The Binomial Block Bootstrap Estimator for Evaluating Loss on Dependent Clusters

Matt Barnes
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
mbarnes1@cs.cmu.edu

Artur Dubrawski
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
awd@cs.cmu.edu

Abstract

In this paper, we study the non-IID learning setting where samples exhibit *dependency within latent clusters*. Our goal is to estimate a learner's loss on new clusters, an extension of the out-of-bag error. Previously developed cross-validation estimators are well suited to the case where the clustering of observed data is known a priori. However, as is often the case in real world problems, we are only given a noisy approximation of this clustering, likely the result of some clustering algorithm. This subtle yet potentially significant issue afflicts domains ranging from image classification to medical diagnostics, where naive cross-validation is an optimistically biased estimator. We present a novel bootstrap technique and corresponding cross-validation method that, somewhat counterintuitively, injects additional dependency to asymptotically recover the loss in the independent setting.

1 Introduction

The assumption of independent and identically distributed (IID) samples is fundamental to many machine learning algorithms [1, 2]. Some exploration outside this setting has occurred — notably in time-series data, clusters of independent data and less explicitly in active learning [3, 4]. In this paper, we study the setting where samples exhibit dependency both within latent clusters.

To illustrate, consider samples generated according to the simple k-mixture model

$$\begin{aligned} \phi_j &\overset{iid}{\sim} H(\gamma) && \text{for } j = 1, \ldots, k \\ c_i &\overset{iid}{\sim} \text{Categorical}(\pi) && \\ x_i, y_i &\overset{iid}{\sim} G(\phi_{c_i}) && \text{for } i = 1, \ldots, n_x \end{aligned} \tag{1}$$

where ϕ are latent cluster parameters; c are (potentially latent) cluster assignments; $X = x_1, \ldots, x_{n_x}$ are n_x samples; y are the corresponding labels; H is some distribution over cluster parameters; γ, π, k, n_x are model parameters and π is in the k-dimensional probability simplex. This includes, for example, many mixture models and topic models. Note that without conditioning on the latents ϕ, samples within the same cluster are dependent while samples in different clusters are independent. Our goal in this setting is to find a learner $f : \mathcal{X} \to \mathcal{Y}$ which performs well on new clusters, i.e. has small out-of-cluster loss $\mathbb{E}_{x', y'} \ell(y', f(x' \mid X_{1:n_x}, y_{1:n_x}))$, where ℓ is a continuous loss function, $x', y' \sim G(\phi')$ and $\phi' \sim H(\gamma)$.

To address this problem, previous work has considered the case where the partition c is observed and the leave-one-cluster-out (LOCO[1]) estimator is used for cross-validation [1, 2, 5, 6],

$$\widehat{\text{Err}}_{\text{LOCO}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j \in c_i^{-1}} \ell(y_j, f(x_j \mid x_{\bar{c}_i^{-1}}, y_{\bar{c}_i^{-1}})), \tag{2}$$

where c_i^{-1} and $\bar{c}_i^{-1}$ denote all sample indices belonging and not belonging to cluster i, respectively. For f, we allow any (possibly stochastic) function, which will realistically be some classifier or regressor. This strategy, which creates independent training and testing folds, is referred to as *conditioning on the partition*. By training and testing on disjoint clusters, LOCO is a very nearly unbiased estimate of the out-of-cluster loss, with a small

[1] also referred to as leave-one-label-out cross-validation, terminology we avoid due to potential confusion between sample and cluster labels

amount of bias due to training on $k-1$ clusters instead of k clusters.

The primary focus of this paper is that the partition c is often unknown a priori or uncertain. Instead of c, we are given an approximation $\hat{c}$, which is usually the result of some unsupervised clustering of X. For example, our original motivation for studying this issue arose from a common problem found in medical applications. Here, medical records correspond to samples X and patients correspond to clusters $i = 1, \dots, k$. In other words, each patient may have multiple medical records. For tasks such as cancer screening and medical imaging, LOCO prevents the learner from overfitting to patient-specific features such as social security number, name, and date-of-birth, which are not useful for prediction on new patients [5, 6, 7]. Better predictors generalize across patients, e.g. unexplained weight loss, fatigue, and tumor image features. This overfitting need not be blatant. An image classifier could learn the shapes of each patient's bone structure to predict whether they have lung cancer, which is not useful for new patients and difficult to inspect for without using LOCO.

The problem in the medical domain is we only observe X and y, and must infer an approximation $\hat{c}$ through clustering or entity resolution of records across hospitals and providers. Using $\hat{c}$ as a surrogate for c in LOCO presents a subtle yet potentially significant issue: cross-validation folds which were previously independent are now dependent due to incorrectly clustered samples. It is equivalent to mistakingly placing testing fold samples in the training folds, and vice versa. These mistakes enable the learner to, once again, overfit to patient-specific features — the exact problem we intended to avoid by using LOCO. We term this phenomenon *dependency leakage* and show that even at small approximation errors in $\hat{c}$, it can cause significant bias in cross-validation results.

Outside of the medical domain, we are familiar with similar problems in the census and counter-human-trafficking communities. At the US Census Bureau, matching persons across censuses is a challenging, imperfect process and the impact of using $\hat{c}$ for demographic, socioeconomic, and other statistical analysis is unclear [8, 9]. Similarly, imperfect record linkage results are used to estimate death counts in Syria and to both estimate and predict human trafficking in the United States [10, 11]. A major concern in these domains is that dependency leakage can bias a learner against certain sub-populations (i.e. clusters). For example, in Section 4 we empirically demonstrate how dependency leakage causes bias against certain demographics in US Census data. This is increasingly relevant as data science plays a greater role in credit and policy decisions [12, 13].

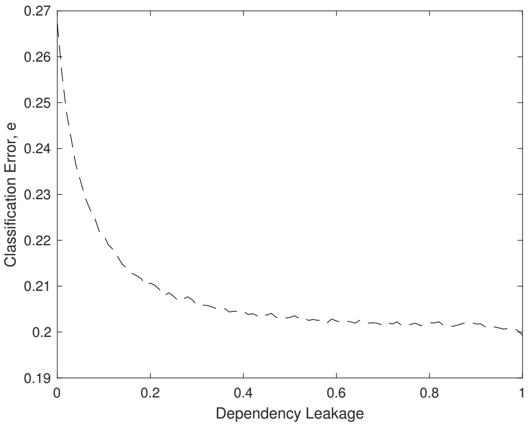


Figure 1: Simulation study on the heart disease dataset showing dependency leakage optimistically biases cross-validation results on clusters of dependent samples by allowing some learning on the test distribution.

In computer vision, consider the task of activity detection from videos. Given many videos of pole vaulters, it is common to have multiple videos of the same event (e.g. the 2016 Olympics). A concern here is the learner may overfit to scene features rather than learning a useful understanding of the activity. Splitting videos by event tags is comparable to using LOCO, which encourages learning activity features which generalize across events. However, event tags and the corresponding clustering are imperfect, and thus this problem may suffer from dependency leakage. One can imagine similar challenges for larger image datasets constructed through un- or semi-supervised means (e.g. knowledge bases).

In this paper, we contribute a novel bootstrap technique for learning on blocks of dependent data, which both estimates and corrects for dependency leakage. This enables learning on clusters of dependent data, where we only observe a noisy approximation of the true clustering, likely the result of some clustering or record linkage algorithm. The key insight is to increase dependency by further corrupting $\hat{c}$, in order to extrapolate an unbiased and consistent estimator for the true c. Simulation studies in the non-asymptotic case show our method significantly outperforms standard cross-validation techniques.

2 Learning in dependent data

The problem of constructing estimators for dependent data has been studied since Singh [14], who provided the first theoretical confirmation of the naive bootstrap's performance with IID data, and also showed its inadequacy for dependent data. Since then, the bootstrap has been extended to both time-series and cluster data. In time-series data, blocks of data are dependent according to some stochastic process [15, 16]. By varying the size and separation of the blocks, these block bootstrap methods can limit the dependency and thus control the bias and variance of the estimator, while sometimes achieving consistency. We refer the reader to [17] for a more thorough overview of the subject.

In cluster data, within-cluster samples are dependent while inter-cluster samples are typically assumed to be independent. This is the same formulation as Eq. 1. Many bootstrap methods have been proposed for variance estimation in the clustering setting, as classical bootstrap estimators will typically be downward biased [18]. Model-based methods assume a parametric model for the within-cluster error correlation. Model-free methods perform post-estimation bias-correction, such as the cluster-robust variance estimator (CRVE) for ordinary least squares [19] and non-linear settings [20]. CRVE suffers from having unbalanced or a small number of clusters, which is addressed in [21]. Field and Welsh [22] provide theoretical asymptotic analysis for several cluster bootstrap techniques, including the randomized cluster bootstrap, two-stage bootstrap [23] and residual bootstrap [24]. Multi-way bootstrap clustering is slightly more general, but still assumes samples belonging to none of the same clusters are independent [25]. Neither these bootstrap techniques nor LOCO cross-validation account for inter-cluster dependency, and will be inadequate for $\hat{c}$ and non-trivial f, ℓ, X and y.

In practice, when the clustering c is latent, researchers choose a coarse clustering $\hat{c}$ to ensure intra-cluster samples are as independent as possible [18]. A coarser clustering decreases bias and increases variance. This approach both lacks guarantees and requires choosing an appropriate clustering coarseness, which is an open problem. The key differentiation of our work is we directly address the issue of inter-cluster dependency due to $\hat{c}$.

Dependency leakage may have a significant impact on cross-validation results and model selection, both in theory and in practice. To illustrate, consider the simulation results depicted in Figure 1. We use heart disease data from hospitals in Cleveland, Hungary, Switzerland and California (see additional details in Section 4) [26]. Each location is a cluster and each sample is a single patient, thus the LOCO estimator generalizes performance at new hospitals. To simulate the effect of using $\hat{c}$ as a surrogate for c, we move samples from the test set into the train set with uniform probability (horizontal axis). Dependency leakage optimistically biases the cross-validation accuracy estimate by more than 25%. We will return to this example in Section 4.

3 The Binomial Block Bootstrap estimator

We now introduce the binomial block bootstrap (B3) class of estimators for cross-validating with dependent blocks of data. First, we begin by formalizing notation to simplify analysis of the core problem. We then proceed with the simplest leakage scenario and gradually build complexity until arriving at our final result. In section 3.2 we begin with the case where samples are moved with known probability in a single direction, from the test blocks to the train blocks or vice versa. Then we show how to solve for the unidirectional dependency leakage in section 3.3 and generalize to the bidirectional case in section 3.4.

3.1 Problem Setup

Broadly, we address the problem of LOCO cross-validation when a noisy approximation of the true latent partition c. For the remainder of the paper, we consider some arbitrary fixed i in Eq. 2 (i.e. a single fold). In the LOCO estimator with known partitioning c, each fold is created such that $\mathcal{T}$ and $\mathcal{V}$ are sets of training and testing samples, respectively, split by the partition, i.e. $c(i) \neq c(j)\ \forall x_i \in \mathcal{T}, x_j \in \mathcal{V}$. Without conditioning on the latent cluster parameters in Eq. 1, samples within the same cluster are dependent while samples in different clusters are independent. Thus, $\mathcal{T}$ and $\mathcal{V}$ are independent.

Using this notation, we think about the core problem as a learner f trained on samples $\mathcal{T}$ drawn IID from distribution $P_{\mathcal{T}}$ and tested on samples $\mathcal{V}$ drawn IID from a related but different distribution $P_{\mathcal{V}}$, a form of transfer learning.

Now, suppose we instead observe noisy datasets $\hat{\mathcal{T}}$ and $\hat{\mathcal{V}}$, where samples have randomly moved between $\mathcal{T}$ and $\mathcal{V}$. This question arises naturally when we only have $\hat{c}$, an approximation of c, likely obtained through clustering. Most importantly, $\hat{\mathcal{T}}$ and $\hat{\mathcal{V}}$ are dependent — which provides additional information to the learner and biases our cross-validation estimator. Our goal then is to answer questions regarding the continuous loss function ℓ evaluated on new clusters, for example $\mathbb{E}_{\mathcal{T}\sim P_{\mathcal{T}}}\mathbb{E}_{(x,y)\sim P_{\mathcal{V}}}\ell(f(x \mid \mathcal{T}), y)$, given only noisy datasets $\hat{\mathcal{T}}$ and $\hat{\mathcal{V}}$.

3.2 Unidirectional leakage with known probability

First, consider the case where samples move with known uniform probability from either $\mathcal{V}$ to $\mathcal{T}$ or vice versa to

[2] $M_{S_1,S_2}(w_1, w_2)$ is a mixture distribution of sets S_1 and S_2, where the probability of sampling from the sets are $w_1 + w_2 = 1$, respectively. Within set samples are drawn uniformly.

Algorithm 1 B3: Unidirectional leakage with known probability

```
procedure KnownUnidirectional($f, \hat{\mathcal{T}}, \hat{\mathcal{V}}, p_0, dir, n', t$)
    $\bar{b} \leftarrow \vec{0}$
    for $p_i$ in $\{p_0, p_0 + \delta, p_0 + 2\delta, \dots, 1\}$ do        ▷ Choose $\delta > 0$ s.t. $|\{p_i\}| > n'$
        $p' \leftarrow \frac{p_i - p_0}{1 - p_0}$
        for $j \leftarrow 1$ to $t$ do
            if $dir$ is $\mathcal{V}$ to $\mathcal{T}$ then
                $\mathcal{T}'_j \overset{n'}{\sim} M_{\hat{\mathcal{T}},\hat{\mathcal{V}}}(1 - p', p')$        ▷ $M$ is a mixture distribution[2]
                $\mathcal{V}'_j \leftarrow \hat{\mathcal{V}} \setminus \mathcal{T}'_j$
            else
                $\mathcal{V}'_j \overset{n'}{\sim} M_{\hat{\mathcal{T}},\hat{\mathcal{V}}}(p', 1 - p')$
                $\mathcal{T}'_j \leftarrow \hat{\mathcal{T}} \setminus \mathcal{V}'_j$
            end if
            $\hat{b}_i \leftarrow \frac{1}{|\mathcal{V}'_j|} \sum_{(x,y) \in \mathcal{V}'_j} \ell(y, f(x \mid \mathcal{T}'_j))$        ▷ $\ell$ is any continuous loss function
            $\bar{b}_i \leftarrow \bar{b}_i + \frac{\hat{b}_i}{t}$
        end for
    end for
    $A_{ij} \leftarrow \mathbb{P}(\text{Binomial}(n', p_i) = j) \quad \forall p_i \in p, j \in \{0, 1, \dots, n'\}$
    $\hat{e}, residual \leftarrow A(A^\intercal A)^{-1} A^\intercal \bar{b}$
    return $\hat{e}_0, residual$
end procedure
```

create $\hat{\mathcal{V}}$ and $\hat{\mathcal{T}}$. Without loss of generality, we consider the case where samples move from $\mathcal{V}$ to $\mathcal{T}$. In other words, $\hat{\mathcal{V}}$ contains only samples from $P_{\mathcal{V}}$ while $\hat{\mathcal{T}}$ contains samples from both $P_{\mathcal{T}}$ and $P_{\mathcal{V}}$. Let p_0 be the fraction of samples in $\hat{\mathcal{T}}$ from $\mathcal{V}$, i.e. $p_0 = \frac{|\hat{\mathcal{T}} \cap \mathcal{V}|}{|\hat{\mathcal{T}}|}$. The analysis for the other direction is identical.

The unidirectional B3 estimator (presented in Algorithm 1) is based on the observation that the number of corrupted samples in a bootstrap sample $\mathcal{T}'$ from $\hat{\mathcal{T}}$ is binomially distributed according to p_0 and $n' = |\mathcal{T}'|$. The bootstrap sample $\mathcal{T}'$ is formed by resampling with replacement n' times from $\hat{\mathcal{T}}$, which we notate as $\mathcal{T}' \overset{n'}{\sim} \hat{\mathcal{T}}$. Let b_0 be the expected bootstrap loss estimate, $b_0 = \mathbb{E}_{(x,y) \sim \hat{\mathcal{V}}} \mathbb{E}_{\mathcal{T}' \overset{n'}{\sim} \hat{\mathcal{T}}} \ell(f(x|\mathcal{T}'), y)$. We can express b_0 as a binomial weighting of the expected error at all numbers of corrupted samples in $\mathcal{T}'$. Formally,

$$b_0 = \langle a_0, e \rangle \tag{3}$$

where a_0 is the probability mass function (pmf) of Binomial(n', p_0), e_i is the expected loss with i corrupted samples in $\mathcal{T}'$ and $\langle \cdot, \cdot \rangle$ denotes the inner product operation. Our goal is to recover e_0, the loss with zero corruption.

At first, this may seem difficult as $b_0 = \langle a_0, e \rangle$ is a very underdetermined system (even assuming we know p_0). To overcome this deficiency, the key insight of our bootstrap technique is to artificially inject additional leakage by further mixing $\mathcal{V}$ into $\hat{\mathcal{T}}$ to create a fully or over defined system. This increases p, alters the binomial pmf a_0, and generates a new linear equality $b_1 = \langle a_1, e \rangle$ where a_1 is the pmf of Binomial(n', p_1). Repeating this process many times results in the linear system

$$\begin{array}{cc} & \begin{array}{ccccc} 0 & 1 & \cdot & \cdot & n' \end{array} \\ \begin{array}{c} p_0 \\ p_1 \\ \cdot \\ \cdot \\ 1 \end{array} & \begin{pmatrix} \leftarrow \text{Binomial pmf} \rightarrow \\ \cdot \\ \cdot \\ \cdot \\ \leftarrow \text{Binomial pmf} \rightarrow \end{pmatrix} \end{array} \begin{pmatrix} \\ \\ e \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ b \\ \\ \\ \end{pmatrix} \tag{4}$$

$$A(p_0) \qquad e \quad = \quad b$$

For any unique choice of $p = (p_0, p_1, \dots, p_m) \in [0, 1]^m$, this system will be well-defined (by Lemma 3.1) and can be readily solved for e_0. A somewhat similar clustering randomization idea is used in [27] for estimating treatment effects, though their formulation is quite different than Eq. 4.

Lemma 3.1. *Let matrix A be defined such that*

$$A_{ij} = \mathbb{P}(\text{Binomial}(n', p_i) = j). \tag{5}$$

Then A has full rank for any choice of unique parameters $p = (p_0, p_1, \dots, p_m) \in [0, 1]^m$.

Proof. See Appendix A.1. □

Algorithm 2 B3: Unidirectional leakage with unknown probability

```
procedure UnknownUnidir($f, \hat{\mathcal{T}}, \hat{\mathcal{V}}, dir, n', t$)
    $residual^* \leftarrow \infty$
    $n \leftarrow |\hat{\mathcal{T}}|$
    for $\hat{p}_0$ in $\left\{\frac{0}{n}, \frac{1}{n}, \dots, \frac{n-1}{n}\right\}$ do
        $\hat{e}_0, residual \leftarrow$ KnownUnidir($f, \hat{\mathcal{T}}, \hat{\mathcal{V}}, \hat{p}_0, dir, n', t$)
        if $residual < residual^*$ then
            $\hat{e}_0^* \leftarrow \hat{e}_0$
            $\hat{p}_0^* \leftarrow \hat{p}_0$
            $residual^* \leftarrow residual$
        end if
    end for
    return $\hat{e}_0^*, \hat{p}_0^*$
end procedure
```

Roughly speaking, Algorithm 1 is estimating the loss at increasing levels of dependency leakage, and then extrapolating the loss at zero dependency. It is possible to achieve reasonable results in practice because we know the true formulation to be a binomial weighted regression problem and thus know matrix A exactly. Further, the extrapolation does not extend far beyond the known range for practical clusterings $\hat{c}$ with small p_0.

The estimator $\hat{e}_0$ in Algorithm 1 is consistent, unbiased and has variance decreasing linearly with respect to the number of bootstrap samples t.

Theorem 3.2. *The estimator $\hat{e}_0$ in Algorithm 1 satisfies*

1. *Consistent:* $\hat{e}_0 \xrightarrow{p} \mathop{\mathbb{E}}_{\mathcal{T}' \overset{n'}{\sim} P_{\mathcal{T}}} \mathop{\mathbb{E}}_{(x,y)\sim P_{\mathcal{V}}} \ell(y, f(x \mid \mathcal{T}'))$ *as* $t, |\hat{\mathcal{T}}|, |\hat{\mathcal{V}}| \to \infty$
2. *Unbiased:* $\mathbb{E}[\hat{e}_0] = e_0$ *for finite* t *and infinite* $|\hat{\mathcal{T}}|, |\hat{\mathcal{V}}|$.
3. $\mathrm{Var}(\hat{e}_0) =$

$$\sum_{i=0}^{n'} \left[\frac{\sum\limits_{\substack{0 \le m_0 < \cdots < m_{n'-1} \le n' \\ m_0, \dots, m_{n'-1} \neq i}} p_{m_0} \cdots p_{m_{n'-1}}}{\prod\limits_{0 \le m \le n', m \neq i} (p_m - p_i)} \right]^2 \frac{\sigma_{b_i}^2}{t}$$

where $\sigma_{b_i}^2$ is the variance of $\hat{b}_i$ in Algorithm 1, which is a function of f, ℓ and the data.

Proof. See Appendix A.2. □

Remark For classification error ℓ, note $\sigma_{b_j}^2 \le \frac{1}{4}$ by Popoviciu's inequality. Generally speaking, there exists a variance tradeoff when choosing $p_0, \dots, p_{n'}$ — we can expect lower variance as the values are spaced further apart (larger denominator) and when they are closer to p_0 (smaller numerator), which are competing choices.

Remark The quality of the clustering $\hat{c}$ plays an important role in the performance of our estimator. As p_0 increases, the estimator remains unbiased but the variance increases according to Statement 3.

3.3 Unidirectional leakage with unknown probability

We now extend the unidirectional leakage scenario from Section 3.2 to the situation where p_0 is unknown a priori. The general strategy is to minimize the residual $||A(\hat{p}_0)e - \bar{b}||$ over $\hat{p}_0$ and show that a unique minimum exists and it is always the true leakage probability p_0. The most basic optimization procedure detailed in Algorithm 2 searches over the discrete set of possible solutions, though one can imagine other optimization procedures. The search space will be, at most, the one dimensional line defined by $\left[0, \frac{n-1}{n}\right]$ where $n = |\hat{\mathcal{T}}|$.

Our optimization routine in Algorithm 2 converges to the true leakage probability p_0 if the following assumption holds

Assumption 1. *b is independent of the columns of $A(\hat{p}_0)$ (except, obviously, at $\hat{p}_0 = p_0$).*

Remark This is a weak assumption when choosing $m >> n'$: it is unlikely the loss vector b happens to fall in the column space of A.

Theorem 3.3. *If Assumption 1 holds, then the estimators $\hat{p}_0^*$ and $\hat{e}_0^*$ in Algorithm 2 are consistent, i.e. $\hat{p}_0^* \xrightarrow{p} p_0$ and $\hat{e}_0^* \xrightarrow{p} e_0$ as $t, |\mathcal{T}|, |\mathcal{V}| \to \infty$ and for $p_0 < 1$.*

Proof. Without loss of generality, we prove the case where samples move in the direction from $\mathcal{V}$ to $\mathcal{T}$. We begin by proving the convergence of $\hat{p}_0^*$. Let $n = |\hat{\mathcal{T}}|$. In Algorithm 2, $p_0^{*(t)} = \arg\min_{p_0 \in \left\{\frac{0}{n}, \frac{1}{n}, \dots, \frac{n-1}{n}\right\}} g^{(t)}(p_0)$, where the function

$g^{(i)}(p_0) = ||A(p_0)(A^\intercal(p_0)A(p_0))^{-1}A^\intercal(p_0)\bar{b}^{(i)} - \bar{b}^{(i)}||_2^2$ if $p_0 \in \left[0, \frac{n-1}{n}\right]$ and else infinity. We use $\bar{b}^{(i)}$ to denote the mean estimator $\bar{b}$ in Algorithm 1 after $t = i$ samples. Let $g(p_0) = ||A(p_0)(A^\intercal(p_0)A(p_0))^{-1}A^\intercal(p_0)b - b||_2^2$ if $p_0 \in \left[0, \frac{n-1}{n}\right]$ and else infinity.

Both g and the sequence of functions $\{g^{(0)}, g^{(1)}, \dots\}$ are level-bounded, lower semi-continuous and proper. By Lemma 3.4, $g^{(i)} \xrightarrow{e} g$ where $\xrightarrow{e}$ denotes convergence in epigraph. Thus, $residual = \min_{p_0 \in \left[0, \frac{n-1}{n}\right]} g^{(t)}(p_0) \xrightarrow{p} \min_{p_0} g(p_0)$ [28]. We know at least one perfect solution $g(p_0) = 0$ exists, that this solution is unique (by Assumption 1) and that this solution is in $\left\{0, \frac{1}{n}, \dots, \frac{n-1}{n}\right\}$. Thus, $p_0^* \xrightarrow{p} p_0$ and $residual \xrightarrow{p} 0$. □

Lemma 3.4. *Let*

$$g^{(i)}(p_0) = \begin{cases} ||A(p_0)(A^\intercal(p_0)A(p_0))^{-1}A^\intercal(p_0)\bar{b}^{(i)} - \bar{b}^{(i)}||_2^2 \\ \qquad\qquad\qquad\qquad \textit{if} \quad p_0 \in \left[0, \frac{n-1}{n}\right] \\ \infty \qquad\qquad\qquad\quad\;\; \textit{else} \end{cases}$$

$$g(p_0) = \begin{cases} ||A(p_0)(A^\intercal(p_0)A(p_0))^{-1}A^\intercal(p_0)b - b||_2^2 \\ \qquad\qquad\qquad\qquad \textit{if} \quad p_0 \in \left[0, \frac{n-1}{n}\right] \\ \infty \qquad\qquad\qquad\quad\;\; \textit{else} \end{cases}$$

Then $g^{(i)} \xrightarrow{e} g$, where we use $\xrightarrow{e}$ to denote convergence in epigraph.

Proof. Recall, $g^{(i)} \xrightarrow{e} g$ if and only if at each point e

$$\liminf_i g^{(i)}(e^{(i)}) \geq g(e), \text{ for every } e^{(i)} \to e \qquad (6a)$$

$$\limsup_i g^{(i)}(e^{(i)}) \leq g(e), \text{ for some } e^{(i)} \to e \qquad (6b)$$

Let $\mathcal{N}_\infty^\# = \{N \in \mathbb{N} | N \text{ is infinite}\}$ be all infinite sets of natural numbers, which we require for cases of periodicity. To establish Eq. 6a, it is sufficient to show that whenever $e^{(i)} \underset{N}{\to} e$ and $f^{(i)}(e^{(i)}) \underset{N}{\to} \alpha$, then $f(e) \leq \alpha$. We consider three cases, when $e \in \left(0, \frac{n-1}{n}\right)$, when $e \notin \left[0, \frac{n-1}{n}\right]$ and when $e \in \left\{0, \frac{n-1}{n}\right\}$. The first case is readily established from the proof of Theorem 3.2, where we showed that $\bar{b}^{(i)} \underset{N}{\to} b \ \forall N \in \mathcal{N}_\infty^\#$, $A(e^{(i)})(A^\intercal(e^{(i)})A(e^{(i)}))^{-1}A^\intercal(e^{(i)}) \underset{N}{\to} A(e)(A^\intercal(e)A(e))^{-1}A^\intercal(e)$, and thus $f^{(i)}(e^{(i)}) \underset{N}{\to} f(e) \ \forall N \in \mathcal{N}_\infty^\#$. In the case where $e \notin \left[0, \frac{n-1}{n}\right]$, $g^{(i)}(e) = \infty$ readily establishes the inequality. In the boundary cases $e \in \left\{0, \frac{n-1}{n}\right\}$, note either $g^{(i)}(e^{(i)}) \underset{N}{\to} \infty$ or $g^{(i)}(e^{(i)}) \underset{N}{\to} g(e)$, respectively. To establish Eq. 6b, choose the sequence $\{e^{(i)}\} = e \ \forall i \in \mathbb{N}$. □

3.4 Bidirectional leakage with unknown probabilities

Lastly, we extend the unidirectional leakage results in Sections 3.2 and 3.3 to the full bidirectional setting, where samples move with unknown uniform probability between $\mathcal{T}$ and $\mathcal{V}$. More specifically, let $p_{\mathcal{T},0} = \frac{|\hat{\mathcal{T}} \cap \mathcal{V}|}{|\hat{\mathcal{T}}|}$ and $p_{\mathcal{V},0} = \frac{|\hat{\mathcal{V}} \cap \mathcal{T}|}{|\hat{\mathcal{V}}|}$ be the probabilities a sample in $\hat{\mathcal{T}}$ and $\hat{\mathcal{V}}$ do not belong in that set, respectively. Similar to the unidirectional case, we independently resample with replacement $n'_\mathcal{T}$ and $n'_\mathcal{V}$ samples from $\mathcal{T}$ and $\mathcal{V}$ to form the bootstrap sample sets $\mathcal{T}'$ and $\mathcal{V}'$, respectively. Thus, the number of corrupted samples in $\mathcal{T}'$ and $\mathcal{V}'$ is drawn according to a joint distribution of two independent binomials. We then formulate a regression problem analogous to Eq. 4,

$$\begin{array}{cc} & \begin{matrix} 0 & 1 & \cdot & \cdot & n' \end{matrix} \\ \begin{matrix} p_{\mathcal{T},0}, p_{\mathcal{V},0} \\ \cdot \\ \cdot \\ \cdot \\ p_{\mathcal{T},n_\mathcal{T}}, p_{\mathcal{V},n_\mathcal{V}} \end{matrix} & \begin{pmatrix} \leftarrow \text{Joint Bin pmf} \rightarrow \\ \cdot \\ \cdot \\ \cdot \\ \leftarrow \text{Joint Bin pmf} \rightarrow \end{pmatrix} \begin{pmatrix} \\ \\ e \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ b \\ \\ \\ \end{pmatrix} \\ & \\ & A(p_{\mathcal{T},0}, p_{\mathcal{V},0}) \qquad\qquad\quad e \ = \ b \end{array}$$

where $n' = (n'_\mathcal{T} + 1)(n'_\mathcal{V} + 1) - 1$. Note since the joint pmf is defined for $(n'_\mathcal{T} + 1)(n'_\mathcal{V} + 1)$ values, we must bootstrap at $(n'_\mathcal{T} + 1)(n'_\mathcal{V} + 1)$ levels of leakage.

In the case where the leakage probabilities $p_{\mathcal{T},0}$ and $p_{\mathcal{V},0}$ are unknown, we again minimize the residual. The resulting methods for the bidirectional leakage scenario with known and unknown probabilities are presented in Algorithms 3 and 4 (see Appendix B), respectively.

Here, we show the full rank and consistency results for Algorithms 1 and 2 extend to Algorithms 3 and 4. The main difference is we consider the *joint* binomial matrix A, which is also full rank and thus the regression problem is well defined.

Lemma 3.5. *Joint binomial matrix A has full rank for any choice of unique parameters $p_\mathcal{T} = (p_{\mathcal{T},0}, p_{\mathcal{T},1}, \dots, p_{\mathcal{T},m}) \in [0,1]^m$ and $p_\mathcal{V} = (p_{\mathcal{V},0}, p_{\mathcal{V},1}, \dots, p_{\mathcal{V},m'}) \in [0,1]^{m'}$.*

Likewise, the consistency results in Theorems 3.2 and 3.3 extend to the bidirectional leakage scenario.

Theorem 3.6. *For $p_{\mathcal{T},0}, p_{\mathcal{V},0} < 1$ in Algorithm 3, e_0 converges to the expected error on uncorrupted distributions $\mathcal{T}$ and $\mathcal{V}$, $\hat{e}_0 \to \mathbb{E}_{\mathcal{T}' \overset{n'_\mathcal{T}}{\sim} P_\mathcal{T}} \mathbb{E}_{(x,y)\sim P_\mathcal{V}} \ell(y, f(x \mid \mathcal{T}'))$ as $t, |\mathcal{T}|, |\mathcal{V}| \to \infty$.*

Theorem 3.7. *For $p_{\mathcal{T},0}, p_{\mathcal{V},0} < 1$ in Algorithm 4,*

$p^*_{\mathcal{T},0} \xrightarrow{p} p_{\mathcal{T},0}$, $p^*_{\mathcal{V},0} \xrightarrow{p} p_{\mathcal{V},0}$ *and* $\hat{e}^* \xrightarrow{p} e^*$ *as* $t, |\mathcal{T}|, |\mathcal{V}| \to \infty$.

Proofs are in Appendices A.3-A.5.

4 Simulation study

Thus far, we have appealed to asymptotic theory and bias-variance analysis. This is not uncommon for bootstrap and cross-validation analysis, and like others, we now turn to empirical arguments. In this section, we present simulation study results which demonstrate our core method in Algorithm 1 significantly outperforms conventional methods. For all experiments, we consider the more difficult direction where samples move from $\mathcal{V}$ to $\mathcal{T}$.

Our estimators are unbiased and consistent, but they may have large variance (see Theorem 3.2). When practically implementing these estimators, it is beneficial to add a small amount of regularization to achieve a better bias–variance tradeoff. Although we know from Lemma 3.1 and 3.5 that matrix A is full rank, it may be ill-conditioned. Adding regularization helps to improve the condition number of matrix A. Evidence suggests this is a tradeoff worth making. Specifically, in the linear system objective function within Algorithms 1 and 4 we instead solve some variation of

$$\begin{aligned} &\underset{\hat{e}}{\text{minimize}} \quad ||A\hat{e} - \bar{b}||_2^2 + \lambda R(\hat{e}) \\ &\text{subject to} \quad \hat{e}_{j-1} \geq \hat{e}_j \geq 0 \quad \text{for all } j = 1, \ldots, n'. \end{aligned}$$

where λ is a regularization constant and R is some regularization function. We choose the trend filter regularizer $R(\hat{e}) = ||D\hat{e}||_2^2$ to ensure $\hat{e}$ is smooth [29]. For a second-order filter, which regularizes the second derivative of $\hat{e}$, D is the difference matrix

$$D = \begin{bmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{bmatrix}$$

where unshown entries are zero. Matrices D for higher order trend filters follow similarly. Intuitively, we expect the estimator error to degrade both monotonically (the constraint) and somewhat smoothly (the regularizer). Later results validate these assumptions.

Experiment I For the synthetic simulation study, we use a partition model with $k = 2$ parts and n sufficiently large such that duplicate resamples are improbable, a subsample of which is depicted in Figure 2. We choose the number of corruption levels $m = 2n$ following the arguments in Assumption 1 and perform $t = 1000$ bootstrap samples at each p_i. For f, we use a linear regression model and set the loss ℓ as the mean squared error. To simulate the effects of noisy clusters $\hat{c}$, we move samples between the two parts $\mathcal{T}$ and $\mathcal{V}$ with initial uniform probability $p_0 = 0.1$. All experiments use either a second (T2), third (T3) or fourth (T4) order trend filter with $\lambda = 0.1$ and/or a monotonic constraint. For baselines, we compare against naive IID k-fold and the current state-of-the-art LOCO cross-validation.

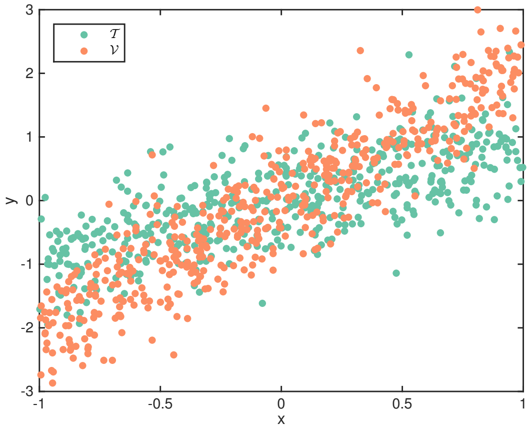
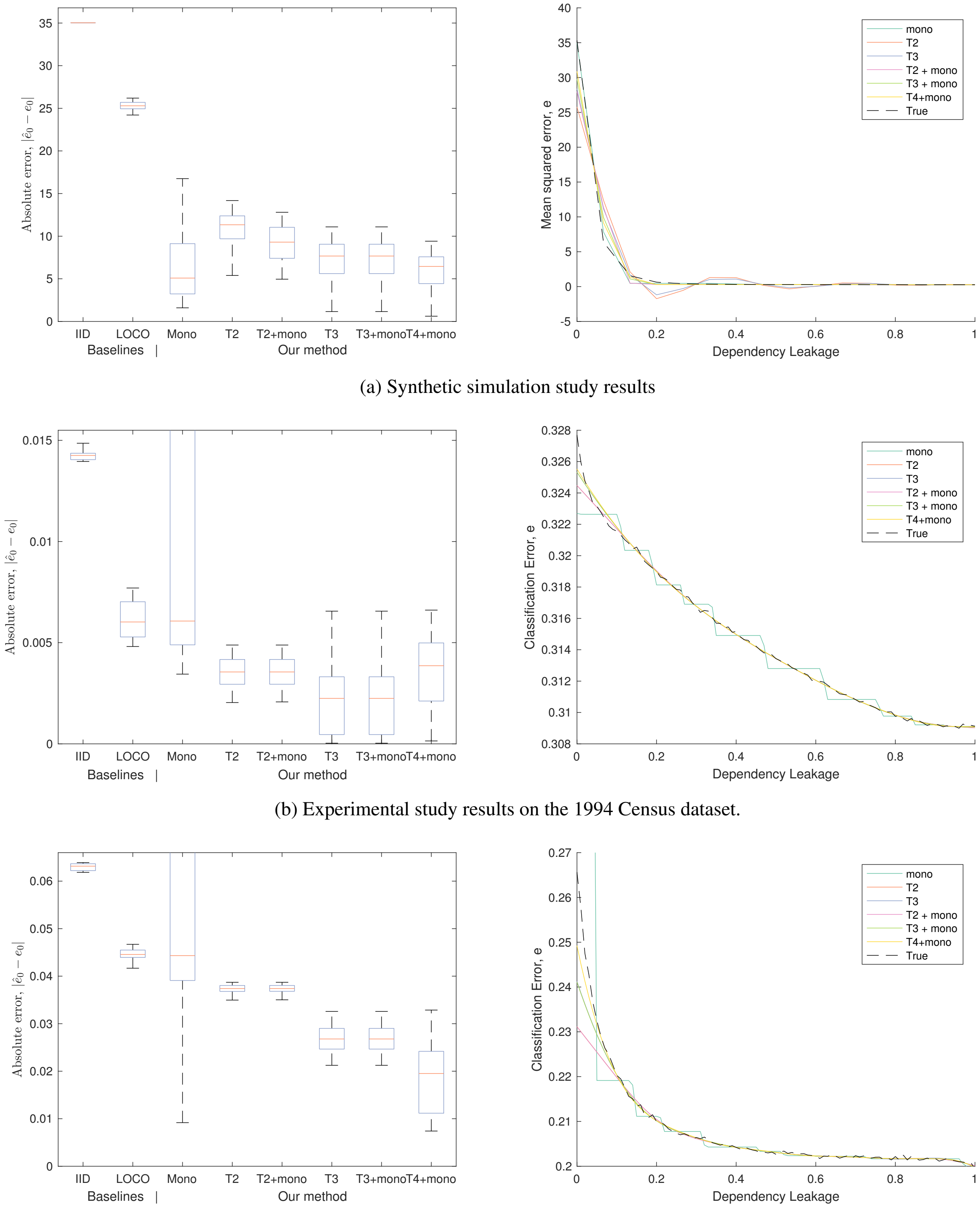


Figure 2: Subsample of data used in the synthetic Experiment I.

Our main synthetic results are presented in the Figure 3a boxplot. LOCO outperforms traditional IID cross-validation, which suggests blocking on the corrupted clusters $\hat{c}$ partially limits the effects of dependency leakage. However, even at $p_0 = 0.1$, LOCO is still unacceptably biased. Our methods, with various forms of regularization, all significantly outperform both existing estimators. Figure 3a also suggests a bias-variance trade-off among all the tested methods. IID cross-validation has high bias and low variance, whereas our methods have low bias and higher variance. Ultimately, this tradeoff allows our methods to achieve lower MSE by choosing an appropriate form and strength of regularization.

An interesting consequence of our method is that in addition to recovering the independent partition performance e_0, we also recover the performance $e_1, e_2, \ldots$ at all levels of dependency leakage, as depicted in Figure 3a. The true loss e (dashed black line) decays monotonically and smoothly, which justifies our regularization choices.

Experiment II In the second experiment, we use data from the 1994 US Census to validate our claim that conventional cross-validation introduce bias against subpopulations due to dependency leakage [26]. Empirical



(a) Synthetic simulation study results

(b) Experimental study results on the 1994 Census dataset.

(c) Experimental study results on the heart disease dataset.

Figure 3: **Left** Estimating the generalization loss e_0. Our class of B3 estimators, with various forms of regularization (monotonic; second, third or fourth-order trend filter) outperform existing estimators. Baseline cross-validation methods are biased against the sub-populations we studied, and our class of B3 estimators help correct this bias. **Right** The B3 estimator recovers the full loss vector e. Empirically, the true loss decays monotonically and smoothly in practice, justifying our regularization choices.

results show our class of B3 estimators outperform the baseline methods in practical, non-asymptotic situations.

Here, we consider the task of predicting a person's income given their demographic, educational and occupational information. Our training set consists of samples from certain origin countries and we wish to train a learner which performs well for people of all countries. In other words, we minimize the LOCO generalization loss, where clusters correspond to origin countries. For this experiment, we use 30368 persons from the United States, El Salvador, Germany, Mexico, Philippines and Puerto Rico for training set $\mathcal{T}$ and validate with $\mathcal{V}$ on 221 immigrants from India and Canada. We set $p_0 = 0.1$, $\lambda = 10$, $n' = 100$, $m = 2n'$ and $t = 2500$. For features, we consider their age, years of education, work hours per week, race, and occupation. We trained an SVM classifier to predict whether their yearly income is greater than US$50k per year.

The boxplot in Figure 3b shows small error in origin country causes the learner to be biased against Indian and Canadian immigrants, due to dependency leakage. In other words, the classifier is rewarded for learning attributes specific to the training countries, even though they do not generalize across all countries. Similarly to the synthetic study, the B3 estimators outperform the baseline methods and accurately recover the full loss vector e, which decays monotonically and smoothly.

Experiment III In the third experiment, we use heart disease data collected from Cleveland, USA; VA Long Beach, USA; Switzerland and Hungary [26]. The task is to predict whether a patient has heart disease, given their demographic information and vital signs. We need to train a classifier which performs well at new hospitals – given data from only these 4 locations. Thus, clusters correspond to hospital location and we use LOCO to estimate the generalization error. Training clusters correspond to 479 patients in Cleveland, Long Beach and Switzerland, testing clusters correspond to 262 patients in Hungary. All other experimental details are the same as Experiment II. The results are shown in Figure 3c.

5 Extensions

This work poses several additional questions, some of which we briefly address now. For example, we have extended these methods from estimating the expected loss e to estimating an expected loss histogram E in Eq. 4. To do so, one can simply store the empirical bootstrap histogram $\bar{B}$ in lieu of the empirical bootstrap mean $\bar{b}$. The downside is estimating the additional information in E increases the variance by a linear factor according to the number of histogram bins.

To improve the numerical solution in Algorithm 1 in the direction where samples move from $\mathcal{T}$ to $\mathcal{V}$, note that e will be a linear vector, i.e. $e_{i+1} - e_i = \beta \ \forall i \in \{0, \ldots, n'-1\}$. This is because the training set $\mathcal{T}'$ has zero corruption, the expected number of corrupted samples in $\hat{\mathcal{V}}$ varies linearly with p_i for fixed δ, and the empirical loss is a mean loss of the samples in $\hat{\mathcal{V}}$. Enforcing this constraint on $\hat{e}$ would improve the solution quality for the direction where samples move from $\mathcal{T}$ to $\mathcal{V}$. We always considered the more difficult $\mathcal{V}$ to $\mathcal{T}$ leakage direction, where we have no prior knowledge of e.

The question of unbalanced clusters for CRVE was addressed in [21]. In our cross-validation method, small Var($|\hat{\mathcal{T}}|$) and Var($|\hat{\mathcal{V}}|$) across the cross-validation folds improves convergence. With unbalanced clusters, instead of leaving one cluster out, we could leave multiple clusters out such that $|\hat{\mathcal{T}}|$ and $|\hat{\mathcal{V}}|$ have lower variance even with high variance cluster sizes. CRVE also suffers from having a small number of clusters k [18]. Our estimator will be nearly unbiased but have high variance with a small number of clusters, due to the same properties as LOCO (see Section 1).

Though we have shown asymptotic convergence of our methods, there are several open questions. Notably, we use a naive discrete optimization routine in Algorithms 2 and 4 to solve for $p_{\mathcal{T},0}$ and $p_{\mathcal{V},0}$. The functions $g(i)(p_0)$ are non-convex, but they are smooth with finite support and faster convergence may be possible.

6 Conclusions

In this paper, we addressed the issue of evaluating a learner on blocks of dependent data. Unlike existing bootstrap methods, which assume a perfect clustering, we allow for imperfect clusterings $\hat{c}$ such that inter-cluster samples may be dependent. Real world applications ranging from medical diagnostics to computer vision fall into this class of problems. Empirical evidence on synthetic data, the 1994 US Census and heart disease data shows dependency leakage biases cross-validation results and thus affects model selection. We presented the B3 class of estimators, which significantly outperform existing cross-validation methods in this setting. The key insight of our bootstrapping methods is that by injecting additional dependency, we can extrapolate an unbiased and asymptotically consistent estimator of the performance on independent clusters.

Acknowledgements

This work has been partially supported by DARPA under awards FA8750-14-2-0244 and FA8750-12-2-0324.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 6. Springer, 2013.

[3] Daniel Joseph McDonald. *Generalization Error Bounds for Time Series*. PhD thesis, Carnegie Mellon University, 2012.

[4] Burr Settles. *Active Learning*. Morgan & Claypool, 2012.

[5] Chao Liu, Hernando Gomez, Bridget Deasy, Brian Zuckerbraun, Srinivasa G Narasimhan, Artur Dubrawski, and Michael R Pinsky. Micro-vascular Video Analysis for Critical Care. *IEEE Transactions on Medical Imaging*, 2017.

[6] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. Voodoo Machine Learning for Clinical Predictions. *bioRxiv*, 2016.

[7] Pavel Krepelka and Petr Drexler. A Fiber Optic Biosensor Improved with NIR Spectroscopy for Bacterial Identification. In *International Conference on Near Infrared Spectroscopy*, 2015.

[8] William E. Winkler and Yves Thibaudeau. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census. Technical report, U.S. Census Bureau, 1990.

[9] William E. Winkler. Overview of record linkage and current research directions. Technical report, U.S. Census Bureau, 2006.

[10] Peter Sadosky, Anshumali Shrivastava, Megan Price, and Rebecca C Steorts. Blocking Methods Applied to Casualty Records from the Syrian Conflict. *arXiv*, 2015.

[11] Artur Dubrawski, Kyle Miller, Matt Barnes, Benedikt Boecking, and Emily Kennedy. Leveraging Publicly Available Data to Discern Patterns of Human-Trafficking Activity. *Journal of Human Trafficking*, 1:65–85, 2015.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2011.

[13] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[14] Kesar Singh. On the Asymptotic Accuracy of Efron's Bootstrap. *The Annals of Statistics*, 9(6):11877–1195, 1981.

[15] Peter Hall. Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20(2):231–246, 1985.

[16] Regina Y. Liu and Kesar Singh. Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*, pages 225–248. Wiley-Interscience, 1992.

[17] Soumendra Nath Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2003.

[18] R. C. Cameron and Douglas L. Miller. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372, 2015.

[19] Halbert White. *Asymptotic theory for econometricians*. Academic Press, 1984.

[20] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[21] James G. MacKinnon and Matthew D. Webb. Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics*, 32:233–254, 2017.

[22] C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(3):369–390, 2007.

[23] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge University Press, 1997.

[24] Michael K Andersson and Sune Karlsson. Bootstrapping error component models. *Computational Statistics*, 16(2):221–231, 2001.

[25] Douglas Miller, A. Cameron, and Jonah Gelbach. Robust Inference with Multi-way Clustering. *Journal of Business & Economic Statistics*, 29(2), 2011.

[26] Moshe Lichman. UCI Machine Learning Repository, 2013.

[27] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph Cluster Randomization: Network Exposure to Multiple Universes. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 329–337, 2013.

[28] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 2009.

[29] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. L1 Trend Filtering. *Society for Industrial and Applied Mathematics (SIAM) Review*, 51(2):339–360, 2009.

[30] Nathaniel Macon and Abraham Spitzbart. Inverses of Vandermonde matrices. *The American Mathematical Monthly*, 65(2):95–100, 1958.