Chapter 20
Neural Networks for Propensity Score Estimation: Simulation Results and Recommendations

Bryan Keller, Jee-Seon Kim, and Peter M. Steiner

Abstract Neural networks have been noted as promising for propensity score estimation because they algorithmically handle nonlinear relationships and interactions. We examine the performance neural networks as compared with main-effects logistic regression for propensity score estimation via simulation study. When the main-effects logistic propensity score model is correctly specified, the two approaches yield almost identical mean square error. When the logistic propensity score model is misspecified due to the addition of quadratic terms and interactions to the data-generating propensity score model, neural networks perform better in terms of bias and mean square error. We link the performance results to balance on observed covariates and demonstrate that our results underscore the importance of checking balance on higher-order covariate terms.

Keywords Propensity score analysis • Neural networks • Logistic regression • Data mining • Covariate balance

20.1 Introduction

The goal of propensity score analysis is to correct for bias due to confounding in a non-randomized experiment. The *propensity score* is defined as the probability of assignment to the treatment group—we assume a dichotomous treatment variable—given the observed covariates (Rosenbaum and Rubin 1983). The application of propensity score analysis involves (1) estimating the propensity score for each participant and (2) conditioning on the estimated propensity scores to estimate an average treatment effect.

B. Keller (✉)
Teachers College, Columbia University, New York, NY, USA
e-mail: keller4@tc.columbia.edu

J.-S. Kim • P.M. Steiner
University of Wisconsin-Madison, Madison, WI, USA
e-mail: jeeseonkim@wisc.edu; psteiner@wisc.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer
Proceedings in Mathematics & Statistics 140, DOI 10.1007/978-3-319-19977-1_20

In practice, propensity scores are most often estimated by logistic regression. However, data-mining techniques that algorithmically handle nonlinear relationships have been noted as promising for propensity score estimation because they are able to adapt to complex response surfaces in their naive implementations (Westreich et al. 2010).

To our best knowledge the performance of neural networks for propensity score estimation has been examined via simulation in only one study (Setoguchi et al. 2008). The results of that study suggested that neural networks are a viable alternative to main-effects logistic regression for propensity score estimation, though the authors caution that more work is needed over a broader range of scenarios.

We respond to the need for empirical evaluation by contributing a simulation study which examines the performance of neural networks as compared with main-effects logistic regression for propensity score estimation. Although we describe the simulation study in detail below, two aspects deserve particular attention. First, we use a *weight decay* smoothing parameter to inhibit over-fitting with neural networks. Second, we generate data from a pair of models: the propensity score data-generation model and the outcome data-generation model. The unique aspect here is that we consider the effect of nonlinear terms in the outcome data-generation model. In fact, we hypothesize that it is precisely when there are confounding higher-order terms in *both* the propensity score data-generation model and the outcome data-generation model that neural networks will have the potential to most drastically outperform main-effects logistic regression in terms of bias and mean square error.

The remainder of the paper is organized as follows: in the remainder of this section we describe propensity score analysis and assumptions required to estimate the average treatment effect for a population. In the next section we describe logistic regression and neural networks for propensity score estimation. We then discuss the method used to condition on the propensity score: optimal full matching. We then describe the design and results of the simulation study and conclude with some recommendations.

20.1.1 The Average Treatment Effect

The *potential outcomes* notation is based on the Neyman–Rubin framework for causal inference (Holland 1986). Let $Z_i = 1$ if the ith unit was assigned to the treatment group and $Z_i = 0$ otherwise. Let Y_i^j be a response variable such that each experimental unit has two potential outcomes, Y_i^1 and Y_i^0, depending on assignment Z_i.

Two causal quantities which are most commonly of interest are the overall population average treatment effect τ and the population average treatment effect for the treated τ_T (Imbens 2004; Schafer and Kang 2008; Steiner and Cook 2013):

$$\tau = E(Y_i^1 - Y_i^0) = E(Y_i^1) - E(Y_i^0) \tag{20.1}$$

and

$$\tau_T = E(Y_i^1 - Y_i^0 | Z_i = 1) = E(Y_i^1 | Z_i = 1) - E(Y_i^0 | Z_i = 1). \quad (20.2)$$

We focus on the average treatment effect on the treated, τ_T, in the simulation study because it provides an estimate of the effect of treatment on those who received it, which is often more interesting than the overall treatment effect (Morgan and Winship 2007).

20.1.2 Assumptions for Identifying and Estimating the Average Treatment Effect

The propensity score (PS) is defined as the conditional probability of assignment to the treatment group given the observed covariates $X = (X_1, \ldots, X_p)'$ (Rosenbaum and Rubin 1983). That is,

$$\text{PS}(X) = P(Z = 1|X). \quad (20.3)$$

Propensity scores may be conditioned upon in an application such as matching, stratification, or weighting in order to restore covariate balance across groups to what would have been expected from a randomized experiment. In order for the propensity score to be effective in eliminating bias some assumptions are necessary. First, the treatment assignment must be *strongly ignorable* (Rosenbaum and Rubin 1983; Rubin 1978). Strong ignorability specifies (a) that the potential outcomes are independent of the treatment assignment given the observed covariates and (b) that each experimental unit in the population has a true propensity score that lies strictly between zero and one. That is,

$$Y^1, Y^0 \perp\!\!\!\perp Z|X \quad (20.4)$$

and

$$0 < P(Z = 1|X) < 1. \quad (20.5)$$

In practice, strong ignorability is satisfied when all of the confounding covariates (i.e., those that are associated with both treatment assignment and the outcome) are observed, there is overlap between the propensity score distributions of the treatment and control groups, and the covariates are measured reliably (Steiner et al. 2011). When there is a lack of overlap, τ_T is only identified for the subpopulation of overlapping units.

Second, it is assumed that there is only one version of the treatment and that the value of each potential outcome is independent of the particular assignment pattern in Z. These two assumptions are referred to collectively as the stable unit treatment value assumption (SUTVA; Rubin 1978, 1980).

Finally, since true propensity scores are not known in observational study settings, they must be estimated. Assuming strong ignorability and SUTVA hold, an additional analytic assumption required for consistent estimation is that the propensity score estimates are adequate for bias removal. Although the necessary and sufficient conditions for estimation of adequate propensity scores depend on the method used to condition on them (Waernbaum 2010), one way to ensure adequate propensity score estimates is to correctly specify the relationship between selection Z and covariates X in a parametric model.

In practice, however, the model is never exactly correct. Thus, the focus of the propensity score estimation literature is on the proposal and evaluation of methods that attempt to approximately satisfy this analytic assumption.

20.1.3 Covariate Balance

If strong ignorability and SUTVA are satisfied, the propensity score is a *balancing score* for X (Rosenbaum and Rubin 1983); that is,

$$X \perp\!\!\!\perp Z | \mathrm{PS}(X). \tag{20.6}$$

As a result, the extent to which covariate distributions are balanced across treatment groups may be used as a diagnostic tool for checking the adequacy of the propensity score estimates.

Balance measures based on means are easy to calculate and have been shown to outperform other methods in simulation studies (Ali et al. 2014; Belitser et al. 2011), thus, we measure covariate balance with standardized mean differences. The standardized mean difference for covariate X is

$$d = \frac{\bar{X}_T - \bar{X}_C}{\hat{\sigma}_T} \tag{20.7}$$

where $\bar{X}_T$ and $\bar{X}_C$ are the means of treated and control units, respectively, and $\hat{\sigma}_T$ is the estimated standard deviation for X for treated units. We divide by the standard deviation of the treated cases instead of the pooled standard deviation across groups because the value of $\hat{\sigma}_T$ is not affected by propensity score weighting when estimating τ_T (McCaffrey et al. 2004).

To summarize balance over multiple covariates, both measures can be extended by taking averages. For covariates $X_1, X_2, \ldots, X_p$, the average standardized absolute mean difference (ASAMD) is

$$\mathrm{ASAMD} = \frac{1}{p} \sum_{i=1}^{p} |d_i| \,. \tag{20.8}$$

20.2 Propensity Score Estimation

Logistic regression is the most frequently used method for estimating propensity scores. For dichotomous outcome Z and covariates $X_1, X_2, \ldots, X_p$, each vectors of length N, the multiple logistic regression model is

$$log\left[\frac{P(Z=1|X_1,X_2,\ldots,X_p)}{1-P(Z=1|X_1,X_2,\ldots,X_p)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \quad (20.9)$$

We refer to the model in Eq. (20.9) as the *main-effects logistic regression* because the model contains one first-order term for each covariate.

The single-layer feed-forward neural network consists of an input layer of p observed covariates and a constant term, an output layer containing a single unit for dichotomous classification, and one hidden layer of M unobserved variables and a constant term (see Fig. 20.1).

The hidden units ($\mathbf{H} = H_1, \ldots, H_M$ in Fig. 20.1) are created by forming weighted linear combinations of the input variables and then applying the logistic function $f(t) = 1/(1 + e^{-t})$.

The dichotomous exposure variable Z is then used as the outcome in a logistic regression on the hidden units in the final step. The weights of the network are similar to regression coefficients in a traditional regression analysis in that larger weights indicate sharper changes in the slope of the response surface predicted by the model.

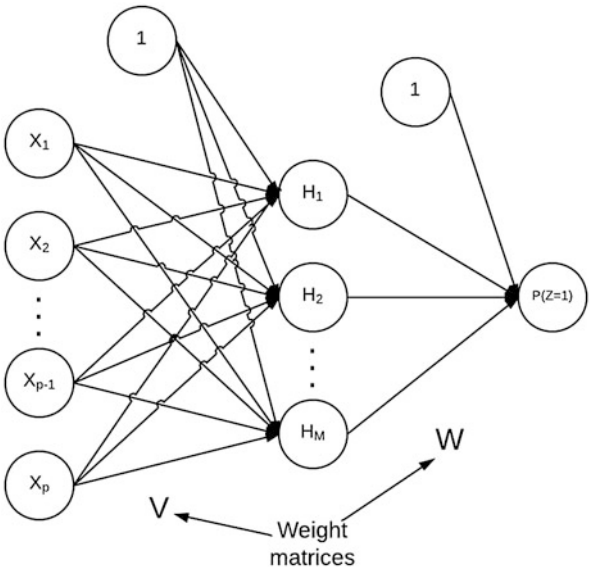

Fig. 20.1 Neural network with p inputs, M hidden nodes, and 1 classification output. The 1s represent intercepts

The model may be expressed as follows:

$$P(Z = 1|\mathbf{X}) = f(f(\mathbf{X}\mathbf{V}^T)\mathbf{W}^T), \tag{20.10}$$

where $\mathbf{X}_{N\times(p+1)}$ is the matrix of predictors, augmented to include a column of ones, and $\mathbf{V}_{M\times(p+1)}$ and $\mathbf{W}_{1\times(M+1)}$ are weight matrices containing the coefficients for the network. The hidden layer $\mathbf{H} = f(\mathbf{X}\mathbf{V}^T)$ is also augmented to include a column of 1s before being multiplied by $\mathbf{W}^T$. These augmentations are analogous to including the constant term for the intercept in the design matrix of a multiple regression and are represented by encircled 1s in Fig. 20.1.

The size of the hidden layer (M) determines how many parameters the model will have and, thus, how flexible the network will be in modeling the relationship between the predictors and the output. Increased flexibility, however, comes at the cost of an increased risk of overfitting random noise in the data. *Weight decay* is a technique which imposes penalties on large weights in the network, as in ridge regression for linear models, thereby smoothing boundaries and preventing over-adaptation to the particularities of the data (Hastie et al. 2009; Ripley 1996).

20.3 Propensity Score Application

The goal of the propensity score application step is to condition the outcome on the estimated propensity score, thereby restoring balance to the observed covariates and allowing for unbiased estimation of the average treatment effect. Matching techniques aim to accomplish this goal by identifying groups of individuals from the treatment and control groups that are as alike as possible according to the logit of the estimated propensity score. This may be done in a one-to-one, one-to-many, or many-to-many fashion; *full matching* refers to the latter case.

The goal of *optimal* full matching is to define S mutually exclusive strata each containing at least one treated and one control unit such that the configuration minimizes a global measure of distance (Rosenbaum 2002). Because optimal matching minimizes an overall measure of distance, it avoids the problem of different results based on matching order which occurs with other matching algorithms such as nearest neighbor matching.

After groups have been formed by optimal full matching, τ_T may be estimated by taking the difference of weighted averages across treatment and comparison groups. For τ_T, weights are calculated as follows:

$$\lambda_i = Z_i + (1 - Z_i)\frac{N_C n_T}{N_T n_C}, \tag{20.11}$$

where N is the overall sample size, N_T and N_C are the number of treated and comparison units, respectively, and n_T and n_C are the number of treated and comparison units in the subclass to which unit i belongs, respectively.

Let T be the set of indexes assigned to the treated condition and let C be the set of indexes assigned to the comparison condition. Then the estimator for τ_T is

$$\hat{\tau}_T = \frac{\sum_{i \in T} \lambda_i Y_i}{\sum_{i \in T} \lambda_i} - \frac{\sum_{i \in C} \lambda_i Y_i}{\sum_{i \in C} \lambda_i}. \tag{20.12}$$

20.4 Simulation Study

The purpose of the simulation is to examine the effect of propensity score estimation method on bias and mean square error of the treatment effect estimates and on balance on observed covariates.

20.4.1 Data Generation and Simulation Design

Twelve covariates were independently generated from a standard normal distribution. A correlation structure was induced via Cholesky decomposition so that $\rho_{ij} = 0.3$ for all $i \neq j$, where ρ is the Pearson product-moment correlation coefficient. The main-effects propensity score model is

$$\text{PS}_1 = (1 + \exp\{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{12} X_{12}\})^{-1}. \tag{20.13}$$

The complex propensity score model,

$$\begin{aligned} \text{PS}_2 = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{12} X_{12} + \\ &\beta_{13} X_1 X_{12} + \beta_{14} X_2 X_{11} + \beta_{15} X_2 X_{10} + \beta_{16} X_4 X_{12} + \beta_{17} X_1 X_8 + \\ &\beta_{18} X_2^2 + \beta_{19} X_5^2 + \beta_{20} X_8^2 + \beta_{21} X_{11}^2)\})^{-1}, \end{aligned} \tag{20.14}$$

includes five two-way interaction terms and four quadratic terms in addition to the main-effects in PS_1. The regression coefficients for the propensity score models were specified as follows:

$$\begin{aligned} \beta_0, \cdots, \beta_6 &= -1.00 \quad -0.49, \quad -0.18, \quad -0.40, \quad -0.26, \quad -0.16, \quad 0.51, \\ \beta_7, \cdots, \beta_{12} &= \quad -0.84, \quad 0.08, \quad -0.31, \quad 0.73, \quad -0.04, \quad -0.34, \\ \beta_{13}, \cdots, \beta_{17} &= \quad -0.42, \quad -0.26, \quad 0.16, \quad -0.36, \quad 0.31, \\ \beta_{18}, \cdots, \beta_{21} &= \quad -0.50, \quad 0.46, \quad 0.30, \quad 0.36. \end{aligned}$$

The regression models used to generate the continuous outcome are shown in Eqs. (20.15) and (20.16). The main-effects outcome model is

$$\text{Y}_1 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{12} X_{12} + \gamma Z. \tag{20.15}$$

The complex outcome model is

$$\begin{aligned} Y_2 =& \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{12} X_{12} + \\ & \alpha_{13} X_1 X_{12} + \alpha_{14} X_1 X_{11} + \alpha_{15} X_2 X_{11} + \alpha_{16} X_3 X_{12} + \alpha_{17} X_4 X_{12} + \\ & \alpha_{18} X_2^2 + \alpha_{19} X_3^2 + \alpha_{20} X_5^2 + \alpha_{21} X_{11}^2 + \alpha_{22} X_{12}^2 + \gamma Z. \end{aligned} \tag{20.16}$$

For each case i, the dichotomous selection variable Z_i was generated by comparing the propensity score to a random uniform draw from [0,1]. If the uniform draw was less than or equal to the propensity score for case i, Z_i was assigned to be 1; otherwise, Z_i was assigned to be 0. In both models, the selection variable was multiplied by the constant treatment effect $\gamma = -0.40$. The regression coefficients for the outcome models were specified as follows.

$$\begin{aligned} \alpha_0, \cdots, \alpha_6 &= 1.00, \quad 0.24, \quad 0.38, \quad -0.50, \quad 0.40, \quad -0.60, \quad -0.30, \\ \alpha_7, \cdots, \alpha_{12} &= \quad 0.06, \quad -0.66, \quad 0.58, \quad 0.34, \quad -0.58, \quad -0.40, \\ \alpha_{13}, \cdots, \alpha_{17} &= \quad -0.21, \quad -0.14, \quad -0.49, \quad 0.11, \quad 0.22, \\ \alpha_{18}, \cdots, \alpha_{22} &= \quad -0.30, \quad 0.41, \quad 0.31, \quad 0.26, \quad -0.20. \end{aligned}$$

The two PS models were crossed with the two outcome models to create four data-generation conditions. One thousand data sets were simulated and analyzed for each of the four conditions based on a sample size of 2000. Table 20.1 displays the standardized initial biases and the probability of treatment assignment for each of the four scenarios. The standardized initial bias was calculated as the unadjusted mean difference (treatment minus control) minus the true treatment effect of −0.4 divided by the standard deviation of the treatment group. The probability of assignment to the treatment group is simply the proportion of simulated participants in the population assigned to the treatment group.

Table 20.1 Population standardized initial bias and probability of assignment to treatment for each of four data generation scenarios

Scenario	Standardized initial bias	P(Z = 1)
$PS_1 \times Y_1$	0.257	0.332
$PS_1 \times Y_2$	0.214	0.332
$PS_2 \times Y_1$	0.231	0.418
$PS_2 \times Y_2$	0.462	0.418

Note: PS_1 and PS_2 represent the linear and nonlinear PS data-generating models, respectively; Y_1 and Y_2 represent the linear and nonlinear outcome data-generating models, respectively.

20.4.2 Analysis

Logistic regression was run with main effects only for $X_1, \cdots, X_{12}$, as in Eq. (20.9); neural networks were fit with eight hidden nodes and the weight decay tuning parameter set at $\lambda = 0.10$ for the scenarios with linear propensity score model and $\lambda = 0.13$ for the nonlinear propensity score models. For the last two data-generation scenarios (both with PS_2) we also estimated propensity scores with the correctly specified model, displayed in Eq. (20.14), in order to have a baseline for comparison with the other methods.

In practice, with a single data set, an analyst would select optimal tuning parameter values for a data mining method by searching over a grid of many possible choices and settling on the combination which produced the best cross-validated prediction or the best balance. In order to avoid the prohibitive computational cost of running a cross-validated grid search at each iteration, we ran such a grid search on five data sets generated from the linear PS and five data sets generated from the nonlinear PS and used the results to select sensible values. Our approach for selecting the value of weight decay (λ) for the neural networks was motivated by the usual design-based recommendations for propensity score model fitting: we selected the value of λ that was associated with the best covariate balance (though ten-fold cross validation based on prediction yielded similar results).

To assess covariate balance we used a weighted composite of the ASAMD on first-order terms and the ASAMD on second-order terms. These were weighted equally in order to assign the same conceptual importance to the class of first-order terms as the class of second-order terms in determining the resultant balance. For each dataset, as the value of λ increased, the balance improved for a period and then began to decrease. For the linear propensity score model optimal balance was attained at about $\lambda = 0.10$; for the nonlinear propensity score model, optimal balance was attained at about $\lambda = 0.13$. Thus these values were used throughout all 1000 simulation replications.

For each replication and for each propensity score estimation method, cases in the treatment or control group with no counterpart in the opposite group within 0.1 pooled standard deviations of the propensity score logit were considered non-overlapping and discarded from the analysis. After discarding cases, propensity scores were re-estimated on the remaining cases and those values were used going forward. For PS_1, both methods resulted in about 4 % of cases being discarded due to lack of overlap. For PS_2, 1 and 7 % of cases were discarded for main-effects logistic regression and neural networks, respectively.

20.4.3 Results

For the first and second scenarios, the main-effects logistic regression model (see Eq. (20.13); abbreviated MELR in Table 20.2) was the correctly specified model.

Table 20.2 Performance metrics averaged over 1000 replications; optimal full matching was used to estimate the average treatment effect on the treated

Metric	Method	Scenario			
		$PS_1 \times Y_1$	$PS_1 \times Y_2$	$PS_2 \times Y_1$	$PS_2 \times Y_2$
Bias and MSE					
|Bias (%)|	MELR	0.07	0.03	25.14	173.80
	NN	3.58	3.08	12.73	13.90
	LR-20.14	NA	NA	0.85	1.94
Bias	MELR	0.000	0.000	0.101	0.695
	NN	−0.014	−0.012	−0.051	0.056
	LR-20.14	NA	NA	0.003	0.008
SE	MELR	0.002	0.004	0.002	0.003
	NN	0.002	0.004	0.003	0.004
	LR-20.14	NA	NA	0.003	0.004
MSE	MELR	0.006	0.018	0.014	0.495
	NN	0.006	0.016	0.012	0.020
	LR-20.14	NA	NA	0.012	0.018
Covariate balance					
ASAMD on 1st-order terms	MELR	0.042	0.042	0.047	0.048
	NN	0.044	0.044	0.074	0.073
	LR-20.14	NA	NA	0.053	0.055
ASAMD on 2nd-order terms	MELR	0.067	0.067	0.118	0.118
	NN	0.059	0.060	0.069	0.069
	LR-20.14	NA	NA	0.071	0.071

Note: PS_1 and PS_2 represent the simple and complex propensity score data-generating models, respectively; Y_1 and Y_2 represent the simple and complex outcome data-generating models, respectively; MELR: main-effects logistic regression as in Eq. (20.13); NN: neural networks; LR-20.14: logistic regression as in Eq. (20.14); ASAMD: average standardized absolute mean difference across the covariates (see Eq. (20.8)); SE: simulation standard error; and MSE: simulation mean square error

Note that biases associated with MELR were both within two simulation standard errors of zero, indicating they are not significantly different from zero. For the third and fourth scenarios, the data-generating model (see Eq. (20.14); abbreviated LR-20.14 in Table 20.2) was also used to estimate propensity scores. The estimates based on LR-20.14 for the last two scenarios were also within two simulation standard errors from zero. Thus, when the propensity score model was correctly specified, estimates based on optimal full matching were not significantly biased.

Estimates based on neural networks were associated with lower mean square error than main-effects logistic regression for all four scenarios, including the first two scenarios, for which the main-effects logistic model was correctly specified. This finding is not altogether surprising because the feed-forward neural network can be thought of as a generalization of logistic regression. In particular, by setting the coefficients in the matrix $\mathbf{V}$ and vector $\mathbf{v_0}$ all equal to zero, the feed-forward neural network described above is identical to main-effects logistic regression.

When the data-generating propensity score model was complex, estimates based on neural networks were far less biased than those based on main-effects logistic regression. In the last scenario, in which the propensity score and outcome data-generation models both contained second-order terms, estimates based on main-effects logistic regression were biased by 174 % of the magnitude of the treatment effect, while neural networks yielded about 14 % residual bias. Across all four scenarios, propensity scores based on neural networks resulted in less than 14 % bias.

Regarding covariate balance on first-order terms, the balance attained by the main-effects logistic regression was better than that attained by neural networks across the board. On second-order terms, however, the opposite held true, with greater disparities evident when the data-generating propensity score model was complex.

20.4.4 Discussion

The results of the simulation study suggest that if the relationship between covariates and selection involves only first-order terms, it does not make much difference in terms of bias or mean square error whether main-effects logistic regression or neural networks is used to estimate propensity scores. For the first two scenarios, both methods were less than 4 % biased, with nearly identical mean square error.

If the true selection model involves more than just linear terms, misspecification of the logistic propensity score estimation model by way of omitting higher-order terms creates the potential for bias, the magnitude of which depends on the relationship between the covariates and the outcome. If nonlinear terms omitted from the propensity score estimation model are also related to the outcome, as was the case in scenario 4 (note the common terms in Eqs. (20.14) and (20.16)), the bias may be very large because the omitted terms act as confounding variables that have not been accounted for.

Importantly, we found that balance checks on first-order terms did not help in diagnosing this problem. The last column of Table 20.2 reveals that model selection based exclusively on first-order balance would have favored the main-effects logistic model over both neural networks and the correctly specified logistic model.

While these results clearly highlight the importance of checking balance on higher-order terms, they also raise questions. First, in practice, what is the highest-degree covariate transformation on which balance should be assessed? Second, how should balance measures on higher-order terms be weighted when comparing propensity score estimation models or techniques? For example, for an analysis with 10 covariates there are 10, 55, and 220 possible first, second, and third-order terms, respectively whereas, for an analysis with 20 covariates there are 20, 210, and 1540 possible first, second, and third-order terms, respectively. Further research aimed at addressing these questions would be useful.

20.5 Conclusion

Propensity scores are most often estimated by logistic regression in practice because it is familiar, available in most statistical software packages, and easy to implement. The most challenging aspect associated with its use is the need for iterative respecification of the model based on balance checking, which, with many covariates, is tedious at best and untenable, due either to exhaustion of degrees of freedom or exhaustion of the analyst, at worst.

Neural networks are promising for propensity score estimation because they algorithmically deal with nonlinearities in the selection surface, making iterative respecification unnecessary. We found, through simulation, that propensity scores estimated by neural networks resulted in better balance on second-order terms than those estimated by main-effects logistic regression. In practice, the analyst will not know which higher-order terms (if any) are actually predictive of selection. The most useful algorithmic approach for propensity score estimation is one which automatically detects such terms and accounts for them in the propensity score estimates, which is what neural networks did here.

There are some potential challenges with the implementation of neural networks as well. First, while the selection of optimal tuning parameter values related to weight decay and the number of hidden units can be carried out automatically using packages designed to do cross-validation (we used package `caret` Kuhn 2014 in R Core Team 2014), the process is computationally expensive, ranging anywhere from several seconds to several hours of computational time, depending on the size of the problem and the speed of the computer. Second, if the neural network results in poor covariate balance even after selecting optimal tuning parameters, there is no guidance as to how an analyst should alter the model to improve the balance. For this second point, however, neural networks are flexible enough such that, if tuning parameters are carefully selected, this should be a relatively rare occurrence which might suggest a problem with the suitability of the data for propensity score analysis, rather than a problem with the neural network specification. Finally, even with the use of the weight decay smoothing parameter it is possible that with many noisy covariates neural networks may still overfit the data.

Although neural networks performed favorably relative to main-effects logistic regression, further research is needed to determine if neural networks continue to perform well in cases with many weak predictors, when coupled with other approaches for conditioning on the estimated propensity scores, and compared with other data-mining methods.

Finally, while we compared neural networks to main-effects logistic regression because it is the approach most often used in practice, the undiscerning use of the main-effects logistic model for propensity score estimation is not recommended. In practice, an analyst using a logistic regression framework for modeling selection would experiment with various formulations of the model in an iterative process aimed at maximizing covariate balance. While this approach is difficult to mimic in a simulation study, the performance of neural networks could be compared with a custom logistic model created by an experienced analyst in a case study setting.

Acknowledgements This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D120005. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S. V., Roes, K. C. B., Hoes, A. W., et al. (2014). Propensity score balance measures in pharmacoepidemiology: A simulation study. *Pharmacoepidemiology and Drug Safety, 23*, 802–811.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H. H., de Boer, A., & Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety, 29*, 1115–1129.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, NY: Springer.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–970.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*, 4–29.

Kuhn, M. (2014). Caret: Classification and regression training. R package version 6.0-35. http://CRAN.R-project.org/package=caret.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*, 403–425.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. New York, NY: Cambridge University Press.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rubin, D. B. (1978) Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6*, 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*, 591–593.

Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*, 279–313.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546–555.

Steiner, P. M., & Cook, T. D. (2013). Matching and propensity scores. In T. Little (Ed.), *Oxford handbook of quantitative methods*. Oxford: Oxford University Press.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36*, 213–236.

Waernbaum, I. (2010). Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference, 140*, 1948–1956.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology, 63*, 826–833.