ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Learning Counterfactual Densities via Marginal Contrastive Discrimination

Aminata Ndiaye, Katia Meziani and Madalina Olteanu *

Paris Dauphine University PSL - CEREMADE
Place du Maréchal de Lattre de Tassigny, Paris, France.

Abstract. Estimating counterfactual densities provides a richer understanding of causal effects than traditional estimators based for instance on average treatment effects (ATE). However, reliable conditional density estimation remains challenging, especially in high-dimensional settings. We propose to use Marginal Contrastive Discrimination (MCD), a recent method that reframes conditional density estimation as a generalised contrastive learning task, enabling the use of supervised learning. The result is a new framework which delivers accurate counterfactual density estimates, illustrated through simulations. The latter show that the proposed technique handles high-dimensional data while improving state-of-the-art, providing new potential for nuanced and robust causal analysis.

1 Introduction

Causal inference is increasingly applied in research and industrial fields, including medicine, economics, and sociology, and is more and more intertwined with machine learning approaches. Aimed at identifying and estimating the effect of an intervention on a target variable from observational or experimental data, it provides an essential tool for decision making and highlights the limitations of predictive machine learning based solely on correlations. Two main frameworks structure the field: the potential outcomes framework, introduced by Neyman [1923] and extended by [Rubin, 1974, Imbens, 2000, Flores et al., 2012], and the structural causal model framework of Pearl [2000], which uses directed acyclic graphs and the *do-calculus* to analyse interventions.

Traditionally, causal effects have been summarized through the average treatment effect (ATE), defined as the expected difference in outcomes between the treated and the control units. While simple and intuitive, the ATE captures only an incomplete picture of the differences between counterfactual distributions and may hide important information. This limitation is well documented; for example, DiNardo et al. [1996] highlight that understanding wage inequality requires looking beyond means. For this reason, in the past years, a substantial amount of research focused on characterizing treatment effects in more refined ways. The first results were focused on quantiles or cumulative distribution functions (CDF) as in [Abadie et al., 2002, Firpo, 2007, Chernozhukov et al., 2013]. The issue of directly estimating interventional densities or distances between counterfactual densities has been considered only recently. The latter includes using normalizing flows [Melnychuk et al., 2023], kernel embeddings [Muandet

*The authors are grateful to Renault Company for providing the PhD funding for A. Ndiaye.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

et al., 2021] or semi-parametric and non-parametric estimates [Kennedy et al., 2023, Kim et al., 2024]. The present contribution is situated within [Pearl, 2000] causal framework and focuses on estimating causal effects by directly targeting the counterfactual density. To this end, we work within a fairly standard setting, by assuming *backdoor admissibility*. In this framework, estimating causal densities roughly amounts to estimating conditional densities. For the latter, we adopt *Marginal Contrastive Discrimination* (MCD) [Meziani et al., 2026], a recent method that reframes conditional density estimation as a generalized contrastive learning task. MCD was shown to have strong empirical performance, especially in high-dimensional settings where classical density estimators struggle. Our objective is to investigate whether, and to what extent, using MCD for a causal density learning task improves the results of the state-of-the-art.

The rest of the manuscript is organized as follows: Section 2 discusses the conditional density estimation and briefly describes the MCD method [Meziani et al., 2026]. In Section 3, we present the causal framework, derive the expression of counterfactual densities under the backdoor criterion and explain how to use the MCD as a plug-in estimate. Section 4 provides numerical illustrations comparing the proposed approach with state-of-the-art on simulated data.

2 Learning conditional densities with MCD

Let us start by summarizing the principles of *Marginal Contrastive Discrimination* (MCD) and explain how it transforms the problem of conditional density estimation into a binary classification task. Throughout, we consider an output feature $Y \in \mathbb{R}$ and a (potentially high-dimensional) input feature $X \in \mathbb{R}^d$. The goal is to estimate the conditional density $f_{Y|X=x}(y)$. MCD relies on the *marginal contrast function*,

$$q_r(x,y) = \frac{r f_{(X,Y)}(x,y)}{r f_{(X,Y)}(x,y) + (1-r) f_X(x) f_Y(y)}, \tag{1}$$

where $r > 0$ is a mixing parameter. The key property is that $q_r(x,y)$ summarizes how much more likely (x,y) was drawn under the joint distribution than under the product of marginals. A direct manipulation shows that the conditional density can be recovered from $q_r(x,y)$:

$$f_{Y|X=x}(y) = f_Y(y) \frac{q_r(x,y)}{1-q_r(x,y)} \times \frac{1-r}{r}. \tag{2}$$

Thus, estimating the conditional density reduces to estimating two quantities: the marginal density f_Y, and the contrast function q_r. The MCD estimate consists in plugging estimates $\hat{f}_Y$ and $\hat{q}_r$ into Eq. (2). The marginal $\hat{f}_Y(y)$ may be obtained using standard kernel density estimators (KDE). The main challenge remaining is estimating $\hat{q}_r(x,y)$. A crucial observation is that the contrast function admits the probabilistic representation $q_r(x,y) = \mathbb{P}(Z = 1 \mid W = (x,y))$, where $Z \sim \text{Ber}(r)$ and W such that $f_{W|Z=1}(x,y) = f_{(X,Y)}(x,y)$,

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

$f_{W|Z=0}(x,y) = f_X(x)f_Y(y)$, and $f_W(x,y) = rf_{(X,Y)}(x,y) + (1-r)f_X(x)f_Y(y)$. With this is mind, estimating $q_r(x,y)$ is a *supervised binary classification* task: one must discriminate samples drawn from the joint distribution $f_{(X,Y)}$ from samples drawn from the product of marginals $f_X f_Y$.

Constructing the contrastive dataset. In practice, one has an i.i.d. dataset from the joint distribution only, $\mathcal{D}_n^{X,Y} = \{(x_i, y_i)\}_{i=1}^n$. To perform classification, one must augment it with samples representing $f_X f_Y$. When enough data is available, one may for example split $\mathcal{D}_n^{X,Y}$ into two disjoint subsets: assign $Z = 1$ to the first, and assign $Z = 0$ to the second after randomly permuting the y_i's. Other strategies for building augmented data together with their theoretical guarantees are detailed in Meziani et al. [2026]. The resulting dataset is $\mathcal{D}_N^{W,Z} = \{(w_i, z_i)\}_{i=1}^N$, on which one may train any supervised learning algorithm to obtain an estimate $\hat{q}_r(x,y)$. In the present manuscript, multilayer perceptrons (MLPs) and gradient boosting (CatBoost) are used.

3 Causal density learning

For estimating the interventional density, we use the causal framework defined by Pearl [2000] and allowing to model causal relations between features as a directed acyclic graph (DAG). In this representation, each vertex represents a feature, and each directed edge a causal relation. Consider a setting where one has access to an iid sample $(V_1, ..., V_n)$ of $V = (X, T, Y)$, where $X \in \mathbb{R}^d$ represents the covariates or *confounders*, $T \in \mathbb{R}^k$ is a multidimensional treatment or exposure, and $Y \in \mathbb{R}$ is the univariate output. The setting is illustrated in Figure 1: the confounder X is influencing both T and Y, opening a backdoor path that must be adjusted for.

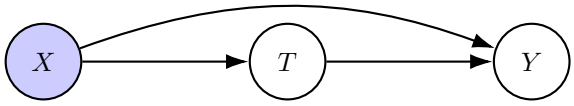

Fig. 1: Causal DAG between treatment T, outcome Y, and confounder X that influences both T and Y.

Furthermore, the *backdoor admissibility* is assumed: X blocks all backdoor paths from T to Y, and X contains no descendant of T. Backdoor admissibility ensures the identification of causal effects. We will also assume that the causal graph is shaped exactly as in Figure 1, that X includes all common causes of T and Y, and that there are no other problematic paths (colliders, mediators, selection variables). Within this framework, the marginal interventional density $f_{Y|do(T=t)}$ is defined as the density function of Y under an external intervention $T = t$. The marginal density of Y given the intervention $do(T = t)$ is identifiable through the *backdoor adjustment formula*, $f_{Y|do(T=t)}(y) = \int f_{Y|T=t,X=x}(y)\, f_X(x)\, dx$.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Marginal interventional density estimation. In practical applications, the above has to be estimated from data. Actually, two objects have to be estimated: first, the conditional densities under the integral, and second, the marginal interventional density. For the latter, one may use the empirical average or Monte-Carlo (MC) estimate:

$$\hat{f}_{Y|do(T=t)}(y) = \frac{1}{m}\sum_{j=1}^{m} \hat{f}_{Y|T=t,X=x_j}(y), \tag{3}$$

where $\{x_j\}_{j=\overline{1,m}}$ is a sample drawn from the confounder X, independent of the one used for learning the conditional density. The above implies that one will have to carefully split the available data into two subsets: the first to learn the conditional densities $\hat{f}_{Y|T=t,X=x}(y)$, while the second used in the MC estimate computation. The entire procedure is described in Algorithm 1. In Step 2 of the Algorithm, we will plug the MCD estimate.

Algorithm 1 Estimation of Marginal Interventional Density

Require: $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{n}$, split ratio $\alpha \in (0, 1)$ and treatment value t.
1: **Split data:** $\mathcal{D} = \mathcal{D}_{\text{DE}} \cup \mathcal{D}_{\text{MC}}$, so that $|\mathcal{D}_{\text{DE}}| = \lfloor \alpha n \rfloor$, $|\mathcal{D}_{\text{MC}}| = n - \lfloor \alpha n \rfloor$
2: **Learn conditional density estimate:** Fit $\hat{f}_{Y|T=t,X=x}(y)$ using $\mathcal{D}_{\text{DE}}$.
3: **Monte Carlo marginalization:**
4: **for** each x_j in $\mathcal{D}_{\text{MC}}$ **do** Compute $\hat{f}_{Y|T=t,X=x_j}(y)$
5: **end for**
6: **Output:** $\widehat{f}_{Y|do(T=t)}(y) = \frac{1}{|\mathcal{D}_{\text{MC}}|}\sum_{x_j \in \mathcal{D}_{\text{MC}}} \hat{f}_{Y|T=t,X=x_j}(y)$

4 Experimental results

Synthetic data was generated under backdoor admissibility with (T, X) drawn jointly from a multivariate Gaussian with specified mean and covariance matrix. This setup yields a closed-form interventional outcome distribution under any treatment t, enabling exact computation of interventional densities for comparison purposes. Y has been generated under intervention $T = t$ as $Y \mid do(T = t) \sim \mathcal{N}(b^\top t + a^\top \mu_X,\, a^\top \Sigma_X a + \sigma^2)$. The values of the parameters were randomly selected, while the sample size was $n = 1000$. 100 replications were considered. We compare the interventional densities computed with MCD to those obtained using two SOTA estimates: Normalizing Flows (NF) [Rezende and Mohamed, 2016] and Random Forest Conditional Density Estimation (RFCDE) [Pospisil and Lee, 2018]. These two alternatives were selected based on their performance and computation time results on both real and synthetic data. The MCD estimate was learnt either with MLP or with CATBoost algorithms. For each estimate, hyper-parameters were selected via 3-fold cross-validation. A proportion $\alpha = 0.8$ of the data was used to train the conditional density estimate, with the remaining 20% was used for Monte Carlo evaluation. To quantify performance, we computed the Wasserstein distance between the

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

estimated interventional density $\hat{f}_{Y|do(T=t)}$ and its closed-form ground truth. Results are summarised in Figure 2, Figure 3 and Table 1. Across all configurations, MCD:MLP consistently achieves the best accuracy. In low-dimensional settings, MCD:MLP yields substantially lower Wasserstein distance than Normalizing Flows and RFCDE. In higher dimensionality, MCD:MLP performs best. We also compared computation times across all methods. RFCDE remains the fastest due to its efficient tree-based implementation. NF is the most computationally demanding, due to the optimization of deep invertible architectures. MCD:MLP is in-between: although slower than RFCDE, it trains considerably faster than NF while offering superior estimation performance. Overall, MCD:MLP provides an excellent trade-off between accuracy and computational efficiency.

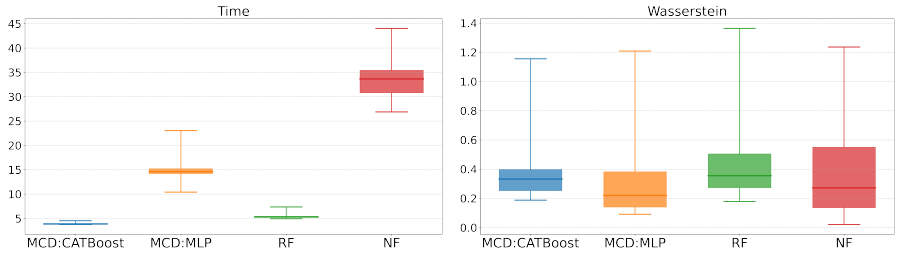

Fig. 2: Computational efficiency in seconds (left) and estimation accuracy using the Wasserstein distance (right).

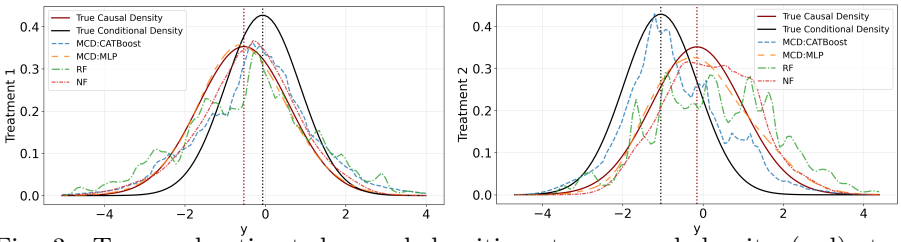

Fig. 3: True and estimated causal densities: true causal density (red), true conditional density (black), estimated interventional densities (dashed).

(d,k)	Metric	Methods			
		MCD:CATBoost	MCD:MLP	NF	RF
(10,3)	Wasserstein	0.36(0.02)	**0.28 (0.03)**	0.36 (0.07)	0.42 (0.05)
	Time	**3.89 (0.02)**	14.91 (3.71)	33.36 (15.05)	5.41 (0.19)
(100,10)	Wasserstein	1.04 (0.02)	**0.79 (0.03)**	0.98 (0.12)	1.06 (0.02)
	Time	**5.50 (0.04)**	13.05 (4.85)	58.47 (37.42)	9.83 (1.02)

Table 1: Mean(sd) Wasserstein-dist. and computation time for estimating causal densities for different (d,k). Best is in **green bold**, second-best is underlined.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

5 Conclusion and perspectives

The present contribution illustrates that MCD provides an effective way to estimate interventional densities under backdoor admissibility, offering a good trade-off between accuracy and computational cost. Several perspectives will be explored next, such as illustrations on real-life data, a broader simulation study with additional comparisons to state-of-the-art methods, and a more detailed discussion of how MCD relates to alternative causal effect estimates. Another natural extension is to investigate different distributional distances for comparing causal effects and to examine how these choices influence interpretation.

References

A. Abadie, J. D. Angrist, and G. W. Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica, 70(1):91–117, 2002.

V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. Econometrica, 81(6):2205–2268, 2013.

J. DiNardo, N. M. Fortin, and Th. Lemieux. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. Econometrica, 64(5):1001–1044, 1996.

S. Firpo. Efficient semiparametric estimation of quantile treatment effects. Econometrica, 75 (1):259–276, 2007.

C. A. Flores, A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann. Estimating the effects of length of exposure to instruction in a training program: The case of job corps. Review of Economics and Statistics, 94(1):153–171, 2012.

G. W. Imbens. The role of the propensity score in estimating dose–response functions. Biometrika, 87(3):706–710, 2000.

E. Kennedy, S. Balakrishnan, and L. Wasserman. Semiparametric counterfactual density estimation. Biometrika, 110(4):875–896, 2023.

K. Kim, J. Kim, and E. H. Kennedy. Causal effects based on distributional distances, 2024. URL `https://arxiv.org/abs/1806.02935`.

V. Melnychuk, D. Frauen, and S. Feuerriegel. Normalizing flows for interventional density estimation. In International Conference on Machine Learning, pages 24361–24397, 2023.

K. Meziani, A. Ndiaye, and B. Riu. MCD: Marginal contrastive discrimination for conditional density estimation. arXiv preprint, 2026.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. Journal of Machine Learning Research, 22:1–71, 2021.

J. Neyman. On the application of probability theory to agricultural experiments. Statistical Science, 1923.

J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000.

T. Pospisil and A. B. Lee. RFCDE: Random Forests for Conditional Density Estimation. May 2018. URL `http://arxiv.org/abs/1804.05753`. arXiv: 1804.05753.

D. J. Rezende and Sh. Mohamed. Variational Inference with Normalizing Flows. June 2016. URL `http://arxiv.org/abs/1505.05770`. arXiv: 1505.05770.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, 1974.