Boosted Kernel Weighting – Using Statistical Learning to Improve Inference from Nonprobability Samples

Christoph Kern[1, 2] Yan Li[3] Lingxiao Wang[3]

[1]University of Mannheim

[2]LMU Munich

[3]University of Maryland, The Joint Program in Survey Methodology (JPSM)

BigSurv20 Conference

Introduction

Non-probability samples	**Probability samples**
Less expensive	
Quick, convenient	Representativeness
Often more detailed info	Inference for population
Large sample sizes	
? Self-selection	? Cost
? Biased estimates	

Methods for improving inference from non-probability samples

- Superpopulation modeling approach (Elliott and Valliant, 2017)
- Mass imputation (Yang and Kim, 2020), ...
- **Propensity-score-based methods**

Propensity-Score-Based Methods

- **Inverse Propensity Score Weighting (IPSW)**
  - Use inverse of propensity score (PS) for constructing pseudo-weights
  - Sensitive to model misspecification, extreme pseudo-weights

- **Propensity Score Adjustment by Subclassification (PSAS)**
  - Use PS as a similarity measure and distribute survey weights equally within subclasses
  - Less variable estimates, but can be ineffective at bias reduction

- **Kernel Weighting** (**KW**, Wang et al. 2020)
  - Distribute survey weights fractionally to non-probability sample units
  - Effective bias reduction while controlling variance

  $\rightarrow$ *Further improve the KW method by pairing it with Machine/Statistical Learning?*

Setting and Notation

Target finite population FP with size N, $i \in (1, \ldots, N)$

Probability sample ($s_p \subset FP$)	Non-probability sample ($s_{np} \subset FP$)
Selection indicator: $\delta_i^{(p)} (= 1 \; if \; i \in s_p)$	$\delta_i^{(np)} (= 1 \; if \; i \in s_{np})$
Inclusion probability: $\pi_i^p \equiv P(i \in s_p \,\vert\, FP)$	$\pi_i^{np} \equiv P(i \in s_{np} \,\vert\, FP)$

Combined sample: $s_p \cup s_{np}$

- Outcome variable of interest y_i
- Vector of covariates $\mathbf{x}_i$
- Indicator of non-probability sample membership: $R_i = 1 \; for \; i \in s_{np}, 0 \; for \; i \in s_p$
- Propensity score: $p(\mathbf{x}_i) = P(i \in s_{np} \,\vert\, s_p \cup s_{np}, \mathbf{x}_i)$

Assumptions

A1. The sample selection is uncorrelated with the variable of interest given the covariates, i.e.
$\pi_i^{(np)} = E_{np}(\delta_i^{(np)} \mid y_i, \mathbf{x}_i) = E_{np}(\delta_i^{(np)} \mid \mathbf{x}_i)$ and
$\pi_i^{(p)} = E_p(\delta_i^{(p)} \mid y_i, \mathbf{x}_i) = E_p(\delta_i^{(p)} \mid \mathbf{x}_i)$ for $i \in FP$
$\rightarrow$ **Exchangeability**

A2. All finite population units have positive inclusion probabilities, i.e.
$\pi_i^{np} > 0$ and $\pi_i^{p} > 0$ for $i \in FP$
$\rightarrow$ **Common support**

KW-ML: Step 1

1. Estimating propensity scores with Machine Learning (ML) methods

   Model-Based Recursive Partitioning (**MOB**)
   - Combines parametric modeling with decision trees (Zeileis et al., 2008)
   - A set of local GLMs with group-specific coefficients

   Conditional Random Forests (**CRF**)
   - Grows Conditional Inference Trees on subsampled data (Strobl et al., 2007)
   - A (large) collection of CTREEs

   Gradient Tree Boosting (**GBM**)
   - Sum-of-trees approach (Friedman et al., 2000)
   - Combines decision trees that are grown in sequence

   Model-Based Boosting (**MBoost**)
   - Boosts small parametric models (Buehlmann and Hothorn, 2007)
   - Base learners can be non-linear, non-additive

KW-ML: Step 1

- Hyperparameter tuning for propensity score estimation

  The KW-ML tuning objective is to optimize covariate balance between the probability (p) and the nonprobability (np) sample:

$$SMD = \frac{(\bar{x}_{np} - \bar{x}_p)}{\sqrt{\frac{\sigma^2_{np} + \sigma^2_p}{2}}}$$

  where $\bar{x}$ and σ^2 are the sample mean and variance,
  and for binary covariates $\sigma^2_{np} = \bar{x}_{np}(1 - \bar{x}_{np})$, $\sigma^2_p = \bar{x}_p(1 - \bar{x}_p)$.

KW-ML: Step 2

2. Constructing pseudo-weights using the KW method based on ML propensity scores

$$w_j^* = \sum_{i \in s_p} d_i \frac{K\left(\frac{p(\mathbf{x}_i) - p(\mathbf{x}_j)}{h}\right)}{\sum_{j \in s_{np}} K\left(\frac{p(\mathbf{x}_i) - p(\mathbf{x}_j)}{h}\right)} \text{ for } j \in s_{np}$$

where $p(\mathbf{x}_i) - p(\mathbf{x}_j)$ measures similarity of $j \in s_{np}$ and $i \in s_p$,
d_i is the sample weight of $i \in s_p$, and
$K(\cdot)$ is an arbitrary kernel function with bandwidth h.

The KW-ML estimate of the population mean is

$$\bar{y}^{KW-ML} = \frac{1}{\widehat{N}} \sum_{j \in s_{np}} w_j^* y_j \text{ with } \widehat{N} = \sum_{j \in s_{np}} w_j^*$$

Simulation Setup

1. Finite population of size $N = 100,000$
   - Four confounders, $X_1, \ldots, X_4$, predictive of outcome variable and sample selection
     - $X_1^*, \ldots, X_4^*$ generated by categorizing $X_1, \ldots, X_4$
     - $X_1^{**}, \ldots, X_4^{**}$ based on linear combinations of X and Y
   - Three design variables, X_5, X_6, X_7, predictive of sample selection only
   - Three outcome predictors, X_8, X_9, X_{10}, predictive of outcome variable only
   - Dichotomous outcome variable generated by a Bernoulli distribution

$$\mu = \{1 + \exp(-\gamma^T \mathbf{X})\}$$

2. Probability proportional to size design to assemble the probability and nonprobability sample ($n_p, n_{np} = 5,000$)
   - Measure of sizes: q^a for s_p and q^b for s_{np}

Simulation Setup

Table 1: Scenarios for probability and nonprobability sample selection

	Level of Non-**A**dditivity	Level of Non-**L**inearity	Level of Misspecifying **V**ariables
Scenario A	0 (additive)	0 (linear)	0 (none)
Scenario B	0 (additive)	1 (moderate)	0 (none)
Scenario C	1 (moderate)	0 (linear)	0 (none)
Scenario D	1 (moderate)	1 (moderate)	0 (none)
Scenario E	2 (strong)	2 (strong)	0 (none)
Scenario F	0 (additive)	0 (linear)	1 (moderate†)
Scenario G	0 (additive)	0 (linear)	2 (strong††)

†: Mis-specifying $x_1^*, \ldots, x_4^*$ by $x_1, \ldots, x_4$.

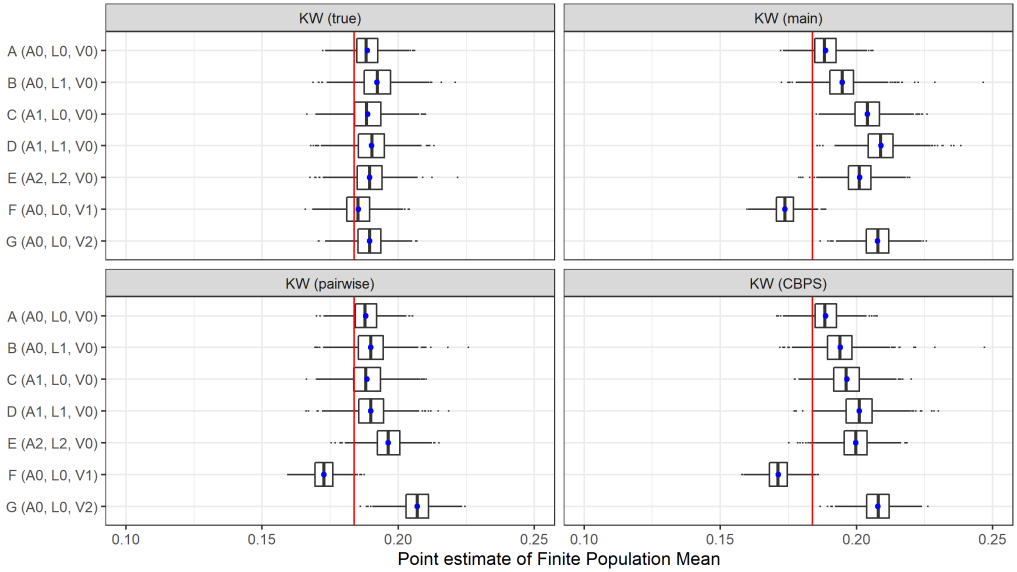
††: Mis-specifying $x_1^{**}, \ldots, x_4^{**}$ by $x_1, \ldots, x_4$.

Simulation Setup

- **Weighting methods**
  - KW-true: True propensity model
  - KW-main: Logistic regression, main effects
  - KW-pairwise: Logistic regression, main effects and interactions
  - KW-CBPS: Covariate balancing propensity scores (Imai and Ratkovic, 2014)
  - KW-ML: KW-MOB, KW-CRF, KW-GBM, KW-MBoost
  - IPSW-main, IPSW-pairwise
- **Analytical Statistic**
  Estimate of prevalence $\bar{Y}$, true $\bar{Y} = 0.184$
- **Criteria**
  Relative bias, empirical variance, mean squared error (MSE), CV of weights, averaged SMD

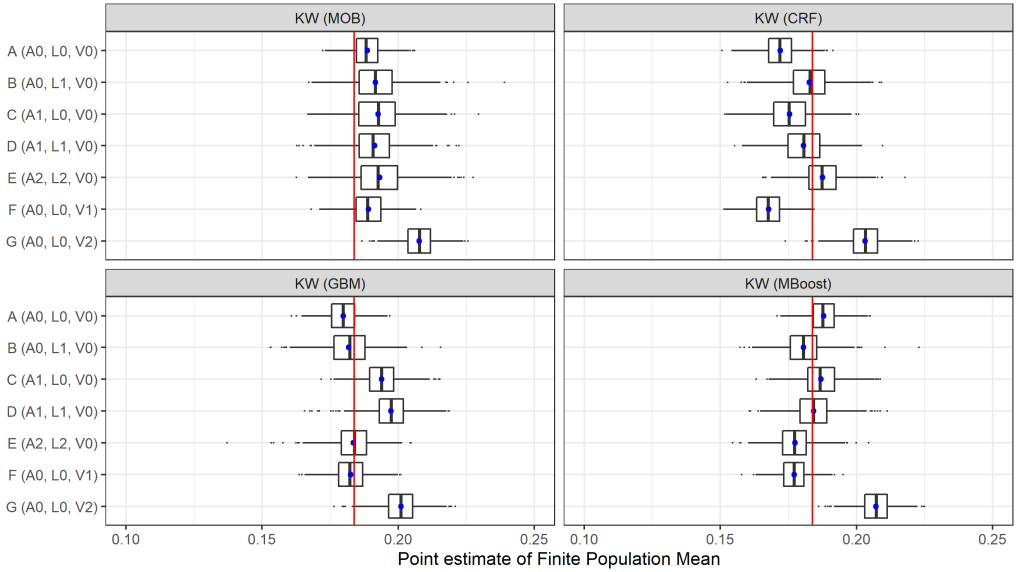
Simulation Results: KW-LG, KW-CBPS

Figure 1: Boxplots of estimates of finite population prevalence from 1,000 simulation runs



Simulation Results: KW-ML

Figure 2: Boxplots of estimates of finite population prevalence from 1,000 simulation runs



Simulation Results

Table 2: Mean Squared Error ($\times 10^4$) of estimates from 1,000 simulation runs

		KW-LG		KW-		KW-ML		
Scenario	True	Main	Pairwise	CBPS	MOB	CRF	GBM	MBoost
A (A0, L0, V0)	0.58	0.58	0.53	0.61	0.57	1.78	**0.52**	**0.52**
B (A0, L1, V0)	1.25	1.75	0.88	1.62	1.50	0.73	0.75	**0.70**
C (A1, L0, V0)	0.75	4.58	0.73	2.08	1.66	1.40	1.49	**0.63**
D (A1, L1, V0)	0.92	6.89	0.90	3.54	1.30	0.85	2.32	**0.57**
E (A2, L2, V0)	0.78	3.40	1.95	2.95	1.90	0.71	**0.54**	0.85
F (A0, L0, V1)	0.42	1.26	1.48	1.81	0.69	2.95	**0.42**	0.75
G (A0, L0, V2)	0.69	6.13	5.73	6.14	6.07	4.20	**3.36**	5.86
Average	0.77	3.51	1.74	2.68	1.95	1.80	**1.34**	1.41

Discussion

Summary

- Proposed approach combines advantages of KW and ML methods
- KW-ML can improve over KW-LG (w. main effects and interactions)
- KW-GBM and KW-MBoost perform best among KW-ML methods

Limitations and Extensions

- KW-ML requires an adequate set of adjustment variables
- ML models are sensitive to hyperparameter settings
- More ML methods could be explored

Thanks!

R package available at
`https://github.com/chkern/KWML`

References

Buehlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science*, 22(4).

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76:243–263.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.

Wang, L., Graubard, B. I., Katki, H. A., and Li, Y. (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, `https://doi.org/10.1111/rssa.12564`.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. `https://arxiv.org/abs/2001.03259`.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Appendix

Table 3: Tuning parameter settings

Method	Hyperparameter	Values
MOB	maxdepth	$[2, \ldots, 10]$
	alpha	0.05
	minsplit	NULL
CRF	mincriterion	$[0.99, 0.95, 0.9]$
	mtry	sqrt(p)
	ntree	100
GBM	interaction.depth	$[1, \ldots, 5]$
	n.trees	$[100, 250, 500, 1000, 2000]$
	n.minobsinnode	10
	shrinkage	0.05
	bag.fraction	1
MBoost	mstop	$[50, 100, 250, 500]$
	nu	0.1

Appendix

Scenario A: A model with additivity and linearity

$q = \exp(\boldsymbol{\beta}^T \mathbf{X})$

Scenario B: A model with additivity and moderate non-linearity

$q = \exp(\boldsymbol{\beta}^T \mathbf{X} + \beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2)$

Scenario C: A model with moderate non-additivity and linearity

$q = \exp(\boldsymbol{\beta}^T \mathbf{X} + \beta_1^* X_1 X_2 + \beta_2^* X_2 X_3 + \beta_3^* X_3 X_4 + \beta_4^* X_4 X_5 + \beta_5^* X_5 X_6 + \beta_1^* X_1 X_3 + \beta_2^* X_2 X_4 + \beta_3^* X_3 X_5 + \beta_6^* X_4 X_6 + \beta_5^* X_5 X_7)$

Scenario D: A model with moderate non-additivity and moderate non-linearity

$q = \exp(\boldsymbol{\beta}^T \mathbf{X} + \beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2 + \beta_1^* X_1 X_2 + \beta_2^* X_2 X_3 + \beta_3^* X_3 X_4 + \beta_4^* X_4 X_5 + \beta_5^* X_5 X_6 + \beta_1^* X_1 X_3 + \beta_2^* X_2 X_4 + \beta_3^* X_3 X_5 + \beta_6^* X_4 X_6 + \beta_5^* X_5 X_7)$

Scenario E: A model with substantial non-additivity and substantial non-linearity

$q = \exp(\boldsymbol{\beta}^T \mathbf{X} + \beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2 + \beta_1^* X_1 X_3 + \beta_2^* X_2 X_4 + \beta_3^* X_3 X_5 + \beta_4^* X_4 X_5 + \beta_5^* X_5 X_6 + \beta_3^{**} X_3^2 X_5^2 + \beta_1^{**} X_1 X_2 X_3 + \beta_4^{**} X_4 X_5 X_7)$

Scenario F: A model with categorized confounder variables

$q = \exp(\sum_{i=1}^{4} \boldsymbol{\alpha}_i \mathbf{X}_i^*)$

Scenario G: A model with confounder variables

$q = \exp(\beta_0 + \beta_1 X_1^{**} + \beta_2 X_2^{**} + \beta_3 X_3^{**} + \beta_4 X_4^{**})$

Appendix

Table 4: Relative bias (%) of estimates from 1,000 simulation runs

		KW-LG		KW-		KW-ML		
Scenario	True	Main	Pairwise	CBPS	MOB	CRF	GBM	MBoost
A (A0, L0, V0)	2.64	2.64	2.32	2.68	2.60	-6.43	-2.15	2.24
B (A0, L1, V0)	4.51	5.99	3.30	5.62	4.27	-0.62	-1.08	-1.79
C (A1, L0, V0)	2.68	11.03	2.55	6.89	4.80	-4.53	5.50	1.75
D (A1, L1, V0)	3.43	13.72	3.29	9.39	4.05	-1.76	7.32	0.22
E (A2, L2, V0)	3.06	9.44	6.82	8.67	5.09	1.99	-0.16	-3.46
F (A0, L0, V1)	0.87	-5.46	-6.05	-6.80	2.83	-8.74	-0.71	-3.64
G (A0, L0, V2)	3.04	13.05	12.59	13.07	12.98	10.54	9.32	12.74
Average	2.89	8.76	5.27	7.59	5.23	4.94	3.75	3.69

Appendix

Table 5: CV of weights from 1,000 simulation runs

		KW-LG		KW-		KW-ML		
Scenario	True	Main	Pairwise	CBPS	MOB	CRF	GBM	MBoost
A (A0, L0, V0)	0.79	0.79	0.80	0.79	0.79	1.01	0.88	0.80
B (A0, L1, V0)	1.06	0.69	0.83	0.70	1.03	1.05	1.17	0.99
C (A1, L0, V0)	0.80	0.62	0.81	0.63	0.80	0.91	0.68	0.81
D (A1, L1, V0)	0.81	0.56	0.70	0.58	0.83	0.81	0.79	0.77
E (A2, L2, V0)	0.94	0.53	0.64	0.53	0.85	0.82	0.96	0.78
F (A0, L0, V1)	0.83	0.35	0.38	0.35	0.73	0.77	0.89	0.48
G (A0, L0, V2)	0.87	0.63	0.64	0.63	0.63	0.72	0.69	0.64
Average	0.87	0.60	0.68	0.60	0.81	0.87	0.87	0.75

Appendix

Figure 3: Boxplots of estimates of finite population prevalence from 1,000 simulation runs