Estimation of Single and Synergistic Treatment Effects under Multiple Treatments with Deep Neural Networks

Yuki Murakami
yuuki.murakami.xg@nttdocomo.com
NTT DOCOMO,INC.
Tokyo, Japan

Kohsuke Kubota
kousuke.kubota.xt@nttdocomo.com
NTT DOCOMO,INC.
Tokyo, Japan

Takumi Hattori
takumi.hattori.zw@nttdocomo.com
NTT DOCOMO,INC.
Tokyo, Japan

Keiichi Ochiai
ochiai.kei.dk@yokohama-cu.ac.jp
Yokohama City University
Kanagawa, Japan

Abstract

The simultaneous application of multiple treatments is increasingly common in many fields, such as healthcare and marketing. In such scenarios, it is important to estimate not only the effect of each single treatment effect, but also the synergistic treatment effects that arise from combinations of treatments. Previous studies have proposed methods that combine a variational autoencoder with a task embedding network, which captures treatment similarities for multi-treatment causal inference. These methods assume the presence of unobserved covariates and regard observed data as proxies for those unobserved covariates. As a result, they may still learn unnecessary latent variables even when the covariates are observed. This model misspecification can lead to misleading estimates of causal effects. To address this issue, we propose a novel deep learning framework that simultaneously captures both single and synergistic treatment effects and mitigates selection bias, using a task embedding network and a representation learning network with the balancing penalty. The task embedding network ensures that similar treatments yield similar representations and outcomes, improving the estimation of both single and synergistic effects. The representation learning network with the balancing penalty directly learns representations from observed covariates and controls distributional differences across treatment patterns using Integral Probability Metrics, thereby reducing the risk of model misspecification due to erroneous latent structures. We evaluate our method using multiple simulation datasets and compare its performance with existing baselines. Our method consistently outperforms baselines by reducing estimation errors in both single and synergistic treatment effects across settings.

CCS Concepts

• **Computing methodologies → Machine learning**.

Keywords

Causal Inference, Multiple Treatments, Deep Neural Networks, Balancing Representations

ACM Reference Format:
Yuki Murakami, Kohsuke Kubota, Takumi Hattori, and Keiichi Ochiai. 2025. Estimation of Single and Synergistic Treatment Effects under Multiple Treatments with Deep Neural Networks. In *Proceedings of KDD'25*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In many fields, such as healthcare and marketing, it is increasingly common for multiple treatments to be applied simultaneously, and estimating both single and synergistic treatment effects is critically important. For example, healthcare strategies often involve the concurrent administration of multiple drugs, potentially resulting in complex interactions and changes in side effects [7, 18, 30]. Similarly, in marketing, companies frequently implement several promotional campaigns simultaneously, where the overall impact cannot be captured by simply summing the effects of each campaign [4, 12, 13, 20]. Misestimating such multiple treatment effects can have a significant impact on decision-making in both domains, highlighting the need for methods that can accurately identify both single and synergistic treatment effects.

Most existing causal inference methods are designed for single treatment settings, and naive extensions to multiple treatments cannot accurately estimate single and synergistic treatment effects [14, 29]. For example, methods that assume each treatment is applied in isolation and do not account for situations where a unit receives multiple treatments simultaneously [8, 19] fail to incorporate the possibility of synergistic effects and, therefore, cannot distinguish between single and synergistic treatment effects. MEMENTO [19] is a deep learning-based method, but its model architecture does not account for the simultaneous application or interaction of treatments. As a result, the estimable causal effects are limited to single treatment effects.

Existing methods for estimating single and synergistic treatment effects still pose structural limitations that may lead to reduced performance in estimating causal effects [1, 22]. Specifically, linear regression with interaction terms can theoretically capture synergistic effects [6], although this approach is highly sensitive to model misspecification and lacks the flexibility to model complex, non-linear interactions between treatments. To overcome this

KDD'25, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

limitation, deep neural network-based methods such as Neural Counterfactual Relation Estimation (NCoRE) [21] and Task Embedding–based Causal Effect Variational Autoencoder (TECE-VAE) [24] have been proposed. NCoRE addresses treatment interactions using separate outcome prediction networks and interaction subnetworks activated only when multiple treatments are simultaneously applied. However, it lacks a structure that supports parameter sharing across similar treatments or combinations, leading to unstable estimates, especially when data is limited. TECE-VAE combines a treatment similarity-aware task embedding network with a VAE. This approach assumes the presence of latent covariates and treats observed covariates as proxies for these latent variables. Even when true covariates are observed, the model still infers latent covariates, increasing the risk of misspecification and degrades estimation performance.

In this study, we propose a novel deep learning framework comprising a task embedding network that captures treatment similarity and a representation learning network with the balancing penalty that mitigates model misspecification and selection bias. The task embedding network learns to assign similar embedding vectors to similar treatments, enabling information sharing across treatment patterns based on their similarity. The representation learning network with the balancing penalty learns representations nonparametrically from the observed covariates while adjusting the representation distributions to be aligned across different treatment patterns, thereby suppressing selection bias. Unlike TECE-VAE, this data-driven flexibility removes the coercion to infer latent covariates, thereby reducing the risk of accuracy degradation caused by model misspecification.

We evaluate our method using three simulation datasets and show that our proposed method consistently outperforms baseline methods in estimating both single and synergistic treatment effects, achieving the lowest estimation errors across all settings. These results demonstrate the robustness and generalizability of our method in diverse data-generating processes.

2 Related Work

Existing deep learning approaches to causal inference can broadly be categorized into two groups. The first group consists of representation learning-based approaches, which learn feature representations directly from observed covariates and predict counterfactual outcomes. The second group comprises deep generative model-based approaches, which assume the existence of latent covariates and model the entire data-generating process using deep generative models.

In the first group, MEMENTO [19] and NCoRE [21] extend models designed for single-treatment settings, such as Treatment-Agnostic Representation Network (TARNet) and Counterfactual Regression (CFR) [25], to the multi-treatment setting by assigning outcome prediction networks to each treatment. TARNet learns a shared representation that is independent of treatment and uses it to predict counterfactual outcomes. CFR extends TARNet by introducing a balancing penalty based on Integral Probability Metrics (IPM) [27], which aligns the representation distributions of the treated and control groups to mitigate selection bias. MEMENTO, based on CFR, employs a separate outcome prediction network for each treatment. NCoRE supports the estimation of both single and synergistic treatment effects by introducing interaction subnetworks that are activated only when multiple treatments are applied simultaneously. Similarly to MEMENTO, this model assigns an independent outcome prediction network to each treatment. Within each treatment arm, interaction subnetworks are incorporated to capture the interactions between the corresponding treatment and other treatments.

However, although these methods have been extended to estimate causal effects in multi-treatment settings, they face challenges when estimating synergistic treatment effects in scenarios where units receive multiple treatments simultaneously. Since MEMENTO does not assume that multiple treatments can be applied simultaneously, it lacks the architectural components necessary to capture interactions among treatments and thus cannot estimate synergistic treatment effects. While NCoRE introduces structures to model such effects, its outcome prediction networks and interaction subnetworks are trained only with samples corresponding to a specific treatment pattern, and there is no parameter sharing across networks. This lack of shared parameters prevents information sharing across similar treatments and often leads to unstable estimation, especially for infrequent treatment patterns.

In the second group of deep generative approaches, TECE-VAE [24] introduces a task embedding network to the models designed for single-treatment settings, Causal Effect VAE [15], allowing the estimation of causal effects with multiple treatments. These methods aim to address selection bias in situations where observed covariates are proxies for unobserved latent covariates. These methods address settings where standard covariate adjustment methods [2], including propensity score matching [28], are inadequate due to the presence of unobserved covariates. Both models assume the existence of latent covariates and that all observed data are proxies for these variables. The VAE decoder is used to generate latent covariates, while the encoder performs an approximate inference of the outcome distribution conditioned on latent covariates. In TECE-VAE, a task embedding network is incorporated into this structure to enable the modeling of synergistic effects.

However, since these approaches inherently assume the presence of latent covariates, this assumption can lead to their models being misspecified if such covariates do not exist. Even when latent covariates are absent or observed data sufficiently explain the covariates, these models still attempt to estimate unnecessary latent covariates. When the observed data structure does not align with the assumptions of the model, this mismatch can lead to poor estimation performance [22].

Our proposed model addresses the primary limitations of both groups by incorporating a task embedding network, which captures treatment similarity, along with a representation learning network equipped with a balancing penalty that mitigates accuracy loss due to model misspecification. From the representation learning perspective, the similarity of the treatment is captured through the task embedding network and exploited via a single outcome prediction network, allowing parameter sharing across treatment patterns. This design improves estimation stability even in data-sparse regimes. From the generative-modeling perspective, the model forgoes explicit latent-covariate assumptions and instead learns balanced representations directly from observed proxies, while the penalty attenuates selection bias. This design mitigates the risk of misspecification-induced accuracy degradation, thereby supporting robust causal inference across various covariate settings.

3 Preliminaries

We formulate the causal inference problem under multiple treatments within the potential outcome framework [23]. The goal is to estimate the single and synergistic treatment effects of multiple binary treatments on a continuous outcome using observed covariates.

We consider N independent units, where $i = 1, \dots, N$. For each unit i, we observe a covariate vector $\boldsymbol{x}_i \in \mathbb{R}^d$ drawn from the covariate space $\boldsymbol{X}$. We assume that the total number of treatments is K and the set of all possible treatment patterns is $\boldsymbol{T} \in \{0,1\}^K$. For a particular treatment vector $\boldsymbol{t} \in \boldsymbol{T}$, the potential outcome for unit i is denoted by $Y_i(\boldsymbol{t}) \in \mathbb{R}$.

To identify causal effects under the potential outcome framework, we adopt the following three assumptions commonly used in observational studies [9]. These assumptions ensure that potential outcomes are identifiable from the observed data.

Assumption 1. (Ignorability.) *For any treatment pattern, the potential outcome is independent of the assigned treatment T given the observed covariates X.*

Assumption 2. (Overlap.) *Every unit has a non-zero probability of receiving any treatment pattern given their observed covariates.*

Assumption 3. (Stable Unit Treatment Value Assumption) *(1) no interference, meaning that the outcome of one unit is unaffected by the treatment assignments of other units; and (2) consistency of treatment, meaning that the potential outcomes correspond to well-defined and unique treatments.*

We define the single average treatment effect (S-ATE) and the synergistic average treatment effect (Sy-ATE) to quantify the individual and combined impacts of multiple treatments. These definitions are mathematically equivalent to the notions of main effects and interaction effects in a 2^K factorial design, respectively [5, 6]. In particular, our definition of the Sy-ATE structurally corresponds to interaction effects in a factorial design, as both are formulated as weighted linear combinations of potential outcomes across treatment patterns. Given covariates $\boldsymbol{x}$ and a treatment vector $\boldsymbol{t}$, we denote the conditional expected outcome by

$$\mu(\boldsymbol{x}, \boldsymbol{t}) = \mathbb{E}[Y(\boldsymbol{t}) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{1}$$

Let $\boldsymbol{t}_{+k}$ denote the one-hot treatment vector whose kth entry equals one, and all other entries equal zero. The S-ATE for treatment k, denoted by $\tau_{\textbf{S-ATE}}(k)$, is defined as

$$\tau_{\textbf{S-ATE}}(k) = \mathbb{E}_{\boldsymbol{x}}\left[\mu(\boldsymbol{x}, \boldsymbol{t}_{+k}) - \mu(\boldsymbol{x}, \boldsymbol{0})\right]. \tag{2}$$

For any subset $S \in \{S' \subseteq \{1, \dots, K\} \mid |S'| \geq 2\}$, the Sy-ATE for the treatment combination S, denoted by $\tau_{\textbf{Sy-ATE}}(S)$, is defined as

$$\tau_{\textbf{Sy-ATE}}(S) = \mathbb{E}_{\boldsymbol{x}}\left[\sum_{Q \subseteq S} (-1)^{|S|-|Q|} \mu\left(\boldsymbol{x}, \boldsymbol{t}_{(+Q)}\right)\right], \tag{3}$$

where $t_{(+Q)}$ is the treatment vector that sets the components indexed by Q to one and all remaining components to zero. For example, $K = 2$ and $S = \{1, 2\}$, $\tau_{\textbf{Sy-ATE}}(\{1,2\}) = \mathbb{E}_{\boldsymbol{x}}[\mu(\boldsymbol{x}, (1,1)) - \mu(\boldsymbol{x}, (1,0)) - \mu(\boldsymbol{x}, (0,1)) + \mu(\boldsymbol{x}, (0,0))]$.

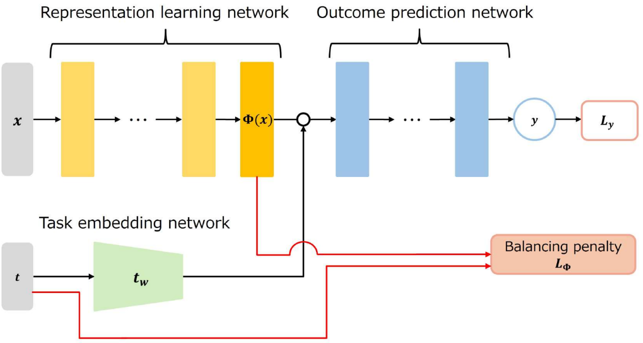

Figure 1: The architecture of the proposed model consists of three components: the representation learning network (yellow), the task embedding network (green), and the outcome prediction network (blue). The latent representation $\Phi(x)$ is concatenated with the task embedding vector $t_w(\boldsymbol{t})$ to predict the outcome y. The model is trained with two loss terms: the prediction loss L_y and the balancing penalty L_Φ (red).

4 Proposed Method

4.1 Model Architecture

The objective of our model is to learn $\mu(\boldsymbol{x}, \boldsymbol{t})$ and compute the causal effects defined in Equations (2) and (3) for any treatment vector $\boldsymbol{t}$. The architecture of the proposed model for estimating $\hat{\mu}(\boldsymbol{x}, \boldsymbol{t})$ is illustrated in Figure 1. The model consists of three main components: (1) a representation learning network with the balancing penalty, (2) a task embedding network, and (3) an outcome prediction network. These components are jointly optimized end-to-end. Based on balanced covariate representations and treatment embeddings, the model aims to accurately estimate counterfactual outcomes under multiple treatments.

First, the representation learning network with the balancing penalty maps the observed covariates $\boldsymbol{x} \in \mathbb{R}^d$ to a latent representation space suitable for estimating causal effects. This network is designed to reduce the influence of selection bias by removing noise and redundant information from the covariates and by encouraging alignment of the representation distributions in different treatment patterns. Specifically, it learns a function $\Phi : \mathbb{R}^d \to \mathbb{R}^p$, where p is a hyperparameter indicating the dimensionality of the learned representations. An IPM-based balancing penalty [25] is applied to the learned representations to minimize the distance between the distributions associated with different treatment patterns, thus mitigating the selection bias arising from treatment assignment.

Next, the task embedding network captures similarities among treatment patterns by mapping the binary treatment vector $\boldsymbol{t}$ into a q-dimensional continuous task embedding vector via a multi-layer perceptron MLP_w. The output is denoted by $t_w(\boldsymbol{t}) = \text{MLP}_w(\boldsymbol{t}) \in \mathbb{R}^q$. This design ensures that treatment patterns with similar content and effects are placed close to each other in the embedding space. Consequently, the model learns to associate similar patterns with similar representations, which allows the outcome prediction network to generalize the learned parameters across similar treatment patterns. As a result, even for treatment combinations with limited data, the model can leverage shared information from related tasks, leading to more efficient and stable estimation of single and synergistic treatment effects.

Finally, the outcome prediction network takes the concatenated vector of $\Phi(\boldsymbol{x})$ and $t_w(\boldsymbol{t})$ as input, forming a $(p+q)$-dimensional representation, and outputs the predicted outcome Y using a neural network $h : \mathbb{R}^{p+q} \to \mathbb{R}$. Unlike previous approaches that use separate networks for each treatment[19, 21], our architecture employs a single shared prediction function h for all treatment patterns. This design enables the sharing of structural parameters, which facilitates consistent learning across diverse treatment patterns.

Our model predicts counterfactual outcomes by directly inputting the treatment vector of interest and estimates causal effects based on the predicted outcomes. Through training, the proposed model learns a function $\hat{\mu}(\boldsymbol{x}, \boldsymbol{t})$ composed of a representation network Φ, a task embedding network t_w, and an outcome prediction network h. Instead of using the actual observed treatment vector $\boldsymbol{t}$, a counterfactual treatment vector $\boldsymbol{t}'$ is fed into $\hat{\mu}(\boldsymbol{x}, \boldsymbol{t})$, and the causal effect is estimated by computing the difference in predicted outcomes according to Equation (2) and (3).

4.2 Objective function

The objective function of our proposed method is designed to simultaneously address two key challenges in causal effect estimation: maximizing outcome prediction accuracy and correcting distributional imbalances caused by selection bias. Optimizing only for the former can result in biased counterfactual predictions while focusing solely on the latter can compromise the expressive capacity of the model. To address this trade-off and overfitting, we design the loss function L as the sum of three components:

$$L = L_y + \alpha L_\Phi(\Phi, \boldsymbol{t}) + \beta \|w\|_2, \tag{4}$$

where the first term L_y represents the outcome prediction error, the second term $L_\Phi(\cdot,\cdot)$ is a balancing penalty that reduces distributional differences in the representation space across treatment patterns, and the third term is an L2 regularization term applied to the network weights. The coefficients α and β are hyperparameters that control the strength of the corresponding components.

Minimizing the outcome prediction error L_y is essential for an accurate estimation of causal effects. However, when the frequency of observed treatment patterns is imbalanced, certain treatment patterns may be underrepresented, resulting in biased predictions. To mitigate this issue, we introduce a correction based on the empirical frequency of each treatment pattern:

$$w_i(\boldsymbol{t}_i) = \frac{1}{2}\left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[\boldsymbol{t}_i = \boldsymbol{t}_j]\right)^{-1}, \tag{5}$$

$$L_y = \frac{1}{N}\sum_{i=1}^{N} w_i(\boldsymbol{t}_i)(y_i - \hat{y}_i)^2, \tag{6}$$

where $\hat{y}_i$ denotes the predicted outcome for unit i, and $w_i(\boldsymbol{t}_i)$ is the inverse of the relative frequency of the treatment $\boldsymbol{t}_i$ in the training data.

The second term L_Φ is designed to suppress selection bias by aligning the representation distributions in all treatment patterns. It is defined by computing the IPM between the representation distributions of each pair of distinct treatment patterns and averaging these distances. The use of IPM allows us to effectively capture non-linear discrepancies between probability distributions, and thus contributing to better generalization in causal effect estimation [25, 26]:

$$L_\Phi = \frac{1}{\binom{|T|}{2}} \sum_{\{\boldsymbol{a},\boldsymbol{b}\}\in T \quad \boldsymbol{a}\neq\boldsymbol{b}} \mathrm{IPM}\left(\{\Phi(\boldsymbol{x}_i)_{i,\boldsymbol{t}_i=\boldsymbol{a}}\}, \{\Phi(\boldsymbol{x}_j)_{j,\boldsymbol{t}_j=\boldsymbol{b}}\}\right), \tag{7}$$

where $\Phi(x)$ is the representation network mapping the input covariates x into a latent representation space, and $\mathrm{IPM}(\cdot,\cdot)$ measures the discrepancy between two distributions. Here, $|T|$ denotes the number of distinct treatment patterns and $\binom{|T|}{2}$ is the total number of pairs of treatment patterns.

5 Experiments

To evaluate the effectiveness of the proposed method, we conducted comparative experiments using simulated datasets. The experiments focused on the performance of estimating single and synergistic treatment effects, and the results were compared with those of several existing methods. This section describes the experimental setup, the hyperparameters for the proposed and baseline models, and the evaluation metrics used in the analysis.

5.1 Simulation datasets

To evaluate the performance of the estimation of single and synergistic treatment effects, we generated simulation datasets under three distinct causal structures. The first scenario assumes that all covariates are directly observable. This scenario allows us to evaluate the estimation performance of the model in an ideal scenario where no unobserved (latent) covariates exist. The second scenario assumes that all covariates influencing treatment and outcome are latent and unobserved, reflecting the assumptions made by deep generative models [15, 24]. Specifically, it considers real-world situations in which true covariates, such as the economic status or lifestyle of units that affect both treatment assignment and outcomes, cannot be measured directly and only proxy variables such as residential area, occupation, or purchase history are observed. This allows a direct comparison with models designed to address latent covariates. The third scenario represents a more realistic hybrid setting where both observed and latent covariates coexist, effectively combining the structures of scenario 1 and scenario 2. This allows us to test the generalizability of our method in complex settings.

In all scenarios, the number of treatments K was set at three and the sample size N was set at 50,000. To ensure a fair comparison with baseline methods and mitigate bias caused by variability in data generation, we generated 100 datasets for each scenario using different random seeds and reported the average results.

Given the true covariates $\boldsymbol{x}_{i,\mathrm{true}}$, treatment assignment and outcome generation follow the same functional form in all scenarios. In the outcome generation function f, treatment patterns that share the same activated components tend to produce similar outcomes since their corresponding interaction terms contribute similarly to the overall effect. This introduces an inherent structural similarity across treatments.

$$t_{i,k} \sim \mathrm{Bern}\left(\sigma\left(\boldsymbol{w}_{t_k}^{\top}\boldsymbol{x}_{i,true}\right) - \delta\right), \quad k = 1, 2, 3,$$

$$f(\boldsymbol{x}_{i,\mathrm{true}}, \boldsymbol{t}_i) = \boldsymbol{w}_x^{\top}\boldsymbol{x}_{i,\mathrm{true}} + (x_{i,\mathrm{true}}^{(1)} + 1)t_{i,1} + 1.2(x_{i,\mathrm{true}}^{(2)} + 1)t_{i,2}$$

$$+0.8(x_{i,\mathrm{true}}^{(3)} + 1)t_{i,3} + (x_{i,\mathrm{true}}^{(4)} + 0.5)t_{i,1}t_{i,2} - 0.5(x_{i,\mathrm{true}}^{(5)} + 1)t_{i,1}t_{i,3}+$$

$$0.1(x_{i,\mathrm{true}}^{(6)} + 1)t_{i,2}t_{i,3} + 0.7x_{i,\mathrm{true}}^{(7)}t_{i,1}t_{i,2}t_{i,3} + 2,$$

$$Y_i \sim N\left(f(\boldsymbol{x}_{i,true}, \boldsymbol{t}_i),\ 1^2\right),$$

where $\sigma(x) = 1/(1 + \exp(-x))$, and δ is a set of bias parameters depending on the simulation setting. Vectors $\boldsymbol{w}_{t_k}$ and $\boldsymbol{w}_x$ are weight vectors whose elements are independently drawn from the uniform distribution $U(-1, 1)$, and used for treatment assignment and outcome generation, respectively.

In all scenarios, only the observed covariates $\boldsymbol{x}_{i,\mathrm{obs}}$ are available to the model during training and evaluation. The true covariates $\boldsymbol{x}_{i,\mathrm{true}}$ are solely used to generate treatments and outcomes and are not accessible at inference time. We vary the structure of the observed covariates $\boldsymbol{x}_{i,\mathrm{obs}}$ and the true covariates $\boldsymbol{x}_{i,\mathrm{true}}$ as follows:

Simulation dataset 1 (observed covariates only).

$$x_{i,n}^{(j)} \sim N(c_n^{(j)}, 1^2), \quad x_{i,u}^{(j)} \sim U(-10, 10), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(p_b^{(j)}),$$

$$\boldsymbol{x}_{i,\mathrm{obs}} = \boldsymbol{x}_{i,\mathrm{true}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}),$$

where $j \in \{1, \ldots, 10\}$ and $c_n^{(j)}$ are drawn from the uniform distribution $U(-1, 1)$. Similarly, each $p_b^{(j)}$ is sampled from $U(0, 1)$. In simulation dataset 1, $\delta = 0.3$.

Simulation dataset 2 (proxy variables only).

$$z_i \sim N(\boldsymbol{0}, I_{10}), \quad x_{i,n}^{(j)} \sim N(\boldsymbol{w}_n^{(j)\top} z_i, 1^2),$$

$$x_{i,u}^{(j)} \sim N(\boldsymbol{w}_u^{(j)\top} z_i, 5^2), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(\sigma(\boldsymbol{w}_b^{(j)\top} z_i)),$$

$$\boldsymbol{x}_{i,\mathrm{obs}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}), \quad \boldsymbol{x}_{i,\mathrm{true}} = z_i,$$

where $j \in \{1, \ldots, 10\}$ and the vectors $\boldsymbol{w}_n^{(j)}$, $\boldsymbol{w}_u^{(j)}$, and $\boldsymbol{w}_b^{(j)}$ are weight vectors whose elements are drawn independently from $U(-1, 1)$. I_d denotes the identity matrix $d \times d$, which is also used as the variance-covariance matrix. In simulation dataset 2, $\delta = 0.2$.

Simulation dataset 3 (observed covariates and latent covariates).

$$z_i \sim N(\boldsymbol{0}, I_{15}), \quad x_{i,p}^{(l)} \sim N(\boldsymbol{w}_p^{(l)\top} z_i, 1^2),$$

$$x_{i,n}^{(j)} \sim N(c_n^{(j)}, 1^2), \quad x_{i,u}^{(j)} \sim U(-10, 10), \quad x_{i,b}^{(j)} \sim \mathrm{Bern}(p_b^{(j)}),$$

$$\boldsymbol{x}_{i,\mathrm{obs}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}, \boldsymbol{x}_{i,p}), \quad \boldsymbol{x}_{i,\mathrm{true}} = (\boldsymbol{x}_{i,n}, \boldsymbol{x}_{i,u}, \boldsymbol{x}_{i,b}, z_i),$$

where $l \in \{1, \ldots, 15\}$ and $j \in \{1, \ldots, 5\}$. Vectors $\boldsymbol{w}_p^{(l)}$ are weight vectors whose elements are drawn independently from $U(-1, 1)$. In simulation dataset 3, $\delta = 0.1$.

5.2 Implementation Details

Our proposed method consists of three neural networks, all of which are built using fully connected (FC) layers with 200 units per hidden layer and leaky ReLU activation [16]. Φ and t_w each consist of two hidden layers, while h has three. The task embedding network outputs a five-dimensional embedding vector. The balancing penalty coefficient α was set to one, and the IPM used in the penalty was the Wasserstein distance [27].

To enable a comprehensive comparison between methods that extend single treatment models to multiple treatment settings and those specifically designed for multiple treatments, we selected four baseline methods: TARNet [25], CFR with Wasserstein-based balancing (CFR-WASS)[25], TECE-VAE[24], and NCoRE [21]. To adapt TARNet and CFR-WASS to the multi-treatment setting, we constructed a separate inference network for each of the 2^K treatment patterns. In CFR-WASS, the penalty coefficient α was also fixed at one. In TECE-VAE, the latent dimension was set to 25, and the task embedding network had two hidden layers with 200 units and ELU activation [3], producing a five-dimensional embedding.

All models were trained with Adam optimizer [11], using a learning rate of 10^{-5}, a batch size of 128, and an L2 regularization of 10^{-5}. Training was carried out for 30 epochs. Each dataset was divided into training sets 70% and test sets 30%, and all evaluations were carried out on the test set.

5.3 Evaluation Metrics

We evaluate the estimation performance of single and synergistic treatment effects using two metrics: the S-ATE error for individual treatment effects and the Sy-ATE error for interaction effects. These are defined analogously to the absolute ATE estimation error commonly used in single-treatment studies [10], computed as the absolute difference between the true and estimated S-ATE or Sy-ATE.

$$\text{S-ATE Error}(k) = |\tau_{\textbf{S-ATE}}(k) - \hat{\tau}_{\textbf{S-ATE}}(k)|, \quad k \in \{1, \ldots K\}, \tag{8}$$

$$\text{Sy-ATE-Error}(S) = \left|\tau_{\textbf{Sy-ATE}}(S) - \hat{\tau}_{\textbf{Sy-ATE}}(S)\right|, \tag{9}$$

where S denotes a subset of treatments with $|S| \geq 2$. $\hat{\tau}_{\textbf{S-ATE}}(\cdot)$ and $\hat{\tau}_{\textbf{Sy-ATE}}(\cdot)$ denote the estimated S-ATE and Sy-ATE.

6 Results

Table 1 presents the estimation errors of the proposed and baseline methods in three types of simulation datasets. In all datasets, the proposed method achieves the lowest estimation errors, demonstrating its superior ability to accurately estimate single and synergistic treatment effects compared to existing methods.

While some baseline methods perform competitively on S-ATE or Sy-ATE estimation, none maintain strong performance on both. In simulation dataset 1, NCoRE demonstrates strong performance in the estimation of S-ATE, while TECE-VAE performs well in the estimation of Sy-ATE. However, NCoRE exhibits large Sy-ATE errors, indicating limitations in capturing synergistic effects without parameter sharing. TECE-VAE, on the other hand, exhibits large S-ATE errors, indicating degraded estimation performance due to model misspecification. TARNet and CFR-WASS consistently underperform in Sy-ATE, suggesting that naive extensions of models designed for single-treatment settings are inadequate for multi-treatment scenarios. In simulation datasets 2 and 3, which are specifically designed to reflect the assumptions of TECE-VAE, TECE-VAE shows relatively low Sy-ATE errors for certain treatment combinations, while its performance degrades in S-ATE estimation. These results highlight the robustness of the proposed method in diverse data-generating processes.

Table 1: Comparison of mean estimation errors and standard deviation for S-ATE and Sy-ATE across multiple simulation datasets. Here, $k \in \{1, 2, 3\}$ indexes individual treatments for S-ATE, and $S \subseteq \{1, 2, 3\}$ (with $|S| \geq 2$) denotes treatment combinations for Sy-ATE.

Sim. No.	Method	S-ATE Error			Sy-ATE Error			
		$k = 1$	$k = 2$	$k = 3$	$S = \{1, 2\}$	$S = \{2, 3\}$	$S = \{1, 3\}$	$S = \{1, 2, 3\}$
1	TARNet	0.11 ± .086	0.12 ± .086	0.11 ± .086	0.17 ± .13	0.17 ± .13	0.17 ± .12	0.25 ± .19
	CFR-WASS	0.14 ± .11	0.13 ± .11	0.12 ± .11	0.21 ± .14	0.19 ± .15	0.19 ± .15	0.27 ± .19
	NCoRE	**0.10 ± .072**	0.12 ± .089	0.11 ± .077	0.13 ± .12	0.16 ± .11	0.14 ± .10	0.19 ± .14
	TECE-VAE	0.23 ± .19	0.26 ± .23	0.20 ± .17	**0.10 ± .086**	0.11 ± .083	0.091 ± .074	0.11 ± .085
	Proposed	**0.10 ± .071**	**0.11 ± .094**	**0.095 ± .078**	**0.10 ± .074**	**0.092 ± .071**	**0.082 ± .060**	**0.094 ± .095**
2	TARNet	**0.17 ± .15**	0.19 ± .16	**0.16 ± .15**	0.19 ± .12	0.18 ± .12	0.17 ± .12	0.29 ± .22
	CFR-WASS	**0.17 ± .14**	0.19 ± .17	0.17 ± .16	0.19 ± .15	0.18 ± .14	0.21 ± .14	0.33 ± .26
	NCoRE	0.19 ± .17	0.21 ± .19	0.19 ± .18	0.23 ± .17	0.22 ± .16	0.20 ± .17	0.28 ± .22
	TECE-VAE	0.19 ± .15	0.22 ± .16	0.20 ± .16	0.32 ± .21	0.16 ± .14	0.17 ± .14	**0.22 ± .17**
	Proposed	**0.17 ± .16**	**0.18 ± .16**	**0.16 ± .14**	**0.18 ± .15**	**0.14 ± .11**	**0.16 ± .11**	**0.22 ± .16**
3	TARNet	0.40 ± .29	0.38 ± .25	0.35 ± .26	0.28 ± .19	0.24 ± .19	0.28 ± .22	0.63 ± .41
	CFR-WASS	0.37 ± .26	0.36 ± .26	0.33 ± .24	0.29 ± .24	0.31 ± .23	0.29 ± .24	0.80 ± .56
	NCoRE	0.43 ± .34	0.34 ± .30	0.31 ± .23	0.31 ± .21	0.31 ± .24	0.31 ± .21	0.77 ± .55
	TECE-VAE	0.33 ± .21	0.40 ± .20	0.27 ± .16	0.27 ± .22	0.25 ± .21	0.22 ± .18	0.69 ± .54
	Proposed	**0.20 ± .16**	**0.23 ± .17**	**0.19 ± .15**	**0.21 ± .16**	**0.22 ± .16**	**0.16 ± .11**	**0.61 ± .41**

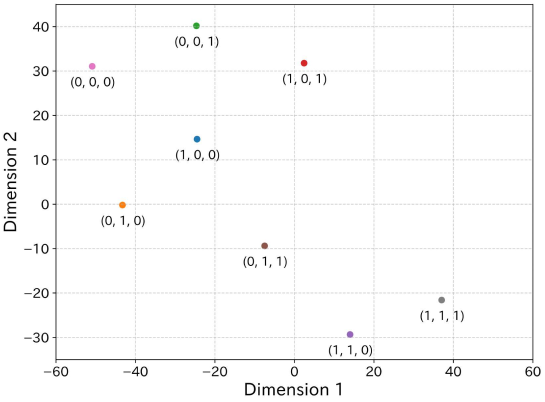

Figure 2: Visualization of task embedding vector using t-SNE

7 Discussion

Based on the simulation results, the proposed model outperformed competing baseline methods in estimating single and synergistic treatment effects under various simulation conditions. The proposed model demonstrates its potential as a viable alternative to existing representation learning-based models and deep generative models.

The results shown in Table 1 suggest that the task embedding network, introduced to capture similarities between treatments, plays an important role in improving the performance of single and synergistic effects estimation. Figure 2 visualizes the task embedding vectors corresponding to different treatment patterns, projected into a two-dimensional space using t-distributed Stochastic Neighbor Embedding (t-SNE) [17]. In the figure, patterns that share common treatment components are located near each other, visually confirming that similar treatments are mapped to similar task embedding vectors. These results imply that the task embedding network effectively captures treatment similarity and that the learned embeddings contribute to improved estimation performance.

8 Conclusion

In this study, we proposed a novel deep learning framework for estimating single and synergistic treatment effects in scenarios involving the simultaneous application of multiple treatments. Our approach combines a task embedding network that captures similarities among treatments with a representation learning network with the balancing penalty that uses IPM to control distributional differences between treatment patterns. This proposed framework enables a stable estimation of causal effects.

Our simulation results demonstrate that the proposed model consistently outperforms competitive baselines in estimating causal effects under a wide range of experimental conditions. The proposed method is expected to be a promising analytical tool for practical applications such as evaluating the combined effects of drugs in medicine and analyzing the combined effects of multiple measures in marketing.

In future work, we plan to pursue three main directions. First, we will explore systematic strategies to select the optimal weight α used in the IPM-based balancing term. Second, we will apply our method to real-world datasets in the medical and marketing domains to verify its practical utility. Third, we aim to extend our framework to a doubly robust estimation setting, such as DR-Learner, by incorporating both propensity score modeling and outcome regression to enhance robustness against model misspecification.

References

[1] Matthew Blackwell and Michael P Olson. 2022. Reducing model misspecification and bias in the estimation of interactions. *Political Analysis* 30, 4 (2022), 495–514.

[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] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* (2015).

[4] Peter J Danaher, André Bonfrer, and Sanjay Dhar. 2008. The effect of competitive advertising interference on sales for packaged goods. *Journal of Marketing Research* 45, 2 (2008), 211–225.

[5] Tirthankar Dasgupta, Natesh S Pillai, and Donald B Rubin. 2015. Causal inference from 2K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 77, 4 (2015), 727–753.

[6] Naoki Egami and Kosuke Imai. 2019. Causal interaction in factorial experiments: Application to conjoint analysis. *J. Amer. Statist. Assoc.* (2019).

[7] Alan H Gradman, Jan N Basile, Barry L Carter, George L Bakris, American Society of Hypertension Writing Group, et al. 2010. Combination therapy in hypertension. *Journal of the American Society of Hypertension* 4, 2 (2010), 90–98.

[8] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[9] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge university press.

[10] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning.* PMLR, 3020–3029.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Katherine N Lemon and Stephen M Nowlis. 2002. Developing synergies between promotions and brands in different price-quality tiers. *Journal of Marketing Research* 39, 2 (2002), 171–185.

[13] Lisan Lesscher, Lara Lobschat, and Peter C Verhoef. 2021. Do offline and online go hand in hand? Cross-channel and synergy effects of direct mailing and display advertising. *International Journal of Research in Marketing* 38, 3 (2021), 678–697.

[14] Michael J Lopez and Roee Gutman. 2017. Estimation of causal effects with multiple treatments: a review and new ideas. *Statist. Sci.* (2017), 432–454.

[15] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems* 30 (2017).

[16] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, Vol. 30. Atlanta, GA, 3.

[17] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[18] Reza Bayat Mokhtari, Tina S Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. 2017. Combination therapy in combating cancer. *Oncotarget* 8, 23 (2017), 38022.

[19] Abhirup Mondal, Anirban Majumder, and Vineet Chaoji. 2022. Memento: Neural model for estimating individual treatment effects for multiple treatments. In *Proceedings of the 31st ACM international conference on information & knowledge management.* 3381–3390.

[20] Prasad A Naik and Kalyan Raman. 2003. Understanding the impact of synergy in multimedia communications. *Journal of marketing research* 40, 4 (2003), 375–388.

[21] Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. 2021. Ncore: Neural counterfactual representation learning for combinations of treatments. *arXiv preprint arXiv:2103.11175* (2021).

[22] Severi Rissanen and Pekka Marttinen. 2021. A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems* 34 (2021), 4207–4217.

[23] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association* 100, 469 (2005), 322–331.

[24] Shiv Kumar Saini, Sunny Dhamnani, Aakash, Akil Arif Ibrahim, and Prithviraj Chavan. 2019. Multiple Treatment Effect Estimation using Deep Generative Model with Task Embedding. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1601–1611. doi:10.1145/3308558.3313744

[25] 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.

[26] Claudia Shi, David M. Blei, and Victor Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. arXiv:1906.02120 [stat.ML] https://arxiv.org/abs/1906.02120

[27] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. 2010. Non-parametric estimation of integral probability metrics. In *2010 IEEE International Symposium on Information Theory.* IEEE, 1428–1432.

[28] 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.

[29] Yingrong Wang, Haoxuan Li, Minqin Zhu, Anpeng Wu, Ruoxuan Xiong, Fei Wu, and Kun Kuang. 2024. Causal Inference with Complex Treatments: A Survey. *arXiv preprint arXiv:2407.14022* (2024).

[30] Rachel M Webster. 2016. Combination therapies in oncology. *Nature reviews. Drug discovery* 15, 2 (2016), 81.