University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 180

Estimating Function Based Cross-Validation and Learning

Mark J. van der Laan* Daniel Rubin†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

http://biostats.bepress.com/ucbbiostat/paper180

Estimating Function Based Cross-Validation and Learning

Mark J. van der Laan and Daniel Rubin

Abstract

Suppose that we observe a sample of independent and identically distributed realizations of a random variable. Given a model for the data generating distribution, assume that the parameter of interest can be characterized as the parameter value which makes the population mean of a possibly infinite dimensional estimating function equal to zero. Given a collection of candidate estimators of this parameter, and specification of the vector estimating function, we propose cross-validation criteria for selecting among these estimators. This cross-validation criteria is defined as the Euclidean norm of the empirical mean over the validation sample of the estimating function at the candidate estimator based on the training sample. We establish a finite sample inequality of this method relative to an oracle selector, and illustrate it with some examples. This finite sample inequality provides us with asymptotic equivalence of the selector with the oracle selector under general conditions. We also study the performance of this method in the case that the parameter of interest itself is path-wise differentiable (and thus, in principle, root-n estimable), and show that the cross-validated selected estimator is typically efficient, and, at certain data generating distributions, superefficient (and thus non-regular). Finally, we combine 1) the selection of sequence of subspaces of the parameter space (i.e., a sieve), 2) the estimating equation as empirical criteria to generate a candidate estimator for each subspace, and 3) estimating function based cross-validation selector to select among the candidate estimators, in order to provide a new unified estimating function based methodology. In particular, we formally establish a finite sample inequality for this general estimator in the case that one uses epsilon-nets as sieve, and point out that this finite sample inequality corresponds with minimax adaptive rates of convergence w.r.t. to the norm implied by the estimating function.

1. Introduction

Suppose that one observes n independent and identically distributed copies of a random variable (referred to as the experimental unit), and that one wishes to learn from this data set a parameter of the distribution of this random variable. It is often possible to generate a (potentially large) set of estimators of this parameter of interest indexed by one or more fine tuning parameters such as quantaties measuring the degree in which the estimator is informed by the actual data (instead of by priors or modelling assumptions). These estimators differ in variance and bias. An important problem in statistics is the construction of a selector based on the data, which selects among these candidate estimators, so that the corresponding data adaptively selected estimator behaves asymptotically well relative to an optimally selected estimator based on actually knowing the truth. The latter type of selector is often referred to as an oracle selector.

Cross-validation is a particular selector which has been extensively studied in the past in the context of density estimation (for example, bandwidth selection for kernel density estimators, and model selection for model-specific maximum likelihood estimators), and regression. As discussed by [4] in the context of dimensionality selection in regression, criteria such as Mallow's C_p, Akaike information's criterion (AIC), and the Bayesian information criterion (BIC), do not account for the data-driven selection of the sequence of models and thus provide biased assessment of prediction error in finite sample situations. Instead, risk estimation methods based on sample reuse have been favored. The main procedures include: leave-one-out cross-validation, V-fold cross-validation, Monte Carlo cross-validation, and the bootstrap (Chapter 3 in [7], [8], [5,6], Chapter 17 in [11], Chapters 7 and 8 in [12], Chapter 7 in [17], Chapter 3 in [21], [23], [24]). Thus, a variety of cross-validation procedures are available for estimating the risk of a predictor. A natural question then concerns the distributional properties of the resulting risk estimators, i.e., their performance as estimators of generalization error, their performance in terms of identifying a good predictor (model selection), and also the impact of the particular cross-validation procedure (e.g., the choice of V in V-fold cross-validation, the use of V-fold vs. Monte Carlo cross-validation). Aside from empirical assessment of different estimation procedures, most of the previous theoretical work has focused primarily on the distributional properties of leave-one-out cross-validation [23,24].

There is a rich literature on leave-one-out cross-validation in nonparametric univariate regression. For example, [22] proposes a fast approximation

of the leave-one out cross-validation method in spline regression. We refer to [14] for an overview on the leave-one-out cross-validation method in kernel regression. In particular, [15,16] establish an asymptotic optimality result for leave-one-out cross-validation for choosing the smoothing parameter in nonparametric kernel regression (see page 158, [14]).

[12] established a finite sample result for the single-split cross-validation selector for the squared error loss function. Their theorem was generalized in [10] to general cross-validation schemes and a general class of loss functions. [10] examine the distributional properties of cross-validated risk estimators in the context of both predictor selection and predictor performance assessment for a general class of loss functions.

Finite sample inequalities and asymptotic optimality results for likelihood-based V-fold cross-validation, or equivalently, cross-validation for the purpose of selection among density estimators of the density of the observed data, are established in [29], in which we also provide an overview of the literature on likelihood based cross-validation (see also [22]), which is omitted here for the sake of space.

These cross-validation methods are focussed on regression or density estimation. In [28] we generalized cross-validation to a selector among candidate estimators of any parameter which is represented as the minimizer over the parameter space of an expectation of a loss function of the experimental unit and a candidate parameter value, where the loss function is possibly indexed by an unknown nuisance parameter of the true data generating distribution. It is illustrated that this unified loss-based cross-validation approach solves a wide range of estimator selection problems, including estimator selection based on censored data, and, in particular, estimator selection in causal inference problems (in causal inference the observed data structure is modelled as a missing data structure on potential counterfactual random variables). The appropriate loss function for censored data structures is defined as the Inverse Probability of Censoring Weighted (IPCW) or double robust IPCW full data loss function, where these mappings from full data functions to functions of the censored data are defined in [32]. Various applications of this unified cross-validation methodology for estimator selection are selection among regression estimators [10], estimator selection with right censored data [18], likelihood-based cross-validation [29], and tree-based estimation and model selection with censored data [20].

In general, for any parameter of interest, we proposed a methodology to construct an appropriate loss function. Our proposed construction of the loss function involves first building a risk function of candidate para-

meter values which is minimized at the true parameter value, where this risk function is defined as a norm of the expectation (under the true data generating distribution) of an estimation function: that is, a function of the experimental unit, parameter of interest, and possibly a nuisance parameter, with expectation zero at the true parameter values. Subsequently, one determines a loss function such that its expectation equals the risk function, and preferably among such loss function we prefer the loss function for which the empirical mean of the loss function actually corresponds with an efficient estimator of the risk function. However, our definition of this risk function in terms of a vector-valued estimating function suggests a direct application of cross-validation to this risk function, instead of constructing a loss function whose expectation equals this risk function. This inspired us to a so called estimating function based cross-validation methodology, presented in this article. For many problems it is easy to generate estimating functions identifying the parameter of interest, while it requires an additional possibly involved step to compute a corresponding loss function. Therefore, this new estimating function based cross-validation methodology provides an important new general method for estimator selection.

The organization of this article is as follows. Firstly, in Section 2 we present our proposed estimating function based cross-validation selector. In Section 3 we establish the asymptotic performance of this cross-validation selector in the case that the parameter of interest is finite dimensional and path-wise differentiable (and thus root-n estimable, in principle). We illustrate the method with a set of examples, and, in particular, show that the cross-validation selector data adaptively under-smooths appropriately so that the resulting estimator is still asymptotically efficient, or at some data generating distributions, is asymptotically superefficient. In Section 4 we derive a general finite sample inequality for its performance relative to the oracle selector, where we note that this result is most relevant for parameters which cannot be estimated at the root-n rate. In Section 5 we provide a variety of corollaries of this finite sample inequality, and in Section 6 we illustrate our method in the classical regression and density estimation example. In Section 7 we provide a general estimating function based estimator based on 1) a sieve on the parameter space, 2) the norm of the estimating equation as empirical criteria to generate a candidate estimator for each element of the sieve, and 3) estimating function based cross-validation to select the element of the sieve. In Section 8 we prove a finite sample inequality for this type of estimator in the case that the sieve is a collection of subspace-specific epsilon-nets corresponding with a sequence

of subspaces, indexed by the subspace and the epsilon. This finite sample inequality shows that our estimator does achieve wished mini-max adaptive rates of convergence w.r.t to an estimating function based dissimilarity measure. We conclude with a discussion. Some fundamental lemmas are presented in the appendix.

2. Estimating function based cross-validation.

Suppose we observe n i.i.d. random variables $O_1, \ldots, O_n$ with common distribution P_0. Let $\mathcal{M}$ be the statistical model: that is, it is known that $P_0 \in \mathcal{M}$. Suppose that $\Psi : \mathcal{M} \to D(\mathcal{S})$ is the parameter of interest, where $D(\mathcal{S})$ denotes a space of real valued functions on a set $\mathcal{S}$. For example, if $\mathcal{S} = \{1, \ldots, d\}$, then $D(\mathcal{S}) = \mathbb{R}^d$ is simply the Euclidean space, but if $\mathcal{S}$ is a Euclidean set in $\mathbb{R}^d$, then $D(\mathcal{S})$ denotes the class of real valued d-variate functions. Let $\mathbf{\Psi} \equiv \{\Psi(P) : P \in \mathcal{M}\} \subset D(\mathcal{S})$ denote the parameter space.

Let $(O, \psi, \gamma) \to D_b(O \mid \psi, \gamma)$ be an estimating function indexed by a b ranging over a countable set $\mathcal{B}$. We note that an estimating function is simply a well defined real valued function on the tensor product of a support of the observed data structure O, the parameter space $\mathbf{\Psi}$, and a nuisance parameter space. Suppose that this set of estimating functions are unbiased in the sense that

$$E_0 D_b(O \mid \Psi(P_0), \Gamma(P_0)) = 0 \text{ for all } b \in \mathcal{B},$$

where $\Gamma : \mathcal{M} \to \{\Gamma(P) : P \in \mathcal{M}\}$ is the nuisance parameter, and $\Gamma(P_0)$ denotes its true value. Let $D(O \mid \psi, \gamma) \equiv (D_b(O \mid \psi, \gamma) : b \in \mathcal{B})$ denote the vector-valued (possibly infinite dimensional) estimating function. The heuristic of our method requires that the estimating functions D_b are appropriately standardized so that for each $b \in \mathcal{B}$

$$P_0 D_b(O \mid \psi, \Gamma(P_0)) = -(\psi_b - \psi_{b0}) + o(\mid \psi_b - \psi_{b0} \mid), \tag{1}$$

for real valued parameters ψ_b of ψ, where ψ_{b0} denotes the true parameter value. That is, formally, $\psi_b = \Phi_b(\psi, P_0)$, and $\psi_{b0} = \Phi_b(\psi_0, P_0)$ for some real valued mapping Φ_b. In the next subsection, we present a general method for construction of such a vector-valued estimating function $D(O \mid \psi, \gamma)$ in which the estimating function D_b is directly derived from the b-specific efficient influence curve for a pathwise differentiable parameter ψ_{b0}, $b \in \mathcal{B}$. In particular, we will point out that (1) can be arranged to hold exactly (no remainder) in a large class of problems in which (e.g.) Ψ_b is a linear parameter on a convex model $\mathcal{M}$, thereby covering many models for censored data structures.

Let $\| \cdot \|$ denote a particular norm on vectors $(x(b) : b \in \mathcal{B})$ with real valued components $x(b)$ of the dimension $| \mathcal{B} |$. If $\mathcal{B}$ is infinite (but countable), then $\| \cdot \|$ denotes a norm on the infinite dimensional Euclidean space R^∞. For example, we could use the weighted euclidean norm

$$\| x \| = \sqrt{\sum_b w(b) x(b)^2}$$

for a known weight function $b \to w(b) \geq 0$ such that $\sum_b w(b) < \infty$. We define a risk function at P_0 as a norm of the expectation (under P_0) of the vector estimating function $D(O \mid \psi, \Gamma(P_0)) = (D_b(O \mid \psi, \Gamma(P_0)) : b)$:

$$\Theta(\psi \mid P_0) \equiv \| P_0 D(O \mid \psi, \Gamma(P_0)) \| .$$

Here we used the notation $P_0 f = \int f(o) dP_0(o)$, and if $f = (f_1, \ldots, f_d)$ is a vector function, then $P_0 f = (P_0 f_1, \ldots, P_0 f_d)$. For example, we can define a risk function at P_0 as the weighted Euclidean norm of the expectation (under P_0) of the estimating function at candidate $\psi \in \mathbf{\Psi}$:

$$\Theta(\psi \mid P_0) \equiv \sqrt{\sum_{b \in \mathcal{B}} w(b) P_0^2 D_b(O \mid \psi, \Gamma(P_0))},$$

where $b \to w(b) \geq 0$ is a known weight function. We note that, if $\mathcal{B}$ is infinite, then $\Theta(\psi \mid P_0)$ needs to be defined as an infinite sum, and thereby as a limit.

Note that $\Theta(\psi_0 \mid P_0) = 0$ so that the corresponding measure of dissimilarity $d(\psi, \psi_0) \equiv \Theta(\psi \mid P_0) - \Theta(\psi_0 \mid P_0)$ between ψ and ψ_0 is indeed always non-negative and minimized at ψ_0:

$$d(\psi, \psi_0) \geq 0, \text{ and } d(\psi, \psi_0) = 0 \text{ if } \psi = \psi_0.$$

In addition, because of property (1), we have for $\psi = \Psi(P)$ close to ψ_0,

$$\Theta(\psi \mid P_0) = \| (\psi_b : b) - (\psi_{b0} : b) + (o(| \psi_b - \psi_{b0} |) : b) \|, \qquad (2)$$

where $\psi_b = \Phi_b(\psi, P)$ and $\psi_{b0} = \Phi_b(\psi_0, P_0)$. For example, if we use the Euclidean norm, then locally one expects to have

$$\Theta(\psi \mid P_0) \approx \sqrt{\sum_{b \in \mathcal{B}} w(b)(\psi_b - \psi_{b0})^2}.$$

In particular, if (1) holds exactly, then we have equality:

$$\Theta(\psi \mid P_0) = \| (\psi_b : b) - (\psi_{b0} : b) \| .$$

This first order expansion of this risk function $\psi \to \Theta(\psi \mid P_0)$ at ψ_0 suggests that $\Theta(\psi \mid P_0)$ is a sensible risk function for the purpose of estimation of $\psi_0 = \Psi(P_0)$, and, in particular, for selecting among candidate estimators of ψ_0.

Specifically, let $B_n \in \{0,1\}^n$ be the random variable defining the cross-validation scheme, where $B_n(i) = 1$ indicates that observation i is a member of the validation sample, and $B_n(i) = 0$ indicates that observation i is a member of the training sample. Let $p \equiv P(B_n(i) = 1)$ denote the proportion of the learning sample which constitutes the validation sample. It is assumed that B_n is independent of the learning sample $(O_1, \ldots, O_n)$. Given a realization b_n of B_n, let P^0_{n,b_n} and P^1_{n,b_n} denote the empirical distribution of the training sample and validation sample, respectively.

Given candidate estimators $P_n \to \hat{\Psi}_k(P_n)$, $k = 1, \ldots, K(n)$, the cross-validated risk function is now defined as

$$\hat{\Theta}_{n(1-p)}(k) \equiv E_{B_n} \parallel P^1_{n,B_n} D(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})) \parallel .$$

For example, in case we use the Euclidean norm, then we have

$$\hat{\Theta}_{n(1-p)}(k) \equiv E_{B_n} \sqrt{\sum_{b \in \mathcal{B}} \left(\sum_{i:B_n(i)=1} D_b(O_i \mid \hat{\Psi}_k(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n}))/np \right)^2 w(b)}.$$

This cross-validated risk function defines our proposed cross-validation selector

$$k_n = K(P_n) \equiv \arg\min_k \hat{\Theta}_{n(1-p)}(k).$$

This finishes the description of our proposed cross-validation selector among candidate estimators of a parameter $\psi_0 = \Psi(P_0)$.

Benchmark selector.

A natural way to benchmark the selector k_n is to define the following true conditional risk function

$$\tilde{\Theta}_{n(1-p)}(k) = E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \gamma_0) \parallel . \tag{3}$$

If we use the Euclidean norm, then this equals

$$\tilde{\Theta}_{n(1-p)}(k) = E_{B_n} \sqrt{\sum_{b \in \mathcal{B}} \left(P_0 D_b(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \gamma_0) \right)^2 w(b)}. \tag{4}$$

We can now define the corresponding oracle selector

$$\tilde{k} = \tilde{k}_{n(1-p)} = \tilde{K}_{n(1-p)}(P_n) \equiv \arg\min_k \tilde{\Theta}_{n(1-p)}(k)$$

for selecting among estimators based on a sample of size $n(1-p)$. In particular, if (1) holds exactly, then

$$\tilde{k} = \arg\min_k E_{B_n} \parallel (\Phi_b(\hat{\Psi}_k(P^0_{n,B_n}), P_0) : b) - (\Phi_b(\psi_0, P_0) : b) \parallel .$$

Finally, we also define the wished oracle selector

$$\tilde{k}_n = \tilde{K}(P_n) \equiv \arg\min_k \tilde{\Theta}_n(k),$$

where

$$\tilde{\Theta}_n(k) \equiv \parallel P_0 D_b(\cdot \mid \hat{\Psi}_k(P_n), \gamma_0) \parallel,$$

which compares estimators based on the whole learning sample P_n. In the case of the Euclidean norm this equals

$$\tilde{\Theta}_n(k) \equiv \sqrt{\sum_{b \in \mathcal{B}} \left(P_0 D_b(\cdot \mid \hat{\Psi}_k(P_n), \gamma_0) \right)^2 w(b)}.$$

2.1. *Method for construction of vector-estimating function.*

In this subsection we present a general method for constructing such a vector estimating function $D(O \mid \psi, \gamma)$. Firstly, one specifies a collection of real valued pathwise differentiable parameters $\Psi_b : \mathcal{M} \to \mathbb{R}$ indexed by $b \in \mathcal{B}$, so that for all $P \in \mathcal{M}$, $(\Psi_b(P) : b \in \mathcal{B})$ identifies $\Psi(P)$ uniquely. For example, if $\Psi(P) = (\Psi_1(P), \ldots, \Psi_d(P))$ is itself already a Euclidean pathwise differentiable parameter, then we would simply define $\Psi_b(P)$ as the b-th component of $\Psi(P)$, $b = 1, \ldots, d$. On the other hand, if $\Psi(P)$ is an infinite dimensional function, then one could, for example, define $\Psi_b(P)$ as the inner product of $\Psi(P)$ with a b-specific basis function. It is assumed that for each $b \in \mathcal{B}$ $\Psi_b(P) = \Phi_b(\Psi(P), P)$ for all $P \in \mathcal{M}$ for some mapping Φ_b. The definition of pathwise differentiability states that Ψ_b is pathwise differentiable at P, relative to a specified set of one-dimensional differentiable submodels $\{P_{\epsilon,s} : \epsilon \in (-\delta, \delta)\} \subset \mathcal{M}$, satisfying $P_{0,s} = P$, with score at $\epsilon = 0$ equal to s, $s \in \mathcal{S} \subset L_0^2(P)$, if for all these submodels

$$\frac{d}{d\epsilon} \Psi_b(P_{\epsilon,s}) \Big|_{\epsilon=0} = \langle \ell_b, s \rangle_P \equiv E_P \ell_b(O) s(O)$$

for some $\ell_b \in L_0^2(P)$. We recall that $L_0^2(P)$ is the Hilbert space of real valued functions of O with mean zero endowed with inner product

$\langle h_1, h_2 \rangle_P \equiv E_P h_1(O) h_2(O)$ being the covariance operator. Here ℓ_b is called a gradient of the pathwise derivative, whose projection onto the tangent space, that is, the closure of the linear span of $\mathcal{S}$ within the Hilbert space $(L_0^2(P), \langle \cdot, \cdot \rangle_P)$, is unique, and this projection is called the canonical gradient. The canonical gradient is also called the efficient influence curve since a regular asymptotically linear estimator with influence curve equal to $D_b(O \mid P_0)$ is asymptotically efficient, by efficiency theory [3]. We refer to [3] for a comprehensive treatment of efficiency theory and illustration with many semiparametric models, and, in the context of censored data, we refer to [32] who provide general representations of the class of all gradients, and, of the canonical gradient/efficient influence curve. Let $D_b(O \mid P_0)$ be a particular gradient, such as this unique canonical gradient. This gives us now a class of functions $D_b(O \mid P_0)$, $b \in \mathcal{B}$, and it is known that $E_{P_0} D_b(O \mid P_0) = 0$ for all $b \in \mathcal{B}$, and all $P_0 \in \mathcal{M}$.

Given a gradient representation $D_b(O, P)$ for all $P \in \mathcal{M}$, it is often not hard to define an actual estimating function $(O, \psi_b, \rho_b) \to D_b(O \mid \psi_b, \rho_b)$ for the parameter ψ_b, possibly depending on a nuisance parameter ρ_b such that $D_b(O \mid \Psi_b(P_0), \rho_{b0}) = D_b(O \mid P_0)$ for all $P_0 \in \mathcal{M}$, where ρ_{b0} denotes the true value of the nuisance parameter. Since a gradient is, by definition, orthogonal to all nuisance scores, that is, scores of one-dimensional submodels for which $\frac{d}{d\epsilon} \Psi_b(P_{\epsilon,s})\big|_{\epsilon=0} = 0$, this estimating function $D_b(O \mid \psi_b, \rho_b)$ for Ψ_b is *minimally dependent on nuisance parameters*: we refer to chapter 1, sections 1.4, Lemma 1.2 and 1.3 in [32] for formal results establishing that the directional derivatives w.r.t. ρ_b are zero, given that ψ_b and ρ_b are variation independent parameters.

In addition, as a consequence of the fact that $D_b(O \mid P_0)$ is a gradient of the pathwise derivative of the parameter $\Psi_b : \mathcal{M} \to \mathbb{R}$, the estimating function will have the property

$$E_0 D_b(O \mid \psi_b, \rho_{b0}) = -(\psi_b - \psi_{b0}) + o(\mid \psi_b - \psi_{b0} \mid). \tag{5}$$

This is a general property for gradients (also called influence curves) of a pathwise derivative, and, in particular, for the canononical gradient (see Lemmas 1.2 and 1.3 in [32]). In particular, for linear parameters Ψ_b on convex models, one obtains an exact equality (see [19], Chapter 2 [25], [26], and [27]):

$$E_0 D_b(O \mid \psi_b, \rho_{b0}) = -(\psi_b - \psi_{b0}). \tag{6}$$

The fact that the derivative at ψ_{b0} equals minus the identity provides us with the motivation for our proposed risk function.

Finally let $(O, \psi, \gamma) \to D_b(O \mid \psi, \gamma)$ be an estimating function satisfying

$$D_b(O \mid \Psi(P_0), \Gamma(P_0)) = D_b(O \mid P_0), \text{ for all } b \in \mathcal{B}, P_0 \in \mathcal{M}.$$

At this step, one will use that $\Psi_b(P) = \Phi_b(\Psi(P), P)$ in order to represent an estimating function in ψ_b (as implied by $D_b(O \mid P_0)$) in terms of an estimating function in terms of ψ.

By now, we have succeeded in deriving a class of unbiased estimating functions $(O, \psi, \gamma) \to D_b(O \mid \psi, \gamma)$, with nuisance parameter γ, indexed by $b \in \mathcal{B}$, satisfying the wished property (2).

3. Euclidean pathwise differentiable parameters.

Consider candidate estimators $\hat{\Psi}_h(P_n)$ of a pathwise differentiable d-dimensional parameter ψ_0 of P_0 indexed by a continuous univariate index h. Let ψ_0 be identified as the parameter which results in a population mean of an estimating function $D(O \mid \psi, \gamma_0) = (D_b(O \mid \psi, \gamma_0) : b = 1, \ldots, d)$ equal to zero, where $\gamma_0 = \Gamma(P_0)$ is the true nuisance parameter value. For example, $\hat{\Psi}_h(P_n)$ could be an integrated kernel density estimator with bandwidth h of a cumulative distribution function ψ_0, which is an example we discuss in detail below. Our results can be generalized to multivariate continuous index h.

Let $\hat{\Gamma}(P_n)$ be an estimator of the nuisance parameter $\Gamma(P_0)$. Since ψ_0 is a nice smooth parameter, under regularity conditions, the solution $\hat{\Psi}_0(P_n)$ of

$$P_n D(\cdot \mid \psi, \hat{\Gamma}(P_n)) = \frac{1}{n} \sum_{i=1}^{n} D(O_i \mid \psi, \hat{\Gamma}(P_n)) = 0 \text{ or } (o_P(1/\sqrt{n})), \qquad (7)$$

will already be asymptotically linear with influence curve $-\frac{d}{d\psi} E_0 D(O \mid \psi, \gamma_0)\Big|_{\psi=\psi_0}^{-1} D(O \mid \psi_0, \gamma_0)$. In particular, if $D(O \mid \psi_0, \gamma_0)$ equals the efficient influence curve, then $\hat{\Psi}_0(P_n)$ is an efficient estimator. It is assumed that $\hat{\Psi}_0(P_n)$ is asymptotically unbiased in the sense that

$$E_P D(O \mid \hat{\Psi}_0(P), \Gamma(P)) = 0 \text{ for all } P \in \mathcal{M}. \qquad (8)$$

That is, $\hat{\Psi}_0$ is an estimator which does not use any smoothing, or more precisely, does not use asymptotically relevant smoothing. The purpose of smoothing (that is, selecting a $h > 0$) is to obtain a finite sample improvement relative to $\hat{\Psi}_0$. For example, if it is known that the true cumulative distribution function is very smooth, then it makes sense to use a smooth

estimator, even though the discrete empirical cumulative distribution function is already asymptotically efficient in a nonparametric model.

Suppose that the parametrization $\hat{\Psi}_h$ in terms of h is such that

$$\frac{d}{dh} P_0 D(\cdot \mid \hat{\Psi}_h(P_0), \gamma_0)\Big|_{h=0} \neq 0. \tag{9}$$

Typically this requires selecting a parametrization such that $h = O(\hat{\Psi}_h(P_0) - \psi_0)$, and $\hat{\Psi}_h(P_0) - \psi_0 = O(h)$ for $h \to 0$. That is, h represents the order of the asymptotic bias of the estimator $\hat{\Psi}_h(P_n)$.

Consider the square of our proposed cross-validated risk function based on the euclidean norm of the estimating function:

$$\hat{\Theta}_{n(1-p)}(h)^2 \equiv E_{B_n} \sum_{b=1}^{d} \left(P^1_{n,B_n} D_b(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})) \right)^2.$$

Note that this represents a slight modification of our general proposal in the sense that we put the E_{B_n} inside the square root (but still outside the squares). This modification does not interfere with the heuristic behind our method in the sense that the corresponding true risk function is still the same, and it just simplifies the algebraic manipulations. Let $h_n = \arg\min_h \hat{\Theta}_{n(1-p)}(h)$ be its minimizer.

Our goal is to show that under general conditions $h_n = o_P(1/\sqrt{n})$, and thereby that $\hat{\Psi}_{h_n}(P_n)$ will be asymptotically equivalent with $\hat{\Psi}_0(P_n)$. That is, our data adaptive procedure for choosing h guarantees enough undersmoothing so that root-n times the bias of the selected estimator converges to zero when sample size converges to infinity. In particular, if $\hat{\Psi}_0(P_n)$ is an asymptotically efficient estimator, the smoothed estimator $\hat{\Psi}_{h_n}(P_n)$ will also be asymptotically efficient. In the next subsection 3.1 we provide the formal theorem and its proof. Subsequently, we show that one of the main conditions of the theorem is indeed a condition one expects to hold under regularity conditions. In the last three subsections we provide two specific and a general class of examples.

3.1. *Theorem.*

The analysis of h_n is based on the derivative $U(h, P_n) = d/dh \hat{\Theta}_{n(1-p)}(h)^2$, which is given by

$$2 \sum_{b=1}^{d} E_{B_n} E_{P^1_{n,B_n}} D_b(O \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})) \frac{d}{dh} E_{P^1_{n,B_n}} D_b(O \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})). \tag{10}$$

Under weak regularity conditions, by definition of h_n, we have that $U(h_n, P_n) = 0$.

Consider now the equation $U(h, P_0) = 0$, where we represent P_0 as the empirical P_n with $n = \infty$, which is thus given by

$$U(h, P_0) = 2\sum_{b=1}^{d} E_{P_0} D_b(O \mid \hat{\Psi}_h(P_0), \gamma_0) \frac{d}{dh} E_{P_0} D_b(O \mid \hat{\Psi}_h(P_0), \gamma_0).$$

Here it is assumed that $\hat{\Gamma}$ is a consistent estimator so that $\hat{\Gamma}(P_{n=\infty}) = \gamma_0$. We note that $h_0 = 0$ is a solution of $U(h, P_0) = 0$. The equations $U(0, P_0) = 0$ and $U(h_n, P_n) = 0$ provides us with a basis for establishing the asymptotic rate of convergence of h_n to $h_0 = 0$.

Firstly, we observe that, by (8), for any P (treated as the empirical distribution from P for $n = \infty$), we have

$$U(0, P) = 2\sum_{b=1}^{d} E_P D_b(O \mid \hat{\Psi}_0(P), \Gamma(P)) \left. \frac{d}{dh} E_P D_b(O \mid \hat{\Psi}_h(P), \Gamma(P)) \right|_{h=0} = 0,$$

uniformly in $P \in \mathcal{M}$. Secondly, again by (8), we also have

$$\left. \frac{d}{dh} U(h, P_0) \right|_{h=0} = 2\sum_{b=1}^{d} \left. \frac{d}{dh} E_{P_0} D_b(O \mid \hat{\Psi}_h(P_0), \gamma_0) \right|_{h=0}^{2}.$$

By (9), this derivative is strictly positive. A standard M-estimator analysis for h_n, following the approach outlined in [?], now suggests that $h_n = o_P(1/\sqrt{n})$ under regularity conditions.

Specifically, such an analysis proceeds as follows:

Step 1: Firstly, we note that

$$U(h_n, P_0) - U(0, P_0) = -\{U(h_n, P_n) - U(h_n, P_0)\}.$$

Step 2: Assume that we already have established that h_n converges to zero in probability. Now, by differentiability of $h \to U(h, P)$ at $h_0 = 0$, and the fact that this derivative is positive, it follows

$$h_n = \left(\left. \frac{d}{dh} U(h, P_0) \right|_{h=0} \right)^{-1} \{U(h_n, P_n) - U(h_n, P_0)\} + o(h_n).$$

Step 3: Assume that $\{U(h_n, P_n) - U(h_n, P_0)\} - \{U(0, P_n) - U(0, P_0)\} = o_P(1/\sqrt{n}) + o_P(h_n)$. Then, it follows that

$$h_n = \left(\left. \frac{d}{dh} U(h, P_0) \right|_{h=0} \right)^{-1} \{U(0, P_n) - U(0, P_0)\} + o_P(h_n) + o_P(1/\sqrt{n}).$$

Step 4: By condition (7), on $\hat{\Psi}_0(P_n)$, under regularity conditions, one will have $U(0, P_n) = U(0, P_n) - U(0, P_0) = o_P(1/\sqrt{n})$: below we provide a template for showing that $U(0, P_n) = o_P(1/\sqrt{n})$, which illustrates that this condition can indeed be expected to hold.

Now, it follows that

$$h_n = o_P(h_n) + o_P(1/\sqrt{n}),$$

which implies the wished result $h_n = o_P(1/\sqrt{n})$.

This proves the following general theorem.

Theorem 1: Consider candidate estimators $\hat{\Psi}_h(P_n)$ of a pathwise differentiable d-dimensional parameter ψ_0 of P_0. Suppose that $\hat{\Psi}_0(P_n)$ is based on an unbiased representation $\hat{\Psi}_0$ in the sense that

$$E_P D_b(O \mid \hat{\Psi}_0(P), \Gamma(P)) = 0 \text{ for all } P \in \mathcal{M}, \tag{11}$$

and that

$$U(0, P_n) - U(0, P_0) = o_P(1/\sqrt{n}), \tag{12}$$

where $U(h, P_n)$ is defined as

$$2\sum_{b=1}^{d} E_{B_n} P^1_{n,B_n} D_b(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})) \frac{d}{dh} P^1_{n,B_n} D_b(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})),$$

and $U(h, P_0)$ is given by

$$2\sum_{b=1}^{d} P_0 D_b(\cdot \mid \hat{\Psi}_h(P_0), \Gamma(P_0)) \frac{d}{dh} P_0 D_b(\cdot \mid \hat{\Psi}_h(P_0), \Gamma(P_0)),$$

Assume

$$\frac{d}{dh} P_0 D_b(O \mid \hat{\Psi}_h(P_0), \Gamma(P_0))\Big|_{h=0} \neq 0 \; b = 1, \ldots, d. \tag{13}$$

Define the cross-validated risk function as

$$\hat{\Theta}_{n(1-p)}(h) \equiv \sqrt{E_{B_n} \sum_{b=1}^{d} \left(P^1_{n,B_n} D_b(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n})) \right)^2}.$$

Let $h_n = \arg\min_h \hat{\Theta}_{n(1-p)}(h)$ be its minimizer. Consider the derivative $U(h, P_n)$ of $\hat{\Theta}_{n(1-p)}(h)^2$. Suppose that h_n satisfies $U(h_n, P_n) = 0$, and that $h_n = o_P(1)$.

Assume the derivative of $h \to U(h,P)$ at $h_0 = 0$ exists and (is thus) given by

$$\frac{d}{dh}U(h,P_0)\Big|_{h=0} = 2\sum_{b=1}^{d}\frac{d}{dh}P_0 D_b(\cdot \mid \hat{\Psi}_h(P_0),\gamma_0)\Big|_{h=0}^2.$$

Assume that

$$\{U(h_n,P_n) - U(h_n,P_0)\} - \{U(0,P_n) - U(0,P_0)\} = o_P(1/\sqrt{n}) + o_p(h_n). \tag{14}$$

Then $h_n = o_P(1/\sqrt{n})$.

3.2. *Why is (12) a reasonable condition?*

Firstly, we note that

$$U(0,P_n) = 2\sum_{b=1}^{d} E_{B_n}P^1_{n,B_n}D_b(O \mid \hat{\Psi}_0(P^0_{n,B_n}),\hat{\Gamma}(P^0_{n,B_n}))D_b(P_0)$$

$$+2\sum_{b=1}^{d} E_{B_n}P^1_{n,B_n}D_b(O \mid \hat{\Psi}_h(P^0_{n,B_n}),\hat{\Gamma}(P^0_{n,B_n}))\{D_b(B_n,P_n) - D_b(P_0)\},$$

where

$$D_b(B_n,P_n) = \frac{d}{dh}P^1_{n,B_n}D_b(O \mid \hat{\Psi}_h(P^0_{n,B_n}),\hat{\Gamma}(P^0_{n,B_n}))\Big|_{h=0}$$

$$D_b(P_0) = \frac{d}{dh}P_0 D_b(O \mid \hat{\Psi}_h(P_0),\gamma_0))\Big|_{h=0}.$$

Regarding the second term, under regularity conditions, $D_b(B_n,P_n) - D_b(P_0)$ can be shown to be $O_P(1/\sqrt{n})$, while, for n converging to infinity, $P^1_{n,B_n}D_b(O \mid \hat{\Psi}_0(P^0_{n,B_n}),\hat{\Gamma}(P^0_{n,B_n}))$ should converge to zero in probability (typically, at rate $1/\sqrt{n}$), since $\hat{\Psi}_0$ and $\hat{\Gamma}$ are consistent estimators. As a consequence, the second term can be expected to be $o_P(1/\sqrt{n})$. Since $D_b(P_0)$ does not depend on B_n, the first term would be $o_P(1/\sqrt{n})$ as well, if one can show that

$$E_{B_n}P^1_{n,B_n}D_b(O \mid \hat{\Psi}_0(P^0_{n,B_n}),\hat{\Gamma}(P^0_{n,B_n})) = o_P(1/\sqrt{n}). \tag{15}$$

We will now show that the latter condition (15) can indeed be expected to hold if $\hat{\Psi}_0(P_n)$ solves $P_n D(\cdot \mid \psi,\hat{\Gamma}(P_n)) = 0$, and $\hat{\Psi}_0(P_n)$ is (as a consequence) asymptotically linear with influence curve $c^{-1}D(O \mid \psi_0,\gamma_0)$, where $c \equiv -d/d\psi_0 P_0 D(O \mid \psi_0,\gamma_0)$.

Consider the one-step estimator based on initial estimator $\hat{\Psi}_0(P_n)$, using sample splitting, defined as

$$\hat{\Psi}_1(P_n) \equiv E_{B_n}\hat{\Psi}_0(P^0_{n,B_n}) + E_{B_n}P^1_{n,B_n}c_n^{-1}D(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})),$$

where c_n denotes the empirical counterpart of the derivative matrix c. This estimator is also asymptotically linear with influence curve $c^{-1}D(O \mid \psi_0, \gamma_0)$, under even weaker regularity conditions than needed to establish this asymptotic linearity result for $\hat{\Psi}_0(P_n)$ (see e.g. [19], [25], page 44). As a consequence, $\sqrt{n}(\hat{\Psi}_1(P_n) - \psi_0) - \sqrt{n}(\hat{\Psi}_0(P_n) - \psi_0)$ converges to zero in probability: that is, $\hat{\Psi}_1(P_n) - \hat{\Psi}_0(P_n) = o_P(1/\sqrt{n})$.

However, by definition,

$$\hat{\Psi}_1(P_n) - E_{B_n}\hat{\Psi}_0(P^0_{n,B_n}) = E_{B_n}P^1_{n,B_n}c_n^{-1}D(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})).$$

Thus, condition (15) follows if we can show that $\hat{\Psi}_0(P_n) - E_{B_n}\hat{\Psi}_0(P^0_{n,B_n}) = o_P(1/\sqrt{n})$. The latter is a simple consequence of the asymptotic linearity of $\hat{\Psi}_0(P_n)$, as we will show now.

By the asymptotic linearity of $\hat{\Psi}_0(P_n)$ with influence curve $IC(O)$ applied to the sample P^0_{n,B_n} for a fixed B_n, we have

$$\hat{\Psi}_0(P^0_{n,B_n}) - \psi_0 = (P^0_{n,B_n} - P_0)IC + o_P(1/\sqrt{n}),$$

which implies that (assuming B_n has finite number of realizations, for convenience)

$$E_{B_n}\hat{\Psi}_0(P^0_{n,B_n}) - \psi_0 = (P_n - P_0)IC + o_P(1/\sqrt{n}).$$

Here we use that $E_{B_n}(P^0_{n,B_n} - P_0)IC = (P_n - P_0)IC$. Similarly, by asymptotic linearity of $\hat{\Psi}_0(P_n)$, we also have

$$\hat{\Psi}_0(P_n) - \psi_0 = (P_n - P_0)IC + o_P(1/\sqrt{n}).$$

The last two asymptotic linearity results yield the wished condition (15) $E_{B_n}\hat{\Psi}_0(P^0_{n,B_n}) - \hat{\Psi}_0(P_n) = o_P(1/\sqrt{n})$.

3.3. *Example: Estimation of the mean with shrinkage estimators.*

Let $X \sim f_0$, the model is nonparametric, the parameter of interest is $\psi_0 = E_0 r(X)$ for some function r, and suppose we observe n i.i.d. observations $X_1, \ldots, X_n$ of X. The estimating function for ψ_0, derived from the efficient influence curve $D(X \mid P_0) = r(X) - \psi_0$ at P_0, is given by $D(X \mid \psi) =$

$r(X) - \psi$. Suppose that our candidate estimators are $\hat{\Psi}_\rho(P_n) = \rho P_n r$ for $\rho \in [0,1]$. Our cross-validated estimating function criteria is given by:

$$\hat{\Theta}_{n(1-p)}(\rho) = E_{B_n}\left(P^1_{n,B_n} r - \rho P^0_{n,B_n} r\right)^2.$$

This is a convex function in ρ (i.e., second derivative is positive). Consider the solution of setting the derivative of $\hat{\Theta}_{n(1-p)}(\rho)$ equal to zero:

$$\rho_n = \frac{E_{B_n} P^1_{n,B_n} r P^0_{n,B_n} r}{E_{B_n}(P^0_{n,B_n} r)^2}.$$

If $\rho_n \in (0,1)$, then this is the minimum. If $\rho_n > 1$, then 1 is the minimum, and if $\rho_n < 0$, then 0 is the minimum. Let $\rho_n^* \in [0,1]$ be this minimizer of $\hat{\Theta}_{n(1-p)}(\rho)$. If $P_0 r \neq 0$, then it follows:

$$\begin{aligned}\rho_n - 1 &= \frac{E_{B_n}(P^0_{n,B_n} - P_0) r (P^1_{n,B_n} - P_0) r)}{E_{B_n}(P^0_{n,B_n} r)^2} \\ &= O_P(1/n),\end{aligned}$$

where we assume that $E_0 r^2 < \infty$ and, for simpicity, that B_n has finite support uniformly in n (as in V-fold cross-validation). This shows that $\rho_n^* - 1 = O_P(1/n)$, and thereby that $\hat{\Psi}_{\rho_n^*}(P_n)$ is asymptotically equivalent with the efficient sample mean. If $E_0 r = 0$, then ρ_n converges in distribution to a random variable with mean zero and large variance. As a consequence, in this case we would obtain an estimator which equals a random factor between $[0,1]$ times the sample mean, and is therefore a superefficient estimator at a P_0 with $P_0 r = 0$.

3.4. *Example: Shrinkage of linear regression estimators.*

Let $O = (Y, W) \sim P_0$, and assume the linear regression model $E_0(Y \mid W) = \beta_0^\top W$, $W \in \mathbb{R}^d$. Let β_0 be the parameter of interest, and suppose we observe n i.i.d. observations $O_1, \ldots, O_n$. Let $\hat{\Psi}_0(P_n) = \arg\min_\beta \sum_i (Y_i - \beta^\top W_i)^2$ be the standard least squares estimator. Let $\hat{\Psi}_h(P_n) = (1-h)\hat{\Psi}_0(P_n)$, $h \in [0,1]$. These estimators shrink the least squares estimator to zero. Consider the estimating function

$$D(O \mid \beta, c_0) = c_0^{-1} D(O \mid \beta) = c_0^{-1} W(Y - \beta W),$$

where $c_0 = -\frac{d}{d\beta} E_0 D(O \mid \beta)\Big|_{\beta=\beta_0}$. We note that $c_0 = E_0 W W^\top$. Let $\hat{C}(P_n) = 1/n \sum_i W_i W_i^\top$ be the estimator of c_0. Our proposed estimating

function based cross-validation criteria is given by

$$\hat{\Theta}_{n(1-p)}(h)^2 = E_{B_n} \sum_{j=1}^{d} \left\{ P^1_{n,B_n} D_j(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})) \right\}^2.$$

Indeed, we now have $d/d\beta E_0 D(O \mid \beta, c_0)|_{\beta=\beta_0} = -I$ equals minus the identity matrix so that $E_0 D(O \mid \beta, c_0) \approx -(\beta - \beta_0)$ in first order. Consequently, the corresponding target criterian $\tilde{\Theta}_{n(1-p)}(h)$ approximates the squared euclidean norm of the difference $\hat{\Psi}_h(P^0_{n,B_n}) - \psi_0$. Let h_n be the minimizer of $\hat{\Theta}_{n(1-p)}(h)$. Taking the derivative w.r.t. h yields the equation $U(h, P_n) = 0$ with $U(h, P_n)$ given by

$$E_{B_n} \sum_{j=1}^{d} \left(P^1_{n,B_n} D_j(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})) \right) P^1_{n,B_n} D^*_j(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})),$$

where $D^*_j(O \mid \psi, c) \equiv [c^{-1} W \psi(W)]_j$. The solution h^*_n of $U(h, P_n) = 0$ is given by the following closed form expression

$$\frac{E_{B_n} \sum_{j=1}^{d} P^1_{n,B_n} D^*_j(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})) \left(P^1_{n,B_n} D_j(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})) \right)}{E_{B_n} \sum_{j=1}^{d} \left(P^1_{n,B_n} D^*_j(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})) \right)^2}.$$

If $h^*_n \in [0, 1]$, then $h_n = h^*_n$, if $h^*_n < 0$, then $h_n = 0$, and, if $h^*_n > 1$, then $h_n = 1$. If $\psi_0 \neq 0$, then the denominator of h^*_n converges to some positive number. Consider now the j-term of the sum in the numerator, and denote the factor in front of the term within brackets with $\Phi_j(P^1_{n,B_n}, P^0_{n,B_n})$. Now, write this factor as $\{\Phi_j(P^1_{n,B_n}, P^0_{n,B_n}) - \Phi_j(P_0, P_0)\} + \Phi_j(P_0, P_0)$. The first difference is $O_P(1/\sqrt{n})$ so that the corresponding term will be $o_P(1/\sqrt{n})$. The second term results in the following contribution to the numerator of h^*_n

$$\sum_{j=1}^{d} \Phi_j(P_0, P_0) \left(E_{B_n} P^1_{n,B_n} D_j(\cdot \mid \hat{\Psi}_0(P^0_{n,B_n}), \hat{C}(P^0_{n,B_n})) \right).$$

The latter term is precisely (15), so that it is shown to be $o_P(1/\sqrt{n})$ in the same manner as outlined in general under (15). This shows that, if $\psi_0 \neq 0$, then, under minor conditions, $h_n = o_p(1/\sqrt{n})$. If ψ_0 happens to be zero, then the data adaptively shrinked estimator will be superefficient, as in our other examples.

3.5. ***Example: Convex combination of estimators.***

Let $X \sim f_0$, the model is nonparametric, the parameter of interest is $\psi_0 = E_0 r(X)$ for some function r, and suppose we observe n i.i.d. observations $X_1, \ldots, X_n$ of X. The estimating function for ψ_0, derived from the efficient influence curve $D(X \mid P_0) = r(X) - \psi_0$ at P_0, is given by $D(X \mid \psi) = r(X) - \psi$. Let $\hat{\Psi}_0(P_n) \equiv P_n r$ be the efficient estimator in the nonparametric model, and let $\hat{\Psi}_1(P_n)$ be an estimator of ψ_0 based on a submodel. For example, $\psi_0 = F_0(t)$ is the cumulative distribution function at a point t, $\hat{\Psi}_0(P_n)$ is the empirical cumulative distribution function, and $\hat{\Psi}_1(P_n)$ is an estimator based on a parametric model. Let $\hat{\Psi}_h(P_n) = h\hat{\Psi}_1(P_n) + (1-h)\hat{\Psi}_0(P_n)$, $h \in [0,1]$, be the convex combination of the two estimators. Our cross-validated estimating function criteria is given by:

$$\hat{\Theta}_{n(1-p)}(h) = E_{B_n}\left(P^1_{n,B_n} r - \hat{\Psi}_h(P^0_{n,B_n})\right)^2.$$

This is a convex function in h (i.e., second derivative is positive). Consider the solution of setting the derivative of $\hat{\Theta}_{n(1-p)}(h)$ equal to zero:

$$\begin{aligned} h_n &= \frac{E_{B_n}\{P^1_{n,B_n} - P^0_{n,B_n}\} r(\hat{\Psi}_1 - \hat{\Psi}_0)(P^0_{n,B_n})}{E_{B_n}\left\{(\hat{\Psi}_1 - \hat{\Psi}_0)(P^0_{n,B_n})\right\}^2} \\ &= \frac{E_{B_n}\{(P^1_{n,B_n} - P_0) - (P^0_{n,B_n} - P_0)\} r\left\{(\hat{\Psi}_1 - \hat{\Psi}_0)(P^0_{n,B_n}) - (\hat{\Psi}_1 - \hat{\Psi}_0)(P_0)\right\}}{E_{B_n}\left\{(\hat{\Psi}_1 - \hat{\Psi}_0)(P^0_{n,B_n})\right\}^2}, \end{aligned}$$

where we used that $E_{B_n}(P^1_{n,B_n} - P^0_{n,B_n}) r = 0$.

If $h_n \in (0,1)$, then this is the minimum. If $h_n > 1$, then 1 is the minimum, and if $h_n < 0$, then 0 is the minimum. Let $h_n^* \in [0,1]$ be this minimizer of $\hat{\Theta}_{n(1-p)}(h)$. If $\hat{\Psi}_1(P_n)$ is inconsistent and $\hat{\Psi}_1(P_n) - \hat{\Psi}_1(P_0) = o_P(1)$, then it follows immediately that $h_n^* = o_P(1/\sqrt{n})$. Here we also assume that $E_0 r^2 < \infty$ and, for simpicity, that B_n has finite support uniformly in n (as in V-fold cross-validation). This shows that under these conditions $\hat{\Psi}_{h_n^*}(P_n)$ is asymptotically equivalent with the efficient empirical cumulative distribution function. In the special case that $\hat{\Psi}_1(P_n)$ is also a consistent estimator for ψ_0, then h_n converges in distribution to a random variable with mean zero and support $[0,1]$. As a consequence, $\hat{\Psi}_{h_n^*}(P_n)$ is a superefficient estimator at any P_0 for which $\hat{\Psi}_1$ is consistent.

3.6. *Example: Substitution estimators based on kernel density estimators.*

Let $X \sim f_0$, the model is nonparametric, the parameter of interest is $\psi_0 = E_0 r(X)$ for some function r, and suppose we observe n i.i.d. observations $X_1, \ldots, X_n$ of X. The estimating function for ψ_0, derived from the efficient influence curve $D(X \mid P_0) = r(X) - \psi_0$ at P_0, is given by $D(X \mid \psi) = r(X) - \psi$. Let $\hat{\Psi}_h(P_n)$ be the mean of $r(X)$ w.r.t. a kernel density estimator $1/nh \sum_i K((X_i - \cdot)/h)$, with density kernel K, and let $\hat{\Psi}_0(P_n) = \int r(x) dP_n(x)$ be the empirical mean of $r(X)$. Suppose that the kernel K does not have mean zero so that the bias of the estimators $\hat{\Psi}_h(P_n)$ is linear in h: if we work with orthogonal kernels, we would define h as a power of the bandwidth so that h still represents the bias of the kernel density estimator. In this case, the estimators $\hat{\Psi}_h(P_n)$ have an asymptotic bias $O(h)$, which thus only disappears at $\sqrt{n}$-rate if $h = o(1/\sqrt{n})$. Our cross-validation criteria is defined as

$$\hat{\Theta}_{n(1-p)}(h)^2 = E_{B_n} \left(\hat{\Psi}_0(P^1_{n,B_n}) - \hat{\Psi}_h(P^0_{n,B_n}) \right)^2, \tag{16}$$

and the cross-validation selector h_n of h is its minimizer. For example, if B_n represents a leave-one out cross-validation scheme, then this criteria would resemble standard leave-one out cross-validation, except where the outcome is replaced by r:

$$\hat{\Theta}_{n(1-p)}(h)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(r(X_i) - \hat{\Psi}_h(P_{n,-i}) \right)^2.$$

Note that (16) represents a slight modification of our general proposal in the sense that we put the E_{B_n} inside the square root (but still outside the squares). It simplifies the algebraic manipulations needed to establish the wished result for general pathwise differentiable parameters (see our accompanying technical report). We have $\tilde{\Theta}_{n(1-p)}(h)^2 = E_{B_n} \left(\psi_0 - \hat{\Psi}_h(P^0_{n,B_n}) \right)^2$. That is, the oracle selector $\tilde{k}_{n(1-p)}$ corresponds with selecting the estimator whose training sample realizations are closest to the true value ψ_0.

In this example $\hat{\Psi}_0$ is an estimator which does not use any smoothing. The purpose of smoothing in this example (that is, selecting a $h > 0$) is to obtain a finite sample improvement relative to $\hat{\Psi}_0$. For example, if it is known that the true cumulative distribution function is very smooth, then it makes sense to use a smooth estimator, even though the discrete empirical cumulative distribution function is already asymptotically efficient in the

nonparametric model. However, in order to remain efficient the bandwidth will have to be chosen so that the asymptotic bias is of smaller order than $1/\sqrt{n}$: that is, we wish to show that $h_n = o_P(1/\sqrt{n})$.

We note

$$\begin{aligned}\hat{\Psi}_h(P_n) &= \frac{1}{nh}\sum_{i=1}^{n}\int r(x)K((X_i - x)/h)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int r(X_i + yh)K(y)dy,\end{aligned}$$

and the derivative of $\hat{\Theta}_{n(1-p)}(h)^2$ w.r.t. h is given by

$$U(h, P_n) = 2E_{B_n}\left(\hat{\Psi}_0(P^1_{n,B_n}) - \hat{\Psi}_h(P^0_{n,B_n})\right)P^0_{n,B_n}\frac{d}{dh}\int r(\cdot + yh)K(y)dy.$$

We will follow the proof of Theorem 1 in order to illustrate it in this example. We have $U(h_n, P_n) = 0$, and we note that $U(h_0 = 0, P_0) = 0$, where $U(h, P_0)$ is defined by replacing P^1_{n,B_n} and P^0_{n,B_n} by P_0. The equations $U(0, P_0) = 0$ and $U(h_n, P_n) = 0$ provides us with a basis for establishing that $h_n = o_P(1/\sqrt{n})$, and thereby that $\hat{\Psi}_{h_n}$ is still asymptotically efficient.

Firstly, we note that

$$U(h_n, P_0) - U(0, P_0) = -\{U(h_n, P_n) - U(h_n, P_0)\}.$$

Since

$$U(h, P) = 2\left(\hat{\Psi}_0(P) - \hat{\Psi}_h(P)\right)P\frac{d}{dh}\int r(\cdot + yh)K(y)dy,$$

it follows that

$$\frac{d}{dh}U(h, P_0)\Big|_{h=0} = -2\left(P_0\,\frac{d}{dh}\int r(\cdot + yh)K(y)dy\Big|_{h=0}\right)^2,$$

which verifies that the derivative of $h \to U(h, P_0)$ at $h = 0$ is bounded away from zero. Below, we will show that h_n converges to zero in probability. Then, it follows

$$h_n = \left(-\frac{d}{dh}U(h, P_0)\Big|_{h=0}\right)^{-1}\{U(h_n, P_n) - U(h_n, P_0)\} + o(h_n).$$

Below, we will also show that

$$\{U(h_n, P_n) - U(h_n, P_0)\} - \{U(0, P_n) - U(0, P_0)\} = o_P(1/\sqrt{n}) + o_P(h_n). \tag{17}$$

Then, it follows that

$$h_n = \left(-\frac{d}{dh} U(h, P_0)\Big|_{h=0} \right)^{-1} \{U(0, P_n) - U(0, P_0)\} + o_P(h_n) + o_P(1/\sqrt{n}).$$

Now, we note that

$$U(0, P_n) = E_{B_n} \left(P^1_{n,B_n} r - P^0_{n,B_n} r \right) P^0_{n,B_n} r_1,$$

where $r_1 \equiv \frac{d}{dh} \int r(\cdot + yh) K(y) dy \big|_{h=0}$. Write $P^0_{n,B_n} r_1 = (P^0_{n,B_n} - P) r_1 + P r_1$, and note that the term resulting from $P r_1$ equals exactly zero. Thus,

$$U(0, P_n) = E_{B_n} \left((P^1_{n,B_n} - P_0) r - (P^0_{n,B_n} - P_0) r \right) (P^0_{n,B_n} - P_0) r_1,$$

which is indeed $o_P(1/\sqrt{n})$ if $P_0 r_1 < \infty$ and $P_0 r^2 < \infty$. To conclude, this shows that $h_n = o_P(h_n) + o_P(1/\sqrt{n})$, and thus the wished result $h_n = o_P(1/\sqrt{n})$.

Convergence of h_n to 0: We need to verify that $h_n = o_P(1)$. If r has compact support, it follows that $h_n \leq M$ for some $M < \infty$. As a consequence, by compactness of $[0, M]$, for each subsequence of h_n, there exists a subsequence (say) h_k which converges to a h_∞ for k converging to infinity. Since $U(h_k, P_k) = 0$ and $U(h_k, P_k) - U(h_k, P_0)$ converges to zero as k converges to infinity under the already needed Donsker class condition specified below, it follows that $U(h_k, P_0)$ converges to zero. In addition, it also follows that $U(h_k, P_0)$ converges to $U(h_\infty, P_0)$, which shows that $U(h_\infty, P_0) = 0$. This shows that $h_\infty = 0$. This proves that h_n converges to zero a.s., and thus, in particular, in probability.

Verification of (17): This second order difference can be written as a sum with the following four terms:

$$\begin{array}{l} E_{B_n}(P^0_{n,B_n} - P_0)(R_{h_n} - R_0) P^0_{n,B_n} r_{1,h_n} \\ E_{B_n} \left((P^1_{n,B_n} - P_0) R_0 - (P^0_{n,B_n} - P_0) R_{h_n} \right) P^0_{n,B_n} (r_{1,h_n} - r_{1,0}) \\ E_{B_n} P^0_{n,B_n} (R_{h_n} - R_0)(P^0_{n,B_n} - P_0) r_{1,h_n} \\ E_{B_n} \left(P^1_{n,B_n} R_0 - P^0_{n,B_n} R_{h_n} \right) (P^0_{n,B_n} - P_0)(r_{1,h_n} - r_{1,0}), \end{array}$$

where $R_h(X) \equiv \int r(X + yh) K(y) dy$, and $r_{1,h}(X) = \frac{d}{dh} \int r(X + yh) K(y) dy$. If $\int (R_{h_n} - R_0)^2(x) dP_0(x)$ converges to zero in probability (as follows from $h_n = o_P(1)$ and continuity of r), and if $\{R_h - R_0 : h \in [0, 1]\}$ is a P_0-Donsker class, then it follows ([?]) that this $(P^0_{n,B_n} - P_0)(R_{h_n} - R_0) = o_P(1/\sqrt{n})$. Examples of P_0-Donsker classes are provided in [?]: e.g., if R_h has variation smaller than a universal $M < \infty$, then $\{R_h - R_0 : h\}$ is a P_0-Donsker class.

We also assume that $\{r_{1,h} : h > 0\}$ is a Glivenko-Cantelli class so that $\sup_h (P^0_{n,B_n} - P_0) r_{1,h} = o_P(1)$. This proves that the first three terms are

$o_P(1/\sqrt{n})$: for technical convenience, we assume a V-fold cross-validation scheme so that E_{B_n} only yields a sum of V terms, which each can be analyzed seperately. Regarding the fourth term, first write

$$P^1_{n,B_n}R_0 - P^0_{n,B_n}R_{h_n} = (P^1_{n,B_n} - P_0)R_0 - (P^0_{n,B_n} - P_0)R_{h_n} + P_0(R_0 - R_{h_n}).$$

The terms resulting from the first term is $o_P(1/\sqrt{n})$ since $(P^1_{n,B_n} - P_0)R_0 = O_P(1/\sqrt{n})$ and $(P^0_{n,B_n} - P_0)(r_{1,h_n} - r_{1,0}) = o_P(1)$. Similarly, this follows for the term resulting from the second term. The term resulting from $P_0(R_0 - R_{h_n})$ is given by:

$$P_0(R_0 - R_{h_n})(P_n - P_0)(r_{1,h_n} - r_{1,0}).$$

It follows that $P_0(R_{h_n} - R_0) = O_P(h_n)$ so that the last term is $o_P(h_n)$.

Thus under these empirical process conditions on $r_{1,h}$, we have proved that $h_n = o_P(1/\sqrt{n})$, and, consequently, that the substitution estimators based on an integrated kernel density estimator using bandwidth h_n will be asymptotically efficient. We will state this as a result. Inspection of the proof shows that we can replace the condition that $\{r_{1,h} : h \in (0, M]\}$ is a P_0-Glivenko-Cantelli class by $(P_n - P_0)r_{1,h_n} = o_P(1)$, which allows sharper results, as illustrated in the next subsection.

Theorem 2: Let X be a real valued random variable with density f_0 with compact support contained in $[0, M]$, and, given a function r which is continuous F_0-a.e., let $\psi_0 = E_0 r(X)$ be its parameter of interest. Suppose we observe n i.i.d. observations $X_1, \ldots, X_n$ of X. Let $\hat{\Psi}_b(P_n)$ be the mean of $r(X)$ w.r.t. a kernel density estimator $1/nb\sum_i K((X_i - \cdot)/b)$, with density kernel K, and let $\hat{\Psi}_0(P_n) = \int r(x)dP_n(x)$ be the empirical mean of $r(X)$. Let $h \to b(h)$ be a 1-1 parametrization with inverse $b \to h(b)$ satisfying

$$\frac{d}{dh}\, P_0 \int r(\cdot + yb(h))K(y)dy\Big|_{h=0} \neq 0.$$

That is, $h(b)$ represents the order of the bias of the integrated kernel density estimator with bandwidth b. For example, if the kernel K does not have mean zero, then one can choose $h = b$.

Let

$$\hat{\Theta}_{n(1-p)}(h)^2 = E_{B_n}\left(\hat{\Psi}_0(P^1_{n,B_n}) - \hat{\Psi}_h(P^0_{n,B_n})\right)^2,$$

and h_n is its minimizer over the interval $[0, M]$. Let B_n correspond with V-fold cross-validation for a fixed V. Let $r_{1,h}(X) \equiv d/dh \int r(X + yb(h))K(y)dy$, and $R_h(X) \equiv \int r(X + yh)K(y)dy$. Assume that $\{r_{1,h} : h \in (0, M]\}$ is a P_0-Glivenko-Cantelli class or the weaker assumption

$(P_n - P_0)r_{1,h_n} = o_P(1)$, and $\{R_h : h \in [0, M]\}$ is a P_0-Donsker class. For example, the latter holds if R_h has variation smaller than a universal $C < \infty$.

Then $h_n = o_P(1/\sqrt{n})$.

Special case: Smoothing of the empirical cumulative distribution.

As a challenging example, consider the case $\psi_0 = F_0(x_0)$, which corresponds with the choice $r(X) = I(X \leq x_0)$. In order to be explicit, let $K(x) = I_{[-1,1]}(x)$ be the uniform kernel density. In this case, $r_{1,b}(X) = r_{1,h(b)} = I(\{X - x_0\}/b < 1)/(2b)$ is a uniform density over $[x_0 - b, x_0 + b]$, and typically $h(b) = b^2$. Though, $\sup_{b\in[0,M]} P_0 r_{1,b} < \infty$, the class of functions $\{r_{1,b} : b\}$ does not have an integrable envelope, which shows that it is not a Glivenko-Cantelli class: see page 125 [?]. So, though our Theorem 2 is very close to also being able to include non-smooth r's, our conditions are too strong for showing that our smoothed empirical distribution at x_0 is an asympotically efficient estimator of $F_0(x_0)$. However, inspection of the proof of Theorem 2 shows that we can replace the Glivenko-Cantelli class condition on $r_{1,b}$ by the condition that $(P_n - P_0)r_{1,b_n} = o_P(1)$, where we already know $b_n \to 0$ a.s. That is, we need that b_n is such that the kernel density estimator with bandwidth b_n is consistent at x_0. Intuitively, we certainly expect this to hold, since consistency of kernel density estimators for non-random bandwidth b_n only requires $nb_n \to \infty$.

Formally, the following type proof has been used to establish consistency of kernel density estimators for data dependent bandwidth b_n for which $nb_n \to \infty$ a.s (or in probability): personal communication by Aad van der Vaart. For notational convenience, let $x_0 = 0$ and consider the uniform kernel. For fixed $\delta > 0$ consider the class of all functions $x \to f_b(x) \equiv I_{(-b,b)}(x)(2b)^{-1}$ with $b > \delta$. This is a VC class with envelope $F_\delta(x) = \min((2|x|)^{-1}, (2\delta)^{-1})$. Consequently, by empirical process theory ([?]) $E \sup_{b>\delta}(P_n - P_0)f_b \lesssim n^{-1/2}J(1)(PF_\delta^2)^{1/2}$, where $J(1)$ denotes the uniform entropy relative to the envelope, a finite number for a VC-class. Since the $1/2\delta$ in the envelope only contributes to the integral over the interval $[0, \delta]$, it follows that $PF_\delta^2 = O(1/\delta)$, where we assume that the true density f_0 is bounded. This yields an upper bound of the order $(n\delta)^{-1/2}$ for small δ. We now proceed as follows. Let M_n be a sequence such that $Pr(b_n > M_n/n) \to 1$, which exists since $nb_n \to \infty$. For this sequence M_n

we have

$$Pr\left(\mid (P_n - P)f_{b_n} \mid > \epsilon\right) \leq Pr\left(\sup_{b > M_n/n} \mid (P_n - P)f_b \mid > \epsilon\right) + o(1).$$

By the above VC-class empirical process result, the expectation of $\sup_{b>M_n/n} \mid (P_n - P)f_b \mid$ is bounded by $O(1/\sqrt{M_n}) = o(1)$. Since convergence in expectation implies convergence in probability this shows that $Pr\left(\mid (P_n - P)f_{b_n} \mid > \epsilon\right)$ converges to zero.

Thus, it remains to prove that $1/(nb_n) = o_P(1)$. Though, we expect this to hold, actually proving that there exist no subsequence of b_n which converges too fast does not seem to be easy to us. Therefore, we propose the following to obtain a formal result. We redefine our bandwidth as $b_n^* = \max(b_n, c_n/n)$ for a given sequence $c_n \to \infty$ with $c_n/\sqrt{n} = o(1)$, and aim to prove that the corresponding bias $h_n^* = o_P(1/\sqrt{n})$. Consider now the subsequence $b_{n(m)}$, $m = 1, 2, \ldots$, defined by deleting each element of b_n^* which does not equal b_n. In order to prove that $h_n^* = o_P(1/\sqrt{n})$ it suffices to prove that $h_{n(m)} = o_P(1/\sqrt{n(m)})$. For this subsequence, we have $U(h_{n(m)}, P_{n(m)}) = 0$, and we can apply Theorem 2, where now, by definition of $h_{n(m)}$, we know that $n(m)b_{n(m)} \to \infty$, so that we have $(P_{n(m)} - P_0)r_{b_{n(m)}} = o_P(1)$. This proves the following formal result.

Theorem 3: Let X be a real valued random variable with bounded density f_0 with compact support contained in $[0, M]$, let the parameter of interest $\psi_0 = F_0(x_0) = \int I_{(0,x_0]}(x)dF_0(x)$ be the cumulative distribution function at $x_0 \in (0, M)$. Suppose we observe n i.i.d. observations $X_1, \ldots, X_n$ of X. Let $\hat{\Psi}_b(P_n) = \int I_{(0,x_0]}(x)f_{n,b}(x)dx$, where $f_{n,b}(x) = 1/nb \sum_i K((X_i - \cdot)/b)$ is a kernel density estimator with density kernel K and bandwidth b. Let $\hat{\Psi}_0(P_n) = P_n I_{(0,x_0]}$ be the empirical cumulative distribution function. Let $h \to b(h)$ be a 1-1 parametrization with inverse $b \to h(b)$ satisfying

$$\frac{d}{dh}\, P_0 \int r(\cdot + yb(h))K(y)dy\Big|_{h=0} \neq 0.$$

That is, $h(b)$ represents the order of the bias of the integrated kernel density estimator with bandwidth b.

Let

$$\hat{\Theta}_{n(1-p)}(b)^2 = E_{B_n}\left(\hat{\Psi}_0(P^1_{n,B_n}) - \hat{\Psi}_b(P^0_{n,B_n})\right)^2,$$

where B_n correspond with V-fold cross-validation for a fixed V. Let b_n be its minimizer over the interval $[0, M]$. Let $b_n^* = \max(b_n, c_n/n)$ for a given sequence $c_n \to \infty$ with $c_n/\sqrt{n} = o(1)$.

Then $h_n^* = o_P(1/\sqrt{n})$.

3.7. *Example: Smooth estimation of a survival function.*

Let $O = (\tilde{T} \equiv \min(T, C), \Delta \equiv I(T \leq C))$, where T is a survival time of interest with cumulative distribution function F_0, C is a right-censoring variable independent of T with cumulative distribution function G_0. Let $\bar{F}_0(t) = P(T > t)$ be the survival function of T, and $\bar{G}_0(t) = P(C > t)$ be the cumulative distribution function of C. We will assume that $\bar{G}_0(T) > 0$ F_0-a.e. Let $\psi_0 = E_0 r(T)$. For example, if $r(T) = I(T \geq t)$, then $\psi_0 = S_0(t)$ is the survival function at t. We observe n i.i.d. observations $O_1, \ldots, O_n$ of O, and we are concerned with smooth estimation of ψ_0. Smoothing Kaplan-Meier estimator has received considerable attention in the literature: see e.g. [1] and [9]. However, though it has been recognized that undersmoothing is essential in order to obtain an asymptotically efficient estimator, a data adaptive method for selecting such a bandwidth has not been presented.

Let $\hat{\Psi}_h(P_n) = \int r(s) f_{nb}(s) ds$, where

$$f_{nb}(s) = \frac{1}{nb} \sum_{i=1}^{n} K((T_i - s)/b) \frac{\Delta_i}{\bar{G}_n(T_i)}$$

is an inverse probability of censoring weighted kernel density estimator of the true density $f_0(s)$ of T with kernel K and bandwidth $b = b(h)$. Here $\bar{G}_n(t) = 1 - G_n(t)$ denotes the Kaplan-Meier estimator of $\bar{G}_0(t) \equiv 1 - G_0(t)$ based on $(\tilde{T}_i, 1 - \Delta_i)$, $i = 1, \ldots, n$. We note that

$$\hat{\Psi}_0(P_n) = \frac{1}{n} \sum_{i=1}^{n} r(T_i) \frac{\Delta_i}{\bar{G}_n(T_i)},$$

which is known to be an efficient estimator of ψ_0 (Chapter 3, [32]). In particular, if $r(T) = I(T > t)$, then $\hat{\Psi}_0(P_n)$ equals the Kaplan-Meier estimator of the survival function $S(t)$ at time t.

Let $h \to b(h)$ be a 1-1 parametrization with inverse $b \to h(b)$ satisfying

$$\frac{d}{dh} P_0 \int r(\cdot + yb(h)) K(y) dy \Big|_{h=0} \neq 0.$$

That is, $h(b)$ represents the order of the bias of the integrated kernel density estimator with bandwidth b. For example, if he kernel K does not have mean zero, then one can choose $h = b$.

Let $D(O \mid \psi, G) \equiv r(T) \frac{\Delta}{\bar{G}(T)} - \psi$ be the inverse probability of censoring weighted full-data estimating function $r(T) - \psi$ for ψ_0. The actual efficient

influence curve based estimating function for ψ_0 is given by

$$D(O \mid \psi, G, S) \equiv r(T)\frac{\Delta}{\bar{G}(T)} - \psi + \int \frac{S(\max(t, u-))}{S(u-)} \frac{dM_G(u)}{\bar{G}(u-)},$$

where $dM_G(u) = I(\tilde{T} \in du, \Delta = 0) - I(\tilde{T} \geq u)\frac{dG(u)}{\bar{G}(u-)}$. Here S denotes a candidate survival function and $D(O \mid \psi_0, G_0, S_0)$ equals the efficient influence curve for ψ_0 (see Chapter 3, [32]).

Consider now our proposed criterian based on the estimating function $D(O \mid \psi, G)$:

$$\begin{aligned}\hat{\Theta}_{n(1-p)}(h)^2 &= E_{B_n}\left\{P^1_{n,B_n} D(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), G^0_{n,B_n})\right\}^2 \\ &= E_{B_n}\left\{\hat{\Psi}_h(P^0_{n,B_n}) - \frac{1}{np}\sum_{i=1}^{n} I(B_n(i) = 1) r(\tilde{T}_i)\frac{\Delta_i}{\bar{G}^0_{n,B_n}(\tilde{T}_i)}\right\}^2,\end{aligned}$$

where G^0_{n,B_n} denotes the Kaplan-Meier estimator of G_0 based on P^0_{n,B_n}. Similarly, one defines the criterian based on the estimating function $D(O \mid \psi, G, S)$. Let h_n be its minimizer.

We also note that the corresponding target criterian for both estimating functions is given by

$$\begin{aligned}\tilde{\Theta}_{n(1-p)}(h)^2 &= E_{B_n}\left\{P_0 D(\cdot \mid \hat{\Psi}_h(P^0_{n,B_n}), G_0, S_0)\right\}^2 \\ &= E_{B_n}\left\{\hat{\Psi}_h(P^0_{n,B_n}) - \psi_0\right\}^2.\end{aligned}$$

That is, the comparable oracle selector $\tilde{k}_{n(1-p)}$ corresponds with selecting the estimator whose training sample realizations are closest to the true value ψ_0.

Following the template as laid out in Theorem 1, we can establish that, if there exists a $\delta > 0$ s.t. $\bar{G}(\tau) > \delta > 0$ and r satisfies the conditions of Theorem 2, then $h_n = o_P(1/\sqrt{n})$. As pointed out after Theorem 2, our conditions are too strong for the indicator function $r(T) = I(T > t)$ so that it remains to be shown that the smoothed Kaplan-Meier estimator $\hat{\Psi}_{h_n}(P_n)$ is an asymptotically efficient estimator of $S(t)$.

3.8. *A class of general examples.*

Consider candidate estimators $\hat{\Psi}_h(P_n)$ of a pathwise differentiable d-dimensional parameter ψ_0 of P_0 indexed by a continuous univariate index h. Let $\psi_0 = \Psi(P_0) \in \mathbb{R}^d$ be pathwise differentiable euclidean parameter of the data generating distribution P_0 in a model $\mathcal{M}$, which is identified as the

solution of the d-dimensional vector equation $P_0 D(\cdot \mid \psi, \gamma_0) = 0$ for ψ ranging over the parameter space of $\mathbf{\Psi}$. For example, D could be the optimal estimating function derived from the efficient influence curve representation of $\Psi : \mathcal{M} \to \mathbb{R}^d$. Let $\hat{\Psi}_0(P_n)$ be defined as the solution of the corresponding estimating equation $0 = P_n D(\cdot \mid \psi, \hat{\Gamma}(P_n))$, which, for example, could be an efficient estimator of ψ_0 in model $\mathcal{M}$. One can now imagine a variety of general methods for constructing candidate estimators of ψ_0 to which we can then apply the cross-validation method based on estimating function D. Firstly, one might have a sequence of submodels $\mathcal{M}_h \subset \mathcal{M}$ indexed by a fine tuning parameter h, and $\hat{\Psi}_h(P_n)$ would be an efficient estimator of the true ψ_0 under the assumption that $P_0 \in \mathcal{M}_h$. We conjecture that, if the true data generating distribution happens to be an element of submodel $\mathcal{M}_h$ for a $h > 0$, then the resulting cross-validation selected estimator $\hat{\Psi}_{h_n}$ might be superefficient (and thus be a non-regular estimator). In our previous examples, this could be explicitly shown. Secondly, one might have a particular submodel of interest of $\mathcal{M}_1 \subset \mathcal{M}$, and let $\hat{\Psi}_1(P_n)$ be an estimator of ψ_0 under the assumption that $P_0 \in \mathcal{M}_1$. One can now define candidate estimators as convex combinations $\hat{\Psi}_h(P_n) = (1-h)\hat{\Psi}_0(P_n) + h\hat{\Psi}_1(P_n)$, where $h \in [0, 1]$.

A particular example of this type is treated in an upcoming paper [2] in which the goal is to estimate the causal effect of treatment on an outcome in a randomized trial with non-compliance, assuming a particular functional form of the causal effect (e.g. through a structural nested mean model as studied in [31]). In these applications, using the randomization arm as an instrumental variable, it is possible to construct unbiased estimating functions, and thereby regular asymptotically linear estimators, for the causal parameter without making the commonly made assumption that the actual treatment is not subject to unmeasured confounding. However, these estimators are typically extremely variable. On the other hand, if one is willing to make the assumption of no unmeasured confounding then one has access to estimators with much smaller variance. This suggest to let $\mathcal{M}$ be the big model only assuming randomization of the treatment arm (which is known to be true), and let $\mathcal{M}_1$ be the submodel which also assumes that treatment is not confounded by unmeasured variables. Our cross-validation methodology allows us now to compute this range of estimators $\hat{\Psi}_h(P_n)$, $h \in [0, 1]$, relying on h-specific degree on the no-unmeasured confouding assumption, and simply let the data present us with the appropriate choice of h, while still being at least as efficient as the estimator $\hat{\Psi}_0$ in the big model $\mathcal{M}$. For a detailed treatment of this important application of our estimating function

methodology we refer to [2]. As the reader can imagine, it is not hard to come up with a large range of very interesting and important applications of this methodology, which allows us to trade off bias and variance w.r.t. to a parameter of interest within the context of a given possibly large model.

4. General finite sample result.

Recall $\tilde{\Theta}_{n(1-p)}(k) \equiv E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \gamma_0) \parallel$, $\tilde{k} = \arg\min_{k=1,\ldots,K(n)} \tilde{\Theta}_{n(1-p)}(k)$, $\hat{\Theta}_{n(1-p)}(k) \equiv E_{B_n} \parallel P^1_{n,B_n} D(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})) \parallel$, and $k_n = \arg\min_{k=1,\ldots,K(n)} \hat{\Theta}_{n(1-p)}(k)$. For notational convenience, we define $D^0_{n,B_n}(O \mid \psi) \equiv D(O \mid \psi, \hat{\Gamma}(P^0_{n,B_n}))$ and $D(O \mid \psi) \equiv D(O \mid \psi, \gamma_0)$. We also define $G^1_{n,B_n} = \sqrt{np}(P^1_{n,B_n} - P_0)$ as the centered empirical process based on the validation sample identified by the sample split B_n, and recall the notation $Gf \equiv \int f(o) dG(o)$.

The following theorem provides us with our most general finite sample result. It compares our cross-validation selector with the oracle selector in terms of the criterian $\tilde{\Theta}_{n(1-p)}$ which measures in first order the norm of the estimator minus the true parameter. Our subsequent results are derived by establishing bounds on the remainder terms in this theorem.

Theorem 4: We have

$$\begin{aligned}\tilde{\Theta}_{n(1-p)}(k_n) \leq \tilde{\Theta}_{n(1-p)}(\tilde{k}) \qquad (18)\\ +\frac{1}{\sqrt{np}} E_{B_n} \left\{ \parallel G^1_{n,B_n} D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel + \parallel G^1_{n,B_n} D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel \right\}\\ +\frac{1}{\sqrt{np}} E_{B_n} \left\{ \parallel G^1_{n,B_n} \left(D^0_{n,B_n} - D\right)(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \right\}\\ +2 \max_{k \in \{1,\ldots,K(n)\}} E_{B_n} \parallel P_0 (D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n})) \parallel .\end{aligned}$$

Proof: Firstly, we note that by the triangle inequality property of a norm, we have

$$\begin{aligned}E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel \leq E_{B_n} \parallel P^1_{n,B_n} D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel\\ +\frac{1}{\sqrt{np}} E_{B_n} \parallel G^1_{n,B_n} D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n}) \parallel .\end{aligned}$$

The left-hand side equals $\tilde{\Theta}_{n(1-p)}(k_n)$, and the last term on the right-hand side represents one of the empirical process terms on the right-hand side of the inequality (18) to be proved. We will now study the other term, and bound it by a sum of five terms, which results in the inequality

(18). Repeated application (first and third inequality below) of the triangle inequality property of a norm, and the fact that by definition of k_n $\hat{\Theta}_{n(1-p)}(k_n) \leq \hat{\Theta}_{n(1-p)}(\tilde{k})$ (second inequality below), provides us with the following series of inequalities:

$$\begin{aligned} E_{B_n} \parallel P^1_{n,B_n} D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel &\leq E_{B_n} \parallel P^1_{n,B_n} D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel \\ &+ E_{B_n} \parallel P^1_{n,B_n}(D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel \\ &\leq E_{B_n} \parallel P^1_{n,B_n} D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \\ &+ E_{B_n} \parallel P^1_{n,B_n}(D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel \\ &\leq E_{B_n} \parallel (P^1_{n,B_n} - P_0) D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \\ &+ E_{B_n} \parallel P_0 D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \\ &+ E_{B_n} \parallel (P^1_{n,B_n} - P_0)(D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel \\ &+ E_{B_n} \parallel P_0(D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel . \end{aligned}$$

Finally, again by the triangle inequality property, we have that the second term of this sum of 4 terms can be bounded as follows:

$$\begin{aligned} E_{B_n} \parallel P_0 D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \leq{}& E_{B_n} \parallel P_0(D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \\ &+ E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \end{aligned}$$

Collection of all 5 terms yields the proof of the theorem. □

5. Corollaries of Theorem 4.

We will present corollaries of Theorem 4 for the following three norms.

Definition 5: For a countable sequence of real numbers a_i and weights $w_b \geq 0$ with $\sum_{b=1}^{\infty} w_b = 1$, define the following norms:
$\|(a_1, a_2, ...)\|_1 = \sum_{b=1}^{\infty} w_b |a_b|$.
$\|(a_1, a_2, ...)\|_2 = \sqrt{\sum_{b=1}^{\infty} w_b |a_b|^2}$.
$\|(a_1, a_2, ...)\|_\infty = sup_{b \geq 1} |a_b|$.

The following corollary of Theorem 4 establishes that our cross-validation selector performs as well as the oracle selector $\tilde{k}$ up till a term of order $\log K(n)/np$ and a term r_n due to the estimation of the nuisance parameter.

Corollary 6: *Let*

$$r_n \equiv 2 \max_{k \in \{1,...,K(n)\}} E_{B_n} \parallel P_0(D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n})) \parallel .$$

Assume that $\sup_{b \in \mathcal{B}} |D_b(\cdot|\cdot)| \leq M < \infty$ *a.s. Define* $h_{n,K(n)}$ *as the maximum of the covering numbers* $\int_0^\infty \sup_Q \sqrt{\log N(\epsilon M^2, \mathcal{F}_n, L_2(Q))} d\epsilon$ *corresponding with the following choices of function classes:* $\mathcal{F}_n = \{D_b(\cdot|\hat{\Psi}_k(P^0_{n,B_n}))\}$,

$\mathcal{F}_n = \{(D^0_{n,B_n})_b(\cdot|\hat{\Psi}_k(P^0_{n,B_n}))\}$, *and* $\mathcal{F}_n = \{(D^0_{n,B_n} - D)_b(\cdot|\hat{\Psi}_k(P^0_{n,B_n}))\}$, *for* $1 \leq b < \infty$, $1 \leq k \leq K(n)$.

Then for $\|\cdot\| = \|\cdot\|_1$ *or* $\|\cdot\| = \|\cdot\|_2$, *we have for a universal constant* c *(only depending on* M*)*

$$E[\tilde{\Theta}_{n(1-p)}(k_n)] \leq E[\tilde{\Theta}_{n(1-p)}(\tilde{k})] + c\sqrt{\log K(n)}/\sqrt{np} + E[r_n].$$

For $\|\cdot\| = \|\cdot\|_\infty$, *we have*

$$E[\tilde{\Theta}_{n(1-p)}(k_n)] \leq E[\tilde{\Theta}_{n(1-p)}(\tilde{k})] + M^2 h_{n,K(n)}/\sqrt{np} + E[r_n].$$

Proof: Firstly, we take expectations on both sides of the inequality in Theorem 4. For the second, third, and fourth terms on the right side of the inequality (call them $E_{n,B_n}[U_{n,i}]$, i=1,2,3), we note that by Fubini's theorem,

$$E[E_{n,B_n}U_{n,i}] = E[U_{n,i}] = E_{B_n,P^0_{n,B_n}}E_{P^1_{n,B_n}}U_{n,i}.$$

Finally, we apply the inequalities from Lemma 10 below to the inner expectation for these three terms to obtain the desired result. □

5.1. *Lemmas for Corollary 6.*

Lemma 7: *Let* $f_1, ..., f_{K(n)}$ *be functions with the same domain and range dimension as* $O \to D(O|\psi, \gamma)$, *where* f^b_k *represents the* b^{th} *component of the* k^{th} *function. Assume that* $|f^b_k| \leq M < \infty$ *for* $1 \leq b, k < \infty$.
Then $E[\max_{f^b_k \in \{f^b_1,...,f^b_{K(n)}\}} |G^1_{n,B_n} f^b_k|] \leq cM\sqrt{\log K(n)}$, *for a universal contant* c.

Proof: This is trivially implied by formula (2.5.5) in vdVW. □

Lemma 8: *Under the assumptions and notation of Lemma 7, we have* $E[\max_{f^b_k \in \{f^b_1,...,f^b_{K(n)}\}} |(P^1_{n,B_n} - P)f^b_k|^2] \leq c \log K(n)/n$.

Proof: Firstly, we use the Bonferroni inequality to bound the tail probability (probability of exceeding s) of the quantity inside the expectation by $K(n) \max_k Pr[|(P^1_{n,B_n} - P)f^b_k|^2 \geq s]$. Thge latter equals $K(n) \max_k Pr[|(P^1_{n,B_n} - P)f^b_k| \geq \sqrt{s}]$, which can be bounded by $K(n) \exp(-sn/c)$ for some constant c, by Bernstein's inequality. Finally, by Lemma 16, this bound implies the desired result. □

Lemma 9: *Let* $N(\cdot, \cdot, \cdot)$ *denote the covering number as defined in* [?]. *Consider the assumptions and notation of Lemma 7. Let* $\mathcal{F}_n \equiv \{f^b_k : 1 \leq k \leq$

$K(n), 1 \leq b \leq \infty\}$, *and* $h_{n,K(n)} \equiv \int_0^\infty \sqrt{\log \sup_Q N(\epsilon M^2, \mathcal{F}_n, L_2(Q))} d\epsilon$. *Then*

$$E[\sup_{f \in \mathcal{F}_n} |G^1_{n,B_n} f|] \leq cM^2 h_{n,K(n)} \text{ for some universal constant } c.$$

proof. See Chapter 2 in van der Vaart, Wellner (1996). □

Lemma 10: *Under the assumptions and notation of Lemma 7, Lemma 8, and Lemma 9, we have*

$$\frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \|G^1_{n,B_n} f_k\|_1] \leq c\sqrt{\log K(n)}/\sqrt{np}.$$
$$\frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \|G^1_{n,B_n} f_k\|_2] \leq c\sqrt{\log K(n)}/\sqrt{np}.$$
$$\frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \|G^1_{n,B_n} f_k\|_\infty] \leq M^2 h_{n,K(n)}/\sqrt{np}.$$

Proof:
$\frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \|G^1_{n,B_n} f_k\|_1]$
$= \frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \sum_{b=1}^\infty w_b |G^1_{n,B_n} f^b_k|]$
$\leq \frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\sum_{b=1}^\infty w_b \max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} |G^1_{n,B_n} f^b_k|]$
$= \frac{1}{\sqrt{np}} \sum_{b=1}^\infty w_b E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} |G^1_{n,B_n} f^b_k|]$,
The desired result then follows from Lemma 7.

$\frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \|G^1_{n,B_n} f_k\|_2]$
$= \frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \sqrt{\sum_{b=1}^\infty w_b |G^1_{n,B_n} f^b_k|^2}]$
$\leq \frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\sqrt{\sum_{b=1}^\infty w_b \max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} |G^1_{n,B_n} f^b_k|^2}]$
$\leq \frac{1}{\sqrt{np}} \sqrt{\sum_{b=1}^\infty w_b E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} |G^1_{n,B_n} f^b_k|^2]}$, by Jensen's inequality.
Then desired result then follows from Lemma 8.

$\frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\max_{f_k \in \{f_1,\ldots,f_{K(n)}\}} \|G^1_{n,B_n} f_k\|_\infty]$
$= \frac{1}{\sqrt{np}} E_{P^1_{n,B_n}}[\sup_{f \in \mathcal{F}_n} |G^1_{n,B_n} f|]$, for $\mathcal{F}_n \equiv \{f^b_k : 1 \leq k \leq K(n), 1 \leq b \leq \infty\}$
The desired result then follows from Lemma 9. □

5.2. *Asymptotic implications.*

The next corollary shows that the cross-validation selector k_n is asymptotically equivalent to the oracle selector $\tilde{k}_{n(1-p)}$ in the case that 1) the rate of

convergence achieved by the oracle selector is worse than the almost parametric rate $\log K(n)/n$, and 2) the number $K(n)$ of candidate estimators is polynomial in n.

Our observation is based on the following trivial lemma.

Lemma 11: *Suppose that* $a_n, b_n, c_n \in R$ *are such that* $0 < a_n \leq b_n \leq a_n + c_n$, *and that* $\limsup \frac{c_n}{a_n} = 0$. *Then* $\lim_{n\to\infty} a_n/b_n = 1$.

Proof: Dividing by a_n in the inequality, we obtain,
$0 < 1 \leq \liminf b_n/a_n \leq \limsup b_n/a_n \leq 1 + \limsup c_n/a_n = 1$
$\Longrightarrow \lim b_n/a_n = 1$
$\Longleftrightarrow \lim a_n/b_n = 1$.

Corollary 12: *Let* r_n *and* $h_{n,K(n)}$ *be defined as in Corollary 6, and assume that, in addition to the assumptions of Corollary 6, we have*

$$\frac{\max(r_n, \sqrt{\log K(n)/np})}{E[\tilde{\Theta}_{n(1-p)}(\tilde{k})]} \to 0 \text{ for } n \to \infty. \tag{19}$$

Then

$$\lim \frac{E[\tilde{\Theta}_{n(1-p)}(\tilde{k})]}{E[\tilde{\Theta}_{n(1-p)}(k_n)]} = 1.$$

Proof: This is immediate from Corollary 6 and Lemma 11. □

We note that the proportion p in Corollary 12 can be selected to converge to zero with sample size at a rate $p(n)$ so that $\log K(n)/np(n)$ remains of smaller order than $E[\tilde{\Theta}_{n(1-p)}(\tilde{k})]$. In this case, Corollary 12 provides also conditions under which the cross-validation selector is asymptotically equivalent with the oracle selector $\tilde{k}_n$.

5.3. *The oracle selector in convex linear models.*

Finally, we state here that our target criterian $\tilde{\Theta}_{n(1-p)}(\psi)$ actually reduces to the norm of the difference $\psi_0 - \psi$ in the case that the parameter Ψ is linear, and the parameter space is convex, and various other cases. This was made explicit in our examples in Section 3.

Corollary 13: *Suppose that*

$$P_0 D_b(O \mid \psi, \gamma_0) = \psi_{b0} - \psi_b \text{ for all } b \in \mathcal{B}.$$

This holds, in particular, if $D_b(O \mid \psi_0, \gamma_0)$ *is a gradient of a real valued* linear *parameter* $\Psi_b : \mathcal{M} \to \mathbb{R}$ *at* P_0 *(for all* $P_0 \in \mathcal{M}$*),* $\mathcal{M}$ *is convex, and*

the regularity conditions of Theorem 2.2 in [25] *hold. Let* ψ *be representing the vector* $(\psi_b : b \in \mathcal{B})$*, so that* $\| \psi - \psi_0 \| \equiv \| (\psi_b : b) - (\psi_{b0} : b) \|$ *is defined as the norm* $\| \cdot \|$ *of the corresponding vector.*

Then in Corollaries 6 and 12, $E[\tilde{\Theta}_{n(1-p)}(\tilde{k})]$ *and* $E[\tilde{\Theta}_{n(1-p)}(k_n)]$ *may be replaced by* $E\|\hat{\Psi}_{\tilde{k}}(P^0_{n,B_n}) - \Psi(P_0)\|$ *and* $E\|\hat{\Psi}_{k_n}(P^0_{n,B_n}) - \Psi(P_0)\|$*. Here* $\tilde{k}$ *is the minimizer over* $k \in \{1, 2, ..., K(n)\}$ *of the random quantity* $E_{B_n}\|\hat{\Psi}_k(P^0_{n,B_n}) - \Psi(P_0)\|$*.*

Proof: This is trivial. □

6. Estimating function based cross-validation for density/hazard estimation

In this section we present estimating function based cross-validation for density estimation and hazard estimation by specifying a particular choice of estimating function. In particular, we illustrate that the corresponding target criterian satisfies the exact difference between a candidate density and the true density. Our example would be completely analogue in the regression case, and is therefore omitted, and left to the interested reader. Density estimation and regression allow the application of loss-based cross-validation based on the minus log loss function and squared error loss function, respectively. Since density estimation and regression are two well studied problems in the literature, we decided that it is of interest to illustrate our method in these cases as well. We remark, however, that the need for estimating function based cross-validation was motivated by problems in which loss-based cross-validation is not easily available.

6.1. *Example: Density estimation.*

Let $X \sim P_0$ be a univariate random variable with density f_0. Suppose that the model is nonparametric, the parameter of interest is the density itself: $\Psi(P) = f$, $\psi_0 = \Psi(P_0) = f_0$. Let ϕ_b, $b = 1, \ldots$ be a countable orthonormal basis in the Hilbert space $L^2(dx)$ of square integrable functions w.r.t Lebesgue measure, where dx denotes the Lebesgue measure. Let $\Psi_b(P) = P\phi_b = \int \phi_b(x) f(x) dx$. Now, the efficient influence curve of $\Psi_b(P)$ is given by $D_b(X \mid P_0) = \phi_b(X) - E_0\phi_b(X)$. A corresponding estimating function is given by $D_b(X \mid \psi) = \phi_b(X) - \int \phi_b(x)\psi(x)dx$. Thus $\hat{\Theta}_{n(1-p)}(k) = E_{B_n} \| (P^1_{n,B_n}\phi_b - \int \phi_b(x)\hat{\Psi}_k(P^0_{n,B_n})(x)dx : b) \|$. Since

$P_0 D_b(\cdot \mid \psi) = \int \phi_b(x)(\psi_0 - \psi)(x)dx = \psi_{b0} - \psi_b$, we have

$$\tilde{\Theta}_{n(1-p)}(k) = E_{B_n} \parallel \left(\int \phi_b(x)(\hat{\Psi}_k(P^0_{n,B_n}) - \psi_0)(x)dx : b \right) \parallel .$$

If we use the Euclidean norm, then

$$\tilde{\Theta}_{n(1-p)}(k) = \sqrt{\int \left\{ \hat{\Psi}_k(P^0_{n,B_n})(x) - \psi_0(x) \right\}^2 dx}$$

is simply the $L^2(dx)$-norm between the candidate density estimator $\hat{\Psi}_k(P^0_{n,B_n})$ and the true density ψ_0.

6.2. *Example: Hazard estimation.*

Let $X \sim P_0$ be a univariate random variable with density f_0, survival function S_0 and hazard $\lambda_0 = f_0/S_0$. Suppose that the model is nonparametric, and the parameter of interest is the hazard: $\Psi(P) = \lambda = f/S$, $\psi_0 = \Psi(P_0) = f_0/S_0$. Let ϕ_b, $b = 1, \ldots$ be a countable orthonormal basis in $L^2(dx)$. Let $\Psi_b(P) = \int \phi_b(x)\lambda(x)dx$, where λ denotes the hazard corresponding with probability distribution P. Now, the efficient influence curve of the real valued parameter $\Psi_b(P)$ at P_0 is given by

$$D_b(X \mid P_0) = \frac{\phi_b(X)}{S_0(X)} - \int_0^X \frac{\phi_b(x)}{S_0(x)} \lambda_0(x)dx.$$

A corresponding estimating function for a candidate hazard ψ in thus defined as

$$D_b(X \mid \psi, S_0) = \frac{\phi_b(X)}{S_0(X)} - \int_0^X \frac{\phi_b(x)}{S_0(x)} \psi(x)dx,$$

which is thus indexed by a root-n estimable nuisance parameter S_0. Thus, the cross-validation criteria $\hat{\Theta}_{n(1-p)}(k)$ is defined as

$$E_{B_n} \parallel \left(P^1_{n,B_n}\phi_b/S_0 - \int P^1_{n,B_n} I(x < \cdot) \frac{\phi_b(x)}{S^0_{n,B_n}(x)} \hat{\Psi}_k(P^0_{n,B_n})(x)dx : b \right) \parallel,$$

where S^0_{n,B_n} denotes an estimator of the survival function based on the training sample P^0_{n,B_n}. Since $P_0 D_b(\cdot \mid \psi, S_0) = \int \phi_b(x)(\psi_0 - \psi)(x)dx = \psi_{b0} - \psi_b$, we have that the target criterian equals

$$\tilde{\Theta}_{n(1-p)}(k) = E_{B_n} \parallel \left(\int \phi_b(x)(\hat{\Psi}_k(P^0_{n,B_n}) - \psi_0)(x)dx : b \right) \parallel .$$

If we use the Euclidean norm, then

$$\tilde{\Theta}_{n(1-p)}(k) = \sqrt{\int \hat{\Psi}_k(P^0_{n,B_n})(x) - \psi_0(x)\}^2 dx}$$

is simply the $L^2(dx)$-norm between the candidate hazard estimator $\hat{\Psi}_k(P^0_{n,B_n})$ and the true hazard ψ_0. Finally, we note that our cross-validation selector for selection among hazard estimators would in first order not be affected by the nuisance parameter since the nuisance parameter S_0 can be estimated at a parametric rate, while the minimax rates for hazard estimation is worse than the root-n rate for all smoothness classes.

7. Unified estimating function based learning.

In this article we have focussed on estimating function based cross-validation methodology for selecting among a class of estimators. However, in this section, we show that the estimating function itself, in combination with a sieve, can be used to construct such candidate estimators, and thereby, in combination with cross-validation, results in an estimator of the parameter of interest. This definition of an estimator follows the complete analogue of the unified loss based estimation methodology presented in [28] and subsequent application papers by us, but with the empirical mean of the loss function replaced by the norm of the empirical mean of the estimating function.

7.1. *Road Map.*

Specifically, this unified estimating function based estimation methodology can be represented by the following road map.

Parameter of interest: Let $O_1, \ldots, O_n$ be i.i.d. observations of $O \sim P_0$, where it is known that $P_0 \in \mathcal{M}$ for a statistical model $\mathcal{M}$. Let $\Psi : \mathcal{M} \to D(\mathcal{S})$ be the parameter of interest, where $D(\mathcal{S})$ denotes the class of real valued functions on $\mathcal{S}$ (e.g $\mathcal{S} = \{1, \ldots, d\}$, or $\mathcal{S} = \mathbb{R}^d$). Let $\psi_0 = \Psi(P_0)$ be the true parameter value we aim to learn from the data.

Estimating function: Let $D(O, \psi \mid \upsilon) = (D_b(O \mid \psi, \upsilon) : b \in \mathcal{B})$ be an unbiased estimating function for ψ_0 with nuisance parameter Υ. That is, for each $b \in \mathcal{B}$, $(O, \psi, \upsilon) \to D_b(O \mid \psi, \upsilon)$ is a real valued function on the cartesian product of a support of P_0, the parameter space $\boldsymbol{\Psi}$ of Ψ, and the nuisance parameter space $\{\Upsilon(P) : P \in \mathcal{M}\}$

of a particular nuisance parameter Υ on model $\mathcal{M}$. In addition, it is assumed that the estimating function is unbiased in the sense that $P_0 D(\cdot \mid \psi_o, \Upsilon(P_0)) = 0$. In Section 2 it was shown how to construct such an estimating function in terms of the b-specific efficient influence curves of b-specific pathwise differentiable real valued parameters.

Nuisance parameter estimator: Let $P_n \to \hat{\Upsilon}(P_n)$ be an estimator of $\upsilon_0 = \Upsilon(P_0)$.

Criteria: Norm of estimating equation: Given a particular norm $\| \cdot \|$ on vectors $(x(b) : b \in \mathcal{B})$ with real valued components, we define the following empirical criteria on the parameter space $\mathbf{\Psi}$ of Ψ:

$$\hat{\Theta}(P_n)(\psi) \equiv \| P_n D(\cdot \mid \psi, \hat{\Upsilon}(P_n)) \| .$$

For example, if we use the euclidean norm with weights $w(b)$, we obtain

$$\hat{\Theta}(P_n)(\psi) \equiv \sum_{b \in \mathcal{B}} w(b) \left(P_n D_b(\cdot \mid \psi, \hat{\Upsilon}(P_n)) \right)^2 .$$

Sieve on parameter space: Let $\mathbf{\Psi}_s \subset \mathbf{\Psi}$ be a subspace of the parameter space indexed by $s \in \mathcal{A}_n$.

Subspace-specific estimators: For each subspace, we can define the estimator as the minimizer of the norm of the estimating equation:

$$\hat{\Psi}_s(P_n) = \arg \min_{\psi \in \mathbf{\Psi}_s} \hat{\Theta}(P_n)(\psi).$$

Cross-validation selector: Let

$$\hat{S}(P_n) = \arg \min_{s \in \mathcal{A}_n} E_{B_n} \| P^1_{n,B_n} D(\cdot \mid \hat{\Psi}_s(P^0_{n,B_n}), \hat{\Upsilon}(P^0_{n,B_n})) \| .$$

For example, if we use the Euclidean norm, we can choose (here we put E_{B_n} within square root, but outside squares):

$$\hat{S}(P_n) = \arg \min_{s \in \mathcal{A}_n} E_{B_n} \sum_{b \in \mathcal{B}} w(b) \left(P^1_{n,B_n} D_b(\cdot \mid \hat{\Psi}_s(P^0_{n,B_n}), \hat{\Upsilon}(P^0_{n,B_n})) \right)^2 .$$

Estimator: We estimate ψ_0 with

$$\hat{\Psi}(P_n) \equiv \hat{\Psi}_{\hat{S}(P_n)}(P_n).$$

This estimating equation methodology generalizes the estimating equation methodology as currently used for euclidean pathwise differentiable parameters such as (locally efficient) generalized estimating equations for repeated measures regression, and the locally efficient estimating function methodology for censored data as presented in [32].

8. Minimax adaptive estimating function based learning.

In [28] we proved that, for epsilon-net sieves $\mathbf{\Psi}_{s,\epsilon} \subset \mathbf{\Psi}_s$ indexed by both s *and* ϵ (where $\mathbf{\Psi}_{s,\epsilon}$ denotes a finite set of elements which approximates each element in $\mathbf{\Psi}_s$ within distance ϵ), this type of estimator based on the empirical mean of a loss function (instead of the norm of the estimating equation) satisfies a finite sample inequality, which implies that the estimator is minimax adaptive w.r.t. to the sequence of subspaces $\mathbf{\Psi}_s$, $s \in \mathcal{A}_n$. In this section we will establish the analogue of this finite sample inequality for the estimating function based estimator presented in the previous section.

Let $\mathbf{\Psi}_{s,\epsilon} = \{\psi_j^{s,\epsilon} : j = 1, \ldots, N_s(\epsilon)\} \subset \mathbf{\Psi}_s$ be a finite subset of size $N_s(\epsilon)$, $(s, \epsilon) \in \mathcal{A}_n$, where the size of $\mathcal{A}_n$ is denoted with $K(n)$. We view $\mathbf{\Psi}_{s,\epsilon}$ as a discrete approximation of $\mathbf{\Psi}_s$, where the approximation error is an increasing function in ϵ, and the approximation error converges to zero for $\epsilon \to 0$. If we set $\mathbf{\Psi}_{s,\epsilon}$ equal to an ϵ-net of $\mathbf{\Psi}_s$, that is, a set of points such that each element of $\mathbf{\Psi}_s$ is within a distance ϵ of a point in $\mathbf{\Psi}_{s,\epsilon}$, then our finite sample inequality below implies minimax adaptive rates of convergence for our estimator.

For each $(s, \epsilon) \in \mathcal{A}_n$, we define the estimator

$$\hat{\Psi}_{s,\epsilon}(P_n) = \arg \min_{\psi \in \mathbf{\Psi}_{s,\epsilon}} E_{B_n} \parallel P^1_{n,B_n} D(\cdot \mid \psi, \hat{\Gamma}(P^0_{n,B_n})) \parallel .$$

We select the subspace-index s and the resolution ϵ with estimating function based cross-validation:

$$(s_n, \epsilon_n) = (\hat{S}(P_n), \hat{E}(P_n)) = \arg \min_{s,\epsilon} E_{B_n} \parallel P^1_{n,B_n} D(\cdot \mid \hat{\Psi}(P^0_{n,B_n}), \hat{\Gamma}(P^0_{n,B_n})) \parallel .$$

Our estimator is given by

$$\hat{\Psi}(P_n) \equiv \hat{\Psi}_{s_n,\epsilon_n}(P_n).$$

The next theorem presents a finite sample inequality for this estimator.

Theorem 14: Define

$$r(n) \equiv 2 \sup_{\psi \in \mathbf{\Psi}} E_{B_n} \parallel P_0 \left\{ D(\cdot \mid \psi, \hat{\Gamma}(P^0_{n,B_n})) - D(\cdot \mid \psi, \gamma_0) \right\} \parallel .$$

We will assume that

$$Er(n) \leq r_{nuis}(n(1-p)),$$

where $r_{nuis}(n)$ is a particular function of n converging to zero. We also assume that

$$\sup_{o,\psi,v,b} \mid D_b(o \mid \psi, v) \mid \leq M < \infty,$$

where the supremum is taken over the cartesian product of a support of P_0, the parameter space for Ψ, and the nuisance parameter space for Υ.

Let

$$B_0(\epsilon, s) \equiv \min_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \parallel P_0 D(\cdot \mid \psi, \gamma_0) \parallel .$$

Then,

$$E \parallel P_0 D(\cdot \mid \hat{\Psi}_{s_n,\epsilon_n}(P^0_{n,B_n}), \gamma_0) \parallel \leq \min_{s,\epsilon} \left\{ B_0(\epsilon, s) + c(M) \frac{\log N_s(\epsilon)}{n(1-p)p} \right\}$$
$$+ r_{nuis}(n(1-p)^2) + r_{nuis}(n(1-p)) + c(M) \frac{\log K(n)}{np}.$$

Proof: In this proof we will use the same short-hand notation as introduced at the start of Section **??**.

Given a collection of candidate estimators $\mathcal{E} \equiv \{\hat{\Psi}_k : k\}$ whose realizations are in $\mathbf{\Psi}$, and the distribution F_{B_n} of the cross-validation scheme B_n, we define the following function of the empirical distribution function P_n:

$$R_e(P_n \mid \mathcal{E}, F_{B_n}) \equiv \frac{1}{\sqrt{np}} E_{B_n} \left\{ \parallel G^1_{n,B_n} D^0_{n,B_n}(\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \right\} +$$
$$\frac{1}{\sqrt{np}} E_{B_n} \left\{ \parallel G^1_{n,B_n} D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n})) \parallel + \parallel G^1_{n,B_n} \left(D^0_{n,B_n} - D \right) (\cdot \mid \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) \parallel \right\}.$$

Given a collection of candidate estimators $\mathcal{E} \equiv \{\hat{\Psi}_k : k \in \mathcal{A}_n\}$, and the distribution F_{B_n} of the cross-validation scheme B_n, we also define

$$R_{nuis}(P_n \mid \mathcal{E}, F_{B_n}) \equiv 2 \max_{k \in \{1,\ldots,K(n)\}} E_{B_n} \parallel P_0 (D^0_{n,B_n} - D)(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n})) \parallel .$$

Theorem 4 states that

$$E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n}), \gamma_0) \parallel \leq \min_k E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \gamma_0) \parallel$$
$$+ R_e(P_n \mid \mathcal{E}, F_{B_n}) + R_{nuis}(P_n \mid \mathcal{E}, F_{B_n}). \quad (20)$$

Under the condition that $\sup_{o,\psi,\gamma,b} \mid D_b(o \mid \psi, \gamma) \mid \leq M < \infty$, by Lemma 7, we have that

$$E R_e(P_n \mid \mathcal{E}, F_{B_n}) \leq c(M) \frac{\log \mid \mathcal{E} \mid}{np},$$

where $c(M)$ only depends on M. That is, this upper bound on the expectation of $R_e(P_n \mid \mathcal{E}, F_{B_n})$ only depends on the class of estimators through the actual number of estimators.

We can bound the R_{nuis} term as follows:

$$R_{nuis}(P_n \mid \mathcal{E}, F_{B_n}) \leq 2 \sup_{\psi \in \mathbf{\Psi}} E_{B_n} \parallel P_0(D^0_{n,B_n} - D)(\cdot \mid \psi) \parallel$$
$$\equiv R_{nuis}(P_n \mid F_{B_n}),$$

where the latter quantity does not depend on the class of estimators anymore. In the theorem we assumed that

$$ER_{nuis}(P_n \mid F_{B_n}) \leq r_{nuis}(n(1-p)),$$

where $r_{nuis}(n)$ is a particular function of n converging to zero.

To simplify notation, we will denote $R_e(P_n \mid \mathcal{E}, F_{B_n})$ with $R_e(P_n \mid K(n), p)$, since in our analysis only the expectation bounds on these random variables matter, and as shown above, these only depend on the number of estimators $K(n)$ and the proportion p constituting the validation sample. Similarly, we will denote $R_{nuis}(P_n \mid F_{B_n})$ with $R_{nuis}(P_n \mid p)$. Using this notation, the finite sample inequality (**??**) reads as:

$$E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{k_n}(P^0_{n,B_n}), \gamma_0) \parallel \leq \min_k E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_k(P^0_{n,B_n}), \gamma_0) \parallel$$
$$+ R_e(P_n \mid K(n), p) + R_{nuis}(P_n \mid p). \quad (21)$$

The proof of the Theorem is essentially a double application of this inequality (**??**), as we will show now.

Application of this inequality (**??**) to the estimators $\hat{\Psi}_{\epsilon,s}$, $(s, \epsilon) \in \mathcal{A}_n$, yields

$$E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{s_n,\epsilon_n}(P^0_{n,B_n})) \parallel \leq \min_{s,\epsilon} E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{s,\epsilon}(P^0_{n,B_n})) \parallel$$
$$+ R_e(P_n \mid K(n), p) + R_{nuis}(P_n \mid p). \quad (22)$$

Application of (**??**) to the constant estimators $\hat{\Psi}^{s,\epsilon}_j(P_n) = \psi^{s,\epsilon}_j$, $j = 1, \ldots, N_s(\epsilon)$, yields

$$\parallel P_0 D(\cdot \mid \hat{\Psi}_{s,\epsilon}(P_n)) \parallel \leq \min_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \parallel P_0 D(\cdot \mid \psi) \parallel$$
$$+ R_e(P_n \mid N_s(\epsilon), p) + R_{nuis}(P_n \mid p).$$

Application of the latter inequality to the empirical distribution P^0_{n,B_n} of the B_n-specific training sample gives us:

$$\parallel P_0 D(\cdot \mid \hat{\Psi}_{s,\epsilon}(P^0_{n,B_n})) \parallel \leq \min_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \parallel P_0 D(\cdot \mid \psi) \parallel$$
$$+ R_e(P^0_{n,B_n} \mid N_s(\epsilon), p) + R_{nuis}(P^0_{n,B_n} \mid p).$$

Substitution of this inequality in (**??**), and noting that $E_{B_n} R_{nuis}(P^0_{n,B_n} \mid p)$ does not depend on (ϵ, s), yields:

$$E_{B_n} \parallel P_0 D(\cdot \mid \hat{\Psi}_{s_n,\epsilon_n}(P^0_{n,B_n})) \parallel \leq \tag{23}$$

$$\min_{s,\epsilon} \left\{ \min_{\psi \in \mathbf{\Psi}_{\epsilon,s}} \parallel P_0 D(\cdot \mid \psi) \parallel + E_{B_n} R_e(P^0_{n,B_n} \mid N_s(\epsilon), p) \right\}$$

$$+ E_{B_n} R_{nuis}(P^0_{n,B_n} \mid p) + R_e(P_n \mid K(n), p) + R_{nuis}(P_n \mid p). \tag{24}$$

Taking expectations on both sides of the latter inequality provides us with

$$\begin{aligned} E \parallel P_0 D(\cdot \mid \hat{\Psi}_{s_n,\epsilon_n}(P^0_{n,B_n})) \parallel &\leq \min_{s,\epsilon} \left\{ B_0(\epsilon, s) + E R_e(P^0_{n,B_n} \mid N_s(\epsilon), p) \right\} \\ &+ E R_{nuis}(P^0_{n,B_n} \mid p) + E R_e(P_n \mid K(n), p) + E R_{nuis}(P_n \mid p) \\ &\leq \min_{s,\epsilon} \left\{ B_0(\epsilon, s) + c(M) \frac{\log N_s(\epsilon)}{n(1-p)p} \right\} \\ &+ r_{nuis}(n(1-p)^2) + r_{nuis}(n(1-p)) + c(M) \frac{\log K(n)}{np}, \end{aligned}$$

where we applied the previously stated bounds on the expectations of the R_e and R_{nuis} terms. This proves the theorem. □

Asymptotic interpretation of finite sample inequality. In order to interpret this finite sample inequality one should first notice that if the number $K(n)$ of candidate sets $\mathbf{\Psi}_{\epsilon,s}$ is polynomial in sample size, which is neither a theoretical (for achieving optimal rates of convergence) or practical limitation, then the term $\log K(n)/np = O(\log n/n)$. Secondly, the contribution due to the estimation of γ_0 is bounded by the rate $r_{nuis}(n)$. By selecting the estimating function in the manner described in Section 2, one expects that $\gamma \to P_0 D(\cdot \mid \psi_0, \gamma)$ has directional derivatives equal to zero. Consequently, in that case, the rate $r_{nuis}(n)$ might be mainly a function of a second order difference between $\hat{\Gamma}(P_n)$ and γ_0. In the situation that $r_{nuis}(n)$ is relatively small (in relation to the first term) and $K(n)$ is polynomial in sample size, the driving term is the minimum over ϵ, s of the sum of the approximation error $B_0(\epsilon, s)$ for the sieve $\mathbf{\Psi}_{s,\epsilon}$ and the logarithm of the covering number $N_s(\epsilon)$ divided by n. Using known covering numbers for a variety of sieves and an estimating function so that $\parallel P_0 D(\cdot \mid \psi, \gamma_0) \parallel$ correponds with a standard norm between ψ and ψ_0 (e.g., see regression and density estimation example in Section **??**), it follows that this trade-off corresponds with a minimax adaptive rate of convergence in the classical regression and density examples. We refer to [28] and [30] for such illustrations.

9. Summary

In this article we introduced a cross-validation method for selecting an estimator among a class of estimators, which only requires specification of an (possibly infinite dimensional) estimating function for the parameter of interest. We provide a method, based on canonical gradients of a collection of real valued path-wise differentiable parameters, for constructing an estimating function so that its expectation at a candidate parameter value equals in first order a difference between the candidate and the true parameter value. In this manner, the estimating function based cross-validation selector is aimed at selecting the estimator which performs best w.r.t. to the parameter of interest. This is formally established by showing that 1) if the parameter is path-wise differentiable, our cross-validation selector data adaptively under-smooths so that it results in an asymptotically efficient (or superefficient) estimator, and 2) if none of the candidate estimators achieve the parametric rate, then the cross-validation selector is asymptotically equivalent with the oracle selector. We also provided various examples indicating the wide range of applications for this methodology. In particular, we show that our methodology solves a long standing problem of how to *data adaptively* smooth an empirical cumulative distribution function (or Kaplan-Meier estimator) while preserving the asymptotic efficiency. Finally, we generalize estimating function methodology for path-wise differentiable parameters to a completely general sieve based estimating function methodology, thereby generalizing estimating function methodology fo0r pathwise differentiable parameters and "machine learning" for regression and density estimation, to learning of general parameters. We formally prove that, if one augments a given sieve with the inclusion of epsilon-nets within each element of the sieve, then this estimator satisfies a finite sample inequality, which shows it is mini-max adaptive w.r.t. the sieve and w.r.t. to the estimating function based norm. In the future we plan to investigate and apply this general estimating function methodology to a variety of examples.

Appendix.

Some useful lemmas

Our proof of finite sample results is based on Bernstein's inequality, which we state here as a lemma for ease of reference. A proof is given in Lemma A.2, p. 564 in [12].

Lemma 15: Bernstein's inequality. *Let* Z_i, $i = 1, \ldots, n$, *be independent*

real valued random variables such that $Z_i \in [a,b]$ *with probability one. Let* $0 < \sum_{i=1}^n VAR(Z_i)/n \leq \sigma^2$. *Then, for all* $\epsilon > 0$,

$$Pr\left(\frac{1}{n}\sum_{i=1}^n (Z_i - EZ_i) > \epsilon\right) \leq \exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right).$$

This implies

$$Pr\left(\frac{1}{n} \mid \sum_{i=1}^n (Z_i - EZ_i) \mid > \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right).$$

We have the following immediate corollary of Bernstein's inequality, which allows us to obtain the wished tail probabilities for products Z_{1n}^2, $Z_{1n}Z_{2n}$.

Lemma 16: Bernstein's inequality. *Given arbitrary random variables* (Z_{1n}, Z_{2n}), *we have*

$$P(Z_{1n}Z_{2n} \geq s) \leq P(Z_{1n} \geq \sqrt{s}) + P(Z_{2n} \geq \sqrt{s}).$$

This allows us to obtain explicit tail probabilities from tail probabilities derived for Z_{1n} *and* Z_{2n} *separately. In particular, if* (Z_{1i}, Z_{2i}), $i = 1, \ldots, n$, *are independent bivariate random variables such that* $Z_{ji} \in [a,b]$, $j = 1,2$, *with probability one,* $0 < \sum_{i=1}^n VAR(Z_{ji})/n \leq \sigma^2$, $j = 1,2$, *then, for all* $\epsilon > 0$,

$$Pr\left(\frac{1}{n}\sum_{i=1}^n (Z_{1i} - EZ_{1i})\frac{1}{n}\sum_{i=1}^n (Z_{2i} - EZ_{2i}) > \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon}{\sigma^2 + \sqrt{\epsilon}(b-a)/3}\right).$$

This implies

$$Pr\left(\mid \frac{1}{n}\sum_{i=1}^n (Z_{1i} - EZ_{1i})\frac{1}{n}\sum_{i=1}^n (Z_{2i} - EZ_{2i}) \mid > \epsilon\right) \leq 4\exp\left(-\frac{1}{2}\frac{n\epsilon}{\sigma^2 + \sqrt{\epsilon}(b-a)/3}\right).$$

These bounds can be directly translated into bounds on the corresponding expectations. In particular, we can apply the following simple lemma.

Lemma 17: *Let* Z_n *be a random variable satisfying that* $P(Z_n \geq s) \leq C(n)\exp(-ns/c)$ *for all* $s \geq 0$. *Then* $EZ_n \leq \frac{c(\log C(n)+1)}{n}$.

Acknowledgments

This work was supported by NIH grants NIH R01 GM67233 and NIH R01 GM071397. We thank Rob Strawderman and Aad van der Vaart for helpful email discussions.

References

1. P. K. Andersen, O. Borgan, R.D. Gill, N. Keiding, Statistical models based on counting processes, Springer-Verlag, New York (1992).
2. E. Bein, A.E. Hubbard, and M.J. van der Laan, Estimating function based cross-validation to estimate the causal effect in randomized trials with noncompliance, Technical report, Division of Biostatistics, University of California, Berkeley (2005).
3. P.J. Bickel, C.A.J. Klaassen, Y.Ritov, and J. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Springer Verlag, (1997).
4. L. Breiman, The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error, *Journal of the American Statistical Assocation* **87**(419), 738–754 (1992).
5. L. Breiman, Heuristics of instability and stabilization in model selection, *Annals of Statistics* **24** (6), 2350–2383 (1996a)
6. L. Breiman, Out of bag estimation, Technical report, Department of Statistics, University of California, Berkeley (1996b).
7. L. Breiman, J.H. Friedman, R. Olshen, C.J. Stone, *Classification and regression trees*, The Wadsworth Statistics/Probability series, Wadsworth International Group (1984).
8. L. Breiman and P. Spector, Submodel selection and evaluation in regression: the x-random case, *International Statistical Review* **60**, 291–319 (1992).
9. S.R. Cosslett, Efficient semiparametric estimation of censored and truncated regressions via a smoothed self-consistency equation, *Econometrica*, **72**(4), 1277–1284.
10. S. Dudoit and M.J. van der Laan, Asymptotics of cross-validated risk estimation in model selection and performance assessment, Technical Report **126**, Division of Biostatistics, University of California, Berkeley, Feb. 2003, URL: www.bepress.com/ucbbiostat/paper126, to appear in The Indian Journal of Statistical Methodology (2003).
11. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC (1993).
12. L. Györfi, M. Kohler, A. Krzyźak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York, 2002.
13. L. Gyorfi, D. Schäfer, and H. Walk, Relative Stability of global errors of nonparametric function estimators, *IEEE Transactions of Information Theory*, **48**(8), 2230: 2242 (2002).
14. W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press (1993).
15. W. Härdle and J.S. Marron, Asymptotic Equivalence of some bandwidth selectors in nonparametric regression, *Biometrika*, **72**, 481–484 (1985a).
16. W. Härdle and J.S. Marron, Optimal bandwidth selectin in nonparametric regression function estimation, *Annals of Statistics*, **13**, 1465–1481 (1985b).
17. T. Hastie, R. Tibshirani, and J.H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag (2001).
18. S. Keles, M.J. van der Laan, and S. Dudoit, Asymptotically optimal model

selection method with right-censored outcomes, Technical Report 124, Division of Biostatistics, University of California, Berkeley, Sept. 2003, URL: www.bepress.com/ucbbiostat/paper124, to appear in *Bernoulli*.

19. C.A.J. Klaassen, Consistent estimation of the influence function of locally asymptotically linear estimators, *Annals of Statistics*, **15**, 1548–1562 (1987).
20. A.M. Molinaro, S. Dudoit, and M.J. van der Laan, Tree-based multivariate regression and density estimation with right-censored data, In S. Dudoit, R. C. Gentleman, and M. J. van der Laan (eds), Special Issue on Multivariate Methods in Genomic Data Analysis, *Journal of Multivariate Analysis*, Vol. 90, No. 1, p. 154-177.
21. B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge, New York, 1996.
22. B.W. Silverman, A fast and efficient cross-validation method for smoothing parameter choice in spline regression, *Journal of the American Statistical Association*, **79** (387), 584–589 (1984).
23. M. Stone, Cross-validatory choice and assessment of statistics predictions, *Journal of the Royal Statistical Society, Series B*, **36** (2), 111–147 (1974).
24. M. Stone, Asymptotics for and against cross-validation, *Biometrika*, **64** (1), 29–35 (1977).
25. M.J. van der Laan, Efficient and Inefficient Estimation in Semiparametric Models. CWI-tract **114**, Centre for Mathematics and Computer Science, Amsterdam, the Netherlands.
26. M.J. van der Laan, An Identity for the Nonparametric Maximum Likelihood Estimator in Missing Data and Biased Sampling Models. *Bernoulli* **1(4)**, pp. 335–341 (1995).
27. M.J. van der Laan, Identity for NPMLE in Censored Data Models, *Lifetime Data Models* **4**, 83–102 (1998).
28. M.J. van der Laan, S. Dudoit, *Unified Cross-Validation Methodology For Selection among Estimators, and a General Cross-validated Adaptive epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples*, Working Paper #130, Division of Biostatistics, UC Berkeley.
29. M. J. van der Laan, S. Dudoit, S. Keles (2004), Asymptotic Optimality of Likelihood-Based Cross-Validation, *Statistical Applications in Genetics and Molecular Biology* Vol. 3: No. 1, Article 4. http://www.bepress.com/sagmb/vol3/iss1/art4.
30. M.J. van der Laan, S. Dudoit, A.W. van der Vaart (2004), The Cross-Validated Adaptive Epsilon-Net Estimator, U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 142. http://www.bepress.com/ucbbiostat/paper142.
31. M.J. van der Laan, A. Hubbard, N.P. Jewell, Estimation of Treatment Effects in Randomized Trials with Noncompliance and a Dichotomous Outcome, technical report , Division of Biostatistics, http://www.bepress.edu/ucbbiostat, submitted for publication in the Journal of the Royal Statistical Society B (2004).
32. M.J. van der Laan, J.M. Robins (2002), *Unified Methods for Censored Longitudinal Data and Causality*, Springer Verlag, New York.

33. A.W. van der Vaart, J.A. Wellner, *Weak Convergence and Empirical Processes: with Applications to Statistics*, Springer Verlag, New York (1996).