Causally Informed Factorization Machines

ABSTRACT

Factorization machines (FMs) are a class of general predictors for sparse data. One major benefit from FMs is their ability to capture the interactions across features when making recommendations. In this paper, we note that the interactions captured by existing FMs generally represent correlations in the data and we argue that such correlations, unless informed by the true causality structures underlying the data, may be spurious and may result in unwanted bias. To tackle this challenge, we propose a *Causally-Informed Factorization Machine (CIFM)* model that introduces a novel *causal injection* mechanism. CIFM leverages *a priori* causal knowledge, described in the form of a *causal graph*, to boost the representational ability of FMs and achieve better predictions. Specifically, given a (potentially learned) causal graph which describes the causal relationships among features, CIFM distills this structural information into a *pairwise causal impact matrix* and guides the learning process to ensure that the learned representations capture those relationships that are consistent with the causal relationships. Extensive evaluations of CIFM, along with its integrations with NeuralFM and DeepFM, conducted with synthetic and real-world data sets, demonstrate effectiveness of causal injection in generating better recommendations.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**.

KEYWORDS

Factorization Machines, Causality

ACM Reference Format:

. 2018. Causally Informed Factorization Machines. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Supervised learning is a fundamental tasks in understanding key patterns in data, constrained by user provided labels. The goal is to infer a function that predicts a target label, given data features as input. When the target labels are continuous, this is referred to as a regression task and when they are categorical, this is known as a classification task. Supervised learning has widespread applications in data-driven decision making, including recommendation systems [5] and online advertising [13].

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Factorization machine [25] is a supervised learning method that has been shown to be effective in handling the sparse data: Each feature is represented as a k-dimensional feature vector and the learned function takes into account contributions of the features as well as their pairwise interactions. Several FMs variants have been proposed to enhance the expressive power of the function being learned: neural FM (NFM [9]) utilizes multi-layer perceptron (MLP) to capture non-linear correlations among features; convolutional FM (CFM [35]) introduces an interaction cube by stacking outer product results from feature embeddings to capture such correlations; attentional FM (AFM [34]) leverages an attention mechanism to discriminate the importance of different feature interactions; while W2FM [15] abstracts interactions as additional affine transformations.

1.1 Shortcomings of Existing FM Models – Spurious Interactions

While different FM variants use different approaches to extract feature interactions, in all these cases, the learned interactions represent some form of correlation across features. *Correlation, however, is not causation* and the literature is full of examples (such as *Simpson's Paradox* [11]), where decision making purely based on correlations, without accounting for the underlying causal structures, can result in faulty outcomes. Therefore, in this paper, we argue that this *correlation-centric* approach of existing FM models is a weakness: unless informed by *causal* structures underlying the data, correlations may be spurious and may result in unwanted bias.

1.2 Causality to the Rescue – Causally-Informed Factorization Machine (CIFM)

Causality (which can better answer the question *why?* than correlation can), may not only be needed to better explain the recommendations made by FMs based on available data, but also to ensure that these recommendations are not based on faulty (spurious) statistics. Based on this premise, in this paper, we propose the concept of *Causally-Informed Factorization Machine (CIFM)* model, which introduces a novel *causal injection* mechanism. CIFM leverages *a priori* causal knowledge, described in the form of a *causal graph*, to boost the representational ability of FMs and achieve better predictions. Specifically, given a causal graph which describes the known causal relationships across features, CIFM distills this information into a *pairwise causal impact matrix* and guides the FM process to assure that the learned representations include those that are consistent with the underlying causal relationships.

1.3 Summary of Contributions

The main contributions of proposed *Causal Informed factorization machine (CIFM)* are as follows:

- To the best of our knowledge, this work is the first to disclose the causal interpretation of factorization machines (FMs) – in particular, in Section 4, we show that FMs can be interpreted as inherently discovering and taking into account hidden confounders while generating their recommendations.

- Based on this causal interpretation, we propose *Causally-Informed Factorization Machine (CIFM)* model to augment the expressiveness of FMs with the help of a causal graph – in this paper, we causally augment vanilla FM [25]; neural FM (NFM [9]) and deep FM (DFM [8]). Given a causal graph, CIFM measures pairwise causal relationships between features taking into account two pivotal considerations: a) *causal distance*, and b) *causal reinforcement*. Building upon these fundamental principles, we introduce a novel metric called *causal graph implied causal impact (CI for short)* to assess the pairwise causal interactions among variables *implied by a given causal graph*. CI accounts for not only the causal distance and causal reinforcement between pairs of variables, but it also takes into account the existence of certain causal structures (such as causal colliders) that might impose certain statistical anomalies unless they are properly accounted for.
- We propose *indirect* (Section 6.1) and *direct* (Section 6.2) causal injection mechanisms, which leverage the concept of pairwise *causal graph implied causal impact* information (introduced in Section 5) in different ways: **(a) indirect causal injection** incorporates causal knowledge by weighting the causal interactions learned by the FM in a post-operative fashion, while **(b) direct causal injection** modifies the constraints and objective functions underlying the FM training process that takes into account causal graph implied causal impact among pairs of variables.

By employing these and other causality-aware mechanisms, CIFM ensures that only meaningful interactions among variables have been accounted for, thus significantly enhancing predictive capabilities of the factorization machine. Empirical results, presented in Section 7[1], on synthetic and real world data sets show the effectiveness of CIFM.

2 RELATED WORK

2.1 Factorization Machines

Factorization machines (FMs) [25] were proposed as a way to take into account feature interactions to better handle sparse data; they consolidate advantages of support vector machines (SVMs) with factorization models. FMs are general predictors, widely used in recommendation systems. Various variants enhance FMs' expressiveness; these include Deep FM (DFM[8]) and Neural FM (NFM [9]), which deepen FMs under the neural framework to learn high-order feature interactions. Field-aware FM (FFM [13]) associates multiple embedding vectors for each feature to differentiate its interaction with other features for different fields. PNFM [3] recovers polynomial networks [18] and helps obtain higher order feature interactions. Attentional FM (AFM [34]) method utilizes a neural attention network to learn the importance of each feature interaction. Input-aware factorization machine (IFM [36]) enhances FMs by refining the weight and embedding vector of each feature taking into account different instances and, consequently, enhances nonlinearity.

For efficient training of high-order FMs, HOFM [2] leverages a dynamic programming algorithm for evaluating the ANOVA kernel and computing gradients. The recent Holographic FM (HFM [29]) approach replaces the inner product in FMs with a holographic reduced representation, whereas the convolutional FM (CFM [35]) introduces an interaction cube by stacking outer products of feature embeddings to capture correlations.

Note that, despite their various differences, all these models rely on correlation-based interactions among features and, as we argue and illustrate in this paper, this potentially has a negative impact on the prediction results as "*correlation is not causation*."

2.2 Causality

Causality is the relationship between the cause (treatment) and an effect (outcome) that gives rise to it [24] and, as such, is a topic of increasing interest in big data analysis [7], machine learning [12] [19], and reinforcement learning [1], among others. Most of the recent research on causality focuses on finding *causation* within data. Generally speaking, *causal inference* is the process of identifying the cause of a phenomenon, by establishing covariation of cause and effect [27] [4]. *Causal discovery* is finding *causal structure* (*causal graph*) by analyzing statistical properties of purely observational data [28] [30] [20]. Pearl has shown that a priori knowledge *causal graphs or causal structures* is critical in analyzing data[23], which capture such causal knowledge, can be used to avoid statistical pitfalls that correlation based techniques would face. Although we expect causal graphs depict causal relations within observed features in the data set, *unobserved confounder(s)* still might exist. An unobserved confounder is a variable that both affects the causal variables and outcome variables in the causal graph. Most of causal inference methods have strong assumption that we observe all confounders, but this assumption is untestable[10]. Deconfounder[32] is proposed to resolve the *unobserved confounder(s)* issue – it combines unsupervised learning and predictive model checking to use the dependencies among multiple causes as indirect evidence for some of the unobserved confounders.

2.3 Causal Recommender Systems

Recently, several researchers aimed to pose the recommendation task from a causal perspective, posing the recommendation problem as a causal inference problem. [6], for example, considered item exposure as treatment and user ratings as outcomes. Deconfounder [32, 33] extended this exposure-rating concept by modelling the exposure and using it as a substitute for unobserved confounders. In [16], user's social relations are leveraged to estimate the exposure along with propensity score to estimate exposure and reduce selection bias. [37] proposed an approach to leverage the good aspects of popularity bias and deconfound the bad aspects for improving recommendations. CASTLE [14] regularize recommendation process jointly learning the causal structure among variables.

One orthogonal work [17] proposes a personalized causal FM to address i.i.d. assumption violation in real world data and provide a robust recommendation by using confounder balancing regularization. But the work adopts the unconfounderness assumption [26] that there are no unobserved confounders in the data. Most of the above works concentrate on, and therefore, qre designed specifically for user-item relationships; in contrast, in this paper, we propose a general FM-based model to leverage arbitrary causal relationship graphs for the recommendation task. Secondly, most of them focus on causal discovery (as they formulate the recommendation task as a

[1] Implementations and datasets: https://anonymous.4open.science/r/cifm-DD5A/README.md

discovery task), while our aim to bring in *a priori* causal knowledge, whether discovered or expert provided, into the FM learning process.

3 PRELIMINARIES AND KEY NOTATIONS

In this section, we present the preliminary knowledge necessary for the development of causally-informed factorization machines (CIFM) and introduce key notations.

3.1 Factorization Machines

Factorization Machines (FMs [25]) enhance a linear prediction model learning by capturing pairwise interactions among features[2]. In this respect, they are similar to polynomial kernels (PKs):

$$\hat{F}_{PK}(X) = \underbrace{\vec{\rho}_0 + \vec{w}X^T}_{lin.\ predictor} + \underbrace{diag(XAX^T)}_{interaction\ term}, \tag{1}$$

where $X \in \mathbb{R}^{n\times m}$ is a data matrix where the n rows are the (transpose of) m-dimensional data vectors, $\vec{\rho}_0 \in \mathbb{R}^n$ is a vector where all entries have the same value $w_0 \in \mathbb{R}$, $\vec{w} \in \mathbb{R}^m$ is a weight vector, and $A \in \mathbb{R}^{m\times m}$ is a (symmetric) matrix that describes the degree of feature interaction, and $diag(X)$ represent the diagonal elements in matrix X:

- the first half of the model represents a linear predictor, consisting of a global bias and a linear transformation applied on the input data; whereas
- the second term shifts the prediction for each individual data point by an amount representing the interactions between its features – the effective result is that the boundary gets *warped* in a non-linear manner.

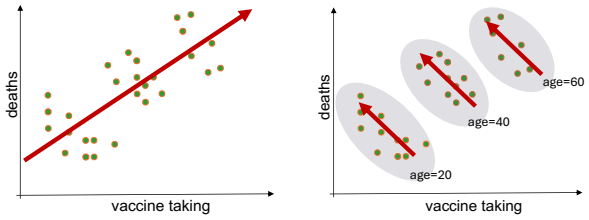
Note that, when considering each data vector individually, the above model can be written as

$$\hat{F}_{PK}(\vec{x}) = \underbrace{w_0 + \sum_{i=1}^{m} w_i x_i}_{lin.\ predictor} + \underbrace{\sum_{i=1}^{m}\sum_{j=i+1}^{m} A_{(i,j)} x_i x_j}_{warping\ term}. \tag{2}$$

Here x_i are the individual components of a real-valued input vector, $\vec{x} \in \mathbb{R}^m$; $x_i = 0$ when the i-th feature does not exist in the observation. The output of $\hat{F}_{PK}(\vec{x})$ is a scalar, representing the estimated target.

Unfortunately, the number of model parameters in Equation 2 is quadratic in the number of dimensions of the feature space and, consequently, PK models may be ineffective when the data is sparse – in particular, when the data is sparse, only few cross feature observations may exist in the data. To address this sparsity issue, FMs assume that the $m \times m$ feature interaction matrix A is low-rank and can be decomposed into $A \simeq VV^T$, where $V \in \mathbb{R}^{m\times k}$. Relying on this assumption, FMs factorize pair-wise feature interaction matrix to capture hidden interactions within features. This means that instead of representing the feature interactions as a single monolithic $m \times m$ matrix, FMs associate a k-dimensional vector $\vec{v}_i \in \mathbb{R}^k$ (where $k \ll m$) to each component i, such that $A_{(i,j)} \simeq \langle \vec{v}_i, \vec{v}_j \rangle$.

[2]Note that FMs can be generalized to higher degrees of feature interactions. In this paper, without loss of generality, we focus on pairwise FMs, which have been shown to be generally effective and, thus, make up the most commonly used approach for FMs – details can be found in [25].



(a) Simpson's paradox **(b) Simpson's paradox solved using causal information**

Figure 1: (a) Simpson's paradox is a statistical fluke that (b) disappears when we account for the underlying causal structure: an apparent positive correlation between vaccine taking and deaths are explained away, when considering confounding variable age which simultaneously affects both vaccine taking and death processes

$$\hat{F}_{FM}(x) = \underbrace{w_0 + \sum_{i=1}^{m} w_i x_i}_{lin.\ predictor} + \underbrace{\sum_{i=1}^{m}\sum_{j=i+1}^{m} \langle \vec{v}_i, \vec{v}_j \rangle x_i x_j}_{low\ rank\ warping}, \tag{3}$$

or, also considering that $\hat{y}_{FM}(\vec{x}) = \hat{F}_{FM}(x)$ and that $\langle \vec{v}_i, \vec{v}_j \rangle = v_i^T v_j$, we can also rewrite this as

$$\hat{y}_{FM}(\vec{x}) = w_0 + \sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m}\sum_{j=i+1}^{m} v_i^T v_j \cdot x_i x_j \tag{4}$$

$$= w_0 + \vec{w}^T\ \vec{x} + \vec{x}^T\ V\ V^T\ \vec{x}. \tag{5}$$

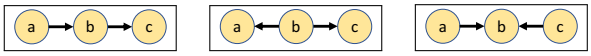
Note that the feature interaction term in Equation 4 can be reformulated [25] as:

$$\sum_{i=1}^{m}\sum_{j=i+1}^{m} v_i^T v_j x_i x_j = \frac{1}{2}\sum_{f=1}^{k}((\sum_{j=1}^{m} v_{j,f} x_j)^2 - \sum_{j=1}^{m} v_{j,f}^2 x_j^2), \tag{6}$$

where $v_{j,f}$ denotes the f-th element in v_j. As we see here, this significantly reduces the number of model parameters from quadratic to linear in the number of dimensions of the feature space and, thus, supports more effective learning when the data is sparse. In particular, while the time complexity of Equation 4 is $O(km^2)$, with the reformulation the time complexity drops to $O(km)$.

3.2 Statistical Flukes, Causal Paradoxes, and Causal Graphs

3.2.1 Statistical Problems and Causal Paradoxes. As we see above, factorization machines seek to leverage statistical patterns underlying the given data to discover mappings from the inputs to the output variables, while also accounting for the interactions among the input variables. Unfortunately, though, decision making purely based on (data-driven) statistics may lead to poor outcomes due to potential flukes. A well-known example of this is the Simpson's paradox, illustrated in Figure 1: in this example, (possibly unobserved) confounders causally impact multiple variables in the data in such a way that the statistics learned from the data may be misleading:



(a) chain (b) fork c) collider

Figure 2: Basic causal structures

in this example, "age" simultaneously increases the likelihood of "vaccine taking" and "death", giving the illusion that deaths increase with vaccine taking.

3.2.2 Causal Graphs. To avoid the negative impact of such statistical flukes, we need to account for the causal structure underlying the data. As we mentioned in the related work section and discussed in the above example (see Figure 1(b)), causal relationships among variables are commonly represented using a directed acyclic causal graph, $G = (\mathcal{V}, \mathcal{E})$, which describes the causal effects between variables. Here, $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. In a causal graph, each node represents a random variable that could be the treatment, the outcome, or some other variables. A directed edge $X \rightarrow Y$ denotes a direct causal effect of variable X on variable Y. Figure 9 presents two causal graph examples.

As we discussed in the the related work section, discovering causal graphs is an active research area, refered to as *causal discovery* [28]. Our work's primary objective and contribution is to leverage causal graphs to guide the learning process in the FMs. So, without loss of generality, we assume that a causal graph is already available – either provided by the domain expert [21] or pre-extracted from the data using existing causal discovery techniques [20, 28, 30].

3.2.3 Confounders, and other Causal Structures. In addition to confounders, which played a crucial role in Simspon's paradox (Figure1), we are interested in three other basic causal structures [23] as they can also contribute to statistical flukes (Figure 2):

- In the *chain* structure, a causally affects c through its influence on b. Note that in a chain structure, if one fixes the value of b, the variables a and c appear to be independent.
- In the *fork* structure, b is a common cause of both a and c – note that, in this case, a and c are causally related but there is no causation between them. If one fixes the value of b, the variables a and c are render independent.
- In the *collider*, both a and c independently cause b; once again, a and c are causally related through b, but there is no causation between them. If one fixes the value of b, a and c appear to be negatively correlated.

In the *confounder* structure, as in the fork, b is the common cause of both a and c; but, in this case, a is also a direct cause of c. As we have seen in Figure 1, confounders lead to statistical paradoxes, such as the Simpson's paradox, unless they are accounted for. In short, causal structures impact the statistical relationships among the variates and therefore they need to be considered during data analysis to make sure that the learned recommendations are not based on poor statistics.

4 CAUSAL INTERPRETATION OF FACTORIZATION MACHINES

In this section, we show that factorization machines can be re-interpreted as engines that inherently discover and take into account hidden confounders in the recommendation process. More specifically, we argue that the component of the factorization machines that seeks interactions between pairs of variables can be interpreted as seeking hidden confounders in the underlying causal graph.

Let us remember from Section 3.1, the Equation 6 underlying FMs can be written as:

$$\hat{y}_{FM}(\vec{x}) = w_0 + \vec{w}^T \vec{x} + \vec{x}^T V V^T \vec{x},$$

where $\vec{w} \in \mathbb{R}^m = \{w_1, \cdots, w_m\}$ and $V \in \mathbb{R}^{m \times k} = \{\vec{v_1}; \cdots; \vec{v_m}\}$. and $\vec{v_i} \in \mathbb{R}^k$. As we have discussed in Section 3.1, the third term in Equation 4.2 shifts the prediction for each data point by an amount representing the interactions between its features (in the terminology of [15], it warps the space to account for feature interactions). In the rest of the section, we show that this last term can be interpreted as accounting for unobserved confounders in the data. This interpretation will enable us to develop our proposed Causally Informed Factorization Machine (CIFM) formalism; but first we need to introduce the concept/theory of *deconfounders*.

4.1 Deconfounders

As we discussed in Section 2.3, a common assumption in causally-based inference is *unconfoundedness*, where we assume that there are no *unobserved confounders* in the data; i.e., there are no external variables that have a causal impact on the variables at hand. However, this assumption is generally untestable and may not always hold [26]. [32] proposed a methodology to *deconfound* a given data set under certain specific conditions. As before, let $X \in \mathbb{R}^{n \times m}$ be a data matrix where the n rows are the (transpose of) m-dimensional data vectors, where each of the $1 \leq j \leq m$ dimensions corresponds to a possible cause a_j for a target variable y; i.e., $\vec{y}[j] = Y(\vec{x}_j)$ or equally $\vec{y} = Y(X)$. [32] has shown that, under the assumption that there is no unobserved single-cause confounder that impacts the target variable y along with one of the m individual causes, *the k latent features (z_h for $1 \leq h \leq k$) learned from the factorization of X* can be used as substitutes for the unobserved (multi-cause) confounders in the data. Intuitively, multiple-cause confounders create dependencies among the causes and if the factor model represents the causes' distribution, it can help us extract a latent variable that captures the unobserved multiple-cause confounders. More specifically, [32] has shown that one can use a simple linear function as the outcome model in the presence of a multidimensional latent variable $\vec{z}$. :

$$y = f(\vec{x}, \vec{z}) = \alpha + \beta^T \vec{x} + \gamma^T \vec{z}.$$

4.2 FMs as Implicit Deconfounders

Note that the above equation shows significant similarities to the key FM formulation (i.e., Equation 6),

$$\hat{y}_{FM}(\vec{x}) = w_0 + \vec{w}^T \vec{x} + (\vec{x}^T V)\ (V^T \vec{x}),$$

with several rewrites and constraints:

- α is replaced with w_0 and $\vec{\beta}$ is replaced with $\vec{w}$,
- the latent vector $\vec{z}$ is obtained through the transformation $V^T \vec{x}$, and
- the weight vector γ is further constrained to take the value $V^T \vec{x}$.

Above, two constraints are significant:

- The first constraint, $\vec{z} = V^T \vec{x}$, implies that the transformation V^T should help discover k latent variables that together render the

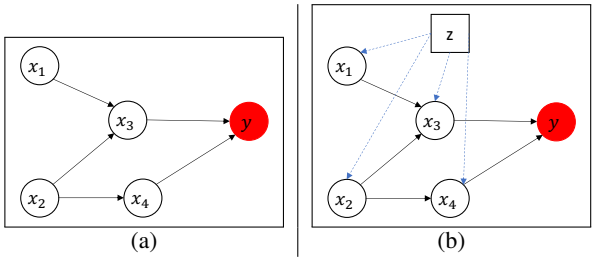
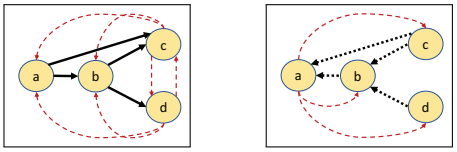


Figure 3: (a) A causal graph with 4 observed variables (features), x_1, x_2, x_3 and x_4 and a target variable y (in red color), (b) substitute confounder discovered through FM process.



(a) A forward RW graph **(b) A backward RW graph**

Figure 4: Forward and backward random walk (RW) graphs - black solid lines depict the direct causation, black dotted lines depict the inverse direct causation, red dashed lines depict extra edges inserted to complete the random walk graph.

causes conditionally independent from each other. While there is no guarantee that this will be true (unless additional constraints are imposed on V in the FM formulation), since V encodes the interactions among the variables in the data, we conjecture that, in practice, the discovered latent variables should be decorrelated (under the common FM assumption that the data is spectrally sparse) – this is because we expect that the learned model captures the unconditional distribution of the causes and thus should render the causes conditionally independent given the (per individual) latent variables.

- The second constraint, $y = \vec{x}^T\ V$, is not necessary from a causal perspective, but does not violate the causal interpretation presented in this paper either – it only helps the FMs to reduce the number of model parameters that need to be learned.

Based on the above, we argue that the FM process can be thought of disclosing the effects of *unobserved confounders* within features. Figure 3 visualizes this: Figure 3(a) presents a causal graph, with 5 variables $(y, x_1, x_2, \cdots, x_5)$, where y is the target variable and the rest are observed features in the data set. Figure 3(b), then, depicts the two latent factors ($k = 2$ for V) discovered in the FMs process that serve as a substitute confounder.

5 STRUCTURAL CAUSAL ATTENTION

There are multiple ways one can measure the causal relationships among variables. Many of these are purely data-driven; examples include average treatment effect (ATE), which quantifies the difference in the outcomes between units assigned to the treatment and units assigned to the control [23]. Since our goal is to take into account causal information which may be provided by the user (rather than being data-driven), in this paper, without loss of generality, we seek non-data-driven measures.

5.1 Desiderata

As we discussed earlier, existing research such as [23] has shown that a priori knowledge, in the form of *causal graphs* can help avoid statistical pitfalls that correlation based techniques would face. In particular, existing work has focused on identifying conditions under which variables are conditionally independent from each other. For instance, [24] introduced the concept of *d-separation*, which helps determine whether a subset, X, of variables would be statistically independent from another subset, Y, of variables under conditioning of a given subset of variables. A related results is that every variable is statistically independent of its graphical non-descendants conditional on its parents [22] and, more generally, the parents, children, and spouses of a variable (also referred to as the *Markov blanket* of a variable) store information about that variable that cannot be obtained from any other variable and, consequently, the Markov blanket is the set of all variables that are dependent on the variable, conditioned on all other variables [22].

While these and other existing results have been shown to be useful in causal discovery tasks, our goal in this paper is different. In particular, we are not seeking to identify the conditions under which variables are independent from each other; rather, we aim to use the given causal graph G to measure a degree of structural causal dependency between pairs of variables, to be leveraged as an attention mechanism when analyzing the data:

- **Desiteratum #1 - Causal Distance:** The closer the two variables are on the causal graph, the stronger the intensity of their causal relationship is. This is because each variable on a causal path may be subject to other causal variables or random noise, which diminishes the overall structural causal dependency on a long chain of variables (Figure 5(a)).
- **Desiteratum #2 - Causal Reinforcement:** The larger the number of causal paths between two variables, the stronger is the intensity of their causal relationships. This is because multiple causal paths from a variable to another may help offset some of the loss due to the pair's causal distance (Figure 5(b)).

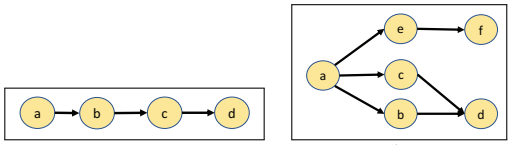
5.2 Random Walks with Causal Graphs

In this section, we argue that the desiderata listed above can be achieved through a random walk on the random walk graphs, carefully constructed from the provided causal graph.

5.2.1 Causal Random Walk Graphs. Remember from Figure 2 that a fork variable distributes causal information, whereas a collider variable gathers causality from multiple sources. To account for these, we create random walk graphs that capture the behaviors of the fork and colliders (Figure 4).

Let $G = (\mathcal{V}, \mathcal{E})$ be a causal graph, where $\mathcal{V}$ is the set of variables and $\mathcal{E}$ is the set of pairwise causations. Let $C \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be a causal adjacency matrix that encodes this graph:

$$C_{(i,j)} = \begin{cases} 1, & \text{if } v_i \text{ directly causes } v_j \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$



(a) (b)

Figure 5: Properties of structural causal dependency on a causal graph - (a) structural causal dependency of pair $\langle a, b\rangle$ is higher than the pair $\langle a, c\rangle$, which is higher than the pair $\langle a, d\rangle$. (b) structural causal dependency of pair $\langle a, d\rangle$ is higher than the structural causal dependency $\langle a, f\rangle$.

where $v_i, v_j \in \mathcal{V}$. Given this matrix, we define both a *forward* adjacency matrix, which captures the cause-to-effect dependencies and accounts for the forks, and a *backward* adjacency matrix, which captures effect-to-cause dependencies and accounts for the colliders: in the input graph G: (Figure 4):

$$M^{forward}_{(i,j)} = \begin{cases} \frac{C_{(i,j)}}{|V^i_{out}|+1}, & \text{if } V^i_{out} > 0, \\ \frac{1}{|\mathcal{V}|} & \text{otherwise,} \end{cases} \tag{8}$$

$$M^{backward}_{(i,j)} = \begin{cases} \frac{C^T_{(i,j)}}{|V^i_{in}|+1}, & \text{if } V^i_{in} > 0, \\ \frac{1}{|\mathcal{V}|}, & \text{otherwise.} \end{cases} \tag{9}$$

Here $V^i_{out/in}$ is the set of variables for which there are outgoing/incoming edges from v_i and C^T is the transpose of C. Note that since the forward graph takes into account forks and backward graph takes into account colliders, we can quantify their relative contributions as to obtain a combined causally-informed adjacency matrix underlying the *causal random walk graph* as

$$M^{fb}_{(i,j)} = \phi \times M^{forward}_{(i,j)} + (1-\phi) \times M^{backward}_{(i,j)}, \tag{10}$$

where ϕ measures the relative impacts of the forward (cause-to-effect) and backward (effect-to-cause) transitions, here we use number of fork nodes (denoted as $\mathcal{F}(G)$) and number of collider nodes (denoted as $C(G)$) in graph G to compute:

$$\phi = \frac{\mathcal{F}(G)}{\mathcal{F}(G) + C(G)}, \tag{11}$$

With the above, we can measure the the structural causal dependency between two variables, taking into account Desideratum 1 and 2, using a random walk with adjacency matrix $M^{fb}_{(i,j)}$ as detailed next.

5.2.2 Causally Informed Random Walks. Given a causal random walk adjacency matrix $M^{fb} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ constructed using the causal graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of variables and $\mathcal{E}$ is the direct causal edges, we define the pairwise causal relationships described by the adjacency matrix M^{fb} relying on a random-walk with restart approach [31]. Intuitively, for each source node v_i, the corresponding causal impact on the rest of the nodes in the graph, measured through a random walk seeded at vertex v_i is described as

$$\vec{\Pi}_{[i]} = \alpha(M^{fb})^T \vec{\Pi}_{[i]} + (1-\alpha)\vec{s}_{[i]}, \tag{12}$$

where $1-\alpha$ is the teleportation rate, $0 < \alpha < 1$ and $\vec{s}$ is a re-seeding vector:

$$\vec{s}_{[i][j]} = \begin{cases} 1 & j = i \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Here, the first component describes the direct transitions from one node to a neighbor following the edges on the graph as described by the transition probabilities in matrix M^{fb}; whereas the second component describes a random jump back to the seed nodes v_i. The parameter α regulates the frequency of the edge transitions vs. seed jumps. Intuitively, the vector $\vec{\Pi}_{[i]}$ describes the ratio of the time a random walker following the above transition/jump process spends on each graph node in the steady state. Consequently, the scores can be said to capture the topological *significance* of the graph nodes with respect to the given seed node – more specifically, the scores get smaller as one gets further away from the node v_i and a node reachable from multiple random walk paths tend to have a higher score due to the resulting reinforcement. In other words, the matrix,

$$\Pi = (I - \alpha(M^{fb})^T)^{-1}(1-\alpha)I. \tag{14}$$

encodes the structural causal dependencies between all pairs of nodes in the input causal graph, G, according to our two desiderata. More specifically, the value of $\Pi_{(i,j)}$ describes the forward causal flow implied by the causal graph, G, from the cause v_i to the effect v_j, whereas the value of $\Pi^T_{(i,j)}$ describes the backward causal flow from the effect node v_i to the cause node v_j.

5.3 Graph Structural Causal Attention

Finally, we are ready to define causal attention implied by the causal graph, G, between two nodes v_i and v_j:

$$\mathfrak{A}(G)_{(i,j)} \quad = \quad \Pi_{(i,j)} + \Pi^T_{(i,j)}, \tag{15}$$

where Π is the structural causal dependencies encoded by the causal graph G. We combine the forward causal flow and the backward causal flow to obtain bi-directional causal attention matrix.

6 CAUSALLY-INFORMED FACTORIZATION MACHINES

In this section, we propose alternative causally-informed factorization machines models to boost the performance of FMs (and its variants) via causal information provided in the form of causal graphs. (a) The first model, CIFM-I, involves enriching FMs through *indirect causal injection*; in CIFM-I, causal attention is applied to the results of the FM model in a post-hoc manner. (b) In the second model, CIFM-D, on the other hand, we are enriching FMs through *direct causal injection*; here causal attention is applied directly on the constraints and objective functions that define the FM model. We discuss these two models next. (c) The hybrid model, CIFM-D, combines these two models.

6.1 FMs with Indirect Causal Injection (CIFM-I)

Let $X \in \mathbb{R}^{n\times m}$ be a data matrix where the n rows are the (transpose of) m-dimensional feature vectors, where the i^{th} dimension corresponds to variable v_i. Let $\vec{y}$ be an n-dimensional vector corresponding to the target variable v_{m+1}. As we discussed earlier, the

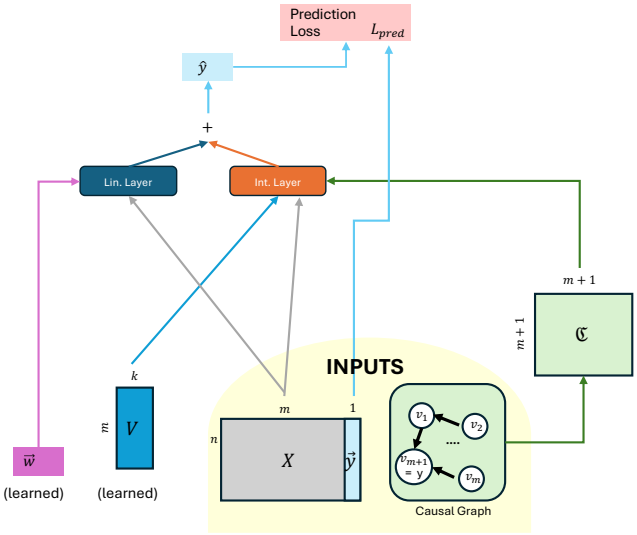


Figure 6: Architecture of the causally informed factorization machine with indirect causal injection (CIFM-I).

basic form of FM can be written as

$$\hat{y}_{FM} = \underbrace{\vec{\rho}_0 + \vec{w}X^T}_{lin.\ predictor} + \underbrace{diag(XAX^T)}_{interaction\ term},$$

such that $A = VV^T$ and the term

$$L_{pred} = \|\vec{y} - \hat{y}_{FM}\|$$

is minimized. The factorization machine model with *indirect causal injection* (CIFM-I), is similarly defined, but taking into account pairwise causal interactions described by the given causal graph, G:

$$\hat{y}_{CIFM-I}(G) = \underbrace{\vec{\rho}_0 + \vec{w}X^T}_{lin.\ predictor} + \underbrace{diag(X(\mathfrak{C} \odot A)X^T)}_{interaction\ term}, \quad (16)$$

where $\mathfrak{C} = \mathfrak{A}(G)$ is the causal attention function as defined in Equation 15 and $\odot$ is the Hadamart (or element-wise) product operation.

Intuitively, the CIFM-I model assumes that a matrix, $A = VV^T$, that describes the feature interactions does exist, but it further posits that this matrix needs to be *causally attentioned*, by multiplying it with the causal attention matrix $\mathfrak{A}(G)$ to causally regulate its contribution on the inferred recommendation (Figure 6).

6.2 FMs with Direct Causal Injection (CIFM-D)

Unlike the CIFM-I model described above, in the direct causal injection model, CIFM-D, we apply the causal attention directly on the constraints and objective functions underlying the FM model.

Let $X \in \mathbb{R}^{n \times m}$ be a data matrix where the n rows are the (transpose of) m-dimensional feature vectors, where the i^{th} dimension corresponds to variable v_i. Let $\vec{y}$ be an n-dimensional vector corresponding to the target variable v_{m+1}. As we discussed earlier, the basic form of FM can be written as

$$\hat{y}_{FM} = \underbrace{\vec{\rho}_0 + \vec{w}X^T}_{lin.\ predictor} + \underbrace{diag(XVV^TX^T)}_{interaction\ term},$$

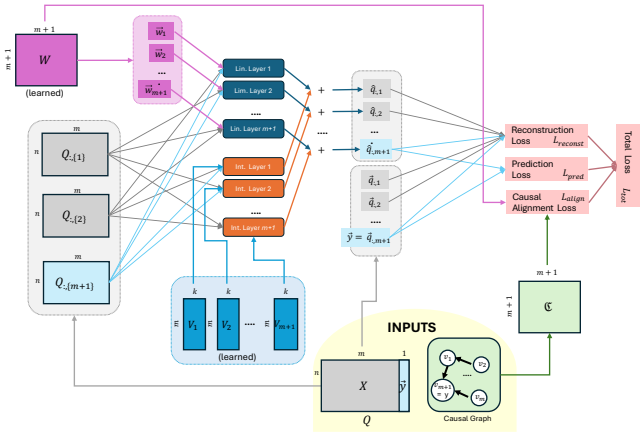


Figure 7: Architecture of the causally informed factorization machine with direct causal injection (CIFM-D)

such that the term

$$L_{pred} = \|\vec{y} - \hat{y}_{FM}\|$$

is minimized. But we can also rewrite this as

$$\hat{q}_{:,m+1} = \underbrace{\vec{\rho}_0 + \vec{w}_{m+1}Q^T_{:,\{m+1\}}}_{lin.\ predictor} + \underbrace{diag(Q_{:,\{m+1\}}V_{m+1}V^T_{m+1}Q^T_{:,\{m+1\}})}_{interaction\ term}, \quad (17)$$

subject to the prediction loss function

$$L_{pred} = \|\vec{q}_{:,m+1} - \hat{q}_{:,m+1}\|,$$

where

- $\vec{w}_{m+1} \in \mathbb{R}^{m+1}$ denotes the connectivity for feature $m+1$ in a graph (incoming nodes point to node $m+1$, includes itself).
- $Q \in \mathbb{R}^{n \times m+1}$ data matrix where the n rows are the (transpose of) $m+1$-dimensional feature vectors, where the i^{th} dimension corresponds to variable v_i and $m+1$ dimensions corresponds to the target variable, v_{m+1};
- $V_{m+1} \in \mathbb{R}^{m \times k}$ denotes the interaction matrix for feature $m+1$ (excludes itself)
- $\vec{q}_{:,i}$ denotes the vector obtained by taking the i^{th} column of the data matrix, Q; and
- $Q_{:,\{i\}}$ denotes the matrix obtained by dropping the i column of the matrix Q.

The above formulation provides us a view of the FM formulation where the target variable and the other variables are represented in one unified data structure, Q.

The CIFM-D model generalizes the FM formulation presented in Equation 17, where the predictive formulation is applied only to the target variable, by applying the predictive formulation to each and every variable in the causal graph:

$$\forall_{1 \le i \le m+1} \quad \hat{q}_{:,i} = \underbrace{\vec{\rho}_i + \vec{w}_i Q^T_{:,\{i\}}}_{lin.\ predictor} + \underbrace{diag(Q_{:,\{i\}}V_i V_i^T Q^T_{:,\{i\}})}_{interaction\ term}, \quad (18)$$

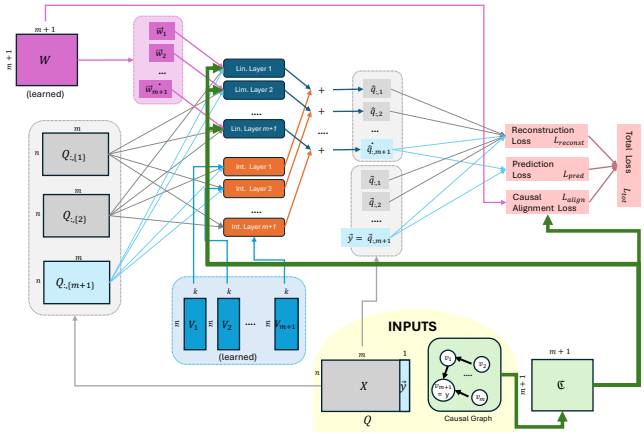


Figure 8: Hybrid causal injection (CIFM-H) modifies the CIFM-D architecture, by applying the causal attention matrix, $\mathfrak{C}$, also in the inference process

subject to the loss functions

$$L_{pred} = \|\vec{q}_{:,m+1} - \hat{q}_{:,m+1}\|, \quad (19)$$

$$L_{reconst} = \sum_{1 \le i \le m+1} \|\vec{q}_{:,i} - \hat{q}_{:,i}\|, . \quad (20)$$

Note that, in this generalized formulation of FM, we not only consider the prediction of the target variable v_{m+1} (Equation 19), but also reconstructions of all variables in the data from the remaining variables, excluding themselves (Equation 20).

We further extend the loss function by adding an alignment loss

$$L_{align} = \Delta(W, \mathfrak{C}), \quad (21)$$

where $\Delta(*,*)$ is a function that measures the misalignment between two matrices, $\mathfrak{C} = \mathfrak{A}(G)$ is a matrix that describes pairwise causal attentions between variables, and $W \in \mathbb{R}^{(m+1)\times(m+1)}$ is an matrix obtained by stacking the $\vec{w}_i$ vectors to represent the connectivity of hidden causal structure. Here we use *cosine distance* between $\mathfrak{C}$ and W as Δ function. The motivation of this alignment loss function is that since $\mathfrak{C}$ leverages prior knowledge about causal structure, we want to utilize it to *guide* the training process to align W with $\mathfrak{C}$.

Finally, total loss is defined as the sum of the three loss components:

$$L_{tot} = L_{pred} + L_{reconst} + L_{align}.$$

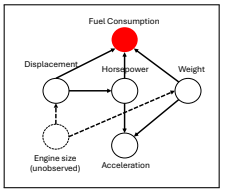
Figure 7 depicts the architecture of the CIFM-D model.

6.3 FMs with Hybrid Causal Injection (CIFM-H)

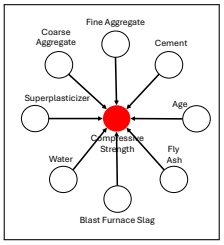
In Section 6.1, we described CIFM-I, with indirect causal injection to leverage causal attention matrix to adjust interaction weights. Above, we described CIFM-D, which directly injects causal attention matrix directly into the loss function. A hybrid solution, CIFM-H, largely follows the CIFM-D architecture, but also leverages causally adjusted weight matrix,

$$W_{causal} = W \odot \mathfrak{C}, \quad (22)$$

in Equation 18, instead of $W = \{\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_{m+1}\}$ (Figure 8).



(a) AutoMPG (with discriminating causal structure)



(b) Concrete (with a non-discriminating causal structure)

Figure 9: Causal graphs for two benchmark datasets [21]: while (a) the AutoMPG data set has a rich and discriminating causal structure, (b) the Concrete data set has a causal structure that does not provide much useful information (as each and every variable has a direct causal relationship with the target variable and there are no other causal relationships in the system)

Hence, in CIFM-H, we not only use alignment loss function to guide the causally-informed training of W, but also adopt indirect causal injection to further causally adjust the weight matrix.

7 EXPERIMENTS

In this section, we conduct experiments to evaluate the proposed CIFM method with synthetic and real-world datasets.

7.1 Datasets

7.1.1 Synthetic Datasets. In order to understand the behavior of the CIFM in diverse, but controlled, scenarios, we generated two synthetic data sets. In particular, each of these data sets are constructed taking a specific (randomly generated) feature interaction as its blueprint – intuitively, each graph corresponds to a *causal structure*. For these experiments, we randomly generated causal interactions graphs with varying *dependency densities* (i.e., the number of edges as percentage of maximum number of edges that would be allowed by a directed acyclic graph). Given a causal interaction graph, we then generated non-linear causal data [14], where each variable is equal to the sum of the sigmoid of its parents plus additive Gaussian noise with a mean of 0 and variance of 1. For each causal density (25%, 50%, 75%, 100%), we randomly generated causal interaction graphs for 10 variables (9 feature variables and 1 target variable) and for each graph, we have generated 100 sets of 1000 instances.

7.1.2 Real-World Datasets. In this paper, we consider two real-world data sets, adapted from [21], with causal structures with varying degrees of discrimination (Figure 9): (a – discriminating causal structure) *AutoMPG* (4 data features, 1 target feature, 392 instances) and (b – non-discriminating causal structure) *Concrete* (8 data features, 1 target feature, 1030 instances). As we see in Figure 9, in the Concrete data set, all features have the same causal relationship with the target variable and, consequently, the causal-structure is non-discriminating for prediction/recommendation purposes and we do not expect much useful information from this causal structure.

Table 1: Median MSE (synthetic data sets - varying causal densities; each density is separately colored; rank=5)

Causal Density (%)	25			50			75			100		
Non-causal Model	FM	NFM	DFM	FM	NFM	DFM	FM	NFM	DFM	FM	NFM	DFM
	0.726	0.567	0.675	0.844	0.536	0.773	0.796	0.404	0.686	0.772	0.349	0.581
CI Model	CIFM	CINFM	CIDFM	CIFM	CINFM	CIDFM	CIFM	CINFM	CIDFM	CIFM	CINFM	CIDFM
Indirect (I)	0.578	0.551	0.528	0.544	0.533	0.453	0.347	0.395	0.288	0.290	0.311	0.249
Direct (D)	0.538	0.547	0.577	0.538	0.457	0.586	0.414	0.348	0.410	0.322	0.298	0.319
Hybrid (H)	0.511	0.531	0.561	0.523	0.459	0.566	0.335	0.328	0.410	0.281	0.300	0.294

Table 2: Median MSE, rank=5 AutoMPG with a discriminating causal structure and Concrete with a non-discriminating causal structure

Dataset	AutoMPG			Concrete		
Non-causal Model	FM	NFM	DFM	FM	NFM	DFM
	0.325	0.299	0.330	0.305	0.332	0.208
CI Model	CIFM	CINFM	CIDFM	CIFM	CINFM	CIDFM
Indirect (I)	0.273	0.295	0.278	0.328	0.323	0.188
Direct (D)	0.289	0.342	0.290	0.414	0.405	0.194
Hybrid (H)	0.285	0.299	0.303	0.417	0.405	0.199

7.2 Baselines

For CIFM, we consider the following three variants:

- CIFM-I: CIFM with indirect causal injection – Section 6.1.
- CIFM-D: CIFM with direct causal injection – Section 6.2.
- CIFM-H: CIFM with hybrid causal injection – Section 6.3.

We compare CIFM with the following FM-based models:

- FM [25]: The original factorization machine implementation.
- NeuralFM (NFM) [9]: Serial coupling of an MLP with vanilla FM. We use the original implementation of NeuralFM, with two-layer MLP with layer sizes set to 16, and dropout rate is set to 0.2.
- DeepFM (DFM) [8]: Parallel coupling of an MLP with vanilla FM. Similar to NFM, a two-layer MLP with layer size 16 and dropout rate 0.2 is used. Unlike NFM, DFM is using an expanded feature representation which has double the number of parameters.
- CINFM and CIDFM: We also consider FM-variants extended with the CIFM model; i.e., replacing FM model in NFM and DFM with the CIFM, to see whether CIFM brings any benefits to these variants. We denote these extended model as CINFM and CIDFM.

7.3 Setup

For FM, we consider interaction rank of 5. α is set to 0.1 as default to compute causal impact matrix in Equation 12. For each dataset, we run each model 100 times and compute the median of mean squared error (MSE). Each model is trained using the Adam optimizer with a learning rate of 10^{-3} for up to 300 epochs. An early stopping regime halts training with a patience of 5 epochs to avoid overfitting. 80% data for training purposes and adopt 10-fold cross validation to select the best model parameters, while allocating the remaining portion for testing. Batch size 64 and λ in L_2 regularization is set to 10^{-5}.

7.4 Results with Synthetic Datasets

Table 1 presents the median MSE values for various baselines and CIFM variants for synthetic data generated with random causal graphs with varying causal densities (as described in Section 7.1.1):

- We see first that, causal injection generally improves accuracies over non-causal models for all causal densities considered.
- Among the non-causal baselines, NMF performs the best; however, even NFM can get a boost from causal injection; hybrid causal injection provides the largest boost for the NFM based models.
- For relatively low causal densities (25%), the best performance is obtained using CIFM variants; in particular, hybrid-injection (CIFM-H) provides the best overall accuracy (even though non-causal FM is the worst overall baseline for this scenario).
- As the causal densities increase, NFM and DFM, receive some boost relative to vanilla FM, indicating that these models are able to implicitly leverage causality to some degree. However, the best accuracies are obtained when we directly inject causal information into the models. In particular, as the causal density increases, the CIDFM variant with indirect injection (CIDFM-I) becomes increasingly advantageous and provides the best overall performance, with a clear difference to its competitors.

7.5 Results with Real-World Datasets

As we see in Tables 2, we consider two real-world data sets, with varying degrees of discriminating power in their causal graphs.

- Once again, causal injection generally improves accuracies over non-causal models.
- For AutoMPG, with a discriminating causal graph, both CIFM and CIDFM provide gains over non-causal baselines and we receive the best overall accuracies with indirect causal injection (CIFM-I).
- As we would expect, for Concrete, with non-discriminating causal graph, causal injection does not work well for FM and NFM models. Interestingly, for this data set, DFM (which has a larger number of parameters than FM and NFM) performs significantly better than FM and NFM and also is able to benefit from causal injection; indeed, the best overall accuracies are obtained with indirect causal injection to the DFM model (CIDFM-I).

8 CONCLUSIONS

In this paper, we argued that FM models, especially neural versions DFM and NFM, can be thought as applying implicit deconfounding on the data during recommendation computation. We further argued that the performance of the FM model (and its variants) can be improved by enriching the computation with an explicitely provided causal graph. Based on this premise, we have provided alternative (indirect, direct, and hybrid) methods to inject causal information into FMs. Experiments on synthetic and real data sets have shown that causal injection provides significant accuracy gains, especially when the causal graph is rich and discriminating.

REFERENCES

[1] Elias Bareinboim, Andrew Forney, and Judea Pearl. 2015. Bandits with Unobserved Confounders: A Causal Approach. In *NIPS*, Vol. 28.
[2] Mathieu Blondel, Akinori Fujino, Naonori Ueda, and Masakazu Ishihata. 2016. Higher-Order Factorization Machines *(NIPS'16)*. 9 pages.

[3] Mathieu Blondel, Masakazu Ishihata, Akinori Fujino, and Naonori Ueda. 2016. Polynomial Networks and Factorization Machines: New Insights and Efficient Training Algorithms, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR.

[4] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. 2021. Spatial-temporal Causal Inference for Partial Image-to-video Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 2 (May 2021), 1027–1035.

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, and et al. 2016. Wide & Deep Learning for Recommender Systems *(DLRS 2016)*. 4 pages.

[6] Laurent Charlin Dawen Liang and David Blei. 2016. Causal Inference for Recommendation. In *Causation: Foundation to Application, Workshop at UAI 2016. AUAI.* 59–67.

[7] Bin Gao and Yuehua Cui. 2015. Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics* 31, 24 (09 2015), 3953–3960.

[8] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 1725–1731.

[9] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics *(SIGIR '17)*. 10 pages.

[10] Paul W. Holland. 1986. Statistics and Causal Inference. *J. Amer. Statist. Assoc.* 81, 396 (1986), 945–960. http://www.jstor.org/stable/2289064

[11] Amir H. Jadidinejad, Craig Macdonald, and Iadh Ounis. 2021. The Simpson's Paradox in the Offline Evaluation of Recommendation Systems. *ACM Trans. Inf. Syst.* 40, 1, Article 4 (sep 2021), 22 pages. https://doi.org/10.1145/3458509

[12] Dominik Janzing. 2019. Causal Regularization. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Vancouver, BC.

[13] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-Aware Factorization Machines for CTR Prediction *(RecSys '16)*. 8 pages.

[14] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. 2020. CASTLE: Regularization via Auxiliary Causal Graph Discovery. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 127, 12 pages.

[15] Mao-Lin Li and K. Selçuk Candan. 2021. W2FM: The Doubly-Warped Factorization Machine. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Springer International Publishing, Cham, 485–497.

[16] Qian Li, Xiangmeng Wang, Zhichao Wang, and Guandong Xu. 2023. Be Causal: De-Biasing Social Network Confounding in Recommendation. *ACM Trans. Knowl. Discov. Data* 17, 1, Article 14 (feb 2023), 23 pages. https://doi.org/10.1145/3533725

[17] Yunqi Li, Hanxiong Chen, Juntao Tan, and Yongfeng Zhang. 2022. Causal Factorization Machine for Robust Recommendation. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 10, 9 pages. https://doi.org/10.1145/3529372.3530921

[18] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. 2014. On the Computational Efficiency of Training Neural Networks. In *NIPS*.

[19] David Lopez Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Leon Bottou. 2017. Discovering Causal Signals in Images. In *CVPR*.

[20] Osman A Mian, Alexander Marx, and Jilles Vreeken. 2021. Discovering Fully Oriented Causal Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 10 (May 2021), 8975–8982.

[21] J.M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research* 17, 32 (2016), 1–102.

[22] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[23] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statist. Surv.* 3 (2009), 96–146.

[24] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.

[25] Steffen Rendle. 2010. Factorization Machines *(ICDM '10)*. IEEE Computer Society, 6 pages.

[26] PAUL R. ROSENBAUM and DONALD B. RUBIN. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (04 1983), 41–55. https://doi.org/10.1093/biomet/70.1.41 arXiv:https://academic.oup.com/biomet/article-pdf/70/1/41/662954/70-1-41.pdf

[27] J. J. Shaughnessy and Zechmeister E. B. 1990. *Research methods in psychology*. McGraw-Hill, Seattle, WA.

[28] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.* (July 2011), 24 pages.

[29] Yi Tay, Shuai Zhang, Anh Tuan Luu, Siu Cheung Hui, Lina Yao, and Tran Dang Quang Vinh. 2019. Holographic Factorization Machines for Recommendation. In *AAAI*.

[30] Jean-François Ton, Dino Sejdinovic, and Kenji Fukumizu. 2021. Meta Learning for Causal Direction. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 11 (May 2021), 9897–9905. https://ojs.aaai.org/index.php/AAAI/article/view/17189

[31] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. [n. d.]. Fast Random Walk with Restart and Its Applications. In *Proceedings of the Sixth International Conference on Data Mining (ICDM '06)*. IEEE Computer Society, USA, 613–622.

[32] Yixin Wang and David M. Blei. 2019. The Blessings of Multiple Causes. *J. Amer. Statist. Assoc.* 114, 528 (2019), 1574–1596. https://doi.org/10.1080/01621459.2019.1686987 arXiv:https://doi.org/10.1080/01621459.2019.1686987

[33] Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. 2020. Causal Inference for Recommender Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 426–431. https://doi.org/10.1145/3383313.3412225

[34] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks *(IJCAI'17)*. 7 pages.

[35] Xin Xin, Bo Chen, Xiangnan He, Dong Wang, Yue Ding, and Joemon Jose. 2019. CFM: Convolutional Factorization Machines for Context-Aware Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 3926–3932. https://doi.org/10.24963/ijcai.2019/545

[36] Yantao Yu, Zhen Wang, and Bo Yuan. 2019. An Input-aware Factorization Machine for Sparse Prediction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 1466–1472. https://doi.org/10.24963/ijcai.2019/203

[37] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3404835.3462875