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

Year 2014 *Paper* 322

Sieve Plateau Variance Estimators: A New Approach to Confidence Interval Estimation for Dependent Data

Molly M. Davies* Mark J. van der Laan†

*University of California, Berkeley, molly_davies@berkeley.edu

†University of California, Berkeley, laan@berkeley.edu

Sieve Plateau Variance Estimators: A New Approach to Confidence Interval Estimation for Dependent Data

Molly M. Davies and Mark J. van der Laan

Abstract

<blockquote>Suppose we have a data set of n-observations where the extent of dependence between them is poorly understood. We assume we have an estimator that is squareroot-consistent for a particular estimand, and the dependence structure is weak enough so that the standardized estimator is asymptotically normally distributed. Our goal is to estimate the asymptotic variance of the standardized estimator so that we can construct a Wald-type confidence interval for the estimate. In this paper we present an approach that allows us to learn this asymptotic variance from a sequence of influence function based candidate variance estimators. We focus on time dependence, but the method we propose generalizes to data with arbitrary dependence structure. We show our approach is theoretically consistent under appropriate conditions, and evaluate its practical performance with a simulation study, which shows our method compares favorably with various existing subsampling and bootstrap approaches. We also include a real-world data analysis, estimating an average treatment effect (and a confidence interval) of ventilation rate on illness absence for a classroom observed over time.</blockquote>

1 Introduction

For the sake of defining the challenge addressed in this paper, let's first suppose one observes realizations of n independent and identically distributed (i.i.d.) random variables. Consider a particular estimator ψ_n of a specified target parameter ψ_0. Statistical inference can now proceed in a variety of ways. Suppose we can prove the estimator is asymptotically linear with a specified influence function, i.e. $\psi_n - \psi_0$ can be written as an empirical mean of the influence function applied to the observation, plus a second order term assumed to converge to zero in probability at a rate faster than $\sqrt{n}$. In that case, it is known that the $\sqrt{n}$-standardized estimator converges to a normal distribution with asymptotic variance σ^2 equal to the variance of the influence function. An estimator of σ^2, which we denote σ_n^2, provides an asymptotic 0.95-confidence interval $\psi_n \pm 1.96\sqrt{\sigma_n^2/n}$. One way to estimate σ^2 is to estimate the influence function and set σ_n^2 equal to the sample variance of the estimated influence function values. We could also use other approaches such as the nonparametric bootstrap or subsampling.

In this paper, we are concerned with obtaining valid statistical inference when the data are known to be dependent, but the precise nature of that dependence is unknown. Specifically, we are interested in a method that will work well for estimators of relatively complex parameters one finds in semiparametric causal inference applications. We assume throughout that the estimator behaves in first order like an empirical mean of dependent random variables. We refer to such an estimator as (generalized) asymptotically linear and these random variables as (generalized) influence functions. In addition, we assume the dependence between influence functions is sufficiently weak so that the $\sqrt{n}$-standardized estimator converges to a normal distribution with mean zero and variance σ_0^2. We focus on time dependence, as it is well-represented in the literature. However,the methods we propose are generally applicable. Dependence could be spatiotemporal, for example, or over a poorly understood network. We discuss such extensions throughout. We limit ourselves to the case of positive covariances, but again, the method has a natural extension to general covariance structures.

Numerous blocked bootstrap and subsampling approaches have been developed to accommodate

unknown time dependence, and there are comprehensive book-length treatments of both (see for example Lahiri [2013] for blocked bootstrap approaches and Politis et al. [1999] for subsampling). These approaches involve estimating a tuning parameter b. In the context of blocked bootstraps, b corresponds to the size of the contiguous blocks resampled with replacement, or to the geometric mean of the block size in the case of stationary block bootstrapping. Blocked bootstraps have been shown to perform well when the optimal b is known or can be effectively estimated from the data. However, these estimators are sensitive to the choice of b, which is frequently difficult to estimate. The bootstrap approach also relies on some important regularity conditions not always met by all asymptotically linear, normally distributed estimators. For example, if the influence function depends on the true data generating distribution through densities, there is a literature warning against the application of a nonparametric bootstrap, and refinements and regularizations will be needed. See Mammen [1992] for a comprehensive discussion and examples.

In subsampling, b corresponds to the size of the contiguous subsample. One of subsampling's most attractive features is that it requires very few assumptions: the size of the subsample must be such that as sample size $n \to \infty$, $b \to \infty$ and $b/n \to 0$; and the standardized estimator must converge to some limit distribution. One need not know the rate of convergence nor the specific limit distribution. However, finite sample performance can be heavily dependent on the choice of b, which must be large enough to capture dependence, yet small enough to adequately approximate the underlying target distribution. Finding an optimal b for any given estimator is a nontrivial undertaking [Politis and Romano, 1993], and for more complex estimators, the sample sizes required in order to adequately estimate ψ_0 in each subsample can be impractically large.

We present a method of learning from sequences of ordered, sparse covariance structures on influence functions, where dependence decreases monotonically with distance. We assume there exists an (unknown) distance threshold $\boldsymbol{\tau}_{0,t}$ for each time point t such that any observation farther than $\boldsymbol{\tau}_{0,t}$ away from observation t is independent from observation t. Our proposed procedure seeks to select a variance estimate close to what we would have obtained had we known the true distance thresholds $(\boldsymbol{\tau}_{0,t} : t = 1, ..., n)$. Assume for a moment this dependence structure is constant across

time, i.e. $\tau_{0,t}$ is equal to some positive integer τ_0 for all t. Theory tells us a variance estimator ignoring this dependence will result in a biased estimate, and the magnitude of this bias will decrease as the number of nonzero covariances used in the variance estimate increases, until all true nonzero covariances are incorporated. Estimates assuming nonzero covariances beyond this will be unbiased, but will become more variable. This simple insight provides the rationale for our proposed approach. Intuitively, Sieve Plateau (SP) variance estimation searches for a 'plateau' in a sequence of variance estimates that assume increasing numbers of nonzero covariances. While our approach requires stronger assumptions than subsampling, its performance does not depend heavily on additional tuning parameters. It can also be used with complex estimators that require substantial sample sizes for proper estimation, and in settings where contiguity of the dependence structure is incompletely understood.

The remainder of this paper is organized as follows. In section 2, we define the formal estimation problem. In section 3, we introduce SP variance estimation and the intuitive rationale behind the approach. In section 4, we provide a more formal justification for why our proposed method works and state conditions upon which our method relies. Section 5 describes several specific implementations of our proposed approach within the context of estimating the variance of a sample mean of a time series. We also present results of an extensive simulation study, which demonstrate our approach works well in practice and consistently outperforms subsampling and blocked bootstrapping in a context where they are known to perform well. In section 6, we present a real data analysis, estimating the Average Treatment Effect (ATE) of ventilation rate on illness absence in an elementary school classroom. We also discuss why subsampling and blocked bootstraps are ill-suited to this particular estimation problem. We conclude with a discussion.

Notation conventions. The distance between two points x and y is denoted $d(x, y)$. Parameters with subscript 0 are features of the true data probability distribution. Subscript n indicates an estimator or some quantity that is a function of the empirical distribution. If f is a function of the observed data and P a possible probability distribution of the data, then Pf is the expectation of f taken with respect to P.

2 Target Parameter

Let $O = (O_1, \ldots, O_n)$ be a data set consisting of n time-ordered observations on a certain random process. Let P_0^n be the probability distribution of O. Let ψ_0 be a real valued feature of P_0^n, i.e. $\psi_0 = \Psi(P_0^n)$ for some mapping Ψ from the statistical model for P_0^n into the real line. We are given an estimator ψ_n based on O. We assume dependence structure in O is sufficiently weak so that ψ_n satisfies an expansion

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} D_i(O;\theta_0) + r_n,$$

where r_n is a second order term assumed to converge to zero in probability at a rate faster than $\sqrt{n}$. The functions $D_i(O;\theta_0)$ are called influence functions. They have expectation zero under P_0^n and depend on the unknown P_0^n through some parameter θ_0. For clarity, we often notate them as $D_i(\theta_0)$, but it is important to remember that they are functions of O. Assume there is a function f that assigns a measure $S = f(O)$ to each unit observation O, and define $s_i = f(O_i)$, $i = 1, \ldots, n$. Assume for all s_i there exists a bounded $\boldsymbol{\tau}_{0,i}$ such that $d(s_i, s_j) > \boldsymbol{\tau}_{0,i}$ implies the covariance $P_0^n D_{0,i} D_{0,j} = 0$. Define $\Omega_{\boldsymbol{\tau}_0,i} = \Omega_{0,i}$ to be the set of j such that $d(s_i, s_j) \leq \boldsymbol{\tau}_{0,i}$. Define

$$\sigma_{0n}^2 \equiv \text{VAR}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(\theta_0)\right\} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\Omega_{0,i}} P_0^n\left\{D_i(\theta_0)D_j(\theta_0)\right\}. \tag{1}$$

In many cases one might also assume $\sigma_{0n}^2 \to \sigma_0^2$ in probability for some fixed σ_0^2, but this is not necessary. We assume that as $n \to \infty$, $\mathbb{E}\big(r_n\sqrt{n}\big)^2 \to 0$ and $\sigma_{0n}^{-1} Z(n) \equiv \sigma_{0n}^{-1}/\sqrt{n}\sum_{i=1}^{n} D_i(\theta_0) \Rightarrow_d N(0,1)$. These assumptions imply that as $n \to \infty$, the standardized estimator converges weakly to a mean zero normal distribution with variance one, i.e.

$$\sigma_{0n}^{-1}\sqrt{n}(\psi_n - \psi_0) = \sigma_{0n}^{-1} Z(n) + o_P(1) \Rightarrow_d N(0,1).$$

If σ_{0n}^2 converges to a fixed σ_0^2, then $Z(n) \Rightarrow_d N(0,\sigma_0^2)$ and $\sqrt{n}(\psi_n - \psi_0) \Rightarrow_d N(0,\sigma_0^2)$. Thus $1/\sqrt{n}$ denotes the rate at which ψ_n converges to ψ_0. This paper is concerned with estimating σ_{0n}^2 so we can construct an asymptotic 0.95-confidence interval $\psi_n \pm 1.96\sqrt{\sigma_n^2/n}$, where σ_n^2 is an estimator of σ_{0n}^2.

Consider estimators of the form

$$\sigma_n^2(\boldsymbol{\tau}) \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{\boldsymbol{\tau}}(i,j)D_i(\theta_n)D_j(\theta_n), \tag{2}$$

where $\boldsymbol{\tau}$ in a parameter space T is an n-dimensional vector defining $\delta_{\boldsymbol{\tau}}(i,j) = I\{d(s_i,s_j) < \boldsymbol{\tau}_i\}$. Our conditions require the size (as measured by entropy) of the parameter space T remains controlled as $n \to \infty$. An extreme example would be $\boldsymbol{\tau}$ such that $\boldsymbol{\tau}_i = \lambda$ for all $i = 1,\ldots,n$ and a constant $\lambda > 0$. More generally, one might parametrize $\boldsymbol{\tau}_i = \boldsymbol{\tau}_i(\lambda)$ for a fixed dimensional parameter λ, or one might assume T is a union of a finite number of such parametric subsets. An influence function (IF) based oracle estimator of (1) is defined as (2) using $\delta_{\boldsymbol{\tau}} = \delta_{\boldsymbol{\tau}_0}$.

It will be convenient for us to define the notion of an arbitrary $\boldsymbol{\tau}$ being 'at least as large as' $\boldsymbol{\tau}_0$. Let $\Omega_{\boldsymbol{\tau}_i} = \{j : d(s_i,s_j) \le \boldsymbol{\tau}_i\}$ be the set defined by $\boldsymbol{\tau}_i$. Define $T_0 \subset T$ as the set of $\boldsymbol{\tau}$ where $\delta_{\boldsymbol{\tau}}(i,j) \ge \delta_{\boldsymbol{\tau}_0}(i,j)$ for all i,j. Thus, T_0 contains $\boldsymbol{\tau}_0$ and all other $\boldsymbol{\tau}$ such that for all i, any element in $\Omega_{\boldsymbol{\tau}_0,i}$ is also in $\Omega_{\boldsymbol{\tau}_i}$. When we say $\boldsymbol{\tau}$ 'contains' $\boldsymbol{\tau}_0$, we mean $\boldsymbol{\tau}$ is an element of T_0. In classical time series where dependence decays over time, it also makes sense to say $\boldsymbol{\tau}$ is at least as large as $\boldsymbol{\tau}_0$ when $\boldsymbol{\tau} \in T_0$.

3 Sieve Plateau Estimators

Suppose we have a collection of vectors $(\boldsymbol{\tau}_{n,k} : k)$, a proportion of which are in T_0. Consider the associated collection of variance estimators $(\sigma_n^2(\boldsymbol{\tau}_{n,k}) : k)$. Under conditions similar to those required for asymptotic normality of ψ_n, estimators based on $\boldsymbol{\tau}_{n,k} \in T_0$ will be unbiased (see Theorem 2). Suppose we could order these variance estimators so that they start out making too many independence assumptions and end up making too few. A smoothed version of this ordered sequence would give a curve with a plateau in the tail. In particular, since we are assuming all true covariances between influence functions are positive, we would expect this curve to be monotone increasing. We propose finding the plateau of this smoothed curve and using this knowledge to estimate σ_{0n}^2. The general approach is as follows.

Construct a sieve of variance estimators.

1. Formulate a set of vectors $\{\boldsymbol{\tau}_{n,k} : k\}$ in T that covers a range of covariance structures starting

with independence of all influence functions or some other appropriate lower bound and ending with covariance structures $\boldsymbol{\tau} \in T_0$. These vectors can reflect true knowledge about the dependence in one's data. For instance, in the case of time dependence, if one knows the dependence lag between all influence functions is constant, a simple sequence of vectors could assume increasing constant dependence lags, starting with 0 and ending with an upper bound $\boldsymbol{\tau}'$. Alternatively, one might believe the dependence structure is either constant or fluctuates smoothly over time according to seasonal trends. Note it is not necessary for $\boldsymbol{\tau}_0$ to be an element of $\{\boldsymbol{\tau}_{n,k} : k\}$. We describe more complex approaches to generating $\boldsymbol{\tau}$ vectors in section 5.

2. For each $\boldsymbol{\tau}_{n,k}$, compute $\sigma_n^2(\boldsymbol{\tau}_{n,k})$ as defined in equation (2).

3. Order the variance estimates so that they become increasingly unbiased for σ_0^2. A valid ordering need not be perfect. For example, we could define a matrix $M_{\boldsymbol{\tau}_{n,k}}$, whose (i,j)-th element is $D_i(\theta_n)D_j(\theta_n)$ if $\delta_{\boldsymbol{\tau}_{n,k}}(i,j) = 1$ and zero otherwise. One could order the variance estimates by L_1 fit (average of absolute values) between its corresponding $M_{\boldsymbol{\tau}_{n,k}}$ and the matrix that assumes independence, starting with the smallest estimated L_1 fit and ending with the largest. We use this ordering in our simulation study. One could also order according to another complexity criterion. In our practical data analysis, for example, we compare L_1 fit ordering with ordering by a rough approximation of the variance of each estimator $\sigma_n^2(\boldsymbol{\tau}_{n,k})$ in the sequence (see equation (3)). One could also order by the number of nonzero pairs directly. This works reasonably well in time series, but could be more problematic in settings with higher dimensional dependence. We implemented this ordering in our practical data analysis, as well.

Find the plateau where variance estimators are unbiased.

4. Estimate a monotone increasing curve from the ordered sequence using weighted isotonic regression. We use Pooled Adjacent Violators (PAV) for this purpose (Turner [2013]; Robertson et al. [1988]), weighting each variance estimate according to the inverse of an approximation

of its variance (see equation (3)). Here we rely on the fact that covariances are known to be positive. Without this assumption, one would have to use other nonparametric regression methods.

5. Find the plateau where the variance estimates are no longer biased. To do this, we estimate the mode of the PAV smoothed curve by first estimating its density using a Gaussian kernel with bandwidth selected according to Silverman [1986] (pg 48, eqn. 3.31). Then we take the maximum of that estimated density. We have found this approach works well under a wide range of sieves and true dependence structures. We can use this estimated mode as a variance estimate in its own right, or we can use it to locate the beginning of the plateau and then take as our estimate either the value of the PAV curve at this location or the value of the closest sample variance estimate to that location. We refer to estimators using the mode as 'mode-based'; those using the value of the PAV curve as 'step-based'; and those using the closest sample variance estimate in the sequence as 'sample-based'. In principle, because the mode-based estimator takes advantage of additional averaging, we would expect it to be less variable than the sample-based estimator. Our simulation results and practical data analysis appear to confirm this intuition. We have observed that the difference between step-based and mode-based estimators is negligible under the three orderings mentioned above.

In practice, the choice of upper bound $\boldsymbol{\tau}'$ is critical to the success of this procedure. It must be true that $\boldsymbol{\tau}'$ approximates elements in T_0 as $n \to \infty$. Preferably, the tail of the ordered sequence contains a number of $\boldsymbol{\tau}_{n,k}$ that approximate elements in T_0. In many applications where one feels comfortable assuming some form of central limit theorem, a sufficiently maximal $\boldsymbol{\tau}'$ will not be difficult to formulate.

3.1 Variance of Variance Estimators

For the sake of determining the weights in the PAV-algorithm, we need a reasonable approximation of the variance of each variance estimator in the sieve. We use the following shorthand notation in this section. Let D_{ij} denote $D_i(\theta)D_j(\theta)$, and $D_{ijk\ell}$ denote $D_i(\theta)D_j(\theta)D_k(\theta)D_\ell(\theta)$. Let Ω_{ij}

be the union of Ω_i and Ω_j. In the context of time-ordered dependence, this would correspond to the union of the two time intervals $[O(i - \boldsymbol{\tau}_i), ..., O(i)]$ and $[O(j - \boldsymbol{\tau}_j), ..., O(j)]$ implied by a candidate $\boldsymbol{\tau}$. Define an indicator function $\gamma(i, j, k, \ell) \equiv I\{\Omega_{ij} \cap \Omega_{k\ell} = \emptyset\}$, which equals one when the intersection between Ω_{ij} and $\Omega_{k\ell}$ is empty. We have found that the following metric works well in practice as an approximation of the variance of variance estimators of the form (2).

$$\frac{1}{n^2} \sum_{i,j,k,\ell} \{1 - \gamma(i, j, k, \ell)\} \delta_{\boldsymbol{\tau}}(i, j) \delta_{\boldsymbol{\tau}}(k, \ell) \tag{3}$$

This choice is inspired by formal calculations approximating the true variance and the use of working models for the components of this approximation. We provide more detail in Appendix A. In general, given that we have assumed the covariance between unit observations is positive and decreases with distance, the variance of variance estimators defined in equation (2) is roughly driven by the number of dependent pairs $\{(i, j) : \delta_{\boldsymbol{\tau}}(i, j) = 1\}$ as well as the number of intersecting non-empty unions Ω_{ij}. We also propose using (3) as an ordering scheme, as it will tend to put the more unbiased estimators in the tail of the sequence.

At first glance, one might assume calculating (3) for a sizable sequence of variance estimates to be computationally impractical. This is certainly true if one uses a naive approach. We have found that careful attention to symmetry, among other things, can reduce computation time by several thousand fold. We discuss this further and provide optimized code suitable for generalized dependence structures in this paper's supplementary materials.

4 Supporting Theory

Theorem 1 establishes conditions under which a generalized SP-variance estimation procedure is consistent for σ^2_{0n}. Beyond an ordering of the sequence concentrating unbiased variance estimators in the tail (a model assumption), the consistency of the SP-variance estimator relies on uniform consistency of $(\sigma^2_n(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$, a process indexed by $\boldsymbol{\tau} \in T$, as an estimator of its limit process $(\sigma^2_{0n}(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$, where $\sigma^2_{0n}(\boldsymbol{\tau}) = \sigma^2_{0n}$ if $\boldsymbol{\tau} \in T_0$. Since this uniform consistency condition is nontrivial, Theorem 2 considers a particular dependence structure on the influence functions for which we formally establish uniform consistency and consistency of the variance estimator under

entropy conditions restricting the size of T.

Theorem 1. *Recall $T = T_0 \cup T_0^c$. Assume the following.*

1. $\sup_{\boldsymbol{\tau} \in T} |\sigma_n^2(\boldsymbol{\tau}) - \sigma_{0n}^2(\boldsymbol{\tau})| \to 0$ *in probability for a $\sigma_{0n}^2(\boldsymbol{\tau})$ process, where $\sigma_{0n}^2(\boldsymbol{\tau}) = \sigma_{0n}^2$ for $\boldsymbol{\tau} \in T_0$.*

2. *We are given a mapping $\hat{\boldsymbol{\tau}}$ that takes as input a process $(\sigma^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ and maps it into an element of T. This mapping has the following properties:*

   (a) $\hat{\boldsymbol{\tau}}(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T) \in T_0$ with probability tending to 1;

   (b) continuity: $\hat{\boldsymbol{\tau}}(\sigma_n^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T) - \hat{\boldsymbol{\tau}}(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ converges to zero in probability.

Then $\boldsymbol{\tau}_n \equiv \hat{\boldsymbol{\tau}}(\sigma_n^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ converges to $\hat{\boldsymbol{\tau}}(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T) \in T_0$ with probability tending to 1, and thus $P(\boldsymbol{\tau}_n \in T_0) \to 1$ and $\sigma_n^2(\boldsymbol{\tau}_n) - \sigma_{0n}^2 \to 0$ in probability as $n \to \infty$.

Proving Theorem 1 is straightforward. Note that $\hat{\boldsymbol{\tau}}$ represents the algorithm that is able to map the limit process $(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ into a desired plateau value $\sigma_{0n}^2(\boldsymbol{\tau}_n^0)$ with $\boldsymbol{\tau}_n^0$ in T_0, where T_0 defines the plateau of $(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$. The additional continuity condition 2b on $\hat{\boldsymbol{\tau}}$, combined with the uniform consistency of $(\sigma_n^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ to $(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$, then establishes that $\sigma_n^2(\boldsymbol{\tau}_n)$ is consistent for σ_{0n}^2. In our concrete example of the SP method, we have that $\hat{\boldsymbol{\tau}}$ is given by the composition of (1) ordering the set of variance estimators $(\sigma_n^2(\boldsymbol{\tau}) : \boldsymbol{\tau})$ in some way, (2) applying the weighted PAV algorithm to this ordered sequence, and (3) finding its plateau value.

The conditions of Theorem 1 highlight an underlying tension between selecting T large enough to ensure it contains a nonempty T_0, but not so large that $(\sigma_n^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ fails to approximate its limit process $(\sigma_{0n}^2(\boldsymbol{\tau}) : \boldsymbol{\tau} \in T)$ *uniformly* in T. It also demonstrates the concrete challenge in coming up with $\hat{\boldsymbol{\tau}}$. One needs to construct an algorithm that can map an unordered set of variance estimators, a subset of which are consistent for σ_{0n}^2, into a consistent variance estimator.

Theorem 2 shows that under a particular type of dependence structure on the influence functions, the crucial uniform consistency assumption in Theorem 1 holds, and $\sigma_n^2(\boldsymbol{\tau}_n)$ is consistent for the true variance σ_{0n}^2 as long as $\boldsymbol{\tau}_n - \boldsymbol{\tau}_0^n \to 0$ for a $\boldsymbol{\tau}_0^n \in T_0$ with probability tending to 1. Theorem

2 also gives a good rate of convergence for the variance estimator (essentially $1/\sqrt{n}$ in many applications). One could weaken the entropy condition stated below while still preserving the asymptotic consistency of the variance estimator, but with a slower rate. However, for the sake of having reliable confidence intervals in finite samples, a good rate of convergence for $\sigma_n^2(\boldsymbol{\tau}_n)$ is important.

Theorem 2. *Let $\mathcal{F} = \{(\boldsymbol{\tau}, \theta) : \boldsymbol{\tau} \in T, \theta \in \Theta\}$. Assume $(\theta_n, \boldsymbol{\tau}_n) \in \mathcal{F}$ with probability tending to 1. Let $\| \cdot \|_{\mathcal{F}}$ be some norm or semimetric on $\mathcal{F}$. Define $\boldsymbol{\tau}^u$ to be such that for all $\boldsymbol{\tau} \in T$, $\boldsymbol{\tau} \leq \boldsymbol{\tau}^u$. Assume $\max_i |\Omega_{i,\boldsymbol{\tau}^u}| < K$ for some $K < \infty$. Define the index set $A = \{(i,j) : i \in \{1, \ldots, n\}, j \in \Omega_{i,\boldsymbol{\tau}^u}\}$. Define*

$$d_n[(\boldsymbol{\tau}, \theta), (\boldsymbol{\tau}_1, \theta_1)] \equiv \frac{1}{n} \sum_{(i,j) \in A} P_0^n \{\delta_{\boldsymbol{\tau}}(i,j) D_{ij}(\theta_n) - \delta_{\boldsymbol{\tau}_1}(i,j) D_{ij}(\theta_1)\}.$$

Let $\sigma_n^2(\boldsymbol{\tau}) = 1/n \sum_{(i,j) \in A} \delta_{\boldsymbol{\tau}}(i,j) D_{ij}(\theta_n)$, and $\sigma_{0n}^2(\boldsymbol{\tau}) = 1/n \sum_{i=1}^n \sum_{j \in \Omega_{i,\boldsymbol{\tau}}} P_0^n \{D_{ij}(\theta_0)\}$. Consider an estimator $\boldsymbol{\tau}_n$ and corresponding variance estimator $\sigma_n^2(\boldsymbol{\tau}_n)$ of σ_{0n}^2.

Assume:

1. $d_n[(\boldsymbol{\tau}_n, \theta_n), (\boldsymbol{\tau}_n^0, \theta_0)] \xrightarrow{n \to \infty} 0$ *in probability, where* $P_0^n(\boldsymbol{\tau}_n^0 \in T_0) \xrightarrow{n \to \infty} 1$.
2. $\| (\boldsymbol{\tau}_n, \theta_n) - (\boldsymbol{\tau}_n^0, \theta_0) \|_{\mathcal{F}} \to 0$ *in probability as* $n \to \infty$.
3. *Sparsity: For each $(\boldsymbol{\tau}, \theta) \in \mathcal{F}$ and $(i,j) \in A \equiv \{(i,j) : \delta_{\boldsymbol{\tau}^u}(i,j) = 1\}$, there exist at most K (universal constant in n) other elements $(k, \ell) \in A$ for which $\delta_{\boldsymbol{\tau}}(i,j) D_{ij}(\theta)$ depends on $\delta_{\boldsymbol{\tau}}(k,\ell) D_{k\ell}(\theta)$.*
4. *Linkage: For each $(\boldsymbol{\tau}, \theta) \in \mathcal{F}$ and all integers $p > 0$, for all $(i,j) \in A$ we have $\{P_0^n (\delta_{\boldsymbol{\tau}}(i,j) D_{ij}(\theta))^p\}^{1/p} \leq C \| (\boldsymbol{\tau}, \theta) \|_{\mathcal{F}}$ for a universal $C < \infty$.*
5. *Bounded entropy: $\exists\, \eta > 0$ so that the entropy integral $\int_0^\eta \sqrt{\log N(\epsilon, \mathcal{F}, \| \cdot \|_{\mathcal{F}})} d\epsilon < \infty$.*

Then,

$$\sigma_n^2(\boldsymbol{\tau}_n) - \sigma_{0n}^2 = O_P\left\{d_n\left[(\boldsymbol{\tau}_n, \theta_n), (\boldsymbol{\tau}_n^0, \theta_0)\right]\right\} + O_P\left\{1/\sqrt{n}\right\},$$

and thus converges to zero in probability as $n \to \infty$.

As a side-result, if we also have $\sup_{\boldsymbol{\tau}} \parallel (\boldsymbol{\tau}, \theta_n) - (\boldsymbol{\tau}, \theta_0) \parallel_{\mathcal{F}} \to 0$ *in probability, and*

$$r(n) \equiv \sup_{\boldsymbol{\tau} \in T} d_n[(\boldsymbol{\tau}, \theta_n), (\boldsymbol{\tau}, \theta_0)] \xrightarrow{n \to \infty} 0 \tag{4}$$

in probability, then

$$\sup_{\boldsymbol{\tau}} \mid \sigma_n^2(\boldsymbol{\tau}) - \sigma_{0n}^2(\boldsymbol{\tau}) \mid = O_P\{r(n)\} + O_P\{1/\sqrt{n}\}.$$

Theorem 2 is proved in Appendix B. The linkage condition is a weak one, as shown in van der Laan [2014]. It allows us to link the entropy condition (5) on $\mathcal{F}$ to the desired entropy of the functions $\delta_{\boldsymbol{\tau}}(i,j)D_{ij}(\theta)$. The sparsity condition (3) essentially corresponds with assuming that $D_i(\theta)$ is independent of $(D_j(\theta) : j \notin \Omega_{\boldsymbol{\tau}^u,i})$, uniformly in θ. Our proof hinges on establishing the asymptotic equicontinuity of a process $(Z_n(\boldsymbol{\tau}, \theta) : (\boldsymbol{\tau}, \theta))$, the mean-zero centered empirical average of $\delta_{\boldsymbol{\tau}}(i,j)D_{ij}(\theta)$. This independence allows us to formally analyze this process. If the sparsity condition does not hold marginally, but holds by conditioning on certain characteristics of the observed data, one could simply apply this theorem with P_0^n being this conditional distribution. In that same spirit, a refined theorem that significantly weakens the sparsity condition can be obtained by first orthogonally decomposing the process $Z_n(\boldsymbol{\tau}, \theta)$ into $Z_n(\boldsymbol{\tau}, \theta) = \sum_{m=1}^{M} \mathbb{E}[Z_n(\boldsymbol{\tau}, \theta) \mid \mathcal{H}(m)] - \mathbb{E}[Z_n(\boldsymbol{\tau}, \theta) \mid \mathcal{H}(m-1)]$, for an increasing sigma-field $\mathcal{H}(m)$ so that $\mathcal{H}(M) = (O_1, \ldots, O_n)$, and $\mathcal{H}(0)$ is empty; and second, applying the proof in Appendix B for weak convergence of $Z_n(\boldsymbol{\tau}, \theta)$ to each m-specific process $\mathbb{E}[Z_n(\boldsymbol{\tau}, \theta) \mid \mathcal{H}(m)] - \mathbb{E}[Z_n(\boldsymbol{\tau}, \theta) \mid \mathcal{H}(m-1)]$ conditional on $\mathcal{H}(m-1)$. Now one only needs the sparsity assumption conditional on $\mathcal{H}(m-1)$. This is the approach used in van der Laan [2014]. The resulting more general theorem assumes the necessary conditions for each m.

5 Simulation Study: Variance of the Sample Mean of a Time Series

We undertook a simulation study to examine the effectiveness of our proposed SP variance estimation approach and compare its performance to other existing methods in estimating the asymptotic variance of the sample mean under several types of time-dependent data generating distributions. Section 5.1 describes the simulated time series. Section 5.2 describes the three main types of

SP variance estimators we implemented and briefly outlines the existing estimators to which we compared their performance. In section 5.3, we discuss our results.

5.1 Simulated Time Series

We simulated three main types of moving average (MA) time series. The first consisted of processes with constant dependence lag structure over time, i.e. $O(t) = u(t) + \beta_1 u(t-1) + \beta_2 u(t-2) + \cdots + \beta_{\tau_0} u(t-\tau_0)$, $u(t) \sim N(0,1)$, $t = 1, ..., n$, where the vector $\beta = (\beta_1, ..., \beta_{\tau_0})$ was constant across all time points. This corresponds to a $\boldsymbol{\tau}_0(t)$ that is constant over time and equal to the length of the vector β. We used three different β vectors: 0.9, (0.9, 0.5, 0.1), and (0.9, 0.7, 0.5, 0.3, 0.1), with associated asymptotic variances of the standardized sample mean of 3.61, 6.25 and 12.25, respectively.

We also simulated two additional types of MA time series where the true dependence lag varied with time. In one simulation, $(\boldsymbol{\tau}_0(t) : t = 1, ..., n)$ had positive linear drift, starting with $\boldsymbol{\tau}_0(1) = 0$ and ending with $\boldsymbol{\tau}_0(n) = 7$. The asymptotic variance of the standardized sample mean for this type of time series is approximately 6.68. In another simulation, $(\boldsymbol{\tau}_0(t) : t)$ had a periodic structure bounded between 1 and 5, with either one or two periods. The associated asymptotic variance of the standardized sample mean for this time series type is approximately 7.12.

We simulated 16,384 instances of each time series type at each sample size of 250, 500, 750, and 1000, for a total of $393,216$ iterations in our simulation study.

5.2 Types of SP Variance Estimators

The idea behind SP variance estimation is quite general: generate an ordered sequence of proposed dependence relationships that can approximate the truth (a 'sieve'), then use smoothing techniques to find the plateau in the ordered variance estimates. We implemented three different SP estimators, all using an L_1 fit ordering.

Model-based $\boldsymbol{\tau}$. Perhaps one knows the dependence structure in one's data adheres to a particular pattern. For instance, one might know the dependence lags increase over time. This estimator takes advantage of such knowledge. When the true dependence lags were constant over time,

$\{\boldsymbol{\tau}_{n,k} : k\}$ consisted of constant-valued n-length vectors of dependence lags, starting with constant values of zero (assuming independence) and ending with an upper bound of constant values of 10, comfortably exceeding the maximum true dependence lag value of 5 for this type of simulated time series. When true dependence lags had positive linear trend, $\{\boldsymbol{\tau}_{n,k} : k\}$ was a collection of randomly generated nondecreasing n-length vectors, each with a unique combination of starting and ending values. The starting values were bounded between 1 and 7, the ending values between 2 and 18, and the maximum span between starting and ending values was 10. When true dependence lags fluctuated periodically, $\{\boldsymbol{\tau}_{n,k} : k\}$ were all periodic n-length vectors, each with a unique combination of number of periods (ranging from 1 to 6), phase shift, and minimum and maximum values (bounded between 1 and 12 in this simulation study).

Constant $\boldsymbol{\tau}$. There are probably more instances where one has limited specific knowledge about the dependence structure in one's data other than some notion of sparsity. One approach in this case is to use a sequence of constant-valued n-length $\boldsymbol{\tau}$ vectors. We found we could improve performance by incorporating some small decision criteria regarding determining a maximum constant value. The algorithm is as follows:

1. For $\{\boldsymbol{\tau}_{n,k} : k\}$ constant in time, start with $\boldsymbol{\tau}_n(t) = 0$ for all time points and compute (2). Continue increasing this constant value and computing the corresponding variance estimate until the first decreasing result or the constant value exceeds some large upper bound (30 in the present study). Let $\boldsymbol{\tau}'$ be the constant dependence lag vector associated with the last variance estimate in the sequence.

2. Estimate the plateau value of the resulting sequence using weighted PAVA and density estimation. Denote this result $\widetilde{\sigma}_n^2$.

3. Let $\boldsymbol{\tau}_{\max}$ be a constant vector of the maximum credible dependence lag we expect to observe (we used 10 in this study). If $\boldsymbol{\tau}' \leq \boldsymbol{\tau}_{\max}$, then $\widetilde{\sigma}_n^2$ is our variance estimate. Otherwise, our estimate is the minimum of $\widetilde{\sigma}_n^2$ and (2) with $\boldsymbol{\tau} = \boldsymbol{\tau}_{\max}$.

Kitchen sink. Another reasonable response to having limited specific knowledge beyond some

degree of sparsity is to generate a diverse set of vectors $\{\boldsymbol{\tau}_{n,k} : k\}$. In our simulation, at each iteration we randomly generated a large collection of about 28,985 unique dependence lag vectors that were constant, linearly increasing, linearly decreasing, periodic, and random walks bounded between zero and five.

We compared these SP variance estimators with the following methods: an oracle benchmark, (2) with $\boldsymbol{\tau} = \boldsymbol{\tau}_0$; variance estimates (2) with constant n-length vector $\boldsymbol{\tau}_{\max}$ equal to 10 for all time points; and an i.i.d. IF-based estimator, as in (2) with constant n-length vector $\boldsymbol{\tau}$ equal to 0 for all time points. We also compared our performance to the stationary block bootstrap (SBB), circular block bootstrap (CBB), and subsampling (SS), with block and subsample sizes selected in three different ways: (1) data-adaptively according to the algorithm in Patton et al. [2009]; (2) enforcing this data-adaptively selected block size to be at least 10; and (3) using a constant proportion as a function of sample size ($b = 0.1n$, SBB and CBB only).

5.3 Simulation Results

Our simulation results confirm our proposed approach works well in practice. All SP estimators outperformed all variants of the SS, CBB and SBB estimators included in our study. This was true across all sample sizes and time series types. It is possible some of this performance advantage can be attributed to less than optimal estimated block sizes. However, estimating an optimal b is difficult, and not necessarily tractable for other parameters of interest. Our simulation can be viewed as a comparison of these algorithms as they are typically used.

Table 1 compares average normalized MSEs, $\left(\sigma_n^2 - \sigma_0^2\right)^2 / \sigma_0^2$, across estimators and sample sizes. Overall, the SP 'kitchen sink' sample-based estimator performed best, with the corresponding mode-based estimator a close second. As expected, the mode version was slightly less variable and more biased than its sample-based counterpart. Notably, both of these estimators had normalized MSEs that were actually smaller on average than the oracle benchmarks at the smallest sample size $n = 250$. The other SP estimators were close competitors, particularly when $n \geq 250$. The modified constant $\boldsymbol{\tau}$ estimator was the least computationally intensive of the SP estimators we implemented, and performed admirably even when the simulated time series did not have constant

τ_0. When the dependence lag exhibited linear drift, for instance, the modified constant τ estimator actually outperformed the (correct) model-based τ estimators. For very large sample sizes and a rich collection of proposed dependence relationships, the time required to compute each variance estimate and its PAV regression weight for a 'kitchen sink'-type estimator could become nontrivial. In such settings, the modified constant τ estimator could be an attractive alternative. However, it may be difficult to implement in settings where underlying dependence is more complex or is not well understood.

estimator	overall	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
oracle, IF-based	0.230 (-0.020)	0.434 (-0.038)	0.224 (-0.020)	0.150 (-0.014)	0.114 (-0.010)
SP 'kitchen sink' (sample)	0.236 (-0.019)	0.426 (-0.019)	0.230 (-0.017)	0.162 (-0.019)	0.127 (-0.020)
SP 'kitchen sink' (mode)	0.248 (-0.073)	0.424 (-0.086)	0.243 (-0.073)	0.179 (-0.069)	0.146 (-0.066)
SP constant τ (IC)	0.263 (-0.015)	0.473 (-0.022)	0.259 (-0.015)	0.179 (-0.013)	0.139 (-0.012)
SP model-based (mode)	0.274 (-0.023)	0.491 (-0.032)	0.272 (-0.022)	0.188 (-0.020)	0.146 (-0.019)
SP model-based (sample)	0.279 (0.006)	0.508 (0.004)	0.276 (0.007)	0.188 (0.006)	0.144 (0.005)
SS, $b \geq 10$	0.403 (-0.110)	0.676 (-0.141)	0.403 (-0.111)	0.294 (-0.098)	0.238 (-0.088)
SS, data-adaptive b	0.407 (-0.115)	0.680 (-0.153)	0.406 (-0.116)	0.298 (-0.101)	0.243 (-0.091)
CBB, $b \geq 10$	0.408 (-0.114)	0.669 (-0.148)	0.409 (-0.116)	0.305 (-0.102)	0.250 (-0.091)
CBB, data-adaptive b	0.413 (-0.120)	0.675 (-0.161)	0.413 (-0.121)	0.309 (-0.105)	0.255 (-0.094)
SBB, $b \geq 10$	0.492 (-0.139)	0.807 (-0.183)	0.490 (-0.140)	0.368 (-0.122)	0.304 (-0.111)
SBB, data-adaptive b	0.499 (-0.151)	0.802 (-0.205)	0.502 (-0.152)	0.379 (-0.130)	0.313 (-0.117)
τ_{max}, IF-based	0.592 (-0.045)	1.109 (-0.086)	0.578 (-0.043)	0.387 (-0.030)	0.295 (-0.022)
CBB, $b = 0.1n$	0.987 (-0.125)	1.028 (-0.146)	0.987 (-0.123)	0.971 (-0.117)	0.961 (-0.112)
SBB, $b = 0.1n$	1.186 (-0.204)	1.251 (-0.226)	1.186 (-0.202)	1.160 (-0.198)	1.149 (-0.192)
iid, IF-based	3.600 (-0.677)	3.625 (-0.679)	3.599 (-0.677)	3.591 (-0.676)	3.584 (-0.675)

Table 1: Simulation results. Normalized MSE with respect to the true variance, $\left(\sigma_n^2 - \sigma_0^2\right)^2 / \sigma_0^2$. Normalized bias with respect to the true variance is in parentheses.

Boxplots in figures 1 and 2 clearly illustrate the performance differences between the SP and the blocked sub/resampling estimators. Subsampling with $b \geq 10$ was the best of the latter group, with normalized MSEs on average $2/3$ larger than those of the SP 'kitchen sink' sample-based estimator. The corresponding CBB estimator had nearly equal performance. Enforcing the block size to be at least 10 did not seem to affect either very much. The corresponding SBB estimators performed substantially less well, with normalized MSEs on average 22% larger than those of subsampling and a little more than twice that of the SP 'kitchen sink' sample-based estimator. Figure 3 examines block sizes as a function of time series type, sample size and normalized bias. Most b were less than 20, but the range of b was substantial, from 2 to 87. Not surprisingly, large

block sizes tended to be associated with negative bias overall. Average b increased with increasing $\max\{\boldsymbol{\tau}_0\}$, a reassuring sign that the algorithm we used to select b was performing reasonably.

The remaining estimators performed substantially less well. As expected, the estimator ignoring dependence did poorly and significantly underestimated the variance. The $\boldsymbol{\tau}_{max}$ estimators also did not fair well. They were not defined in a way that prevented them from giving negative variance estimates, which they did on rare occasions at the smallest sample size of $n = 250$. They improved substantially as sample size increased, and at $n = 1000$, had slightly better average normalized MSEs than the best performing SBB estimator. The CBB and SBB estimators with block size $b = 0.1n$ had very poor performance, a sobering reminder of the sensitivity of these algorithms to block size selection. Clearly, one should not resort to ad hoc methods for selecting b.

Table 2 lists the coverage of the confidence intervals using each of the estimators of the variance of the sample mean, overall and by sample size. With the exception of SBB with $b = 0.1n$, all estimators that assume some form of dependence had adequate coverage. However, SP estimators did have better coverage overall than the sub/resampling estimators, especially at smaller sample sizes. That SS and CBB had reasonable coverage suggests the algorithm we used to estimate b worked as expected.

estimator	overall	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
SP model-based (sample)	0.946	0.942	0.946	0.947	0.948
SP constant $\boldsymbol{\tau}$ (IC)	0.944	0.939	0.943	0.945	0.946
oracle, IF-based	0.943	0.938	0.944	0.946	0.947
SP 'kitchen sink' (sample)	0.943	0.940	0.943	0.944	0.945
SP model-based (mode)	0.942	0.938	0.942	0.944	0.945
SP 'kitchen sink' (mode)	0.936	0.931	0.936	0.938	0.940
$\boldsymbol{\tau}_{max}$, IF-based	0.932	0.914	0.932	0.939	0.942
SS, $b \geq 10$	0.929	0.918	0.928	0.932	0.936
SS, data-adaptive b	0.928	0.917	0.928	0.932	0.935
CBB, $b \geq 10$	0.928	0.918	0.927	0.932	0.935
CBB, data-adaptive b	0.927	0.915	0.927	0.931	0.935
SBB, $b \geq 10$	0.923	0.909	0.923	0.928	0.931
SBB, data-adaptive b	0.921	0.906	0.921	0.927	0.930
CBB, $b = 0.1n$	0.913	0.908	0.912	0.914	0.916
SBB, $b = 0.1n$	0.896	0.890	0.896	0.897	0.900

Table 2: Coverage probabilities

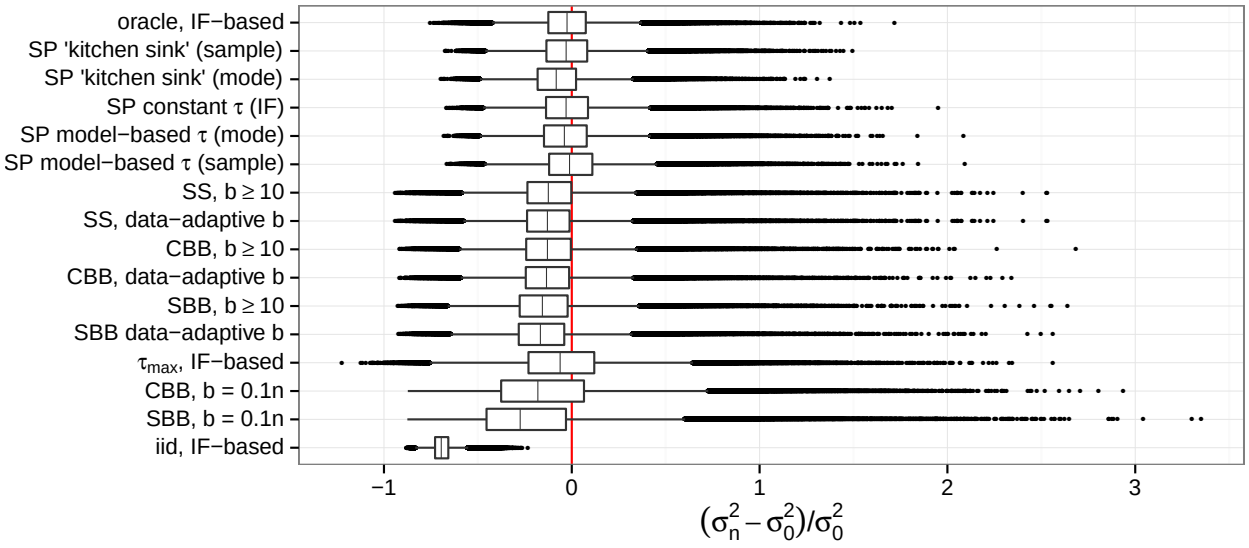


Figure 1: Boxplot of overall standardized bias for a subset of estimators. Boxplots are ordered vertically (top is best, bottom is worst) according to average normalized MSE.

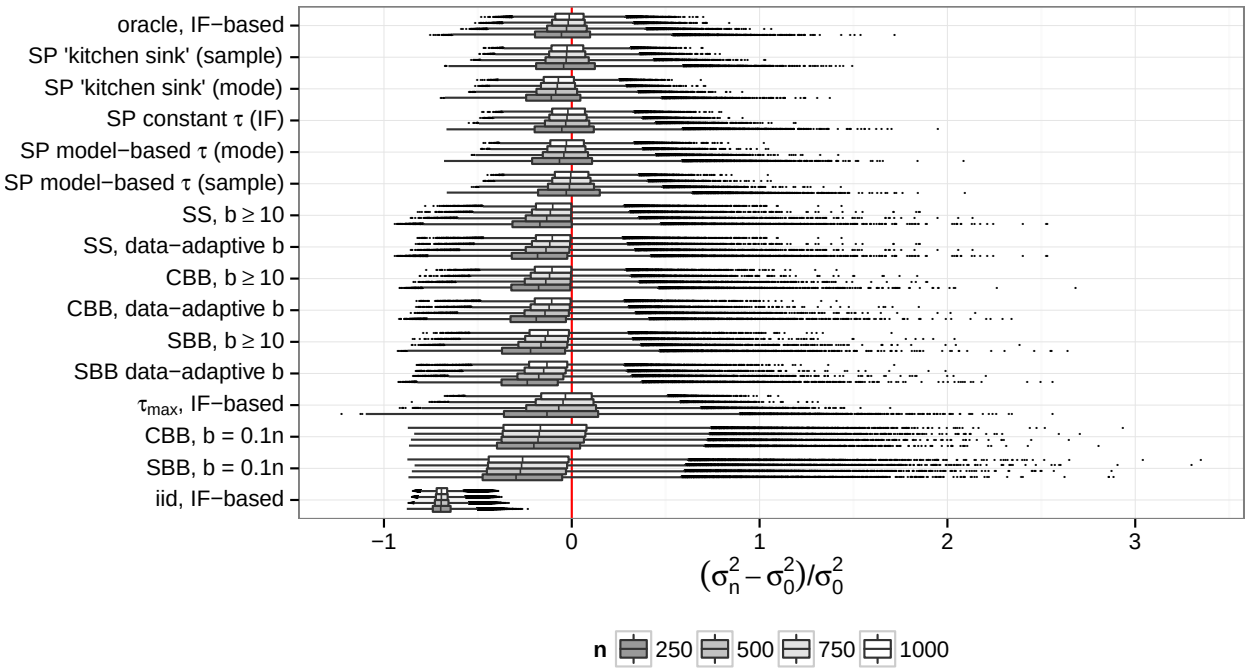
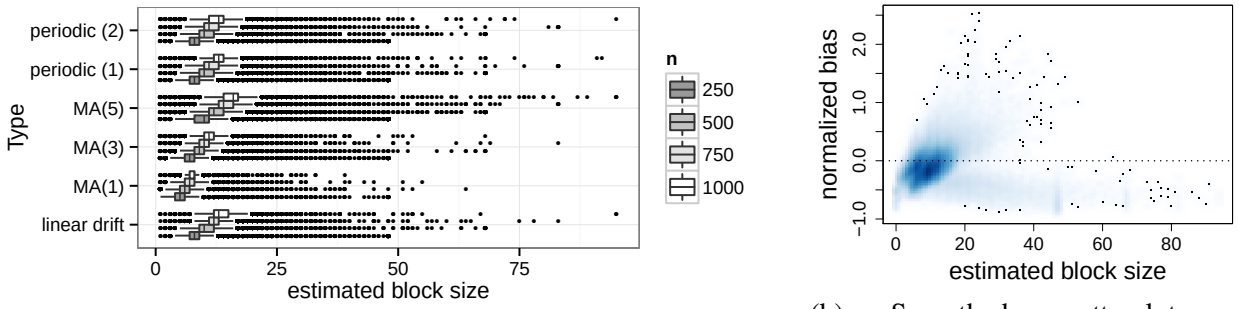


Figure 2: Boxplot of standardized bias by sample size for a subset of estimators. Boxplots are ordered vertically (top is best, bottom is worst) according average normalized MSE.



(a) Boxplot of estimated subsample size by time series simulation type.

(b) Smoothed scatterplot of normalized bias as a function of estimated subsample size.

Figure 3: Diagnostic plots of subsample size for SS estimator with b estimated data-adaptively.

6 Practical Data Analysis: Average Treatment Effect for a time series

Environmental Health applications are an area in which SP variance estimation has the potential to be very useful. Not only are exposures and outcomes frequently dependent in time and/or space, but environmental health studies often inform public policy. Reliable, theoretically sound standard error estimates are crucial in this setting.

In this section, we examine the relationship between ventilation rate (VR, the rate at which outdoor air is brought indoors) and illness absence for a single elementary school classroom followed over two years. These data are a subset of a much larger data set consisting of 162 classrooms in three California school districts. In each room, daily VRs (in units of liters per second per person, L/s/p) were estimated using measurements transmitted every five minutes from Ethernet-wired sensors. Daily counts of the number of students per classroom absent due to illness were also collected, along with other classroom level demographic covariates. These are the largest, most detailed school-based VR data in the scientific literature to date. We refer the reader to Mendell et al. [2013] for technical details regarding the sensors used in this project, how the VRs were estimated, and how the schools were sampled for inclusion in the original study.

6.1 School Ventilation Rates: Public Health Significance

According to the Environmental Protection Agency, the average American (including children) spends about 90% of her time indoors [Klepeis et al., 1996]. Indoor air can contain many irritants

and pollutants that adversely affect human health. These can include chemical emissions from the building and its various components (paint, carpet, cleaning product residue, etc.) as well as potentially infectious emissions from the room's occupants. Proper ventilation decreases the indoor concentrations of these potentially harmful substances. Scientists have found associations between lower VRs and adverse human health outcomes (see for instance Li et al. [2007]; Seppänen et al. [1999]). School-age children spend more time in school than in any other indoor environment besides home [Klepeis et al., 1996]. As of 2010, in the state of California alone there were approximately 6,224,000 students in 303,400 K-12 classrooms [California Department of Education, 2012], all spending on average six to seven hours per day inside classrooms.

Given that a significant percentage of the general population is habitually exposed to the indoor environment of schools and that we know there is a connection between insufficient ventilation and human health, one might assume the current VR guidelines are informed by scientific studies involving human health outcomes. Surprisingly, this is not the case. Historically, VR guidelines were designed to improve the perceived odor of a room [Mendell and Fisk, 2014]. While there have been efforts over the years to incorporate health concerns into recommended guidelines, these efforts have been relatively ad hoc in nature. The larger goal of the study from which these data are drawn was to provide a more rigorous scientific basis for VR threshold guidelines in school classrooms [Mendell et al., 2013].

6.2 Observed Data and Target Parameter

In this subsection, we first familiarize the reader with the observed data. We then introduce the target parameter and its estimator within the context of i.i.d. observations. Next, we outline the assumptions we are willing to make that permit us to use this version of the estimator, even though our data are clearly not i.i.d. Here we highlight the challenges inherent to any estimation problem in this subject matter area. We discuss the limitations of using this estimator, but contend that the implied statistical parameter is still interesting. However, we would like a robust standard error estimate. We use SP variance estimation to do this.

Our outcome of interest, $Y(t)$, is the total number of children absent specifically due to illness

on day t. Figure 4 shows that while illness absence was not a rare event, the majority of the days in our study had no illness absences. The maximum number of illness absences we observed was four.

Let $A(t)$ be a binary indicator equal to one when the average estimated VR taken over the seven calendar days prior to time t is at least 7.1 L/s/p (the current recommended threshold for classrooms, ASH [2010]), and zero otherwise. This is our exposure of interest. VRs on minimum days were excluded from analyses, when they were less likely to reach equilibrium. The classroom in analysis had at least one minimum day per week. Days with CO_2 values considered implausible for a typical occupied room (below 600 or above 7000 ppm) were also excluded. The majority of VRs included in this analysis (78%) did not meet the recommended threshold. A histogram of VRs included in our analysis (figure 4) shows the distribution of VRs varied substantively, with minimum and maximum values of 2.44 and 25.76 L/s/p, respectively. One convenient statistical property of VR in this study (and thus of our exposure of interest) is that daily VRs are independent of past illness absence rates and can be considered conditionally independent of past VRs as well, once one controls for season. The choice of this exposure metric is guided by a sensible evaluation of what little is known with respect to lag times from exposure to disease development for diseases most associated with illness absence in schools. We direct readers to the original paper for a more full discussion. Figure 4 plots $Y(t)$ as a function of seven-day average past VR, where it appear lower exposure values are associated with more non-zero $Y(t)$.

Let $W(t)$ denote a vector of real-valued covariates measured before exposure. In this study, $W(t)$ consisted of two variables: an indicator of winter season, and classroom enrollment count. Season is an important potential confounder (common cause of both exposure $A(t)$ and outcome $Y(t)$) in this study. The observed classroom is naturally ventilated with no air conditioning, thus the primary means by which outside air is brought indoors is via windows, which are more likely to be closed when it is cold outside. We would therefore expect VRs to be lower in winter. There might also be a higher baseline illness rate during winter months, regardless of VR threshold attainment. We control for this potential confounding effect by including an indicator variable equal to one

when t is a winter day, and zero otherwise. The total number of students present each day during the exposure window is another important factor, as this represents the number of children at risk of being absent in the near future who were actually exposed. Unfortunately, we were unable to obtain this information. Because illness absence counts were typically quite low (about 64% of observed days had no illness absences, with the majority of nonzero counts less than 3, see figure 4 for a histogram), we felt a reasonable proxy for the total number at risk was the average number of children enrolled in the previous seven calendar days. Classroom enrollment ranged between 22 and 26 students.

6.3 Estimating an Average Treatment Effect using TMLE

We would like to learn about the effect of the current recommended VR threshold on classroom illness absence. Specifically, we want to estimate the magnitude of the difference in daily illness absence counts when the average VR over the previous seven calendar days always meets or exceeds the current guidelines of 7.1 L/s/p versus when it never meets 7.1 L/s/p. To do this, we estimate an average treatment effect (ATE). The ATE was originally defined within the potential outcomes framework introduced by Rubin (1974) as a missing data problem, but can also be defined using nonparametric structural equation modeling (Pearl, 2010) in terms of interventions on a structural causal model. We use Targeted Maximum Likelihood (TML, van der Laan and Rubin [2006], Gruber and van der Laan [2010], Gruber and van der Laan [2012]) to estimate the ATE and an ensemble machine learning algorithm called Super Learner [Polley and van der Laan, 2012] to estimate the relevant portions of the likelihood.

TML and Super Learner are both components of a general approach called Targeted Learning [Rose and van der Laan, 2011]. Super Learner is a generalization of the stacking algorithm first introduced by Wolpert [1992] within the context of neural networks and later adapted by Breiman [1996] to the context of variable subset regression. It works by combining predictions from a diverse set of competing prediction algorithms using cross-validation, thus eliminating the need to pick a single prediction algorithm a priori. Theory guarantees Super Learner will perform asymptotically at least as well as the best algorithm in the competing set [van der Vaart et al.,

2006, van der Laan et al., 2007]. TML is a procedure for estimating parameters of semiparametric models. They are loss-based defined substitution estimators that work by updating initial estimates in a bias-reduction step targeted toward the parameter of interest instead of the overall density. Provided specific assumptions about the data generating process are met, TML estimators are efficient and unbiased.

We briefly summarize the formal definition of the ATE within the context of TML and i.i.d. data, as this is the version of the estimator we will use. Let $O_i = (O_1, ..., O_n)$ be a data set consisting of n i.i.d. draws from the random variable $O = \{W,\ A,\ Y\}$. O has true probability distribution P_0, contained in the statistical model $\mathcal{M}$. W is one or more covariates, A is a binary treatment or exposure, and Y is a bounded, real-valued outcome. The ATE is defined as the marginal difference in Y if A is set deterministically to 1 versus if A is set deterministically to 0. Under the assumptions of no unmeasured confounding and positivity, the ATE can be written in terms of the observed data distribution P_0 as

$$\psi_0^{ATE} = \mathbb{E}_W\big\{\mathbb{E}_0[Y|A=1,W] - \mathbb{E}_0[Y|A=0,W]\big\}. \tag{5}$$

Let $\overline{Q}_0(A,W) \equiv \mathbb{E}_0[Y|A,W]$, and let $g_0(A,W)$ be the true probability of A given W. The influence function for (5) is

$$D(P_0) = \left\{\frac{I(1)}{g_0(1,W)} - \frac{I(0)}{g_0(0,W)}\right\}\big\{Y - \overline{Q}(A,W)\big\} + \overline{Q}(1,W) - \overline{Q}(0,W) - \psi_0^{ATE}. \tag{6}$$

A TML estimator $\Psi(P_n^*)$ of (5) solves the efficient influence function (i.e. the canonical gradient) of (5), $P_n D(P_n^*) = 0$, which is (6) applied to the empirical distribution that has been updated so as to make the optimal bias variance trade-off, denoted P_n^*. The variance of this efficient influence function gives the asymptotic variance of the TML estimator of (5).

Of course, we do not have i.i.d. data. Rather, we observe a discontinuous time series $O^n = \{O(t) : t\}$ of n observations over a particular time span, with true joint data generating distribution P_0^n. Van der Laan (2014) derived identifiability results and TML estimators for target parameters like the ATE when the data are observed over time on an interconnected network. While these estimators represent a significant advance in semiparametric causal inference for dependent data,

they do require one to have some specific knowledge of the underlying dependence structure and are thus difficult to use in this context. Nevertheless, we believe the dependence in our data is weak; the statistical parameter formulated with respect to an i.i.d. data-generating process is interesting, and an estimate of it will provide scientists with useful information. Below, we summarize what we know about the data generating process we observe, what we are comfortable assuming, and what we are willing to assume for the sake of moving forward.

We believe O^n has a martingale-type dependence structure. Specifically, let $\overline{O}(t) = \{O(t), O(t-1), ..., O(1)\}$, and in an abuse of notation, let $\overline{O}(t-\tau_t) = \{O(t), O(t-1), ..., O(t-\tau_t)\}$ for some t-specific positive integer τ_t that is substantially less than both the sample size n and the total time span covered by the sample. We are willing to assume P_0^n can be factorized as

$$P_0^n = \prod_{t \in T} P_0\{O(t)|\overline{O}(t)\} = \prod_{t \in T} P_0\{O(t)|\overline{O}(t-\tau_t)\},$$

where P_0 is common across all t. We do not know the true dependence lags $(\tau_t : t)$. We feel quite comfortable assuming the time dependence in our likelihood does not come from the distribution of the covariates $W(t)$ or the exposure mechanism $g_0\{A(t), W(t)\}$, for reasons stated above. We believe the randomization assumption holds, since daily VR is not in any way related to past illness absence counts. The problematic relevant portion of the likelihood is $\overline{Q}_0$. Specifically, in order for the i.i.d.-formulated TMLE to be well defined, we must be willing to assume $\overline{Q}_0$ is common across time (i.e. it is not indexed by t); and $\mathbb{E}_0[Y(t)|\overline{A}(t-\tau_t), \overline{W}(t-\tau_t)] = \mathbb{E}_0[Y(t)|A(t), W(t)]$. (Since $W(t)$ varies quite slowly over time, for all practical purposes we can treat it as equivalent to $\overline{W}(t-\tau_t)$.) While we may feel confident that the large majority of $Y(t)$ only depend on VRs captured within $A(t)$, it is not impossible for $Y(t)$ to be a function of VRs in the more distant past for some t. In addition, some $Y(t)$ may be dependent on illness absence counts in the recent past. We have no way of knowing the true dependence lag in either case. However, as we stated above, we still believe an ATE estimator that makes the necessary assumptions about $\overline{Q}_0$ will produce an interesting, useful result, even if we cannot be sure a causal interpretation is appropriate.

The library of learners used in our ATE TML estimation procedure included generalized linear

models (with and without interactions); a step-wise GLM; Generalized Additive Models using cubic smoothing splines, parameterized by equivalent degrees of freedom (2, 4 or 6) [Hastie, 1991]; and multivariate adaptive polynomial spline regression [Stone et al., 1997]. To achieve a uniformly bounded loss function, predicted values for all algorithms were truncated to the range of the observed data. Note that these data are not amenable to traditional time series approaches because of their inherently discontinuous nature.

6.4 Estimating a Standard Error

Because these results could be used to inform future policy, we feel it is important to obtain a robust standard error estimate. We use SP variance estimation to do this. We assume the influence functions in our analysis are dependent in time in a way we don't fully understand. We are comfortable assuming this dependence is bounded in time and that all true covariances are positive. We do not know enough about the underlying dependence to use a model-based $\boldsymbol{\tau}$ approach. The SP approach assuming constant $\boldsymbol{\tau}$ is possible, but less convenient given the temporal discontinuities in these data. We therefore implement the 'kitchen sink' SP variance estimator, using an upper bound on the temporal extent of the dependence of any two influence function values of $\max(\boldsymbol{\tau}) = 21$. This threshold encompasses the vast majority of incubation periods for infectious respiratory diseases [Lessler et al., 2009]. Our 'kitchen sink' sieve consisted of the following sequences of $\boldsymbol{\tau}$ vectors.

1. Winter step functions. Time points in winter may have different dependence lags than in nonwinter: $\boldsymbol{\tau}(t) = \alpha_1 I(t \in \text{winter}) + \alpha_2 I(t \notin \text{winter})$, $(\alpha_1, \alpha_2) \in \{0, 1, \ldots, 21\}^2,\ \alpha_1 \neq \alpha_2$.

2. Seasonally periodic functions. Periodic $\boldsymbol{\tau}$ vectors were generated so that periodicity corresponded with winter season.

3. $\boldsymbol{\tau}(t)$ depends on $Y(t)$. Time points with more extreme average illness absence may be dependent on more past time points than those with less extreme average illness absence (or vice versa).

4. $\boldsymbol{\tau}(t)$ depends on past VR. Days with more extreme VRs may have had more lasting influence on future $Y(t)$ than days with less extreme VRs.

For demonstration purposes, we implemented nine types of SP 'kitchen sink' variance estimators. We used mode, step and sample-based versions, and we ordered by L_1 fit, complexity (defined using equation (3)), and the number of nonzero D_{ij} pairs included in the estimator.

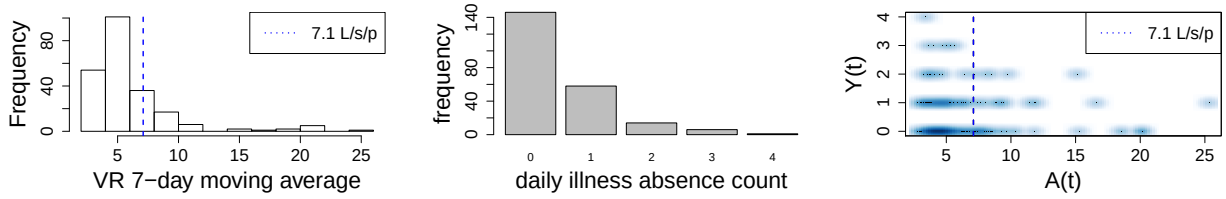
Subsampling and blocked bootstraps are not well suited to this estimation problem for a number of reasons. The computation time required to repeatedly re-estimate the ATE would be considerable. Also, estimating an optimal b would be complicated by the inherently discontinuous nature of the data. Furthermore, one of our confounders, winter, occurs in two large contiguous blocks, one for each year. In order to avoid positivity violations, b would have to be larger than the length of winter. We have no way of knowing whether or not an optimal b this large even exists. Using SP variance estimation is much more convenient and computationally efficient in this context.

6.5 Results

Our TML estimate of the ATE is -0.186. This means that on average, illness absence counts when the average VR over the preceeding seven calendar days at least met the current recommended threshold were 0.186 less than illness absence counts when the preceeding VR failed to meet the recommended threshold of 7.1 L/s/p. Given that the mean illness absence count is 0.48, this is a potentially consequential finding, and may suggest increasing the average VR in this classroom to at least 7.1 L/s/p could substantively reduce illness absence rates. The 0.95 confidence interval ignoring time dependence is $(-0.35, -0.02)$ (p-value = 0.03). The naive practitioner might be tempted to assume the results are both practically and statistically significant. However, our 'kitchen sink' SP estimated standard errors are much larger, providing us with a more realistic sense of the uncertainty of our ATE estimate. Table 3 lists variance estimates, p-values and confidence intervals for each of the SP variance estimators. The variance estimates are all more than three times as large as the i.i.d. IF-based estimate and produce confidence intervals containing zero, and p-values that are no longer significant. They are also in relative agreement with one another. Figure 5 provide a visual illustration of the estimation procedures, where we can clearly see a plateau in each case.

ordering	type	σ_n^2	0.95 CI	p-value
(none)	i.i.d. IF-based	1.592	(-0.351, -0.021)	0.027
number of nonzeros	mode	5.760	(-0.500, 0.128)	0.245
	value of longest plateau	5.774	(-0.500, 0.128)	0.246
	sample value at longest plateau	5.905	(-0.503, 0.132)	0.251
L_1 fit	mode	5.798	(-0.501, 0.129)	0.247
	value of longest plateau	5.825	(-0.501, 0.129)	0.248
	sample value at longest plateau	6.003	(-0.506, 0.134)	0.255
complexity	mode	5.752	(-0.499, 0.127)	0.245
	value of longest plateau	5.763	(-0.500, 0.128)	0.245
	sample value at longest plateau	5.992	(-0.506, 0.134)	0.254

Table 3: ATE variance estimation. Results ignoring dependence, and SP estimators, ordering by number of non-zero elements in the estimator; L_1 fit; and complexity. All SP estimators are of the 'kitchen sink' variety, utilizing $12,956$ unique dependence lag vectors.



(a) (b) (c)

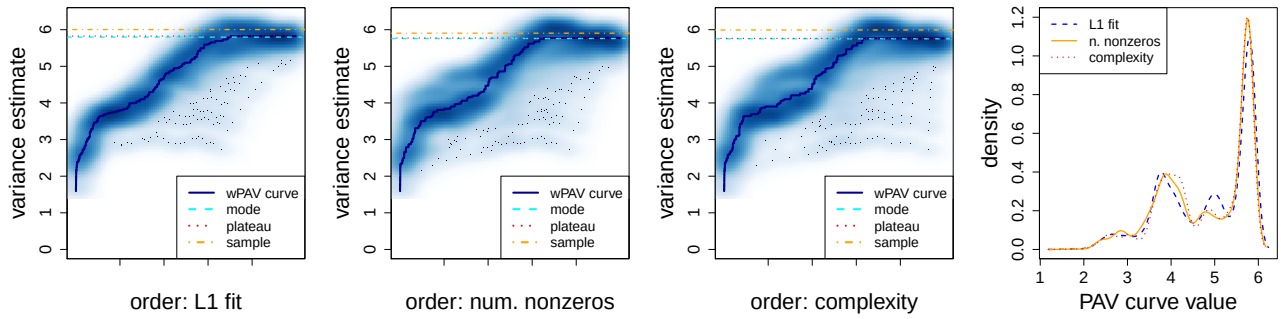
Figure 4: Descriptive plots of (a) 7-day average VRs (L/s/p), (b) daily illness absence counts, and (c) a scatterplot of daily illness absence as a function of prior 7-day average VR.

Overall, the practical significance of our ATE estimate is mixed. Its magnitude suggests that meeting the threshold may have a positive effect on illness absence rates, but our more realistic SP-based standard error estimates should give us pause. Furthermore, collapsing a bounded, continuous exposure into a binary one can make interpretation of results somewhat challenging. A curve examining a range of potential thresholds may be a more informative target parameter. This will require using the full data, which include multiple schools, each with multiple, possibly interdependent classrooms, and is future work.

7 Discussion and Future Directions

Our original goal in our simulation study was to verify our proposed SP variance estimators could do at least as well as subsampling and blocked bootstrap estimators in a setting where these existing estimators have been well-studied. The chief advantage of the SP approach would then be its

Figure 5: Visualizations of SP variance estimation approaches when ordering by (a) L_1fit, (b) complexity, and (c) the number of nonzero D_{ij} pairs included in the estimator. (d) shows the estimated densities of each PAV curve.



(a) (b) (c) (d)

general utility: it can be used in situations where sub/resampling are ill-suited. Our simulations showed that in the familiar setting of sample means of time series, all three SP approaches substantively outperformed their competitors, both in bias and variance.

We also presented a practical data analysis, estimating an ATE for the effect of VR threshold attainment on subsequent illness absence counts in an elementary school classroom. This is an important public health issue, and a good example of a setting where sub/resampling would be difficult to use. We showed that in this case, properly accounting for time dependence gave larger, likely more realistic standard error estimates.

SP variance estimators do rely on more assumptions than subsampling, many of which are not testable, and bootstraps can, in some circumstances, capture second order features that influence function-based methods will miss [DasGupta, 2008]. However, if one feels confident that the assumptions we require are reasonable, our approach appears to be an excellent alternative to blocked resampling and subsampling strategies. It avoids the challenge of estimating the nuisance parameter b and can result in significant computational savings.

Another important advantage of SP variance estimation is that it can be implemented even when one has limited understanding of the ordering of one's data. Dependence can have much more complex structure than what we have presented here. For example, spatiotemporal dependence is common in ecology and environmental health, and applications in internet advertising and

infectious disease epidemiology often involve dependent observations on a poorly understood network. The SP 'kitchen sink' estimator could be particularly useful in these settings.

A limitation of our approach as it is implemented here is that we require monotonic dependence. While monotonic dependence is common, there are applications where this requirement is not met. Health care settings with finite resources, for instance, or plant species distributions where organisms have an inhibitory effect on their immediate surroundings are some examples. Extending the SP variance estimation algorithm to settings with nonmonotonic covariance is an important area for future work.

APPENDIX A. Approximating the Variance of Variance Estimators

We illustrate how we derived (3) by first deriving an approximation of the true variance of a sieve element when this element is in the plateau, and then using a working model on this approximation. We use notation previously defined above in section 3.1.

Consider the variance of an estimator of the form (2) with arbitrary $\boldsymbol{\tau} \in T_0$. Note that if $\Omega_{ij} \cap \Omega_{k\ell} = \emptyset$, then $P_0^n D_{ijk\ell} = (P_0^n D_{ij})(P_0^n D_{k\ell})$. We make the following assumption on the independence structure of the observed data:

$$\begin{aligned}
\mathbb{E}_0\left[\sigma_n^4(\boldsymbol{\tau})\right] &= \frac{1}{n^2}\sum_{i,j,k,\ell}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)P_0^n D_{ijk\ell} \\
&= \frac{1}{n^2}\sum_{i,j,k,\ell}\gamma(i,j,k,\ell)\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)P_0^n D_{ij}P_0^n D_{k\ell} \\
&\quad + \frac{1}{n^2}\sum_{i,j,k,\ell}[1-\gamma(i,j,k,\ell)]\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)P_0^n D_{ijk\ell} \qquad (7) \\
&= \frac{1}{n^2}\sum_{i,j,k,\ell}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)P_0^n D_{ij}P_0^n D_{k\ell} + o(1).
\end{aligned}$$

Specifically, we assume the term (7) is negligible. Given this assumption, we can approximate the variance of (2) as

$$\begin{aligned}\text{Var}\left[\sigma_n^2(\boldsymbol{\tau})\right] &= \mathbb{E}_0\left\{\sigma_n^2(\boldsymbol{\tau})\right\}^2 - \left\{\mathbb{E}_0\sigma_n^2(\boldsymbol{\tau})\right\}^2 \\ &\approx \frac{1}{n^2}\left\{\sum_{i,j,k,\ell}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)P_0^n D_{ijk\ell} - \sum_{i,j,k,\ell}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)P_0^n D_{ij}P_0^n D_{k\ell}\right\} \\ &= \frac{1}{n^2}\sum_{i,j,k,\ell}\left\{1-\gamma(i,j,k,\ell)\right\}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)\left\{P_0^n D_{ijk\ell} - P_0^n D_{ij}P_0^n D_{k\ell}\right\}.\end{aligned}$$

Let $\rho(i,j,k,l) = P_0^n D_{ijk\ell} - P_0^n D_{ij}P_0^n D_{k\ell}$. Suppose for the sake of illustration we define the following working models for ρ: $\alpha^2 = P_0^n D_i^2$, common across i; $\rho_1 = P_0^n D_{ij}$, common across $\{(i,j) : \delta_{\boldsymbol{\tau}}(i,j) = 1, i \neq j\}$, and $\rho_2 = P_0^n D_{ijk\ell}$, common across $\{(i,j,k,\ell) : \delta_{\boldsymbol{\tau}}(i,j) = \delta_{\boldsymbol{\tau}}(k,\ell) = 1,\ \gamma(i,j,k,\ell) = 0\}$. We could estimate these quantities by taking empirical averages across the relevant elements. An estimator of ρ_1 for an arbitrary $\boldsymbol{\tau}$ is given by $\rho_{1,n} = \{\sum_{i,j}\delta_{\boldsymbol{\tau}}(i,j)I(i \neq j)D_{ij}\}/\{\sum_{i,j}\delta_{\boldsymbol{\tau}}(i,j)I(i \neq j)\}$. Similarly, $\alpha_n^2 = 1/n\sum_i D_i^2$, and

$$\rho_{2,n} = \frac{\sum_{i,j,k,\ell}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)(1-\gamma(i,j,k,\ell))D_{ijk\ell}}{\sum_{i,j,k,\ell}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)\{1-\gamma(i,j,k,\ell)\}}.$$

This defines $\rho_n(i,j,k,\ell)$ for each (i,j,k,ℓ) as having three possible values: α_n^2, $\rho_{1,n}$ or $\rho_{2,n}$. The corresponding variance estimator using this working model is

$$\text{Var}\left\{\sigma_n^2(\boldsymbol{\tau})\right\} = 1/n^2\sum_{i,j,k,\ell}\{1-\gamma(i,j,k,\ell)\}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)\rho_n(i,j,k,\ell).$$

Assuming common ρ across all time points is in most cases inappropriate. However, it does show us that the variance of the sieve elements can be viewed as proportional to (3), $1/n^2\sum_{i,j,k,\ell}\{1-\gamma(i,j,k,\ell)\}\delta_{\boldsymbol{\tau}}(i,j)\delta_{\boldsymbol{\tau}}(k,\ell)$. We have found using (3) works well in practice as both a basis for a weight in PAV algorithms and as a sieve ordering.

APPENDIX B. Proof of Theorem 2

We start by proving the last result of the theorem. Define the process $(Z_n(\boldsymbol{\tau},\theta) : (\boldsymbol{\tau},\theta) \in \mathcal{F})$ by

$$Z_n(\boldsymbol{\tau},\theta) = \frac{1}{n}\sum_{i=1}^n\sum_{j\in\Omega_{i,\boldsymbol{\tau}}}\{D_i(\theta)D_j(\theta) - P_0^n D_i D_j(\theta)\}.$$

We have

$$\sigma_n^2(\boldsymbol{\tau}) - \sigma_{0n}^2(\boldsymbol{\tau}) = \{Z_n(\boldsymbol{\tau},\theta_n) - Z_n(\boldsymbol{\tau},\theta_0)\} + Z_n(\boldsymbol{\tau},\theta_0) + \frac{1}{n}\sum_{i=1}^n\sum_{j\in\Omega_{i,\boldsymbol{\tau}}}P_0^n\{D_iD_j(\theta_n) - D_iD_j(\theta_0)\}. \tag{8}$$

By condition (4), the last term in equation (8) converges to zero in probability uniformly in $\boldsymbol{\tau}$. We note that $Z_n(\boldsymbol{\tau}, \theta)$ can be represented as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\Omega_{i,\boldsymbol{\tau}^u}} I(j \in \Omega_{i,\boldsymbol{\tau}})\{D_i(\theta)D_j(\theta) - P_0^n D_i D_j(\theta)\},$$

showing $Z_n(\boldsymbol{\tau}, \theta)$ is a sum over $\mathcal{A} \equiv \{(i,j) : i = 1, \ldots, n, j \in \Omega_{i,\boldsymbol{\tau}^u}\}$. Let n^u be the size of $\mathcal{A}$. Then $Z_n(\boldsymbol{\tau}, \theta)$ can be represented as

$$\frac{1}{n^u}\sum_{k=1}^{n^u} f_k(\boldsymbol{\tau}, \theta) - P_0^n f_k(\boldsymbol{\tau}, \theta),$$

where $f_k(\boldsymbol{\tau}, \theta) = n^u D_{i(k)} D_{j(k)}(\theta)$ and $k \to (i(k), j(k))$ maps the index k into the corresponding index $(i(k), j(k))$ that makes up the index set $A = \{(i,j) : j \in \Omega_{i,\boldsymbol{\tau}^u}\}$. By the sparsity condition (3), linkage condition (4) and entropy condition (5), Theorem 3 in van der Laan [2014] proves that the process $(\sqrt{n}Z_n(\boldsymbol{\tau}, \theta) : (\boldsymbol{\tau}, \theta) \in \mathcal{F})$ is asymptotically equicontinuous, so that, in particular, the following holds: if $\| (\boldsymbol{\tau}_n, \theta_n) - (\boldsymbol{\tau}, \theta_0) \|_{\mathcal{F}} \to 0$ in probability, then $\sqrt{n}\{Z_n(\boldsymbol{\tau}_n, \theta_n) - Z_n(\boldsymbol{\tau}, \theta_0)\} = o_P(1)$, and

$$\sup_{(\boldsymbol{\tau},\theta)\in\mathcal{F}} | Z_n(\boldsymbol{\tau}, \theta) | = O_P(1/\sqrt{n}).$$

As a consequence, we can conclude that if $\sup_{\boldsymbol{\tau}} \| (\boldsymbol{\tau}, \theta_n) - (\boldsymbol{\tau}, \theta_0) \|_{\mathcal{F}} \to 0$ in probability, then $\sqrt{n}\sup_{\boldsymbol{\tau}} |Z_n(\boldsymbol{\tau}, \theta_n) - Z_n(\boldsymbol{\tau}, \theta_0)| = o_P(1)$, and we also have $\sup_{\boldsymbol{\tau}} |Z_n(\boldsymbol{\tau}, \theta_0)| = O_P(1/\sqrt{n})$. Condition (5) thus establishes that $\sqrt{n}\sup_{\boldsymbol{\tau}} |Z_n(\boldsymbol{\tau}, \theta_n) - Z_n(\boldsymbol{\tau}, \theta_0)| = o_P(1)$. This proves $\sup_{\boldsymbol{\tau}} |\sigma_n^2(\boldsymbol{\tau}) - \sigma_{0n}^2(\boldsymbol{\tau})| = r(n) + O_P(1/\sqrt{n})$.

We now prove the consistency of $\sigma_n^2(\boldsymbol{\tau}_n)$ as an estimator of σ_{0n}^2 when $\| (\theta_n, \boldsymbol{\tau}_n) - (\theta_0, \boldsymbol{\tau}_n^0) \| \to 0$ and $d_n[(\boldsymbol{\tau}_n, \theta_n), (\boldsymbol{\tau}_n^0, \theta_0)] \to 0$ in probability for a $\boldsymbol{\tau}_n^0 \in T_0$ (we can act as if the latter holds with probability 1, by a standard simple argument). Firstly, we note that $\sigma_{0n}^2 = \sigma_{0n}^2(\boldsymbol{\tau}_n^0)$. Now we have

$$\sigma_n^2(\boldsymbol{\tau}_n) - \sigma_{0n}^2(\boldsymbol{\tau}_n^0) = Z_n(\boldsymbol{\tau}_n, \theta_n) - Z_n(\boldsymbol{\tau}_n^0, \theta_0) + Z_n(\boldsymbol{\tau}_n^0, \theta_0) + \frac{1}{n}\sum_{(i,j)\in A}\{P_0^n f_{(i,j),\boldsymbol{\tau}_n,\theta_n} - P_0^n f_{(i,j),\boldsymbol{\tau}_n^0,\theta_0}\}. \quad (9)$$

By the asymptotic equicontinuity of Z_n, we have $Z_n(\boldsymbol{\tau}_n, \theta_n) - Z_n(\boldsymbol{\tau}_n^0, \theta_0) = o_P(1/\sqrt{n})$, and $Z_n(\boldsymbol{\tau}_n^0, \theta_0) = O_P(1/\sqrt{n})$. The third term in equation (9) equals $d_n[(\boldsymbol{\tau}_n, \theta_n), (\boldsymbol{\tau}_n^0, \theta_0)]$, which is assumed to converge to zero in probability. Thus, this proves that $\sigma_n^2(\boldsymbol{\tau}_n) - \sigma_{0n}^2 \to 0$ in probability.

□

References

ANSI/ASHRAE Standard 62.1-2010: Ventilation For Acceptable Indoor Air Quality. ASHRAE, Inc., Atlanta, Georgia, 2010.

L Breiman. Stacked Regressions. *Machine Learning*, 24(1):49–64, JUL 1996.

California Department of Education. School facilities fingertip facts, 2012.

A DasGupta. *Asymptotic Theory of Statistics and Probability*, chapter chapter 29: The Bootstrap, pages 468–471. Springer, 2008.

S Gruber and MJ van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.

S Gruber and MJ van der Laan. tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13):1–35, 2012. URL `http://www.jstatsoft.org/v51/i13/`.

T. J. Hastie. *Statistical models in S*, chapter 7: Generalized Additive Models. Wadsworth and Brooks/Cole, 1991.

NE Klepeis, AM Tsang, and JV Behar. Analysis of the national human activity pattern survey (nhaps) respondents from a standpoint of exposure assessment - final epa report. Technical report, Environmental Protection Agency, Washington D.C., 1996.

SN Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2013.

J Lessler, NG Reich, R Brookmeyer, TM Perl, KE Nelson, and Cummings. Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet Infectious Diseases*, 9:291–300, 2009.

Y Li, GM Leung, JW Tang, X Yang, CYH Chang, JZ Lin, JW Lu, PV Nielsen, J Niu, H Qian, AC Sleigh, HJJ Su, J Sundell, and TW Wong. Role of ventilation in airborne transmission of infectious agents in the built environment - a multidisciplinary systematic review. *Indoor Air*, 17:2–18, 2007.

E Mammen. When does the bootstrap work? Asymptotic results and simulations. In *Springer Lecture Notes in Statistics 77*. Singer Verlag, Heidelberg, 1992.

MJ Mendell and WJ Fisk. Developing evidence-based prescriptive ventilation rate standards for commercial buildings in california: a proposed framework. Technical report, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA., 2014.

MJ Mendell, EA Eliseeva, MM Davies, M Spears, A Lobscheid, WJ Fisk, and MG Apte. Association of classroom ventilation with reduced illness absence: a prospective study in california elementary schools. *Indoor Air*, 23(6):515–528, 2013.

A Patton, DN Politis, and H White. CORRECTION TO "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. 2009.

DN Politis and JP Romano. Nonparametric resampling for homogeneous strong mixing random fields. *Journal of Multivariate Analysis*, 47:301–328, 1993.

DN Politis, JR Romano, and M Wolf. *Subsampling*. Springer, New York, 1999.

E Polley and MJ van der Laan. *SuperLearner: Super Learner Prediction*, 2012. URL `http://CRAN.R-project.org/package=SuperLearner`. R package version 2.0-6.

T Robertson, FT Wright, and RL Dykstra. *Order Restricted Statistical Inference*. Wiley, New York, 1988.

S Rose and MJ van der Laan. *Targeted Learning: Casual Inference for Observational and Experimental Data*. Springer, New York, 2011.

O Seppänen, WJ Fisk, and MJ Mendell. Association of ventilation rates and co2 concentrations with health and other responses in commerical and institutional buildings. *Indoor Air*, 9(4):226–252, 1999.

B. W. Silverman. *Density Estimation*. London: Chapman and Hall, 1986.

CJ Stone, MH, C Kooperberg, and YK Truong. The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, 25:1371–1470, 1997.

R Turner. *Iso: Functions to perform isotonic regression*, 2013. URL `http://CRAN.R-project.org/package=Iso`. R package version 0.0-15.

MJ van der Laan. Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2(1):13–74, 2014.

MJ van der Laan and D Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

MJ van der Laan, EC Polley, and AE Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

AW van der Vaart, S Dudoit, and MJ van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics and Decisions*, 24:351–371, 2006.

DH Wolpert. Stacked Generalization. *Neural Networks*, 5(2):241–259, 1992.