Match2: Hybrid Self-Organizing Map and Deep Learning Strategies for Treatment Effect Estimation

Xiao Shou

Institute for Data Exploration and
Applications/Department of
Mathematics, Rensselaer Polytechnic
Institute
Troy, NY
shoux@rpi.edu

Tian Gao, Dharmashankar

Subramanian
Research AI, IBM T. J. Watson
Research Center
{tgao,dharmash}@us.ibm.com

Kristin P. Bennett

Institute for Data Exploration and
Applications/Department of
Mathematics/Computer Science,
Rensselaer Polytechnic Institute
Troy, NY
bennek@rpi.edu

ABSTRACT

Estimating treatment effects from observational data through covariate matching remains an active research area in causal inference. Although existing methods may provide accurate results on simulated datasets, knowing how to tune the parameters to accurately estimate treatments in practice can be a challenge, since the ground truth is not known. We provide an explainable hybrid neural network and self-organizing map (SOM) approach, Match2. Using a supervised learning paradigm, our method simultaneously learns a meaningful latent representation with respect to treatment assignment and a nonlinear neighborhood preserving mapping via SOM in the latent space. To select the appropriate latent dimension, we propose a data-driven strategy based on the minimum validation loss for the treatment classification subproblem. Unlike other matching methods, the hybrid SOM-neural network can be used as the basis for visualizing and quantifying the quality of the matches. The user can understand the quality of the matches to provide confidence in the results and detect any potential problems. We design a novel metric to examine the overall quality of matching along with the visualization. We demonstrate strong performance on four benchmark datasets compared to non-neural-network baselines. Integrating a SOM component may potentially benefit other state-of-the-art neural network models for causal effect estimation by gaining interpretability while retaining prediction/estimation accuracy.

CCS CONCEPTS

- **Applied computing** → **Bioinformatics; Health informatics;**
- **Computing methodologies** → **Neural networks.**

KEYWORDS

explainable clinical decision support, neural networks, covariate matching, causal inference, self-organization maps, average treatment effect on the treated

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM-BCB '21,

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8450-6/21/08...\$15.00
<https://doi.org/10.1145/3459930.3469532>

ACM Reference Format:

Xiao Shou, Tian Gao, Dharmashankar Subramanian, and Kristin P. Bennett. 2021. Match2: Hybrid Self-Organizing Map and Deep Learning Strategies for Treatment Effect Estimation. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21)*, August 1–4, 2021, Gainesville, FL, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3459930.3469532>

1 INTRODUCTION

We examine the problem of estimating the treatment effect of an intervention based on observational data. Causal inference, which estimates the treatment effect on units or subjects after an intervention, has long been studied by economists, political scientists, statisticians, and public health professionals [13, 15, 34, 34]. Causal phenomena are ubiquitous in our daily life. In healthcare, for example, to examine the effectiveness of a new drug requires the causal knowledge learned from clinical trials; to evaluate a complex care management program for high cost and high need patients demands a close monitor of the outcomes before-and-after. Another example is in digital marketing where companies attempt to understand the consequences of placing online ads to their business, which is also causal in nature [3].

The gold standard for examining causal effect is through randomized control trials. Such experimental designs are sometimes prohibitive and impractical due to their potential high costs and ethical concerns. Observational studies using existing data, on the other hand, are more accessible and useful insights can be more readily extracted and learned. If covariates in observational data affect the treatment assignment mechanism and the outcome, then simple estimates of the treatment effect will be biased and potentially inaccurate [29]. In the binary treatment case, the control and treatment groups are likely to differ significantly; treatment effects thus cannot be naively determined as the mean of the difference of two observed outcomes, for example. If, on the other hand, each treatment unit can be matched with a similar unit in the control, the treatment effect can be approximated by estimating the counterfactual with reasonable accuracy. Although other approaches such as regression and weighting are often used, matching offers more interpretability since researchers can easily access the covariate distribution and examine the quality of resulting inferences.

In this study, we focus on the problem of covariate matching to estimate the average treatment effect on the treated (ATT) under the potential outcome framework. In health evaluation studies, it is noted that ATT is more relevant to policy markers than average

treatment effect for the entire population [42]. The interest in this quantity rises from the frequent occurrences of covariate imbalance in observational healthcare and medical data. A good number of classical approaches to estimate ATT via matching are available as discussed in the review [38]. Methods such as matching in Euclidean, Mahalanobis, and propensity score space are widely used. Propensity score matching (PSM) for example proposed to estimate the probability of being treated conditioned on the covariates and remains the most commonly used approach across the social sciences [32]. Classification methods as alternatives to logistic regression such as random forest and neural network can also be used in PSM models [43]. Another category is matching in dimension-reduced subspaces and then apply matching estimators such as nearest neighbor match (NNM). A few examples are principle component analysis (PCA) [18], locality preserving projections (LPP) [12], and projection onto multiple random linear subspaces (RNNM) [24]. The motivation behind these methods is that dimensional reduction filters out features unrelated to either the treatment or the outcome. Other subspace-based models include the Hilbert-Schmidt Independence Criterion (HSIC) [4] and Balanced Nonlinear Representation (BNR) based NNM [23]. Both have demonstrated prominent results over benchmark datasets. However, no models are perfect; Euclidean-NNM and Mahalanobis-NNM may produce poor results because the bias of NNM increases with the dimensionality of data [1], and PSM may suffer from a model misspecification problem in the high-dimensional case [2]. While HSIC-NNM and BNR-NNM leverage the nonlinear representation of data, of which linear dimension reduction methods cannot capture, the computation and tuning involving kernel matrices raise.

Many deep learning methods are available for causal inference problems [10, 17, 26, 35–37, 44, 46], and these have provided state-of-the-art accuracy on simulated datasets where the ground truths are known. When deploying these methods in practice, however, tuning and training can be challenging. They do not directly support examining the quality of matched pairs in real datasets where ground truths are unknown. Furthermore, not many specifically use matching for ATT estimation. For example, a relevant work is the feature selection representation matching (FSRM) which estimates population level average treatment effect (ATE) and the learning of such a meaningful representation proves to be effective [5]. Another recent model, conditional treatment-adversarial learning based matching method (CTAM) is specifically designed for textual covariates [45].

Unlike these methods, our approach is simple yet powerful: Match2 simultaneously learns a meaningful latent representation with respect to the treatment assignment using a hybrid approach which combines a neural network classification treatment model and a nonlinear neighborhood preserving self-organizing map (SOM) which projects the transformed unit in the latent space onto a lower dimensional map; in short, we minimize a joint loss function for classification and energy associated with the SOM. The method includes a strategy for easy tuning. The learned 2D SOM representation supports visualization and quantification of the quality of the match achieved. The data points can be projected onto the learned 2D SOM; similar points are closer while different ones are farther apart. Users such as medical analysts can examine a treatment unit and its closest control(s) in almost real time

Shou, Gao, Subramanian & Bennett which greatly support decision making. Our contributions to causal inference and metric learning via deep learning include:

- Our novel hybrid deep learning/SOM approach, Match2, simultaneously learns a meaningful latent representation with respect to the treatment assignment and the observational covariate space. This approach makes matching more accurate, leading to as good or better estimation of treatment effects as prior approaches. We demonstrate the performance of our model on four benchmark datasets with improved results compared to other subspace baselines.
- Match2 makes matching more transparent and interpretable by capitalizing on the projections of the raw data onto the 2D SOM. We propose to use a global and local score metrics to evaluate the quality of post matching results. This hybrid SOM-neural network strategy, associated visualizations, and score metrics can be easily adapted to other state-of-the-art and future neural network models.
- The method is self-tuning. We propose an auto selection for the size of latent dimensions by finding the minimum loss on a validation holdout, addressing a major weakness of using prior approaches. Furthermore, the resulting match can be examined to determine the quality of the match and any potential problem areas. This innovation is essential for determining ATT on real datasets in practice, since the ground truth is not known.

The structure of this paper is as follows. Section 2 gives an overview of the potential outcome framework and defines the quantity of interest, ATT. Section 3 describes the background of self-organizing map(SOM) and how our model Match2 evolves from SOM. We also give a quality metric for matching under our model. Section 4 illustrates the technique on simulated datasets with known solutions for both health and non-health problems and the additional benefits Match2 enjoys. Section 5 provides an example of interpretability of the model; in section 6 we propose some potential work to adapt SOM to other neural network models.

2 PRELIMINARIES

The potential outcome framework, also known as the Neyman-Rubin Potential Outcomes [33], is a major paradigm for statistical analysis of cause and effect. Consider the binary treatment case in which units (e.g. patients) either receive treatment or control i.e. $T = \{0, 1\}$. Since each unit, i , can either receive the intervention or not, only one of the two outcomes is observed. The observed outcome is called the *factual outcome*, while the unobserved is the *counterfactual outcome*. If treatment is received, for example, then $Y_i(1)$ is factual while $Y_i(0)$ is counterfactual. Thus the individual treatment effect of unit i is $Y_i(1) - Y_i(0)$. Several assumptions in the formulation are as follows:

ASSUMPTION 1. *Stable Unit Treatment Value Assumption (SUTVA). Units do not interfere with each other. The same version of treatment is applied to each unit.*

ASSUMPTION 2. *Consistency. For each unit, potential outcome under treatment T_i equals the observed outcome if actual treatment T_i is received.*

Match2: Hybrid Self-Organizing Map and Deep Learning Strategies for Treatment Effect Estimation

ASSUMPTION 3. Positivity. Each unit with covariates X_i has some chance of getting treated; treatment assignment is not deterministic, i.e. $0 < \Pr(T_i = 1|X_i) < 1$.

ASSUMPTION 4. Ignorability. Given pre-treatment covariates X_i , treatment assignment is independent from potential outcomes. i.e. $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i|X_i$.

Under the above assumptions, matching provides a simple yet effective approach to estimate causal quantities of interest. Compared to regression-based methods, matching methods are more interpretable and less sensitive to model specification. Although they do not permit finding the exact unobserved counterfactual, they do render a reasonable approximation. Alternatively, matching can be viewed as a way to strategically subsample a dataset such that observable covariate distributions in the treated and control groups are balanced [28]. Many distance measures which define closeness and similarity are exploited in matching. Common measures are distance in origin space, in Mahalanobis space and in propensity score space [38]. Linear dimension reduction methods and their applications to matching are also popular. These methods include PCA, LPP, and RNNM which exploit the intermediate dimensional subspaces. Kernel-based methods HSIC and BNR exploit the non-linearity of data. Once a distance measure is selected, a matching estimator can be selected accordingly. We use 1-Nearest Neighbor defined in the latent space determined by the hybrid SOM/Neural to estimate ATT. For each treated unit, NNM finds its nearest neighbor with respect to a defined distance measure in the control group. The true ATT is the expectation of difference of two potential outcomes conditioned on the treated or intervened:

$$ATT = E[Y_i(1) - Y_i(0)|T_i = 1] \quad (1)$$

ATT as an estimator to the ground truth can be calculated as the following:

$$\hat{ATT} = \frac{1}{N_T} \sum_{i:T_i=1} (Y_i(1) - \hat{Y}_i(0)), \quad (2)$$

where $\hat{Y}_i(0)$ is the counterfactual estimated from unit i 's nearest neighbor in the control group and N_T is the number of treatment units.

3 PROPOSED METHODS

3.1 Self Organizing Maps

SOMs were introduced as an unsupervised learning algorithm in the early 1980s [20], and have numerous applications since then [6]. SOMs have many nice properties such as density matching and feature selection [11]; they also create a nonlinear projection that preserves the neighborhood [22]. We describe the SOM algorithm and some key terms to prepare for our discussion in Match2. As an online stochastic process, we first initialize the weights of the K nodes $\{1, 2, \dots, K\}$, also called prototypes in SOM at time t as $m_k(t)$. Then for each input vector x_i , the best match unit (BMU) is defined as the following:

$$c^t(x_i) = \operatorname{argmin}_{k \in \{1, \dots, K\}} \|x_i - m_k(t)\|^2. \quad (3)$$

All the prototypes are then updated via

$$m_k(t+1) = m_k(t) + \epsilon(t) h_{kc^t(x_i)}(t) (x_i - m_k(t)) \quad (4)$$

where $\epsilon(t)$ is the learning rate at t and $h_{kc^t(x)}(t)$ is the neighborhood function. A Gaussian neighborhood function at time t is given as the following:

$$h_{kc^t(x_i)} = \exp\left(-\frac{(k - c^t(x_i))^2}{2\sigma^2(t)}\right) \quad (5)$$

where $\sigma(t)$ is the radius of influence at time t . $\sigma(t)$ decays exponentially:

$$\sigma(t) = \sigma(0) \cdot b^{-\frac{t}{\lambda}} \quad (6)$$

where $\sigma(0)$ is the radius at time 0, b is the logarithmic base, and λ is a time constant, i.e. the total number of iterations. To achieve convergence, the learning rate (similar to one in stochastic gradient descent) should satisfy the Robbins-Monroe condition [31]:

$$\sum_t \epsilon(t) = +\infty, \sum_t \epsilon(t)^2 < +\infty. \quad (7)$$

For some industrial applications, a deterministic batch update version is introduced later [21, 39], which updates as the following with N input data points:

$$m_k(t+1) = \frac{\sum_{i=1}^N h_{kc^t(x_i)} x_i}{\sum_{i=1}^N h_{kc^t(x_i)}} \quad (8)$$

The above process repeats until convergence.

3.2 Match2

To equip for the batch gradient update, we identify the following theorem for SOMs by [30].

THEOREM 1. In the discrete setting where the input space X comprises N data points, the SOM algorithm (described in section 3.1) is a gradient descent algorithm associated with

$$E(m) = \frac{1}{2N} \sum_{i=1}^N \sum_{k=1}^K h_{kc(x_i)} \|m_k - x_i\|^2. \quad (9)$$

where E is the energy function or extended distortion of the stochastic process associated with SOM.

Figure 1: Match2: A hybrid Neural Network/SOM for Covariate Matching. Control and Treatment units (shown in blue and red dots) are fed into a learned Match2 network (a feed forward neural network and 2D SOM). The learned latent representations are projected onto the 2D SOM. A treated unit can be matched with its neighboring control unit. Note both units can be overlapping on the same node.

Inspired by this theorem, we design Match2 to perform classification with neural networks and learning SOM simultaneously as shown in Figure 1. The total loss decomposes as follows:

$$L_{tot}(W, m) = L_{BCE}(W) + \gamma L_{SOM}(W, m) \quad (10)$$

where W represents the set of weights and biases in the neural network, m is the set of weight vectors in the SOM, and γ is the trade off between the two losses. We use the standard binary cross entropy loss for classification. Its mathematical form can be explicitly defined as

$$L_{BCE}(W) = - \sum_{i=1}^N T_i \log(\text{sigmoid}(f_W(x_i))) + (1 - T_i) \log(1 - \text{sigmoid}(f_W(x_i))) \quad (11)$$

where T_i is the binary treatment for unit i (to distinguish from time t), and $f_W(x_i)$ is the nonlinearly transformed version of x_i , the covariates of unit i . Similarly we replace x_i with $f_W(x_i)$ in Equation 9 to obtain the SOM loss in our model:

$$L_{SOM}(W, m) = \sum_{i=1}^N \sum_{k=1}^K h_{kc^t}(f_W(x_i)) \|m_k - f_W(x_i)\|^2. \quad (12)$$

THEOREM 2. *In the discrete setting where the input space X comprises N data points, the energy function of SOM is equivalent to the smoothed max-component (i.e., the component with the largest weighted probability) log-likelihood function of isotropic Gaussian mixtures with tied weights, up to a constant and normalization multiplier.*

We do not provide full proof here and refer the readers to [9]. Alternatively, it is not difficult to see that the L_{SOM} asymptotically approaches k-means loss, i.e.

$$\lim_{t \rightarrow \infty} L_{SOM} = \lim_{t \rightarrow \infty} \sum_{i=1}^N \sum_{k=1}^K h_{kc^t}(f_W(x_i)) \|m_k - f_W(x_i)\|^2 = \sum_{i=1}^N \|m_k - f_W(x_i)\|^2. \quad (13)$$

Therefore the (local) minimizer of L_{tot} approximately minimizes the negative log likelihood of K isotropic Gaussian mixtures with tied weights, or equivalently the distortion loss associated with k-means in the embedded space. Intuitively, each learned node on the 2D SOM grid is a Gaussian mean with the same dimension as the latent embedding in a feed forward neural network. Close inputs in latent space belong to the same Gaussian component or to neighboring components.

3.3 Matching Scores

To quantify quality of matching under our model, we define 2 scores: local quality score τ_l and global quality score τ_g .

DEFINITION 1. *Let X be the covariate units and M be a learned Match2 model. The local score for a treatment unit $x_{i,T=1}$ is defined as $\tau_l(x_{i,T=1}) = \frac{\eta_i}{d_{i+1}}$, where d_i is the nearest neighbor distance, $\min_j \|M(x_{i,T=1}) - M(x_{j,T=0})\|$ (i.e. Euclidean in the latent space) from the unit on the 2D SOM grid, and η_i is the number of neighbors at that distance in the corresponding lattice point(s). The global quality score*

Algorithm 1 Match2

Input: training and validation set; SOM topology; σ_{max} , σ_{min} ; iterations; batchsize; latent dimension d

Output: weights W^* and m^*

```

1: while not converging do
2:    $\sigma(iter) \leftarrow \sigma_{max} (\sigma_{max}/\sigma_{min})^{-iter/iterations}$ 
3:   Compute weight matrix  $h_{kc^{iter}}(x_i)$ 
4:   Fix the weight matrix and train with minibatch
5:   Evaluate  $L_{tot} = L_{BCE} + L_{SOM}$ 
6: end while
7: return optimal  $W^*$  and  $m^*$ 

```

Algorithm 2 Match2 for ATT

Input: Covariates $X \in R^{N \times d}$, Binary Treatment $T \in \{0, 1\}^N$, Factual Outcome $Y \in R^N$

Output: estimated ATT

```

1: Split the (X,T) into train and validation, i.e. 80-20 split
2: Initialize Match2. For a range of latent dimensions, train the classification model; select one with the least validation loss.
3: Train Match2 with the selected dimension until convergence.
4: Map X onto 2D SOM, for each treatment/control unit, find the best match treatment/control unit (BMTU/BMCU) weight vector.
5: Find NNM in BMCU for each BMTU.
6: Perform estimation of ATT. If multiple NNM's in BMCU are found for each BMTU, use the mean of their outcomes  $Y_{NNM}$ .

```

of the entire set is defined as the average of the local scores, $\tau_l(X) = \frac{1}{N_T} \sum_{i=1}^{N_T} \tau_l(x_{i,T=1})$, where N_T is the total number of treatment units.

Intuitively, the local quality score expresses the confidence level of matching for a particular unit of interest; if a treatment unit and its nearest neighbor(s) fall on the same lattice point, i.e. they are in the same Gaussian cluster, then we are fairly confident about the match. If the two are farther away, then the match may be poor. The global score indicates the overall confidence level. We may also define a regional score if we are interested in the subpopulation with a certain age group for example. These scores suggest an alternative for examining the quality of matching and ATT estimation when the ground truth is not known in real life. We demonstrate the use of global quality score in Section 4.

4 COMPUTATIONAL EXPERIMENTS

In this section, we examine the effectiveness of Match2 at computing ATT. We perform a comparison of Match 2 with 7 other approaches on four datasets with known ATT. We examine the effect of varying dimensions, and show that Match2 can effectively self tune the number of latent dimensions and demonstrate decrease variance as sampled size increases. In addition in an ablation study, we show that the hybrid SOM neural network is key; using just the neural network or the SOM alone for matching is not as effective on all four datasets.

4.1 Model Implementation and Training

The base binary classification model is a standard feed forward neural network. This is combined with a SOM layer structure in our model adapted from the deep embedded SOM (DESOM) [8]. The side of a square 2D SOM topology is applied with the heuristic formulae $M \approx \sqrt{5\sqrt{N}}$ [40]. We let exponentially decaying radius $\sigma(0) = \sigma_{max}$ and set to 10; we introduce σ_{min} , and set to 0.1 so that the logarithmic constant $b = \sigma_{max}/\sigma_{min} = 100$. We split each dataset into 80-20 stratified by treatment for training and validation sets. The following are the default settings and are used in all experiments. Relu is used as the activation function; the parameters including the weights, biases, and SOM weights are all initialized with Glorot uniform initializer; Adam is used to train our model [19]. The trade-off parameter γ between the two losses is set to 1 for all runs. The radius of neighbor pair σ_{max} is set to 10, σ_{min} is set to 0.1. In addition, total iteration is set to 3000 for all experiments. The learning rate and number of hidden layers differ slightly from dataset to dataset as listed in Table 1.

Table 1: Training settings for all four datasets. Dims are latent embedding dimensions. * We used varying number neurons for the experiments on synthetic data: the hidden layer and latent embedding dimensions are (10,2), (20,5), (20,10), (40,20), (40,30), (50,40), (50,50), (60,60), (100,70), (100,80), (100,90) and (100,100). ** We used varying number neurons for the experiments on news data: (100,2), (100,5), (100,10), (100), (500) and (1000). Input dimensions for the four datasets are omitted.

Data	learning rate	hidden layers before embedding
Synthetic	0.01	1 HL with varying number neurons*
IHDP	0.01	no HL
News	0.0001	1 HL for low dims; 0 HL for high dims**
LaLonde	0.0001	no HL

For estimating ATT, we first solve the classification subproblem with respect to the treatment, which automatically selects the latent dimension out of a prespecified list of dimensions with the least validation loss. With this selected dimension, we train Match2 to minimize the total loss. Once learning is complete, we map the input data $X \in R^{N \times d}$ onto the 2D SOM. For each treatment unit, a similar unit of control is matched, and ATT can then be estimated. The complete algorithms are given in Algorithm 1 and 2.

4.2 Baselines

We compare our results with those of baseline estimators mentioned above including classical approaches such as Euclidean-NNM, Mahalanobis-NNM and PSM with L1 10 fold-cross validation along with subspace methods PCA-NNM, LPP-NNM, RNNM, and BNR-NNM. We only compare with matching-based methods and not non-matching-based neural network methods since neural network models do not provide explainable insights to the matching. We provide an explainable approach, which can be potentially

adapted to other neural network methods (e.g. counterfactual regression network [17]) by incorporating the 2D SOM into the learned embeddings. Space requires us to leave a fuller investigation of alternative hybrid neural network and SOM architectures as future work. A Python package of comprehensive modules for covariate matching using the hybrid-SOM approach will be provided.¹

4.3 Synthetic Dataset

We simulate our data according to the generating process by [23]. The following 10 basis functions are used to generate each realization: $g_1(x) = x - 0.5$, $g_2(x) = (x - 0.5)^2 + 2$, $g_3(x) = x^2 - 1/3$, $g_4(x) = -2\sin(2x)$, $g_5(x) = e^{-x} - e^{-1} - 1$, $g_6(x) = e^{-x}$, $g_7(x) = x^2$, $g_8(x) = x$, $g_9(x) = I_{x>0}$ (indicator), and $g_{10}(x) = \cos(x)$. Each covariate unit x_i is drawn independently from the multivariate normal distribution $N(0, I)$. When generating the corresponding binary treatment, we define the treatment for each unit $T_i|x_i = 1$, if $\sum_{j=1}^5 g_j(x_{ij}) > 0$, $T_i|x_i = 0$ otherwise; x_{ij} is the j^{th} entry of vector x_i . The outcome for each unit follows $Y_i|x_i, T_i \sim N(\sum_{j=1}^5 g_{j+5}(x_{ij}) + T_i)$. This generating mechanism suggests only the first 10 of the 100 features directly impacts the assignment and outcome in a non-linear way. Our simulation thus mimics a confounding effect while adding irrelevant covariate noise. We draw 5000 samples with 100 features and repeat the process 100 times.

Results and Analysis. As shown in Table 2, Match2 has the best performance over all baselines on the synthetic datasets. Recall the non-classical methods: PCA, PCC< RNNM, BNR, and Match2, have the number of latent dimensions as a hyperparameter. The results show that as the latent dimension increase from 2 to 100, all non-classical baselines show decaying performance; Match2 however stably renders the closest result to the ground truth ATT of 1. The selection of a dimension is important especially for other models because estimated ATT is sensitive to dimension. Match2 automatically selects the appropriate dimension of 20 using validation (shown in parentheses). The best dimensions of other approaches is determined using the ground truth which is typically not known. On the synthetic data, Match2 significantly outperformed both the classical (Euclidean, Mahalanobis, and PSM_L1) and non-classical matching methods.

4.4 Simulations based on real dataset: IHDP

IHDP data is a real randomized experiment dataset collected by the Infant Health and Development Program, to study the effect of high-quality child care and home visits on future cognitive test scores. To simulate an observation study, Hill [2011] simulates the outcomes and artificially introduces imbalance between treated and control subjects by removing a nonrandom subset of the treatment group: all children with non-white mothers. The final dataset consists of 747 subjects (139 treated, 608 control), each with 25 covariates representing attributes of either the child or their mother. The 100 simulations used in our study are based on Shalit et al. 2017.

Results and Analysis. In Table 3, we use the absolute error $\epsilon_{ATT} := |\hat{ATT} - ATT|$ to evaluate the performance of models. Our model selects 2 dimensions as the best result according to the minimum validation loss in the classification subproblem. Again, Match2

¹<https://github.com/xshou1990/match2>

Table 2: Results on synthetic data. Each entry is mean and standard deviation (SD) of estimated ATT over 100 repeated realizations. The ground truth is 1. Results shown for classical match methods: Euclidean, Mahalanobis, and PSM_L1. For non-classical methods (PCA, LPP, RNNM, BNR and Match2, latent dimension (Dim) varies from 2 to 100. The best results for each method is highlighted in bold. (*) indicates the baselines over which Match2 has statistically significant improvement ($p \leq .05$). The dimension for Match2 is automatically selected based on the minimum validation loss in parentheses. The best dimensions of other approaches is determined using the ground truth which is typically not known.

Classical	Dim	PCA	LPP	RNNM	BNR	Match2 (validation loss)
Euclidean: 1.21±0.17*	2	1.06±0.13	1.06±0.13	1.05±0.11	1.06±0.23	1.04±0.12 (.30508)
	5	1.07±0.12	1.07±0.21	1.07±0.11	1.11±0.20	1.05±0.13 (.30162)
	10	1.09±0.12	1.08±0.13	1.08±0.11	1.21±0.22	1.04±0.12 (.30244)
	20	1.11±0.14	1.11±0.15	1.10±0.11	1.35±0.27	1.05±0.13 (.30127)
	30	1.15±0.15	1.11±0.16	1.12±0.11	1.46±0.29	1.04±0.13 (.30282)
Mahalanobis: 1.35±0.14*	40	1.17±0.15	1.13±0.25	1.13±0.11	1.54±0.28	1.05±0.12 (.30354)
	50	1.17±0.16	1.14±0.24	1.14±0.11	1.60±0.29	1.06±0.11 (.30335)
	60	1.19±0.16	1.15±0.17	1.15±0.12	1.66±0.33	1.05±0.13 (.30570)
	70	1.19±0.18	1.16±0.17	1.16±0.12	1.71±0.34	1.04±0.12 (.30643)
PSM_L1: 1.05±0.14	80	1.19±0.17	1.19±0.19	1.16±0.12	1.73±0.36	1.05±0.13 (.30642)
	100	1.21±0.17	1.24±0.28	1.17±0.13	1.82±0.38	1.05±0.13 (.30881)

outperforms all baselines, as the error is much smaller compared to other models.

4.5 Simulations based on real dataset: News

The News dataset is another simulated benchmark based on real data proposed by Johansson et al. [2016]. The dataset contains data on the opinion of media consumers on news items: more specifically, the preference of reading certain topics on a mobile device over on desktop. For each simulated replica, 5000 randomly news articles from the NY Times corpus are selected. For each news item, the counts of words are recorded for each words occurring in a vocabulary of 3477 words, as the union of top 100 most probable words in each topic. Unlike IHDP-based simulations, the news datasets consistently have very sparse and high dimensional covariate matrices. In this study, we use 50 simulations generated according to protocols by the authors [17].

Results and Analysis. Match2 selects 2 dimensions using the validation loss in Table 4, and this model significantly outperforms all baselines except PCA. As shown, the the relative error ϵ_{ATT} 0.06 ± 0.04 is significantly smaller than other models, and better than 0.07 ± 0.04 of PCA.

4.6 Application to real dataset: LaLonde

LaLonde is a widely used benchmark for observational study in the causal inference community. Researchers are interested in determining whether job training will improve employment status which is measured by the raise of income for the participants in the program. The original dataset contains data from a randomized study where unbiased estimation of ATT is available: the group who received job training made $\$886 \pm \448 more than those who did not conditioned on those who received the training. Recently, Imai et al. [2014] combined the original experimental data (297 treated, 425 control) and observational units from the Panel Study of Income

Dynamics, which increases the control size to 2915. The study includes 8 covariates: age, education, race (black, white, or Hispanic), marriage status, high school degree, earnings in 1974, and earnings in 1975. The outcome variable is earnings in 1978. The dataset is accessible online ².

Results and Analysis. The model selects 20 dimensions as it gives the minimum loss for the classification subproblem as shown in Table 5. We use the metric $BIAS\% := (\hat{ATT} - ATT)/ATT$ % to evaluate the performance of models. Our method improves over all baselines along with CBPS [16] and DNN [36] as shown in Table 6. We further perform bootstrapping experiments with 100 different seeds for splitting the data into train and validation. We report the distribution of estimated ATT for different dimensions as shown in Figure 2. We also include the ground truth ATT and the interval within 1 standard error as red-shaded region. Almost all estimated ATT at any dimensions are within 1 standard error of the ground truth. For a real data set, one could then select a latent dimension following the minimum validation loss procedure, and bootstrap the data to report the mean estimated ATT. This procedure is effective for real datasets where repeated simulations are not available.

4.7 Ablation Study

We confirm the strong performance of Match2 is due to a combination of NN representation and SOM learning on all four datasets. Under the same setting, we compare results of two additional baselines 2DLR-NNM and SOM-NNM with Match2. The former removes the SOM component from Match2 and trains the binary classification network with latent dimension equal to 2; the latter removes the classification network and trains the SOM in raw input dimension. Nearest neighbor matching is used for both baselines. The results in Table 7 suggest Match2 achieves significantly better results over its counterparts.

²<https://github.com/kosukeimai/CBPS>

Table 3: Results on IHDP data. Each entry is the mean and SD of estimated ϵ_{ATT} over 100 repeated realizations. The smaller the error is, the more accurate the $\hat{ATT}$ is. The best result for each method highlighted is bold. The star marker (*) indicates the baselines over which Match2 has statistically significant improvement ($p \leq .05$). The latent dimension for Match2 is selected based on the minimum validation loss in the parentheses. The best latent dimensions of other approaches is determined using the ground truth which is typically not known.

Classical	Dim	PCA	LPP	RNNM	BNR	Match2 (validation loss)
Euclidean: 0.38±0.56 Mahalanobis: 0.64±1.16* PSM_L1: 0.97±0.93*	2	0.92±1.79	1.17±2.27	0.76±1.35	0.84±1.22*	0.32±0.41 (.45997)
	5	0.58±1.13	0.66±1.26	0.54±1.00	0.84±1.22	0.50±1.21 (.46043)
	10	0.41±0.56	0.65±1.22	0.42±0.78	0.85±1.22	0.42±0.56 (.46530)
	20	0.46±0.78	0.49±0.79	0.40±0.73	0.47±0.93	0.19±0.82 (.46777)
	25	0.36±0.57	0.68±1.31	0.38±0.71	0.85±1.23	0.44±0.68 (.46009)

Table 4: Results on News data. Each entry is mean and SD of estimated ϵ_{ATT} over 50 repeated realizations. The best result for each method highlighted is bold. * indicates the baselines over which Match2 has statistically significant improvement ($p \leq .05$). The latent dimension for Match2 is selected based on the minimum validation loss in the parentheses.

Classical	Dim	PCA	LPP	RNNM	BNR	Match2 (validation loss)
Euclidean: 0.77±0.54* Mahalanobis: 0.91±0.51* PSM_L1: 0.09±0.05*	2	0.08±0.05	0.10±0.06	0.10±0.05*	0.12±0.08*	0.06±0.04 (.72890)
	5	0.07±0.05	0.10±0.06	0.10±0.05	0.12±0.08	0.07±0.05 (.75048)
	10	0.07±0.04	0.10±0.07*	0.10±0.05	0.12±0.08	0.09±0.05 (.83189)
	100	0.11±0.08	0.11±0.09	0.20±0.17	0.12±0.08	0.09±0.05 (.73618)
	500	0.34±0.20	0.21±0.21	0.50±0.37	0.13±0.08	0.10±0.05 (.83095)
	1000	0.52±0.33	0.41±0.34	0.59±0.44	0.14±0.10	0.10±0.05 (.86300)

Table 5: Dimension selection on LaLonde dataset based on least validation loss on classification subproblem

Dimension	ATT	SD	Validation Loss
2	899.9	403.6	.19344
5	944.9	404.0	.19425
10	1020.5	416.3	.19359
20	860.1	409.6	.19312
30	967.0	408.3	.19315
40	770.0	399.6	.19354
50	853.0	405.1	.19334
60	775.4	423.3	.19581
70	860.4	402.8	.19583
80	853.1	406.6	.19608
90	858.3	413.4	.19611
100	816.8	407.4	.19649

Table 6: Results on Lalonde dataset. BIAS% is the bias in percentage of the true effect.

METHOD	ATT	SD	BIAS(%)
GROUND TRUTH	886	488	N/A
EU-NNM	-565.9	592.8	164%
MAH-NNM	-67.9	526.1	108%
PSM	-947.6	567.9	201%
PCA-NNM	-499.8	592.5	156%
LPP-NNM	-457.1	581.2	152%
RNNM	-557.6	584.9	163%
CBPS	423.3	1295.2	52%
DNN	742.0	N/A	16%
BNR-NNM	783.6	546.3	12%
Match2	860.1	409.6	3%

4.8 Effect of Sample Size

We further examine the performance of our model on different sized datasets. As shown in Figure 3 our estimator improves as sample size increases. For small sample sizes, other methods may have a slightly better performance. For example, CTAM reports slightly better results for IHDP data, and News data with 1000 samples. However, it requires selecting a latent dimension and lacks the SOM-based metric and visualization strengths of the proposed approach.

5 EXPLAINABLE CAUSAL INFERENCE WITH MATCH2

A major challenge in estimating in ATT in observational health studies, is selecting from the many algorithms and hyperparameters. While many methods may work well, it is difficult to be certain since the ground truth is not known. One major advantage of Match2 is that the SOM representation provides a 2D grid clustering that is extremely effective for observing similarities and differences in the data. Powerful SOM visualization methods and associated matching metrics (e.g. Global Score in Section 3.3) can be used to understand

Figure 2: Performance of Match2 with 100 bootstrapped samples at different latent dimensions on LaLonde dataset. The estimated ATT and its standard deviations by Match2 are shown in blue while the ground truth is the redline, and its standard deviation is shaded in red.

Table 7: Results of ablation study. For synthetic, IHDP and News, each entry is the mean and standard deviation of estimated quantity: $\hat{ATT}$ for synthetic data; ϵ_{ATT} for IHDP and News. For Lalonde, each entry is BIAS%. * indicates the baselines over which Match2 has statistically significant improvement ($p \leq .05$).

Data/Model	2DLR-NNM	SOM-NNM	Match2
Synthetic	1.30±0.20*	1.09±0.13*	1.05±0.13
IHDP	0.76±1.35*	0.47±0.80	0.32±0.41
News	0.08±0.05*	0.08±0.04*	0.06±0.04
Lalonde	621%	186%	3%

Figure 3: Effect of sample size on ATT estimation: as sample size increases, ϵ_{ATT} and its variance (error bars) decrease.

the quality in the match in different subpopulations. Thus Match2 can be used to examine the quality of the match across the space.

We demonstrate the visualization of our model on the Lalonde dataset. The learned latent representations were projected onto

the 2D SOM with size 17x17 as plotted in Figure 4. The number of treatment or control units projected onto each grid cluster is represented by size of the marker at each grid point (orange is treatment and blue is control). As shown in the figure, for almost every single treatment unit, there is at least one exact control unit which has zero distance from the treatment, i.e. they share the same coordinate; the one exception is the unit at (2,13). A clinical analyst, for example, can easily click on the treated person of interest in the 2D grid and examine the covariate characteristics of the treated person and its neighboring control counterparts in almost real time to get an intuitive idea of the quality of matching. This interactive visualization feature will be included in *Match Analyzer*, a package for investigating ATT Match quality currently under development.

Figure 4: An illustration of potential matches on the 2D lattice map generated by Match2 on the entire Lalonde dataset. Orange squares are treatment units; blue circles are control units. The size of lattice points is proportional to the number of units (treatment or control) projected on that lattice point. The treatment units are slightly shifted to the right so that they do not completely overlap with control counterparts when they share the same lattice point. Most of the treatment units have matched controls which share the same lattice points. The only exception in our example is the unit on (2,13) whose nearest neighbor is 2 units away if Manhattan distance is used. We are developing a tool *Match Analyzer* that enables the user to examine the treatment and control points within each lattice point, which reports the quality of the match and summary of characteristics of the matched units.

In Figure 5, we compare the match quality on Lalonde for two different trained Match2 models, with latent dimension of 5 and 10, in terms of their global scores and their respective SOM visualizations. The calculated global scores were 237.2 and 149.4 indicating that the former has a better quality than the latter. We can readily see that the number of treatment units which are not exactly matched in the former case is fewer compared to that of the latter case. The integrated SOM visualization and metrics enables the

Match2: Hybrid Self-Organizing Map and Deep Learning Strategies for Treatment Effect Estimation

analyst to understand the quality of the resulting treatment effect estimation. The analyst can select among different model variants and see if there are any subpopulations of treatment units which are difficult to match. Match2's visualizations and quantification of the match quality are especially important for real-world observational studies since the ground truth treatment effect is not known.

These figures represent just one type of the many potential SOM-based visualizations [41] and metrics that could be developed to interpret Match2 ATT results. We are creating a tool *Match Analyzer* for visualizing, quantifying, analyzing, and interpreting the quality of the matching and strength of the treatment effect by exploiting SOM neighborhood structure and visualizations. We will report on *Match Analyzer* in future studies.

6 ADAPTION TO STATE-OF-THE-ART NEURAL NETWORK MODELS

Our hybrid approach can be easily adapted to other state-of-the-art neural network models. Take counterfactual regression network (CFRNET)[17] as an example. Even though it is not a matching model, the learned representations can still be insightful once projected onto SOM. We can iteratively minimize the loss in CFRNET and SOM respectively; the training involves an outer and inner loop. For each outer iteration, update the weights in the CFRNET; the inner loop takes the learned latent representation from a CFRNET model under current outer iteration and optimize the SOM weights over a number of inner iterations. Pseudocodes are provided in 3. The standard inputs for CFRNET can be found online³. It should not be surprising that the learning of SOM parameters does not change the learning of original CFRNET model; therefore in this case, the learned model will also enjoy the benefit of interpretability of SOM, while preserving the prediction and estimation accuracy.

Algorithm 3 CFRNET-SOM

Input: CFRNET inputs ; SOM topology

Output: weights W_{CFR}^* and m_{SOM}^*

```

1: for outer – iteration  $\leftarrow 1$  to cfr – iterations do
2:   update  $W_{CFR}$ , compute latent representation  $h_{cfr}$ 
3:   for inner – iteration  $\leftarrow 1$  to som – iterations do
4:     take  $h_{cfr}$ , update  $m_{SOM}$ 
5:   end for
6: end for
```

7 CONCLUSION

We have proposed a novel hybrid SOM neural matching algorithm to simultaneously learning a latent representation meaningful to treatment and a neighborhood preserving nonlinear mapping for treatment effect estimation. Our model thus leverages the predictive power of deep learning models to accurately estimate counterfactuals while simultaneously providing a 2D map that enables interpretation and evaluation of the quality of the resulting estimation. We believe other state-of-the-art neural network models could also benefit from incorporating SOM; in fact, recent studies of SOM-aided deep learning have demonstrated its strength in different tasks

[25, 27]. This hybrid strategy facilitates the decision making process for an analyst, for example to evaluate a hotspotting program [7]. The strong performance of the model is demonstrated through the experiments on four synthetic and real-world health and business datasets and compared with existing state-of-the-art matching models. While others may work about as well with appropriate hyper-parameter choices such as latent dimension, it is difficult to know if these choices are appropriate when applying in real-world observational studies where the ground truth is not known. Match2's hybrid SOM approach helps overcome this limitation.

Many fruitful extensions are possible. We are creating a tool *Match Analyzer* for visualizing, quantifying, analyzing, and interpreting the quality of the matching and strength of the treatment effect by exploiting SOM neighborhood structure in metrics and visualizations. Additionally, our combined neighborhood mapping and deep learning paradigm is very flexible and can readily be extended. For example, by adding an extra loss term, the method can create an embedded feature representation which is meaningful with respect to both the outcome and treatment. Although the link to Gaussian mixture models provides some theoretical explanation for the performance of our model, we intend to develop a deeper understanding of SOM-aided deep learning and their benefits for explainability in future research.

ACKNOWLEDGMENTS

This work was supported by IBM AI Horizons Network. We thank the reviewers for their valuable feedback and Prof. Malik Magdon-Ismail at RPI for helpful discussion.

REFERENCES

- [1] Alberto Abadie and Guido W Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *econometrica* 74, 1 (2006), 235–267.
- [2] Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22, 1 (2008), 31–72.
- [3] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. 2010. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 7–16.
- [4] Yale Chang and Jennifer G Dy. 2017. Informative Subspace Learning for Counterfactual Inference. In *AAAI*. 1770–1776.
- [5] Zhixuan Chu, Stephen L Rathbun, and Sheng Li. 2020. Matching in Selective and Balanced Representation Space for Treatment Effects Estimation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 205–214.
- [6] Marie Cottrell, Madalina Olteanu, Fabrice Rossi, and Nathalie Villa-Vialaneix. 2018. Self-Organizing Maps, theory and applications. (2018).
- [7] Amy Finkelstein, Annetta Zhou, Sarah Taubman, and Joseph Doyle. 2020. Health care hotspotting—a randomized, controlled trial. *New England Journal of Medicine* 382, 2 (2020), 152–162.
- [8] Florent Forest, Mustapha Lebbah, Hanene Azzag, and Jérôme Lacaille. 2019. Deep Embedded SOM: Joint Representation Learning and Self-Organization. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)* (2019).
- [9] Alexander Gepperth and Benedikt Pfübl. 2020. A Rigorous Link Between Self-Organizing Maps and Gaussian Mixture Models. In *International Conference on Artificial Neural Networks*. Springer, 863–872.
- [10] Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.
- [11] Simon Haykin. 2007. *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc.
- [12] Xiaofei He and Partha Niyogi. 2004. Locality preserving projections. In *Advances in neural information processing systems*. 153–160.
- [13] James J Heckman, Hidehiko Ichimura, and Petra Todd. 1998. Matching as an econometric evaluation estimator. *The review of economic studies* 65, 2 (1998),

³<https://github.com/oddrose/cfrnet>

Figure 5: SOM visualization on the testing subset of Lalonde for two different models. The global matching scores for the left and right figure are 237.2 and 149.4, respectively. The size of the grid points is proportional to the number of units (orange treatment and blue control). For most grid points, there are ample control units to match the treatment in the same cluster. But the left model has better matches than the right one since the number of treatment units which do have not exact matched controls for the right model is greater than that in the left one. The treatment units that are not matched with the same cluster circled in green in both figures.

- 261–294.
- [14] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.
 - [15] Stefano M Iacus, Gary King, and Giuseppe Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political analysis* (2012), 1–24.
 - [16] Kosuke Imai and Marc Ratkovic. 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2014), 243–263.
 - [17] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. 3020–3029.
 - [18] Ian T. Jolliffe. 2002. *Principal Component Analysis*. John Wiley and Sons.
 - [19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
 - [20] Teuvo Kohonen. 1982. Analysis of a simple self-organizing process. *Biological cybernetics* 44, 2 (1982), 135–140.
 - [21] Teuvo Kohonen. 1999. Comparison of SOM point densities based on different criteria. *Neural Computation* 11, 8 (1999), 2081–2095.
 - [22] Martin A Kraaijveld, Jianchang Mao, and Anil K Jain. 1995. A nonlinear projection method based on Kohonen’s topology preserving maps. *IEEE transactions on neural networks* 6, 3 (1995), 548–559.
 - [23] Sheng Li and Yun Fu. 2017. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*. 929–939.
 - [24] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns. In *IJCAI*. 3768–3774.
 - [25] S Lokesh, Priyan Malarvizhi Kumar, M Ramya Devi, P Parthasarathy, and C Gokulnath. 2019. An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications* 31, 5 (2019), 1521–1531.
 - [26] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*. 6446–6456.
 - [27] Weipeng Lu and Xuefeng Yan. 2020. Deep fisher autoencoder combined with self-organizing map for visual industrial process monitoring. *Journal of Manufacturing Systems* 56 (2020), 241–251.
 - [28] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.
 - [29] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
 - [30] Helge Ritter, Thomas Martinetz, Klaus Schulten, Daniel Barsky, Marcus Tesch, and Ronald Kates. 1992. *Neural computation and self-organizing maps: an introduction*. Addison-Wesley Reading, MA.
 - [31] Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics* (1951), 400–407.
 - [32] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
 - [33] Donald B Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics* (1978), 34–58.
 - [34] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
 - [35] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).
 - [36] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.
 - [37] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*. 2507–2517.
 - [38] Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.
 - [39] Kohonen Teuvo. 1995. Self-organizing maps, volume 30 of Springer series in information sciences.
 - [40] Jing Tian, Michael H Azarian, and Michael Pecht. 2014. Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *Proceedings of the European Conference of the Prognostics and Health Management Society*. Citeseer, 1–9.
 - [41] Juha Vesanto. 1999. SOM-based data visualization methods. *Intelligent data analysis* 3, 2 (1999), 111–126.
 - [42] Aolin Wang, Roch A Nianogo, and Onyebuchi A Arah. 2017. G-computation of average treatment effects on the treated and the untreated. *BMC medical research methodology* 17, 1 (2017), 1–5.
 - [43] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology* 63, 8 (2010), 826.
 - [44] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*. 2633–2643.
 - [45] Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. 2019. On the Estimation of Treatment Effect with Text Covariates. In *IJCAI*. 4106–4113.
 - [46] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.