star(P)$ is defined in the statement of the proposition. In particular, Ψ is pathwise differentiable at P wrt the described collection of paths, and $D^\star(P)$ is a gradient of Ψ at P. Since the related tangent space is $L_0^2(P)$ itself, it is necessarily the efficient influence curve.

It remains to prove that $D^\star(P)$ is double-robust. For this purpose, note that

$$\begin{aligned}
&\sigma^2(P')PD^\star(P') - \sigma^2(P)(\Psi(P) - \Psi(P')) \\
&= E_P\{X(\theta(P')(X,W) - \theta(P')(0,W)) - X(\theta(P)(X,W) - \theta(P)(0,W))\} \\
&\qquad + E_P\left\{\left(X - \frac{\mu(P')(W)\mathbf{1}\{X=0\}}{g(P')(0|W)}\right)E_P(Y - \theta(P')(X,W)|X,W)\right\} \\
&= E_P\{X(\theta(P')(X,W) - \theta(P')(0,W)) - X(\theta(P)(X,W) - \theta(P)(0,W))\} \\
&\qquad + E_P\left\{\left(X - \frac{\mu(P')(W)\mathbf{1}\{X=0\}}{g(P')(0|W)}\right\}(\theta(P)(X,W) - \theta(P')(X,W))\right\} \\
&= E_P\left\{X(\theta(P)(0,W) - \theta(P')(0,W)) - \mu(P')(W)\frac{g(P)(0|W)}{g(P')(0|W)}(\theta(P)(0,W) - \theta(P')(0,W))\right\} \\
&\qquad = E_P\left\{(\theta(P)(0,W) - \theta(P')(0,W))\left(\mu(P)(W) - \mu(P')(W)\frac{g(P)(0|W)}{g(P')(0|W)}\right)\right\}.
\end{aligned}$$

Now, the right-hand side expression vanishes as soon as either $\theta(P')(0,\cdot) = \theta(P)(0,)$ or $(\mu(P') = \mu(P)$ and $g(P') = g(P))$. The conclusion readily follows. □

A.3 Proof of Lemma 5

Proof of Lemma 5. Assume for the time being that, for all $W \in \mathcal{W}$, there exists λ_n such that (17) holds with λ_n substituted for λ. Then, for all $W \in \mathcal{W}$, the point with coordinates $(E_{P_n^0}(X|X \neq 0, W), \varphi_{n,\lambda_n}(E_{P_n^0}(X|X \neq 0, W)))$ lies in the convex envelope of the set $\{(X^{(i)}, X^{(i)2}) : i \leq n\} \setminus \{(0,0)\}$. Equivalently, there exist for all $W \in \mathcal{W}$ three non-negative weights p_1, p_2, p_3 summing up to 1 and three different values $x^{(1)}, x^{(2)}, x^{(3)} \in \{X^{(i)} : i \leq n\} \setminus \{0\}$ such that

$$E_{P_n^0}(X|X \neq 0, W) = \sum_{k=1}^3 p_k x^{(k)}, \quad E_{P_n^0}(X^2|X \neq 0, W) = \sum_{k=1}^3 p_k x^{(k)2},$$

the right-hand side expressions being, respectively, the mean and second order moment of the distribution $\sum_{k=1}^3 p_k \text{Dirac}(x^{(k)})$. Thus, there exists $P_n^{00} \in \mathcal{M}$ such that *(i)* and *(ii)* hold.

Set $W \in \mathcal{W}$. Combining (6), (7) and (17) yields that if there exists λ_n such that (17) holds with λ_n substituted for λ, then it must be equal to $\ell_n = (T_n^1 - \sigma^2(P_n^0))/(T_n^1 - T_n^2)$, where $T_n^1 = (m_n + M_n)E_{P_n^0}\{\mu(P_n^0)(W)\} - m_n M_n E_{P_n^0}\{1 - g(P_n^0)(0|W)\}$ and $T_n^2 = E_{P_n^0}\{\mu(P_n^0)(W)^2/(1 - g(P_n^0)(0|W))\}$. In order to conclude, it is therefore sufficient to check that $\ell_n \in [0, 1]$.

By the Jensen inequality, it holds that $E_{P_n^0}(X^2|X \neq 0, W) \geq E_{P_n^0}(X|X \neq 0, W)^2$, which yields in turn with (6) and (7) that $\sigma^2(P_n^0) \geq T_n^2$. Finally, using again (6) and (7), $\sigma^2(P_n^0) - T_n^1$ equals

$$\begin{aligned} &E_{P_n^0}\left\{(1 - g(P_n^0)(0|W))\left(E_{P_n^0}(X^2|X \neq 0, W) - (m_n + M_n)E_{P_n^0}(X|X \neq 0, W) + m_n M_n\right)\right\} \\ &\quad = E_{P_n^0}\left\{(1 - g(P_n^0)(0|W))E_{P_n^0}\left((X - m_n)(X - M_n)|X \neq 0, W\right)\right\} \leq -c^2 P_n^0(X \neq 0), \end{aligned}$$

hence $T_n^2 \leq \sigma^2(P_n^0) < T_n^1$. Thus, $\ell_n \in [0, 1]$, which completes the proof. □

A.4 Proofs of Lemmas 2, 3 and 4

Proof of Lemma 2. It is sufficient to verify that, under the stated assumptions,

$$\limsup_{(\varepsilon,k)\to(0,\infty)} \left| P_n \frac{D^\star(P_n^k)}{1 + \varepsilon D^\star(P_n^k)} - P_n D^\star(P_n^k) \right| = 0.$$

Now, the absolute value above is straightforwardly upper-bounded by

$$\varepsilon M^2 P_n \left|1 + \varepsilon D^\star(P_n^k)\right|^{-1} = \varepsilon M^2 P_n \left(1 + \varepsilon D^\star(P_n^k)\right)^{-1} \leq \varepsilon M^2/(1 - \rho M) = 2\varepsilon M^2.$$

This trivially entails the wished convergence, hence the result. □

Let us introduce, for all $k \geq 0$ and $|\varepsilon| \leq \rho$, $\ell_n^k(\varepsilon) = n^{-1}\sum_{i=1}^n \log P_n^k(\varepsilon)(O^{(i)})$ and

$$A_n^k(\varepsilon) = -P_n \frac{D^\star(P_n^k)^2}{\left(1 + \varepsilon D^\star(P_n^k)\right)^2}.$$

Obviously, the normalized log-likelihood $\ell_n^k(\varepsilon)$ under $P_n^k(\varepsilon)$ is twice differentiable wrt ε, with first derivative at $\varepsilon = 0$ equal to $P_n D^\star(P_n^k)$ and second derivative at ε equal to $A_n^k(\varepsilon)$.

Proof of Lemma 3, first part. Let us first show, by contradiction, that $\lim_{k\to\infty} P_n D^\star(P_n^k) = 0$ under the stated assumptions. Suppose that $P_n D^\star(P_n^k)$ *does not* converge to 0 as $k \to \infty$: there exist $\eta > 0$ and an increasing function $\varphi : \mathbb{N} \to \mathbb{N}$ such that, for all $k \geq 0$,

$$|P_n D^\star(P_n^{\varphi(k)})| \geq \eta > 0. \tag{22}$$

We show that necessarily $\lim_{k\to\infty} \varepsilon_{\varphi(k)} = 0$, hence by Lemma 2 that $\lim_{k\to\infty} P_n D^\star(P_n^{\varphi(k)}) = 0$, contradicting (22).

Set $k \geq 0$. For any $\varepsilon_n'^{\varphi(k)} \in [0, \varepsilon_n^{\varphi(k)}]$, a Taylor expansion of $\ell_n^{\varphi(k)}(\varepsilon)$ yields the existence of $\varepsilon' \in [0, \varepsilon_n'^{\varphi(k)}]$ such that

$$\ell_n^{\varphi(k)}(\varepsilon_n^{\varphi(k)}) - \ell_n^{\varphi(k)}(0) \quad \geq \quad \ell_n^{\varphi(k)}(\varepsilon_n'^{\varphi(k)}) - \ell_n^{\varphi(k)}(0)$$

$$= \quad \varepsilon_n'^{\varphi(k)} P_n D^\star(P_n^{\varphi(k)}) + \frac{(\varepsilon_n'^{\varphi(k)})^2}{2} A_n^{\varphi(k)}(\varepsilon'). \tag{23}$$

By assumption *(iii)* and since

$$-4M^2 \leq \inf_{k'\geq 0} \inf_{|\varepsilon|\leq\rho} A_n^{k'}(\varepsilon) \leq \sup_{k'\geq 0} \sup_{|\varepsilon|\leq\rho} A_n^{k'}(\varepsilon) \leq -\frac{4}{9} \inf_{k'\geq 0} P_n D^\star(P_n^{k'})^2, \tag{24}$$

there exists a constant $\kappa > 0$ (depending on P_n) such that the right-hand side term of (24) is upper-bounded by $-\kappa$, hence $A_n^{k'}(\varepsilon) \leq -\kappa$ simultaneously for all $k' \geq 0$ and $|\varepsilon| \leq \rho$.

The function $\varepsilon \mapsto \frac{\partial}{\partial\varepsilon}\ell_n^{\varphi(k)}(\varepsilon)$ being decreasing and equal to $P_n D^\star(P_n^{\varphi(k)}) \neq 0$ at $\varepsilon = 0$, it necessarily holds that $\varepsilon_n^{\varphi(k)} P_n D^\star(P_n^{\varphi(k)}) > 0$ (*i.e.*, $\varepsilon_n^{\varphi(k)}$ and $P_n D^\star(P_n^{\varphi(k)})$ share the same sign), hence $\varepsilon_n'^{\varphi(k)} P_n D^\star(P_n^{\varphi(k)}) > 0$ too. Furthermore, combining (23) and the left-hand side of (24) yields

$$\begin{aligned} \ell_n^{\varphi(k)}(\varepsilon_n^{\varphi(k)}) - \ell_n^{\varphi(k)}(0) &\geq \varepsilon_n'^{\varphi(k)} P_n D^\star(P_n^{\varphi(k)}) - 2M^2(\varepsilon_n'^{\varphi(k)})^2 \\ &\geq |\varepsilon_n'^{\varphi(k)}|\eta - 2M^2(\varepsilon_n'^{\varphi(k)})^2. \end{aligned} \tag{25}$$

The conclusion is now at hand. Assume that the sequence $\{\varepsilon_n^{\varphi(k)}\}_{k\geq 0}$ *does not* converge to 0: there exist $c > 0$ and another increasing function $\psi : \mathbb{N} \to \mathbb{N}$ such that, for all $k \geq 0$, $|\varepsilon_n^{\psi\circ\varphi(k)}| \geq c > 0$. Note that c can be chosen small enough to guarantee in addition that $c\eta - 2M^2c^2 > 0$. Let us impose now $|\varepsilon_n'^{\psi\circ\varphi(k)}| = c$ for all $k \geq 0$ (this uniquely defines $\varepsilon_n'^{\psi\circ\varphi(k)} \in [0, \varepsilon_n^{\psi\circ\varphi(k)}]$). According to (25), for all $k \geq 0$,

$$\ell_n^{\psi\circ\varphi(k)}(\varepsilon_n^{\psi\circ\varphi(k)}) - \ell_n^{\psi\circ\varphi(k)}(0) \geq c\eta - 2M^2c^2 > 0.$$

Using *(a)* $\ell_n^{k'}(\varepsilon_n^{k'}) - \ell_n^{k'}(0) \geq 0$ for all $k' \geq 0$ and *(b)* $\ell_n^{k'}(0) = \ell_n^{k'-1}(\varepsilon_n^{k'-1})$ for every $k' \geq 1$, one obtains that for all $k \geq 0$,

$$\ell_n^{\psi\circ\varphi(k)}(\varepsilon_n^{\psi\circ\varphi(k)}) - \ell_n^0(0) \geq k(c\eta - 2M^2c^2).$$

This contradicts assumption *(iv)*. So the sequence $\{\varepsilon_n^{\varphi(k)}\}_{k\geq 0}$ *must* converge to 0, Lemma 2 applies, and (22) is contradicted. □

Proof of Lemma 3, second part. For all $k \geq 0$, another Taylor expansion of $\ell_n^k(\varepsilon)$ yields the existence of $\varepsilon_n'^k \in [0, \varepsilon_n^k]$ such that

$$0 \leq \ell_n^k(\varepsilon_n^k) - \ell_n^k(0) = \varepsilon_n^k P_n D^\star(P_n^k) + \frac{(\varepsilon_n^k)^2}{2} A_n^k(\varepsilon_n'^k).$$

We derive from these inequalities that

$$0 \leq (\varepsilon_n^k)^2\kappa \leq (\varepsilon_n^k)^2|A_n^k(\varepsilon_n'^k)| \leq 2\varepsilon_n^k P_n D^\star(P_n^k) \leq \rho|P_n D^\star(P_n^k)|,$$

where the right-hand side converges to 0 as $k \to \infty$ by virtue of the first part of the lemma. This completes the proof. □

Proof of Lemma 4. We first show that the sequence $\{P_n^k\}_{k\geq 0}$ converges in total variation. For this purpose, note that P_n^k is dominated by P_n^0, with a density f_n^k characterized by $f_n^k(O) = \prod_{k'=0}^{k-1}(1+\varepsilon_n^{k'} D^\star(P_n^{k'})(O))$. Since *(a)* the functions $D^\star(P_n^{k'})$ are uniformly bounded by a common constant M, and *(b)* the series $\sum_{k\geq 0}|\varepsilon_n^k|$ converges, the sequence of densities (wrt P_n^0) $\{f_n^k\}_{k\geq 0}$ converges wrt the $\|\cdot\|_\infty$-norm to a limit density (wrt P_n^0) that we denote f_n^*. Density f_n^* gives rise to a data-generating distribution P_n^*, the limit of P_n^k in total variation (hence its weak limit too).

Now, it holds that $\psi_n^k = \Psi(P_n^k) = (E_{P_n^k}\{XY\} - E_{P_n^k}\{X\theta(P_n^k)(0,W)\})/E_{P_n^k}\{X^2\}$. The observed data structure O being bounded, the functions $O=(W,X,Y)\mapsto XY$ and $O=(W,X,Y)\mapsto X^2$ are continuous and bounded, hence $E_{P_n^k}\{XY\}$ and $E_{P_n^k}\{X^2\}$ respectively converge to $E_{P_n^*}\{XY\}$ and $E_{P_n^*}\{X^2\}\geq c$ as $k\to\infty$ by weak convergence. Furthermore, the convergence of f_n^k to f_n^* wrt the $\|\cdot\|_\infty$-norm trivially entails the pointwise convergence of $\theta(P_n^k)$ to $\theta(P_n^*)$, then the wished convergence of $E_{P_n^k}\{X\theta(P_n^k)(0,W)\}$ to $E_{P_n^*}\{X\theta(P_n^*)(0,W)\}$ by the dominated convergence theorem. This completes the proof. □

A.5 Proof of Propositions 2 and 3 and of Corollary 1

Denote $D^\star(\sigma^2,\theta,\mu,g,\psi) = D_1^\star(\sigma^2,\theta,\psi) + D_2^\star(\sigma^2,\theta,\mu,g)$, let $D_1(\sigma^2,\theta)$ be characterized by $D_1(\sigma^2,\theta)(O) = (X(\theta(X,W)-\theta(0,W)))/\sigma^2$, and define $D_1(P) = D_1(\sigma^2(P),\theta(P))$. We use the notation $a\lesssim b$ for "a smaller than b up to a multiplicative constant". Let us start with a useful lemma.

Lemma 8. *Suppose that the assumptions of Proposition 2 are met. There exists $\psi_0\in\mathbb{R}$ such that $\widetilde{\psi}_n^* = \psi_0 + o_P(1)$ (*i.e.*, the TMLE converges in probability). Moreover, it holds that*

$$P_0(D_1^\star(\sigma^2(P_n^{k_n}),\theta(P_n^{k_n}),\widetilde{\psi}_n^*) - D_1^\star(\sigma_0^2,\theta_0,\psi_0))^2 = o_P(1), \tag{26}$$

$$P_0(D_2^\star(P_n^{k_n}) - D_2^\star(\sigma_0^2,\theta_0,\mu_0,g_0))^2 = o_P(1), \quad \textit{hence} \tag{27}$$

$$P_0(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2,\theta_0,\mu_0,g_0,\psi_0))^2 = o_P(1) \tag{28}$$

Proof. Recall that $\|O\|$ is bounded under P_0 and that $\sigma^2(P_n^{k_n}),\sigma_0^2\geq c$. Using repeatedly the telescopic equality $a_1/b_1 - a_0/b_0 = (a_1-a_0)/b_1 - (a_0/b_0)(b_1-b_0)/b_1$ and inequality $(a+b)^2\leq 2(a^2+b^2)$ yields that, under P_0, $(D_1(P_n^{k_n}) - D_1(\sigma_0^2,\theta_0))(O)^2 \lesssim (\theta(P_n^{k_n})-\theta_0)(O)^2 + (\theta(P_n^{k_n})(0,\cdot)-\theta_0(0,\cdot))(O)^2$, and therefore that

$$P_0(D_1(P_n^{k_n}) - D_1(\sigma_0^2,\theta_0))^2 = o_P(1). \tag{29}$$

Similarly, the same tricks as above and the facts that *(a)* both $|(Y-\theta(P_n^{k_n})(X,W))/\sigma^2(P_n^{k_n})|$ and $|(Y-\theta(P_0)(X,W))/\sigma_0^2|$ are upper-bounded under P_0, and *(b)* $g(P_n^{k_n})(0|W), g_0(0|W)\geq c$ imply that, under P_0, $(D_2^\star(P_n^{k_n}) - D_2^\star(\sigma_0^2,\theta_0,\mu_0,g_0))(O)^2 \lesssim (\mu(P_n^{k_n})-\mu_0)(O)^2 + (g(P_n^{k_n})(0|\cdot) - g_0(0|\cdot))(O)^2$, hence (27).

Now, let us rewrite $P_n D^\star(P_n^{k_n}) = o_P(1)$ as

$$\widetilde{\psi}_n^* \frac{E_{P_n}\{X^2\}}{\sigma^2(P_n^{k_n})} = (P_n - P_0)(D_1(P_n^{k_n}) + D_2^\star(P_n^{k_n})) \\ + P_0(D_1(P_n^{k_n}) + D_2^\star(P_n^{k_n}) - D_1(\sigma_0^2, \theta_0) - D_2^\star(\sigma_0^2, \theta_0, \mu_0, g_0)) \\ + P_0(D_1(\sigma_0^2, \theta_0) + D_2^\star(\sigma_0^2, \theta_0, \mu_0, g_0)) + o_P(1) \quad (30)$$

and consider the two first right-hand side terms. Because $D_1(\sigma^2, \theta)(O) = D_1^\star(\sigma^2, \theta, \psi)(O) + X^2\psi/\sigma^2$ and the class $\{O \mapsto X^2\psi/\sigma^2 : (\psi, \sigma^2) \in \mathbb{R} \times [c, \infty]\}$ is P_0-Donsker, it holds that both $D_1(P_n^{k_n})$ and $D_2^\star(P_n^{k_n})$ belong to a P_0-Donsker class with P_0-probability tending to 1, hence so does $D_1(P_n^{k_n}) + D_2^\star(P_n^{k_n})$. Therefore, by (29), (27) and Lemma 19.24 in [25], the first term is $O_P(1/\sqrt{n}) = o_P(1)$. Combining (29) and (27) with the Cauchy-Schwarz inequality yields in turn that the second term is $o_P(1)$. Finally, the law of large numbers and the fact that $\sigma_0^2 \geq c$ entail that $E_{P_n}\{X^2\}/\sigma^2(P_n^{k_n}) = \sigma^2(P_0)/\sigma_0^2 \times (1 + o_P(1))$. Consequently, we deduce from (30) that there exists $\psi_0 \in \mathbb{R}$ such that $\widetilde{\psi}_n = \psi_0 + o_P(1)$.

Because $D_1^\star(\sigma^2, \theta, \psi)(O) = D_1(\sigma^2, \theta)(O) - X^2\psi/\sigma^2$, (26) easily follows from (29) and the convergence in probability of $\widetilde{\psi}_n$ and $\sigma^2(P_n^{k_n})$ to ψ_0 and σ_0^2. Finally (26) and (27) imply (28), thus concluding the proof. □

Proof of Proposition 2. Let us first rewrite $P_n D^\star(P_n^{k_n}) = o_P(1/\sqrt{n})$ as

$$P_0 D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0) = -(P_n - P_0)D^\star(P_n^{k_n}) \\ - P_0(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0)) + o_P(1/\sqrt{n}). \quad (31)$$

Since $D_1^\star(P_n^{k_n})$ and $D_2^\star(P_n^{k_n})$ belong to a P_0-Donsker class with P_0-probability tending to 1, so does $D^\star(P_n^{k_n})$. Therefore, (28) of Lemma 8 and Lemma 19.24 in [25] yield that the first right-hand term in (31) is $O_P(1/\sqrt{n}) = o_P(1)$. Moreover, (28) of Lemma 8 and the Cauchy-Schwarz inequality imply that the second right-hand side term is $o_P(1)$. Consequently, the deterministic quantity $P_0 D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0)$ is equal to 0, and the conditions on (θ_0, μ_0, g_0) ensure that necessarily $\psi_0 = \Psi(P_0)$ *i.e.*, that the TMLE $\widetilde{\psi}_n^*$ is consistent. □

Proof of Proposition 3. Let us resume the previous proof where we left it. The fundamental relationship of this proof, derived from equalities $P_0 D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0) = 0$ and $P_n D^\star(P_n^{k_n}) = o_P(1/\sqrt{n})$, is

$$- P_0(D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0)) = (P_n - P_0)D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*) \\ + (P_n - P_0)(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*)) \\ + P_0(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*)) + o_P(1/\sqrt{n}), \quad (32)$$

where the left-hand side term obviously equals $(\widetilde{\psi}_n^* - \psi_0)\sigma^2(P_0)/\sigma_0^2$. Let us consider now the first right-hand term in (32). Since *(a)* $\{D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi) : \psi \in \mathbb{R}\}$ is a P_0-Donsker class and *(b)* $P_0(D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0))^2 = (\widetilde{\psi}_n^* - \psi_0)^2 E_{P_0}\{X^4\}/\sigma_0^4 = o_P(1)$, it holds that $(P_n - P_0)D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*) = (P_n - P_0)D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \psi_0) + o_P(1/\sqrt{n})$

by Lemma 19.24 in [25]. Regarding the second right-hand side term in (32), note *(a)* that $(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*))(O) = (D_1(P_n^{k_n}) + D_2^\star(P_n^{k_n}))(O) + ((1/\sigma_0^2 - 1/\sigma^2(P_n^{k_n}))X^2\widetilde{\psi}_n^*) - (D_1(\sigma_0^2, \theta_0) + D_2^\star(\sigma_0^2, \theta_0, \mu_0, g_0))(O)$, *(b)* that we have already shown that the first random function between parentheses belongs to a P_0-Donsker class with P_0-probability tending to 1, *(c)* that second random function between parentheses belongs to the P_0-Donsker class $\{O \mapsto (1/\sigma_0^2 - 1/\sigma^2)X^2\psi : (\psi, \sigma^2) \in \mathbb{R} \times [c, \infty]\}$, and *(d)* that the last function of the decomposition is deterministic. Therefore, $D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*)$ belongs to a P_0-Donsker class with P_0-probability tending to 1. Now, by applying repeatedly inequality $(a+b)^2 \leq 2(a^2+b^2)$ we deduce that $P_0(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*))^2 \lesssim P_0(D_1(P_n^{k_n}) - D_1(\sigma_0^2, \theta_0))^2 + P_0(D_2^\star(P_n^{k_n}) - D_2^\star(\sigma_0^2, \theta_0, \mu_0, g_0))^2 + (\widetilde{\psi}_n^*)^2 E_{P_0}\{((1/\sigma_0^2 - 1/\sigma^2(P_n^{k_n}))^2 X^4\}$. But $\|O\|$ is bounded under P_0 and $\sigma^2(P_n^{k_n}), \sigma_0^2 \geq c$, so that $E_{P_0}\{((1/\sigma_0^2 - 1/\sigma^2(P_n^{k_n}))^2 X^4\} \lesssim (\sigma^2(P_n^{k_n}) - \sigma_0^2)^2 = o_P(1)$. This fact combined with (29), (27) and $\widetilde{\psi}_n^* = O_P(1)$ yield that $P_0(D^\star(P_n^{k_n}) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*))^2 = o_P(1)$. Consequently, Lemma 19.24 in [25] implies that the second right-hand side term in (32) is $o_P(1/\sqrt{n})$. Let us turn now to the last right-hand side term in (32). It is easily seen that

$$
\begin{aligned}
&D^\star(P_n^{k_n})(O) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \widetilde{\psi}_n^*)(O) \\
&\quad = D^\star(\sigma^2(P_n^{k_n}), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0))(O) - D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \Psi(P_0))(O) \\
&\qquad - \left(1/\sigma^2(P_n^{k_n}) - 1/\sigma^2(P_0)\right) X^2(\widetilde{\psi}_n^* - \Psi(P_0)),
\end{aligned}
$$

where $(1/\sigma^2(P_n^{k_n}) - 1/\sigma^2(P_0))(\widetilde{\psi}_n^* - \Psi(P_0)) = O_P(1/\sqrt{n}) \times o_P(1) = o_P(1/\sqrt{n})$. Using that $P_0 D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \Psi(P_0)) = 0$, the previous display yields that the third right-hand side term in (32) equals

$$
\begin{aligned}
&P_0 D^\star(\sigma^2(P_n^{k_n}), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0)) + o_P(1/\sqrt{n}) \\
&\qquad = P_0 D^\star(\sigma_0^2, \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0))(1 + o_P(1)) + o_P(1/\sqrt{n}).
\end{aligned}
$$

In summary, we just showed that

$$
\begin{aligned}
(\widetilde{\psi}_n^* - \Psi(P_0))\sigma^2(P_0)/\sigma_0^2 &= (P_n - P_0)D^\star(\sigma_0^2, \theta_0, \mu_0, g_0, \Psi(P_0)) \\
&\quad + P_0 D^\star(\sigma_0^2, \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0))(1 + o_P(1)) + o_P(1/\sqrt{n}),
\end{aligned}
$$

hence the stated relationship. □

Proof of Corollary 1. This result relies on the decomposition:

$$
\begin{aligned}
&P_0 D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0)) \\
&\quad = P_0\left(D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu(P_n^{k_n}), g(P_n^{k_n}), \Psi(P_0)) - D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu_0, g_0, \Psi(P_0))\right) \\
&\qquad + P_0\left(D^\star(\sigma^2(P_0), \theta(P_n^{k_n}), \mu_0, g_0, \Psi(P_0)) - D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0))\right),
\end{aligned}
$$

where we use that $P_0 D^\star(\sigma^2(P_0), \theta_0, \mu_0, g_0, \Psi(P_0)) = 0$. Following the lines of the proof of Lemma 8 and using the Cauchy-Schwarz inequality yield that the first term of the left-hand side decomposition is upper-bounded (up to a multiplicative constant) by square-root of $P_0(\theta(P_n^{k_n})(0,\cdot) - \theta(P_0)(0,\cdot))^2 \times (P_0(\mu(P_n^{k_n}) - \mu_0)^2 + P_0(g(P_n^{k_n})(0|\cdot) - g_0(0|\cdot))^2)$, while the second term equals zero. Thus the latter left-hand side expression is $o_P(1/\sqrt{n})$ by assumption, (32) yields the asymptotic linear expansion, and the central limit theorem completes the proof. □

References

[1] Joseph Andrews, Wendy Kennette, Jenna Pilon, Alexandra Hodgson, Alan B. Tuck, Ann F. Chambers, and David I. Rodenhiser. Multi-platform whole-genome microarray analyses refine the epigenetic signature of breast cancer metastasis with gene expression and copy number. *PLoS ONE*, 5(1):e8665, 01 2010.

[2] Francis S. Collins and Anna D. Barker. Mapping the cancer genome. *Scientific American*, 296(3):50–57, Mar 2007.

[3] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *Misc Functions of the Department of Statistics (e1071), TU Wien*, 2011. URL `http://cran.r-project.org/web/packages/e1071/index.html`. R package version 1.6.

[4] T. Hastie. *Generalized additive models*, 2011. URL `http://cran.r-project.org/web/packages/gam/index.html`. R package version 1.04.1.

[5] P. A. Jones and S. B. Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, Feb 2007.

[6] C. Kooperberg. *Polynomial spline routines*, 2010. URL `http://cran.r-project.org/web/packages/polspline/index.html`. R package version 1.1.5.

[7] C. L. Lawson and R. J. Hanson. *Solving least squares problems*, volume 15. Society for Industrial Mathematics, 1995.

[8] L. M. Le Cam. *Théorie asymptotique de la décision statistique.* Séminaire de Mathématiques Supérieures, No. 33 (Été, 1968). Les Presses de l'Université de Montréal, Montreal, Que., 1969.

[9] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

[10] R. Louhimo and S. Hautaniemi. Cnamet: an r package for integrating copy number, methylation and expression data. *Bioinformatics*, 27(6):887, 2011.

[11] J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P. O Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968, Oct 2002.

[12] E. Polley and M. J. van der Laan. *SuperLearner*, 2011. URL `http://CRAN.R-project.org/package=SuperLearner`. R package version 2.0-4.

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

[14] J. M. Robins and A. Rotnitzky. Comment on Inference for semiparametric models: some questions and an answer, by Bickel, P. J. and Kwon, J. *Statistica Sinica*, 11:920–935, 2001.

[15] J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495, 1992.

[16] T. Speed. From expression profiling to putative master regulators. UC Berkeley Statistics and Genomics Seminar, February 5th, 2009.

[17] Z. Sun, Y. W. Asmann, K. R. Kalari, B. Bot, J. E. Eckel-Passow, T. R. Baker, J. M. Carr, I. Khrebtukova, S. Luo, L. Zhang, et al. Integrated analysis of gene expression, cpg island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, 6(2):e17490, 2011.

[18] The Cancer Genome Atlas (TGCA) research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.

[19] The Cancer Genome Atlas (TGCA) research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

[20] C. Tuglus and M. J. van der Laan. *Targeted Learning: Causal Inference for Observational and Experimental Data*, chapter Targeted methods for biomarker discovery. Springer Verlag, 2011.

[21] M. J. van der Laan. Statistical inference for variable importance. *Int. J. Biostat.*, 2: Article 2, 2006.

[22] M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Verlag, 2011.

[23] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *Int. J. Biostat.*, 2:Article 11, 2006.

[24] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6:Article 25, 2007.

[25] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[26] W. N. van Wieringen and M. A. van de Wiel. Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 5(1):19–29, March 2008.

[27] X. V. Wang, R. G. W. Verhaak, E. Purdom, P. T. Spellman, and T. P. Speed. Unifying gene expression measures from multiple platforms using factor analysis. *PloS one*, 6(3): e17691, 2011.

[28] Z. Yu and M. J. van der Laan. Measuring treatment effects using semiparametric models. Technical report, Division of Biostatistics, University of California, Berkeley, 2003.

(a) full-SL

iteration of the TMLE procedure	$k=0$	$k=1$	$k=2$	$k=3$
gain in relative error	0	0.0469	0.0625	0.0335
gain in relative mean square error	0	0.0365	0.0369	0.0035
Lilliefors test statistic	0.0183	0.0269	0.0298	0.0282
Lilliefors test p-value	0.5718	0.0861	0.0365	0.0582
KS test statistic	0.1566	0.0782	0.0743	0.0786
empirical coverage	–	0.896	0.905	0.898
empirical coverage (optimistic)	–	0.914	0.920	0.916

(b) light-SL

iteration of the TMLE procedure	$k=0$	$k=1$	$k=2$	$k=3$
gain in relative error	0	0.2871	0.2837	0.2866
gain in mean square error	0	0.2352	0.2293	0.2305
Lilliefors test statistic	0.0253	0.0224	0.0218	0.0295
Lilliefors test p-value	0.1251	0.2620	0.2999	0.0400
KS test statistic	0.4227	0.1327	0.1451	0.1377
empirical coverage	–	0.936	0.938	0.929
empirical coverage (optimistic)	–	0.945	0.948	0.941

Table 3: Testing the asymptotic normality of ψ_n^k and the validity of the coverage provided by $[\psi_n^k \pm \xi_{1-\alpha/2} s_n^k/\sqrt{n}]$, with $(s_n^k)^2 = \mathrm{Var}_{P_n} D^\star(P_n^k)(O)$ for $k = 0, 1, 2, 3$, **(a)** for the full-SL procedure and **(b)** for the light-SL procedure. We report the gains in relative error and mean square error (first and second rows), the test statistics and corresponding p-values of Lilliefors tests of normality (third and fourth rows), the test statistics of the KS test of normality with null mean and variance equal to $\psi_B(P^s)$ and $v_B(P^s)$ (fifth rows; the corresponding p-values are all smaller than 10^{-4}), and finally the empirical coverages $c_n^k = \frac{1}{B'} \sum_{b=1}^{B'} \mathbf{1}\{\psi_B(P^s) \in [\psi_{n,b}^k \pm \xi_{1-\alpha/2} s_{n,b}^k/\sqrt{n}]\}$ as well as their *optimistic* counterparts $c_n^{k+} = \frac{1}{B'} \sum_{b=1}^{B'} \mathbf{1}\{[\psi_B(P^s) \pm \xi_{0.975}\sqrt{v_B(P^s)/B}] \cap [\psi_{n,b}^k \pm \xi_{1-\alpha/2} s_{n,b}^k/\sqrt{n}] \neq \emptyset\}$ (sixth and seventh rows).