Computationally Efficient Bayesian Unit-Level Random Neural Network Modeling of Survey Data under Informative Sampling for Small Area Estimation

Paul A. Parker

Department of Statistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA 95064

E-mail: paulparker@ucsc.edu

Scott H. Holan

Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100

E-mail: holans@missouri.edu; scott.holan@census.gov

Summary. The topic of deep learning has seen a surge of interest in recent years both within and outside of the field of Statistics. Neural networks leverage both nonlinearity and interaction effects to provide superior predictions in many cases when compared to linear or generalized linear models. However, one of the main challenges with these approaches is quantification of uncertainty. The use of random weight models, such as the popularized "Extreme Learning Machine," offer a potential solution in this regard. In addition to uncertainty quantification, these models are extremely computationally efficient as they do not require optimization through stochastic gradient descent, which is what is typically done for deep learning. We show how the use of random weights in a neural network model can fit into a likelihood based framework to allow for uncertainty quantification of the model parameters and any desired estimates. Furthermore, we show how this approach can be used to account for informative sampling of survey data through the use of a pseudo-likelihood. We illustrate the effectiveness of this methodology through simulation and with a real survey data application involving American National Election Studies data.

Keywords: ANES; Pólya-Gamma; Pseudo-likelihood; Random weights; Text analysis

1. Introduction

There has recently been a strong interest in collecting novel types of data along with sample surveys. These data types can range from functional data such as the physical activity monitor data contained within the National Health and Nutrition Examination Survey (NHANES; Schuna et al., 2013), free response text data such as those within the American National Election Studies (ANES) surveys (DeBell, 2013), and even complex data from web-based surveys such as mouse movements (Horwitz et al., 2020). These information rich data sources could prove to be useful as covariates, however the rapid

and increased interest in these data types within a survey context has resulted in lagging development of corresponding methodology.

These complex covariates are generally collected at the unit level, and thus, necessitate the need for unit-level modeling strategies. One of the challenges associated with unit-level modeling is accounting for the sampling design under informative sampling mechanisms. A variety of strategies exist for this problem, including the use of pseudo-likelihood modeling (Skinner, 1989; Binder, 1983). In many applications, such as small area estimation (SAE), predictions can be made for every unit in the population and then aggregated as necessary to construct any desired estimates. An exceedingly common solution is that of regression and poststratification, whereby the population is segmented via a set of categorical covariates, and units that are associated with identical covariates are assumed to be independent and identically distributed (Park et al., 2004). In particular, Park et al. (2004) state that they envision this methodology being used for public opinion estimates at the state level. The categorical covariates required for poststratification are generally known for the entire population. This is in contrast to the complex covariates that we consider, which are generally only known for the sampled units.

Another major concern for unit-level models, particularly in a Bayesian setting, is that of computational efficiency. Dependent data models typically rely on Gaussian prior distributions for model parameters (Bradley et al., 2015); however, most survey variables tend to be non-Gaussian, leading to non-conjugate conditional distributions that can be difficult to sample from. This problem is addressed by Parker et al. (2020b) and Parker et al. (2020a) for count data and Binomial data, although they do not consider the further problem of modeling complex data types in a computationally efficient manner.

Herein, we develop a computationally efficient method to model these complex covariates while accounting for informative sampling. Although other applications of this model are possible, we illustrate this method through the problem of SAE. We utilize a neural network structure to handle nonlinear modeling of the complex covariate data, while employing a Bayesian pseudo-likelihood model structure in order to measure uncertainty around our estimates while accounting for informative sampling. The benefits of such models resides in the improvement of the precision of the estimates being produced relative to a design-based estimator and other proposed models. Specifically, as government budgets are often flat or declining, an improvement in the precision of the small area estimates may allow resources to be used to increase the sample size across various under-sampled geographies. In other words, a reduction in MSE may allow for efficient sample allocation, potentially producing better levels of precision while maintaining a fixed or lower cost.

The remainder of this paper is outlined as follows. In Section 2 we introduce the necessary methodological background as well as our model. Section 3 considers an empirical simulation study relying on the use of ANES data. We follow with a full data analysis using the same ANES data in Section 4. Finally, we provide discussion and concluding remarks in Section 5.

2. Methodology

The method that we develop tackles three problems simultaneously. The first issue is that nonlinear modeling is required for the use of complex covariates, without becoming computationally prohibitive. For example, neural network structures typically involve an extremely high-dimensional parameter space. A full Bayesian treatment of these types of models can often require too many computational resources to be fit in a reasonable amount of time. The second problem is that we must account for informative sampling in our model to avoid producing any unnecessary statistical bias. Finally, we require a model that can handle Binomial data types while still accounting for all of the underlying dependencies associated with the data. We explore each of these three problems, and then present our methodology.

2.1. Extreme Learning Machines

The extreme learning machine (ELM) is a type of single layer feed-forward neural network (FNN), introduced by Huang et al. (2006). The key difference between the ELM and traditional FNNs is that the ELM uses random weights (i.e., parameters) drawn from some distribution for the hidden layer nodes. As with other FNNs, the ELM can be used for both regression and classification problems, as well as other types of problems (e.g., unsupervised learning), while allowing for much more flexibility in the mean function than linear or generalized linear models.

The basic ELM considers a nonlinear transformation of the covariate data (features),

$$f(\mathbf{x}_i) = \sum_{j=1}^{N} g_j(\mathbf{a}_j'\mathbf{x}_i + b_j)\beta_j,\ i = 1, \ldots, n$$

where $\mathbf{x}_i$ represents the p-dimensional covariate information for unit i in the sample with size n. The value N represents the number of nodes, where each node considers a unique nonlinear transformation of the data. Each node first applies a linear transformation, with parameters $\mathbf{a}_j = [a_{j1}, \ldots, a_{jp}]'$ and b_j. This linear transformation is followed by a nonlinear transformation, denoted by the function $g_j(\cdot)$, often called an activation function. This is specified a priori, and may be any piecewise continuous function (Huang et al., 2015), but in practice usually consists of a sigmoid function. The output, $f(\mathbf{x}_i)$ is then calculated as a weighted sum of each individual node output, where the weights are denoted by the N dimensional vector $\boldsymbol{\beta}$. Intuitively, this is similar to basis function approaches, in the sense that both techniques may use nonlinear transformations of the input data to construct a flexible nonlinear function for the mean. However, one key difference is that in the case of the ELM, these nonlinear transformations do not need to be selected as they are randomly generated.

The key to ELMs is that for each node, $j = 1, \ldots, N$, the values of $\mathbf{a}_j$ and b_j are randomly drawn. Thus, the only set of parameters that need to be learned or estimated is $\boldsymbol{\beta}$. Common distributional choices for these randomly selected parameters are Normal(0,1) and Uniform(-1,1). Although these hidden layer parameters are only randomly drawn a single time, typically many nodes are used to allow for flexible representation of the function $f(\mathbf{x}_i)$.

More generally, the ELM can be written

$$\boldsymbol{\mu}_i = g_o(\mathbf{B}\mathbf{g}_i)$$
$$\mathbf{g}_i = g(\mathbf{A}\mathbf{x}_i)$$

where the $p \times 1$ dimensional vector $\mathbf{x}_i$ now contains an intercept and the l-dimensional vector of means, $\boldsymbol{\mu}_i$, can now incorporate multivariate responses. Also, $\mathbf{A}$ is an $N \times (p+1)$ dimensional matrix of hidden layer weights, and $\mathbf{B}$ is an $l \times N$ dimensional vector of output weights. The hidden layer activation function is denoted $g(\cdot)$ and the output layer activation function is denoted $g_o(\cdot)$, which will be the inverse of the canonical link function in the case of a GLM.

The above view of the ELM is similar to the generalized linear model and highlights an important strength of ELM. Because the hidden layer parameters are randomly chosen and not estimated, we may view the hidden layer transformations as fixed once these parameters have been generated. This is similar to regression with basis expansions as is often seen when using generalized additive models (GAMs). A key difference with ELM compared to GAMs however, is that the entire vector $\mathbf{x}_i$ is used within each hidden node, which allows for interaction effects. Viewing the random transformations as fixed allows us to extend the entire class of generalized linear models to incorporate nonlinear behavior. Furthermore, pseudo-likelihood approaches may be used in conjunction with the ELM in order to account for informative sampling.

The ELM can be considered a type of reservoir computing (an approach where weights are randomly generated). Random projection is another type of reservoir computing, often used for dimension reduction (Bingham and Mannila, 2001). Under random projection, the original $n \times p$ data matrix $\mathbf{X}$ is "projected" onto a L-dimensional subspace, $\mathbf{X}^* = \mathbf{X}\mathbf{R}$, where $\mathbf{R}$ is a randomly generated $p \times L$ matrix. This is not technically a projection, as the matrix $\mathbf{R}$ is not orthogonal, but due to the random nature of the matrix, it tends to be "approximately" orthogonal. Note that the randomly generated projection matrix could be orthogonalized, but this is not always done in practice due to the computational cost.

Random projection could be used in the context of regression, similar to Principal Components Regression. In this light, it may be seen as a special case of the ELM, where $g_j(\cdot)$ is equal to the identity function for all j. In other words, random projection uses randomly generated parameters for the hidden node linear transformation component, but does not introduce a nonlinear component.

Another common type of reservoir computing is known as the Echo State Network or ESN (Prokhorov, 2005). This is a type of recurrent neural network, where the hidden weights are randomly generated. Recently, the ESN has been used in likelihood-based frameworks for spatio-temporal forecasting (McDermott and Wikle, 2017). The ESN may also be used within a Bayesian model structure in order to give uncertainty quantification (McDermott and Wikle, 2019).

Bayesian methods have been considered in the ELM community as well, beginning with Soria-Olivas et al. (2011). They consider ridge regression fit with an Empirical Bayes procedure. This achieves both regularization as well as uncertainty quantification for the output layer weights and data model variance. They also show that this method tends to give better out of sample predictions compared to the traditional ELM. Chen

et al. (2016) use a variational Bayes approach to fit a Bayesian ELM. By doing so, it is possible to reduce the computational burden of the Bayesian ELM substantially.

2.2. Pseudo-likelihood based SAE

When fitting models with unit-level survey data, it may be the case that there exists a dependence relationship between the unit probabilities of selection, and the response values. This is termed *informative sampling*, and if this relationship is not accounted for, any estimates may be biased (Pfeffermann and Sverchkov, 2007). A thorough review of the modern approaches to handling informative sampling is given by Parker et al. (2019). One popular approach to this problem is the use of a pseudo-likelihood (PL), introduced by Skinner (1989) and Binder (1983). The general idea is to use the reported survey weights to exponentially re-weight the likelihood contribution of each survey unit. Thus, the PL is written as

$$\prod_{i \in \mathcal{S}} f(y_i \mid \boldsymbol{\theta})^{w_i}, \tag{1}$$

where y_i is the response value and w_i is the survey weight for unit i in the sample $\mathcal{S}$. The PL can be maximized in order to make frequentist inference, however Savitsky and Toth (2016) show that in a Bayesian setting, the pseudo-posterior distribution,

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{i \in \mathcal{S}} f(y_i|\boldsymbol{\theta})^{\tilde{w}_i} \right\} \pi(\boldsymbol{\theta}),$$

converges to the population generating distribution, justifying the use of a Bayesian PL for inference on nonsampled units. In this case, $\tilde{w}_i$ represents the survey weights after scaling to sum to the sample size in order to give proper uncertainty quantification. We note that in cases of extreme weights, care must be taken when fitting models through the use of a pseudo-likelihood.

2.3. Logistic Models

Many survey data variables tend to be non-Gaussian at the unit level. For example, the American Community Survey contains a binary indicator of health insurance status as well as many categorical variables such as primary language spoken. In regression frameworks with non-Gaussian responses and Normal prior distributions on any regression parameters, non-conjugate full conditional distributions arise. This may lead to the need for Metropolis steps within the MCMC routine that can be prohibitively difficult to tune.

For the case of logistic models (Binomial, Negative Binomial and Multinomial responses), Polson et al. (2013) introduce a data augmentation scheme that gives rise to conjugate full-conditional distributions. This strategy relies on the use of Pólya-Gamma (PG) random variables. Specifically, they rewrite the Binomial likelihood as,

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) d\omega, \tag{2}$$

where $\kappa = a - b/2$ and $p(\omega)$ is a $\text{PG}(b, 0)$ density. They also show that $p(\omega|\psi) \sim \text{PG}(b, \psi)$. For the linear predictor $\psi = \boldsymbol{x}'\boldsymbol{\beta}$, if we use a Gaussian prior on $\boldsymbol{\beta}$, the full conditional distribution for $\boldsymbol{\beta}$ will also be Gaussian. Furthermore, Parker et al. (2020a) show that under a PL setup, conjugacy is still retained. They develop both a Gibbs sampling algorithm as well as a variational Bayes algorithm for PL-based mixed effects models with Binomial data. In addition to this, they use the stick-breaking representation of the Multinomial distribution in order to extend the algorithms to categorical responses. More specifically, Linderman et al. (2015) show that the Multinomial distribution may be written as a product of independent Binomial distributions,

$$\text{Multinomial}(\boldsymbol{Z}|n, \boldsymbol{p}) = \prod_{k=1}^{K-1} \text{Bin}(Z_k|n_k, \tilde{p}_k), \tag{3}$$

where

$$n_k = n - \sum_{j<k} Z_j, \;\; \tilde{p}_k = \frac{p_k}{1 - \sum_{j<k} p_j}, \;\; k = 2, \ldots, K. \tag{4}$$

Under this view of Multinomial data, $K - 1$ Binomial data models may be fit independently while still accounting for the dependence between categories through the stick-breaking counts and probabilities.

2.4. Proposed Model

We now introduce a Bayesian unit-level random neural network model for informative sampling (BURN). Here, we focus on the case of Binomial and Multinomial data, but note that this approach would be applicable to Gaussian data as well. The Binomial model is written,

$$\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta}, \boldsymbol{\eta} &\propto \prod_{i \in S} \left\{ \text{Bin}\left(Z_i|n_i, p_i\right)^{\tilde{w}_i} \right\} \\
\text{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{g}_i'\boldsymbol{\eta} \\
\boldsymbol{g_i'} &= \frac{1}{1 + e^{-\boldsymbol{A}\boldsymbol{\psi_i}}} \\
\boldsymbol{\eta}|\sigma_\eta^2 &\sim \text{N}_h(\boldsymbol{0}_h, \sigma_\eta^2 \boldsymbol{I}_h) \\
\boldsymbol{\beta} &\sim \text{N}_p(\boldsymbol{0}_p, \sigma_\beta^2 \boldsymbol{I}_p) \\
\sigma_\eta^2 &\sim \text{IG}(a, b) \\
a, b, \sigma_\beta^2 &> 0,
\end{aligned} \tag{5}$$

where Z_i is the Binomial response for unit i in the sample with size n_i and probability p_i. Typically surveys contain Bernoulli data, so for our purposes, $n_i = 1$ for all i. We are using a pseudo-likelihood approach at this data stage of the model in order to account for informative sampling. Within the pseudo-likelihood, we use the scaled survey weights, $\tilde{w}_i$, such that the weight sum to the sample size. The length p vector $\boldsymbol{x}_i$ contains any covariates that do not require nonlinear modeling. The length h vector $\boldsymbol{g}_i$ contains the ELM hidden layer values for unit i. Finally, the length r vector $\boldsymbol{\psi}_i$ contains

the complex covariates that are used within the ELM framework. Note that the $h \times r$ matrix $\boldsymbol{A}$ is sampled and considered fixed before model fitting, so that $\boldsymbol{g}_i$ is determined *a priori*. This allows for the use of generalized linear model fitting procedures rather than custom techniques. Specifically, we use the variational Bayes procedure from Parker et al. (2020a) for all model fitting. This procedure results in an approximate posterior distribution that can be used to draw samples of $\boldsymbol{\zeta}' = (\boldsymbol{\beta}', \boldsymbol{\eta}')$. Specifically, samples of $\boldsymbol{\zeta}$ can be drawn independently from a $\mathrm{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are the outputs of the variational Bayes algorithm. For further details on implementation of this algorithm, see Parker et al. (2020a).

After drawing samples $\boldsymbol{\zeta}^{(r)},\ r = 1, \ldots, R$, we can draw from the predictive distribution to generate $Z_i^{(r)}$ for every unit in the population. Next, we can aggregate units to get a population mean estimate for state s,

$$p_s^{(r)} = \frac{\sum_{i \in c_s} Z_i^{(r)}}{N_s},$$

where c_s is the collection of individuals that belong to state s, and N_s is the population size in the state. Finally, we can use the posterior mean,

$$\hat{p}_s = \frac{1}{R} \sum_{r=1}^{R} p_s^{(r)}$$

as our point estimate of the state population proportion. Similarly, we use the posterior standard deviation as measure of uncertainty around our estimate.

For our purposes, we let $\boldsymbol{x}_i$ consist of any poststratification variables as well as spatial basis functions. These values will typically be known for the full population. The complex covariates contained within $\boldsymbol{\psi}_i$ are not usually known for the full population, and thus must be imputed in order to generate the population posterior predictive distribution necessary for SAE. Our approach revolves around the idea of assigning the observed covariate vectors to all the unobserved population units. Under a simple random sample, a reasonable assumption may be that the observed complex covariates are uniformly distributed throughout the population. However, under an informative sample, the observed covariate vectors are sampled with unequal probability, which should be accounted for when distributing the observed vectors to the population.

For our imputation model, we create imputation cells, similar to poststratification cells, with J total cells. A population unit within imputation cell $j,\ j = 1, \ldots, J$ may only be assigned a vector from the set of observed vectors $(\boldsymbol{\psi}_{j1}, \ldots, \boldsymbol{\psi}_{jn_j})$, where n_j is the sample size within cell j. Rather than sampling from this set with equal probability, we sample with probability proportional to the reported sampling probability, or inversely proportional to the reported sample weight, to account for the original survey sampling scheme. Thus, for population unit i in cell j, the vector of complex covariates is sampled from $(\boldsymbol{\psi}_{j1}, \ldots, \boldsymbol{\psi}_{jn_j})$ with probability proportional to $(1/w_{j1}, \ldots, 1/w_{jn_j})$. This imputation can be done a single time, however we opt to create a separate imputed dataset for each sample from our model based posterior distribution in order to account for the imputation uncertainty within our posterior predictive distribution. For this work, we let the $J = 48$ corresponding to the states where area level estimates are made, however

other choices of imputation cells could be explored. One limitation to this approach is that all imputation cells must have at least one sample unit in order to distribute the sample values within the cell to the population. Note that this imputation procedure is only used for the complex covariates $\boldsymbol{\psi}_i$ and not for the regular covariates $\boldsymbol{x}_i$. We make the typical assumption that $\boldsymbol{x}_i$ is known for the entire population, usually via cell counts.

3. Empirical Simulation Study

To test our methodology, we consider data from the 2012 American National Election Studies (ANES) survey. Specifically, we use the Time Series Study data which measures various responses both pre and post election. We only consider the post election data, which contains a number of free response questions. Our goal is to use the free responses from the question "What are the most important problems facing this country?" in order to improve small area estimates of public opinion.

Figure 1 shows a word cloud of the most frequently occurring words within the ANES data. In many cases, the words will have little meaning on their own, but instead have meaning when paired with other words. For example, the word *security* on its own does not provide much insight, but when paired with either *economic* or *national*, it may indicate the primary concern of the respondent. This suggests the need for a model that can take into account many possible interactions between words rather than considering words individually.

Fig. 1. A word cloud of the top words contained within the ANES data. The size of each word represents the frequency of appearance.

One public opinion question involved in the survey considers whether respondents approve or dissaprove of the way president Barack Obama was handling the job of president at the time. For this simulation, we estimate the proportion of the population within each state that approve. In other words, we consider a binary response. We treat the original ANES sample as our population and take a subsample with probability to proportional size sampling using the Poisson method (Brewer et al., 1984) with an expected sample size of 1,000. For our size variable, we use the original survey weight plus 0.7 if the true response is "approve" in order to explicitly generate an informative sample. We fit the BURN model using $\boldsymbol{\psi}_i = (I(t_{i1}), \ldots, I(t_{i1000}))$ as the input into the ELM, where $I(t_{ij})$ indicates whether or not the jth most frequently occurring word/term appeared in the free response of unit i. For the linear component, $\boldsymbol{x}_i$, we use indicators

of Hispanic ethnicity and gender as poststratification variables, as well as a set of spatial basis functions. Similar to Hughes and Haran (2013), we use the first 25 eigenvectors of the state adjacency matrix as our basis functions, although other basis functions could be substituted here. The state adjacency matrix is defined as the row normalized matrix $\boldsymbol{M}_{S\times S} = (m_{s_1 s_2})$ for $s_1, s_2 = 1, \ldots, S$, where $m_{s_1 s_2}$ is one if states s_1 and s_2 share a border and zero otherwise. Note that we assume a state cannot share a border with itself, thus the diagonal entries of this matrix are zero. These basis functions are at the state level, thus, for any given individual, we use the basis functions from their corresponding state. Hughes and Haran (2013) found that in many scenarios as little as 10% of the available basis functions would suffice. Our choice of 25 basis functions is roughly 50% of the available functions, and thus, intended to be a conservative choice.

We generate our hidden weights in the matrix $\boldsymbol{A}$ from the standard Normal distribution, and then randomly set 10% equal to zero and we use a vague prior distribution by setting $a = b = 0.5$ and $\sigma^2_\beta = 1000$. We have found that in general the model is not overly sensitive to the choice of distribution for the random weights (i.e. the generating distribution for the matrix $\boldsymbol{A}$), however, depending on the application, it may be desirable to select the distribution through cross-validation. Lastly, we set the number of hidden nodes $h = 240$. We have found that in practice, as long as the number of hidden nodes is sufficiently large enough, there is little additional benefit to increasing the number of nodes. A small scale sensitivity analysis confirmed that there is little benefit gained by further increasing the number of nodes above 240 in this case.

We compare estimates under our BURN model when imputation of the text data for nonsampled individuals is required (BURN-IT) as well as when the text data is known for the entire population (BURN-KT). Although the latter scenario would not typically be applicable in practice, this still provides an interesting comparison, especially in terms of the uncertainty introduced through the imputation procedure. We also compare to a model that does not use the text data or the ELM component, which we denote Pseudo-likelihood logistic regression (PLLR),

$$\begin{aligned}
\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S}\left\{\mathrm{Bin}\left(Z_i|n_i,p_i\right)^{\widetilde{w}_i}\right\}\\
\mathrm{logit}(p_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}\\
\boldsymbol{\eta}|\sigma^2_\eta &\sim \mathrm{N}_h(\boldsymbol{0}_h, \sigma^2_\eta \boldsymbol{I}_h)\\
\boldsymbol{\beta} &\sim \mathrm{N}_p(\boldsymbol{0}_p, \sigma^2_\beta \boldsymbol{I}_p)\\
\sigma^2_\eta &\sim \mathrm{IG}(a,b)\\
a, b, \sigma^2_\beta &> 0,
\end{aligned}$$

as well as both weighted and unweighted direct estimators. All weighted direct estimates are constructed using the `mase` package in R (McConville et al., 2018). The two model based approaches both use post-stratification by sampling from the posterior distribution of the parameters, and then generating estimates for the response value of all units in the population. We repeat the sampling and model fitting procedure 100 times.

Table 1 shows the MSE and squared bias of each of the estimators through the simulation. These values are averaged across both the states as well as the simulated

Table 1. Empirical MSE and squared bias of the four estimators based on simulation results. Values are averaged across states and simulated datasets.

Estimator	MSE	Bias^2
BURN-KT	$\mathbf{1.81 \times 10^{-2}}$	7.42×10^{-3}
BURN-IT	1.95×10^{-2}	6.61×10^{-3}
PLLR	2.15×10^{-2}	8.14×10^{-3}
Direct	4.48×10^{-2}	$\mathbf{4.85 \times 10^{-3}}$
UW Direct	4.23×10^{-2}	1.39×10^{-2}

datasets. The first thing to note is that the much higher bias of the unweighted (UW) direct estimator when compared to the direct estimator indicates that the sampling was indeed informative. The two model based approaches were able to handle the informative sampling mechanism through the use of the pseudo-likelihood and improve the MSE dramatically compared to the direct estimates. The BURN-IT model was able to further improve upon the PLLR model by reducing MSE about 10% and reducing squared bias around 19%. It is clear in this case that the inclusion of nonlinear modeling of the text covariates results in better estimates. As one might expect, for the BURN-KT approach, when the population level text data is known, even further reductions in MSE are possible. There may be specific applications where this type of information is available.

In addition to a tabular summary of the simulation results, we compare the distribution of MSE by state for the Direct and BURN-IT estimators in Figure 2. States that fall below the dashed line indicate a reduction in MSE for the BURN-IT approach compared to the direct estimator. Note that only a few states fall above the line, and these are each very slight increases in MSE for states that already had low direct estimate MSE. In contrast, the vast majority of the states fall below the line, with many states seeing very substantial reduction in MSE through the use of the BURN-IT approach.

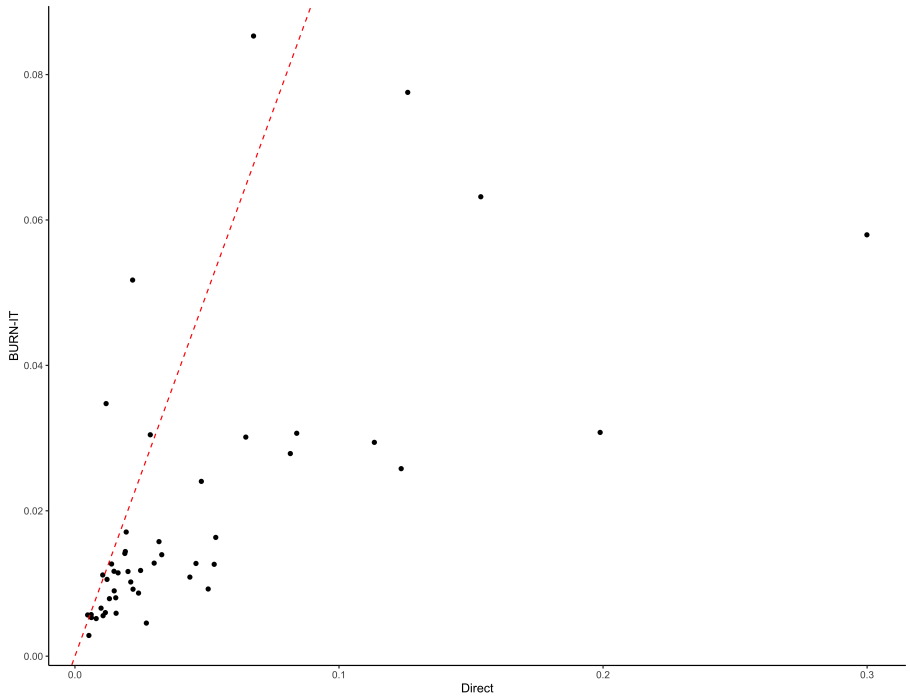

Fig. 2. Empirical MSE by state over simulated datasets for Direct vs. BURN-IT estimates.

Another interesting comparison is that of the standard errors resulting from each estimator. Figure 3 shows the average standard error by state for the model based approaches as well as the direct estimator. The direct estimator results in quite large standard errors, with each of the model based approaches being able to provide significant reductions. Perhaps unsurprisingly, the BURN-KT approach results in the lowest standard errors. The BURN-IT and PLLR approaches both results in strikingly similar standard errors. This indicates that the additional uncertainty resulting from the imputation procedure is offset by additional precision provided by the use of the text covariates.

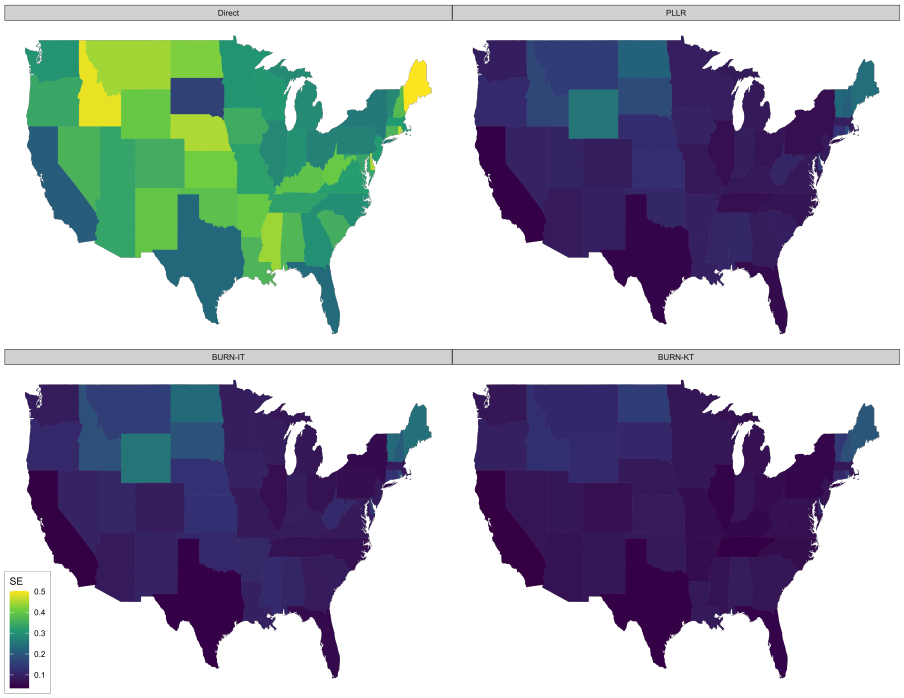

Fig. 3. Simulation based average standard error by state for the model based estimators and the direct estimator.

Finally, in Figure 4, we compare the average ratio of the direct estimate over the BURN-IT estimate to the average sample size by state. For this simulation setting, sample sizes ranged from 1 to 133. It is clear that as sample sizes are small, the direct estimate and the model based estimate can differ to a large degree. However, as the sample size increases, the two estimates become more similar. This coincides with typical behavior seen in other SAE models.

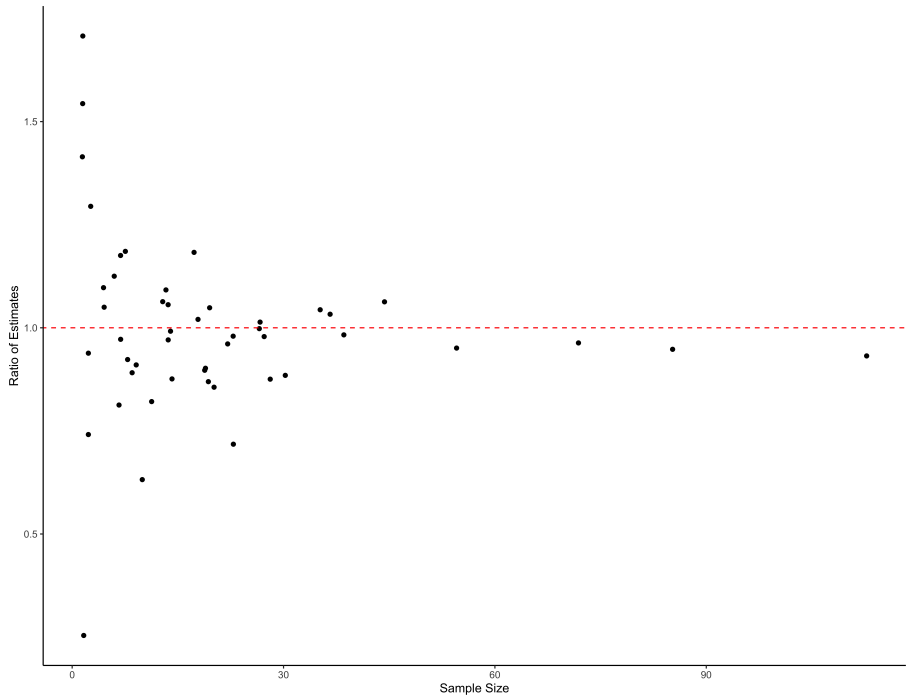

Fig. 4. Simulation based comparison of the average ratio of direct over BURN-IT estimates to average sample size by state.

One limitation of this simulation study is that by nature of subsampling from the original data, we are unable to obtain the same sample sizes as that of the original sample. In this case, for the original ANES sample, many states still had quite small sample sizes. Thus, we still expect SAE to be important for state level estimates. However, for other surveys with much larger sample sizes, SAE may be unnecessary for state level estimates. Nonetheless, larger surveys may allow for more granular estimates such as county or census tract, where SAE techniques would still be required.

4. ANES Data Analysis

In order to illustrate this methodology on a real application, we use the entire 2012 ANES dataset to create estimates under the BURN model. The total sample size for this dataset was 5,878, with state sample sizes ranging from 4 (Wyoming) to 742 (California). Similar to the simulation study considered in Section 3, we estimate the proportion of voting age residents within each state that approved of Barack Obama's job as president at the time the survey was taken. The covariates and hyperparameters were also the same as those considered in the simulation study.

We compare the estimates under the BURN model to the direct estimates in Figure 5. Note that many of the direct estimates fall towards the extremes due to limited sample sizes in some states. In contrast to this, the model based estimates fall in a narrower range due the effect of "borrowing information" across states. For the most

part the spatial pattern under the BURN model is as expected. The more traditionally conservative states in the South and towards the Dakotas tend to have lower estimates of approval than the coastal parts of the country. The Northwest portion of the country has a couple unexpected estimates, most notably Wyoming. The higher than expected estimate for Wyoming is likely due to the limited sample size pulling the estimate upward towards the national average, although an effort to find more suitable spatial basis functions could also aid improvement in this area.

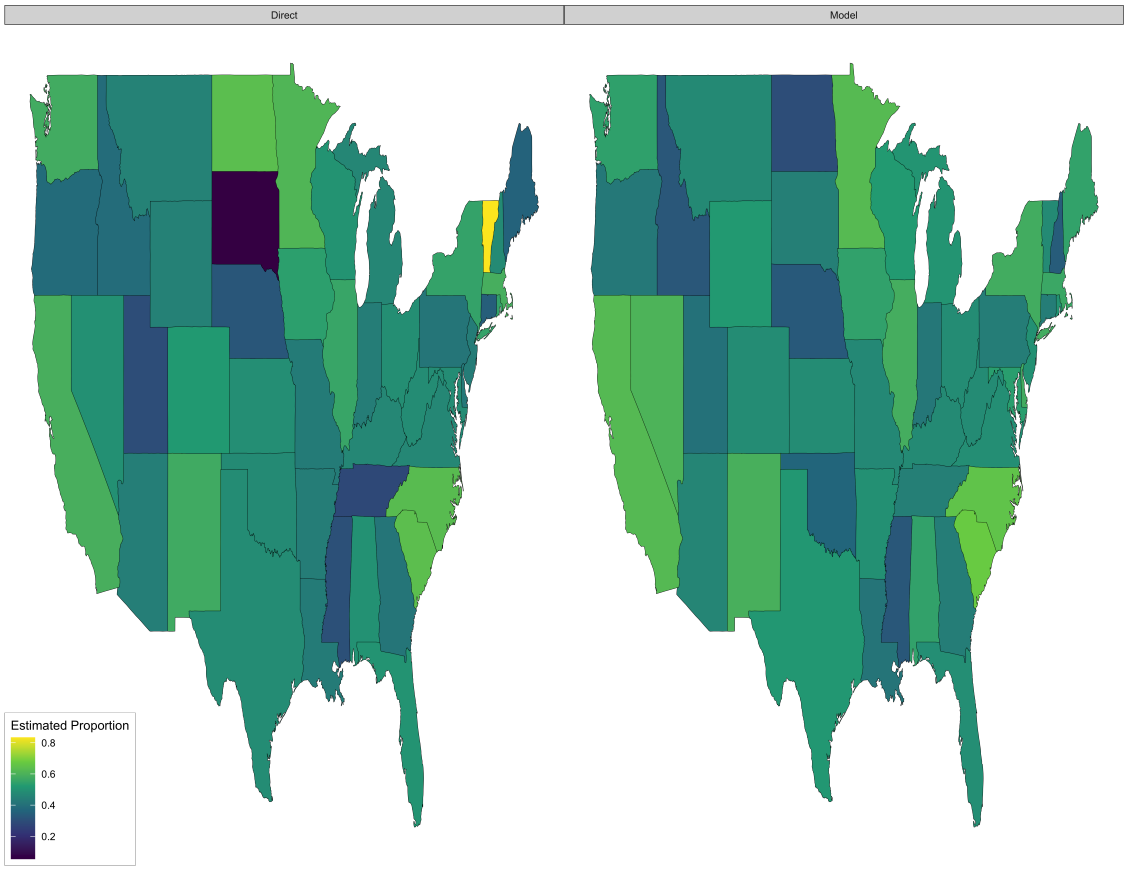

Fig. 5. Comparison of direct estimates to BURN model based estimates for the population proportion using 2012 ANES data.

In addition to the estimates of the population proportion, we also plot the direct and model based standard errors in Figure 6. In nearly every state the model based standard errors are lower than the standard errors for the direct estimate. This effect is most pronounced in states with small sample sizes such as Wyoming and North Dakota. Despite the uncertainty attributed to imputation of the text data for unobserved individuals, the BURN model is able to leverage information from across states through the text data, as well as spatial correlation, in order to reduce the uncertainty around these estimates.

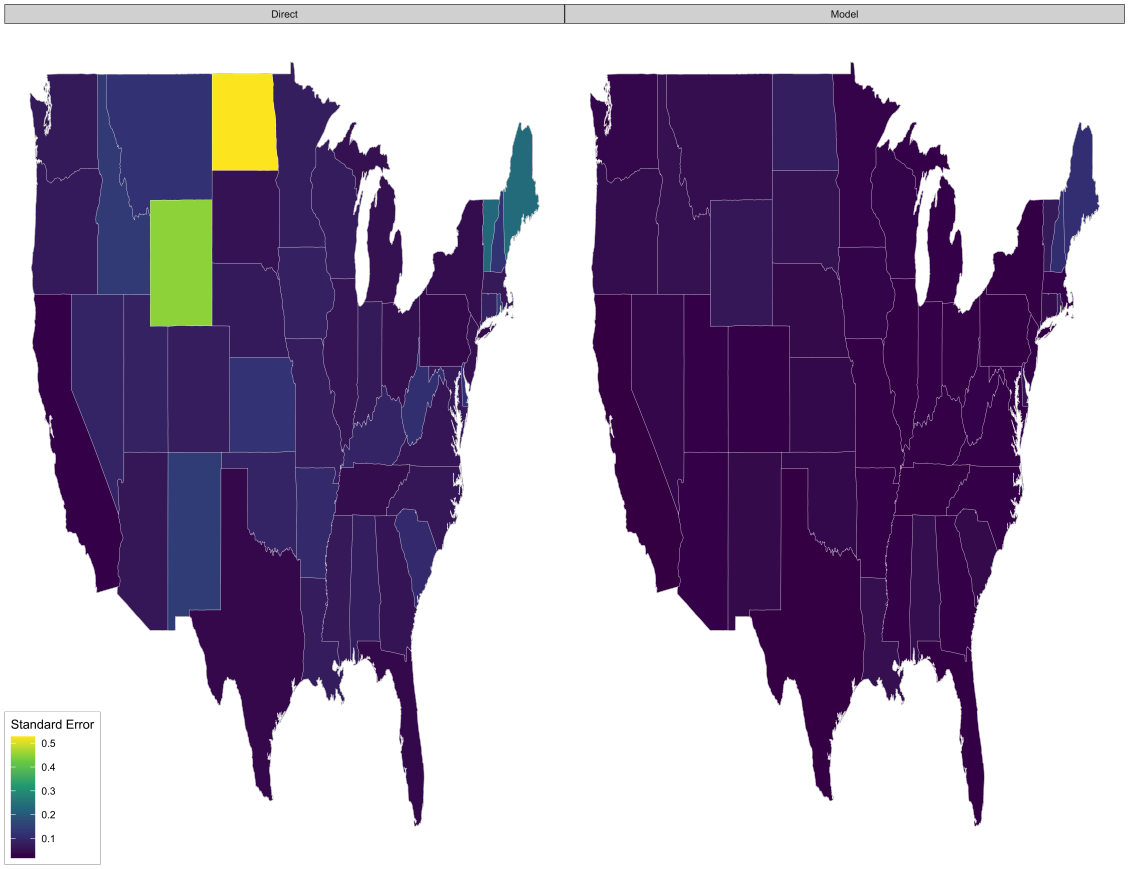

Fig. 6. Comparison of standard errors for direct estimates and BURN model based estimates using 2012 ANES data.

This example emphasizes how the BURN model can be used to construct better estimates of public opinion through the use of complex data types such as free response text. The ANES data set has a sufficient sample size to construct direct estimates at the national level, but many of the state sample sizes are extremely small, leading to very poor estimates. In this specific case, many of the state level direct estimates fall far away from from the national average. In contrast, the BURN model is able to smooth many of these extreme estimates by relying on a model that accounts for spatial dependence between survey respondents as well as the complex free response covariate information. Despite the small state sample sizes, this model is able to yield much more reasonable estimates, such as the general trend of lower approval in the South and higher approval on the East coast. These types of estimates could serve useful for targeting of campaign funds. For instance, in the key states of Florida and Michigan, the model based estimates indicate that Michigan may be more competitive. Furthermore, this example only considers estimates for a single public opinion question, but the ANES survey contains many more public opinion questions that may be of interest to others.

5. Discussion

In order to use complex unit level survey data as a covariate for SAE, we develop a couple important innovations to the PL unit-level model. The first innovation is the use of a neural network to model nonlinear functions of the complex covariates. This is achieved through the use of random weight methodologies, specifically the ELM. By taking this approach we are able to side-step the need for gradient descent techniques that are typically used in deep learning. In addition, this approach is highly computationally efficient, as it is linear in the parameters that are estimated. Further efficiency is gained through the use of a variational Bayes model fitting procedure.

The second innovation is the use of an imputation model that assigns sample covariates to population units while adjusting for the sampling design. Although this approach is relatively straightforward, modeling the population covariates explicitly could be very burdensome for high-dimensional data and this approach provides a path forward. The use of more advanced approaches to this imputation problem is subject to future work.

In addition to the novel modeling approaches explored here, this work highlights the need to collect more complex data types within surveys. Typical surveys include relatively simple data types such as binary and categorical measurements. However this work shows that more complex data types such as text or functional data may be used to improve the precision of survey based estimates. Currently, federal agencies spend significant resources converting open responses into simple categorical variables. Through the use of our proposed model, or extensions thereof, agencies may be able to rely less on these resources while simultaneously extracting more information from the raw data. Additionally, in practice, categories are application and unit (person, household, or establishment) specific. In a production setting it is necessary to provide something automated (i.e., the categories can not be hand picked for each tabulation). Our approach has this feature.

The ANES data considered in our examples was chosen in part because of its public availability, in order to limit the need for disclosure issues. However, our approach could be immediately (or with minor modifications) applicable to other complex survey datasets, such as NHANES physical activity monitor data, or web-based respondent tracking.

Acknowledgements

Support for this research through the Census Bureau Dissertation Fellowship program is gratefully acknowledged. This research was partially supported by the U.S. National Science Foundation (NSF) under NSF grant SES-1853096. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF or U.S. Census Bureau.

References

Binder, D. A. (1983). "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review*, 51, 3, 279–292.

Bingham, E. and Mannila, H. (2001). "Random projection in dimensionality reduction: applications to image and text data." In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 245–250. ACM.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, 9, 4, 1761–1791.

Brewer, K., Early, L., and Hanif, M. (1984). "Poisson, modified Poisson and collocated sampling." *Journal of Statistical Planning and Inference*, 10, 1, 15–30.

Chen, Y., Yang, J., Wang, C., and Park, D. (2016). "Variational Bayesian extreme learning machine." *Neural Computing and Applications*, 27, 1, 185–196.

DeBell, M. (2013). "Harder than it looks: Coding political knowledge on the ANES." *Political Analysis*, 393–406.

Horwitz, R., Brockhaus, S., Henninger, F., Kieslich, P. J., Schierholz, M., Keusch, F., and Kreuter, F. (2020). "Learning from mouse movements: Improving questionnaires and respondents' user experience through passive data collection." *Advances in questionnaire design, development, evaluation and testing*, 403–425.

Huang, G., Huang, G.-B., Song, S., and You, K. (2015). "Trends in extreme learning machines: A review." *Neural Networks*, 61, 32–48.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). "Extreme learning machine: theory and applications." *Neurocomputing*, 70, 1-3, 489–501.

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 1, 139–159.

Linderman, S., Johnson, M. J., and Adams, R. P. (2015). "Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation." In *Advances in Neural Information Processing Systems*, 3456–3464.

McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *mase: Model-Assisted Survey Estimation*.

McDermott, P. L. and Wikle, C. K. (2017). "An ensemble quadratic echo state network for non-linear spatio-temporal forecasting." *Stat*, 6, 1, 315–330.

— (2019). "Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data." *Entropy*, 21, 2, 184.

Park, D. K., Gelman, A., and Bafumi, J. (2004). "Bayesian multilevel estimation with poststratification: State-level estimates from national polls." *Political Analysis*, 375–385.

Parker, P. A., Holan, S. H., and Janicki, R. (2020a). "Computationally Efficient Bayesian Unit-Level Models for Non-Gaussian Data Under Informative Sampling." *arXiv preprint arXiv:2009.05642*.

— (2020b). "Conjugate Bayesian Unit-level Modeling of Count Data Under Informative Sampling Designs." *Stat*, e267.

Parker, P. A., Janicki, R., and Holan, S. H. (2019). "Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions." *arXiv preprint arXiv:1908.10488*.

Pfeffermann, D. and Sverchkov, M. (2007). "Small-area estimation under informative probability sampling of areas and within the selected areas." *Journal of the American Statistical Association*, 102, 480, 1427–1439.

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108, 504, 1339–1349.

Prokhorov, D. (2005). "Echo state networks: appeal and challenges." In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 3, 1463–1466. IEEE.

Savitsky, T. D. and Toth, D. (2016). "Bayesian estimation under informative sampling." *Electronic Journal of Statistics*, 10, 1, 1677–1708.

Schuna, J. M., Johnson, W. D., and Tudor-Locke, C. (2013). "Adult self-reported and objectively monitored physical activity and sedentary behavior: NHANES 2005–2006." *International Journal of Behavioral Nutrition and Physical Activity*, 10, 1, 126.

Skinner, C. J. (1989). "Domain means, regression and multivariate analysis." In *Analysis of Complex Surveys*, eds. C. J. Skinner, D. Holt, and T. M. F. Smith, 80 – 84. Chichester: Wiley.

Soria-Olivas, E., Gomez-Sanchis, J., Martin, J. D., Vila-Frances, J., Martinez, M., Magdalena, J. R., and Serrano, A. J. (2011). "BELM: Bayesian extreme learning machine." *IEEE Transactions on Neural Networks*, 22, 3, 505–509.