Quantifying Common Support between Multiple Treatment Groups Using a Contrastive-VAE

Wangzhi Dai WZHDAI@MIT.EDU
Massachusetts Institute of Technology, MA, USA

Collin M. Stultz CMSTULTZ@CSAIL.MIT.EDU
Massachusetts Institute of Technology
Massachusetts General Hospital, MA, USA

Editors: Emily Alsentzer$^{\otimes}$, Matthew B. A. McDermott$^{\otimes}$, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy‡, Stephanie L. Hyland‡

Abstract

Estimating the effect of a given medical treatment on individual patients involves evaluating how clinical outcomes are affected by the treatment in question. Robust estimates of the treatment effect for a given patient with a prespecified set of clinical characteristics, are possible to obtain when there is sufficient common support for these features. Essentially, features having the greatest common support correspond to regions of significant overlap between the distributions of the different treatment groups. In observational datasets, however, all possible treatment options may not be uniformly represented, and therefore robust estimation of their effect may only be possible for the patients in the overlapping region. In this work, we propose a Contrastive Variational Autoencoder (Contrastive-VAE) to estimate where there is significant overlap between patient distributions corresponding to different treatment options. A Contrastive-VAE exploits shared information between different groups by modeling the shared information as arising from a shared set of latent variables to approximate distributions for treatment options that are not well represented in observational datasets. The result is an improved estimation of the distribution of the groups with a small number of data points. By estimating the likelihood for each group with annealed importance sampling, we are able to quantitatively identify the area of overlap between multiple treatment groups and obtain an effective confidence interval for the estimated individual treatment effect.

Keywords: Common Support; Contrastive Learning; Variational Autoencoder

1. Introduction

Data driven machine learning models are being applied with ever increasing frequency in the clinical domain (Esteva et al., 2019). One fundamental problem that limits their application is that most machine learning models are trained on retro-respective, observational data (Blom et al., 2019; Mohammed et al., 2020). This makes it difficult to identify causal relationships, estimate treatment effects, and make unbiased predictions when the model is deployed in practice (Agniel et al., 2018). Take clinical risk stratification as an example. A risk score typically estimates patient risk using a set of predefined patient characteristics; e.g., predicting death after a heart attack from patient demographics and labs available at admission(Myers et al., 2017). Although such models may have significant discriminatory ability, it is not guaranteed that the chosen patient features are causally related to the outcome of interest. Indeed, the existence of unappreciated confounding factors limits the one's ability to make causal statements from such models. As a case in point, patients with high risk features who receive aggressive therapies may have a lower adverse event rate than many patients with low risk features because the administered treatments are effective at lowering the risk of inimical events (Ambrosino et al., 1995). However, classifying such patients, with high risk features, as low risk is clearly misleading because their outcome is affected by the treatment decisions of their health care providers. The risk provided by such a model is therefore not an unbiased prediction and may not be appropriate for many patients.

Traditional causal inference methods on observational data estimate such treatment effects by reducing the selection bias via simple statistical methods like matching and re-weighting (Rubin, 2006; Rosenbaum and Rubin, 1983). These methods usually depend on strong assumptions such as unconfoundedness (de Luna and Johansson, 2014) and common support(Garrido et al., 2014). Moreover, they typically can only be applied in the setting of a binary treatment decision (Shalit et al., 2017). Real word clinical data, by contrast, are much more sophisticated; e.g., these assumptions are usually hard to meet, and patients are usually given more than one treatment at a time. Modeling such data requires more complex modeling choices that must deal with class imbalance and data scarcity, as some complex treatment decisions may not be well represented in the dataset.

In this paper, we develop a method that estimates both the treatment effect and the common support of this estimate in a multiple treatment group scenario. Furthermore, the approach effectively addresses the class imbalance and data scarcity - common problems that arise when analyzing more than one treatment at a time. By leveraging this knowledge, we obtain insights into the observed data and develop more accurate clinical risk scores that can help guide clinical decision making.

2. Related Work

Common support is a key assumption in treatment effect estimation models; e.g., convariate adjustment and propensity score matching (Pocock et al., 2002; Dehejia and Wahba, 2002). Although a number of strategies have been developed to identify and assess the common support assumption in treated vs. control scenarios, simple methods such as comparing bounds of covariates between groups (Rosenbaum et al., 2010) might fail when the corresponding covariate distributions and their overlap are complex and non-linear. Other methods usually can be viewed as a by-product of causal inference models, for example, by bounding the treatment propensity score (Li et al., 2018), thresholding data points in matching algorithms (Kallus, 2016), or comparing individual-specific posterior distributions for each potential outcome using Bayesian Additive Regression Trees (Hill and Su, 2013). Recently, Johansson et. al. proposed an interpretable assessment by rephrase the problem into finding minimum volume sets subject to coverage constraints with Boolean rule classifiers (Oberst et al., 2020).

However, all of these methods require accurate modeling for each of the treatment group, and this makes it challenging to extend them to more than two treatment groups. Moreover, class imbalance and data scarcity makes it difficult to build separate models for individual treatment groups.

Contrastive learning algorithms provide a way to learn relationships between two or more data sets that share some common information (Severson et al., 2019). A Contrastive-VAE, for example, leverages deep neural network structures to learn nonlinear latent variables for both the shared variation and the unique variation in distinct treatment groups within a given dataset (Abid and Zou, 2019). Contrastive-VAE models can therefore model multiple groups of data simultaneously and yield improved performance relative to individual models, especially in situations of severe class imbalance and data scarcity(Dai et al., 2019).

3. Contrastive-VAE

We propose a Contrastive-VAE to model the distribution of multiple groups of patients features and their outcomes. We assume individual private latent variables exist for each of the groups, in addition to common latent variables that model the shared variation between the different groups. Without loss of generality, suppose there exist two groups of patients who received different treatments $T = 0$ and $T = 1$ respectively. Denote $s \in \mathbb{R}^{d_s}$ as the shared latent variables between classes, and $z^+ \in \mathbb{R}^{d_{z^+}}$ and $z^- \in \mathbb{R}^{d_{z^-}}$ as the private latent variables for each of the two groups. The corresponding generative process for the features $x^+, x^- \in \mathbb{R}^d$ and outcomes $y^+, y^- \in \mathbb{R}$, are shown in Figure 1 (a). We can write the generative distributions of the two classes as follows,

$$p(x^+, y^+) = \int_{z^+} \int_s p(x^+, y^+|s, z^+) p(s) p(z^+) ds dz^+ \quad (1)$$

$$p(x^-, y^-) = \int_{z^-} \int_s p(x^-, y^-|s, z^-) p(s) p(z^-) ds dz^- \quad (2)$$

Here, $p(s)$ and $p(z)$ are the prior normal distribution of the latent variables, and $p(x^+, y^+|s, z^+)$ and $p(x^-, y^-|s, z^-)$ are conditional distributions of the two groups given the latent variables. These conditional distributions are modeled using a shared neural network decoder f_θ, which takes the shared latent variable and the group specific private latent variable as input.

Figure 1 (b) shows the structure of the Contrastive-VAE, where the optimization object is the sum of the evidence lower bounds (ELBO) of each group, i.e., $L(x^+, y^+) + L(x^-, y^-)$, where

$$L(x^+, y^+) \geq \mathbb{E}_{q_{\phi_s}, q_{\phi_{z^+}}}[f_\theta(x^+|s, z^+)] - \quad (3)$$
$$KL(q_{\phi_s}(s|x^+)||p(s)) - KL(q_{\phi_{z^+}}(z|x^+)||p(z^+))$$
$$L(x^-, y^-) \geq \mathbb{E}_{q_{\phi_s}, q_{\phi_{z^-}}}[f_\theta(x^-|s, z^-)] - \quad (4)$$
$$KL(q_{\phi_s}(s|x^-)||p(s)) - KL(q_{\phi_{z^-}}(z|x^-)||p(z^-))$$

The above structure can easily be extended to more than two groups by adding additional private latent variables z for new groups of data.

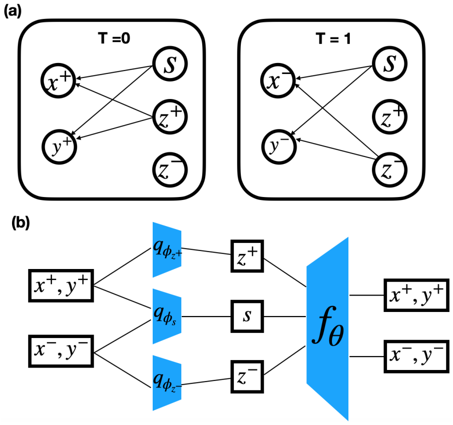

Figure 1: (a) Generative model for two treatment groups (T=0 and T=1). (b) Structure of Contrastive-VAE that learns the generative model.

3.1. Estimating Distribution Overlap

The common support is defined by the distributional overlap between different groups of patients. For a given patient's set of clinical features x, we need to compute

$$\begin{aligned}
&support(x) = \min\{P(x|T=1), P(x|T=0)\} \\
&= \min\{\int_{z^+}\int_s\int_y p(x,y|s,z^+)p(s)p(z^+)dydsdz^+, \\
&\int_{z^-}\int_s\int_y p(x,y|s,z^-)p(s)p(z^-)dydsdz^-\}
\end{aligned} \tag{5}$$

However, direct computation of the data likelihood given an arbitrary generative model is challenging because the associated integral over the entire latent space generally does not have a closed form solution. We therefore used Annealed Importance Sampling (AIS) to estimate the likelihood (Wu et al., 2016). The idea behind AIS is to first find a distribution $p_0(x)$ that we can rigorously compute, and then define K intermediate distributions between $p_0(x)$ and $p_K(x) = p(x)$. By estimating the ratio between each of the intermediate distributions, the desired probability can be obtained by multiplying the estimated ratio and the initial probability:

$$p(x) = \hat{r}p_0(x) \tag{6}$$

where $\hat{r}$ is the ratio estimated by a Markov Chain Monte Carlo procedure:

$$\hat{r} = \frac{1}{M}\sum_{i=1}^{M} w_{AIS}^{(i)} \tag{7}$$

$$w_{AIS} = \frac{p_1(x,z_0)}{p_0(x,z_0)}\frac{p_2(x,z_1)}{p_1(x,z_1)}\cdots\frac{p_K(x,z_K)}{p_{K-1}(x,z_K)} \tag{8}$$

Here M is the number of independent Markov Chains, z_0 is sampled from the initial prior distribution $p_0(z)$, and z_k for $1 \le k \le K$ are sampled from the transition kernel $\mathcal{T}_{k-1}(z_k|z_{k-1})$.

We chose the intermediate distributions to be

$$p_k(x,z) = p_0(x,z)^{1-\beta_k} p_K(x,z)^{\beta_k} \tag{9}$$

where β_0,...,β_K are monotonically increasing numbers from 0 to 1.

3.2. Estimating Individual Treatment Effects

We constructed one Contrastive-VAE to model both the features and the outcome because this allows us to estimate both the treatment effect as well as the corresponding common support for this estimate. Under the assumption of ignorability, the individual treatment effect (ITE) can be computed as,

$$\begin{aligned}
&ITE(x) = \mathbb{E}[y^+|x,T=1] - \mathbb{E}[y^-|x,T=0] \\
&= \int_{y^+,s,z^+} y^+ p(y^+|s,z^+)p(s,z^+|x^+)dydsdz^+ \\
&\quad - \int_{y^-,s,z^-} y^- p(y^-|s,z^-)p(s,z^-|x^-)dydsdz^-
\end{aligned} \tag{10}$$

To estimate this conditional probability of the outcome y given the features x, we used Gibbs Sampling to sample y while keeping x fixed.

3.3. Confidence Interval with Regard to Distribution Overlap

In order to intuitively explain the effect of overlap on the estimated ITE, we introduce a confidence interval with regard to distribution overlap. The vanilla definition of confidence interval is

$$\begin{aligned}
&P\left[\bar{ITE} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{ITE} + 1.96\frac{\sigma}{\sqrt{n}}\right] \\
&= 0.95
\end{aligned} \tag{11}$$

where σ is the standard deviation of the estimated treatment effect, which can be estimated using samples from the Contrastive-VAE. By replacing the number of samples to an effective number of samples in overlap distribution, we get

$$\begin{aligned} &P[I\bar{T}E - 1.96\frac{\sigma}{\sqrt{N\min(P^{+}(x), P^{-}(x))}} \\ &< \mu < I\bar{T}E + 1.96\frac{\sigma}{\sqrt{N\min(P^{+}(x), P^{-}(x))}}] \\ &= 0.95 \end{aligned} \tag{12}$$

where $P^{+}(x)$ and $P^{-}(x)$ are probabilities of the given feature x in the two groups. For continuous variables, we convert the density to probabilities by discretizing the feature space so that the minimum will be a number between 0 and 1.

4. Experiments on Synthetic Data

4.1. Experimental Design

We designed a series of synthetic data experiments to evaluate a Contrastive-VAE's ability to estimate the distributional overlap as well as the treatment effect. For these experiments we construct two groups of patients. The first group receives treatment (T=1) and the second group does not receive the treatment (T=0). We assume that patients data represent samples from 3D Dirichlet distributions. Samples of 3D Dirichlet distributions lie in a 2-simplex, a 2D triangle in 3D space, which mimics the realistic scenario where patient data corresponds a relatively low dimensional manifold in a high dimensional space (Cayton, 2005). We also assume that both treatment groups share some information - which is typically true in practice. For example, it is often of interest to assess the effect of a given therapy on a specific patient population. Although patients in different treatment groups receive different therapies, they nonetheless have the same diagnosis and/or disease. To simulate this situation, we assume one of the marginal distributions to be the same for different treatment groups, i.e.

$$\mathbf{x}^{+} \sim Dir(\alpha_1^{+}, \alpha_2^{+}, \alpha_3^{+}) \tag{13}$$

$$\mathbf{x}^{-} \sim Dir(\alpha_1^{-}, \alpha_2^{-}, \alpha_3^{-}) \tag{14}$$

where, $\alpha_1^{+} = \alpha_1^{-}$ and $\alpha_1^{+} + \alpha_2^{+} + \alpha_3^{+} = \alpha_1^{-} + \alpha_2^{-} + \alpha_3^{-}$. The outcomes for each group is defined by 2 non-linear functions $f^{+} : \mathcal{X} \to \mathcal{Y}$ and $f^{-} : \mathcal{X} \to \mathcal{Y}$, that maps the patients feature $\mathbf{x} \in \mathcal{X}$ to an outcome $y \in \mathcal{Y}$ for either received the treatment ($T = 1$), or dose not received ($T = 0$). We set $\{\mathbf{x}^{+}, y^{+} = f^{+}(\mathbf{x}^{+})\}$ and $\{\mathbf{x}^{-}, y^{-} = f^{-}(\mathbf{x}^{-})\}$ as the observed data, while the counter-factual outcomes $y'^{+} = f^{-}(\mathbf{x}^{+})$ and $y'^{-} = f^{+}(\mathbf{x}^{-})$ were concealed from the model and only used for evaluation of the treatment effect.

In the first experiment, we demonstrate the distributional overlap of the two groups of patients can be reproduced by the Contrastive-VAE. Ground truth probability density was used to evaluate the density estimated by AIS of the trained Contrastive-VAE. We also compared the Contrastive-VAE to Kernel Density Estimation (KDE) and standard VAEs that model the distribution of each group independently. The KDEs used a Gaussian kernel and their bandwidths were decided by a 5-fold cross validation. Both the KDEs and standard VAEs were trained separately for different groups of data, while the contrastive-VAE was trained with the two groups together. We conducted the experiments with different levels of class imbalance.

We did an additional experiment to show that a Contrastive-VAE can model patients with more than two treatment groups. This allows us to extend the model when more than one treatment is given to patients. For example, if we want to compare the effect of treatment A and B, we can find out the overlap of three populations, those who received A, those who received B and those who received nothing. Only for patients with support in all of the three populations, the estimated treatment effect can be thought of as reliable. In this experiment, we sampled three groups of data x_1, x_2 and x_3 from three different Dirichlet distributions. The number of samples from each of the groups are imbalanced to mimic the real observational dataset. A Contrastive-VAE with three private latent space, KDE and three independent VAEs were compared to reconstruct the probability distribution and the overlap of the three groups.

In the second experiment, we trained a Contrastive-VAE on imbalanced data to learn the joint distribution of features $\mathbf{x}$ and outcome y. We then use the trained model to predict the counter-factual outcome for patients of the 2 groups, i.e. y'^{+} for $\mathbf{x}^{+}$ and y'^{-} for $\mathbf{x}^{-}$, and the treatment effects for the 2 groups, i.e. $y^{+} - y'^{+}$ and $y'^{-} - y^{-}$. As the treatment effect estimation is only valid for patients that satisfied the common support assumption, we excluded patients with a overlap probability below a certain threshold. Then the estimated treatment effects were compared with the ground truth simulated by functions f^{+} and f^{-}.

Additionally we demonstrated the confidence interval with regard to common support using samples from 1 dimensional normal distributions. We considered three different situations where the distribution of the two treatment groups are partially overlapped, identical or apart from each other. In each situation, we estimated the treatment effects with a linear function using the samples and computed their confidence interval using equation 12.

4.2. Results of Distributional Overlap for Two Treatment Groups

Figure 2 shows the result of the first experiment with a highly imbalanced training set, 1000 training points for group 1 vs. 10 points for group 2. Probability density for the two groups and their overlap are plotted on the 2D simplex, the plane where all the data exist. It can be seen that the KDE failed to estimate the probability density for both of the 2 groups. An alternate approach is to model each of the two groups separately, using two independent VAEs - one for each treatment group. However, the VAE trained on the group that contains only 10 samples is a poor presentation of the underlying distribution for this group.

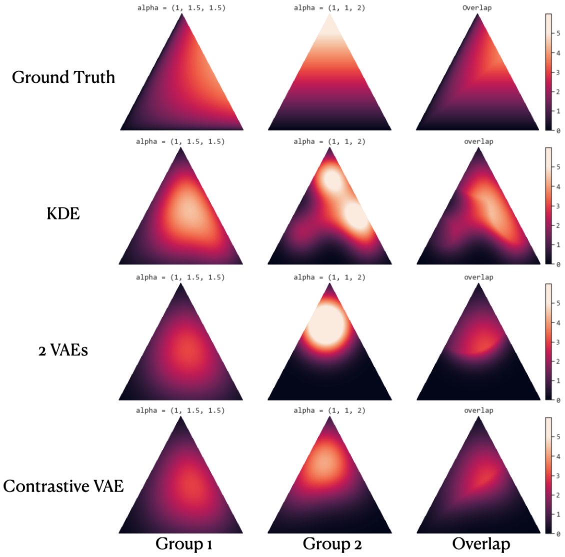

Figure 2: Values of probability densities plotted on the 2D simplex with color. Group 1 and 2 refers to the 2 Dirichlet distributions in the synthetic experiments.

Figure 4 (a) shows the mean squared error of the overlap probability estimated by the above three methods, in different level of class imbalance. Contrastive-VAE gives the significantly lower error, compared to KDE and the standard VAEs, in all situations.

4.3. Distributional Overlap of Three Treatment Groups

Figure 3 shows the results of three groups of data and their overlap. The training set size of the three groups are 1000, 50 and 10 to mimic the class imbalance in real world datasets. Similar to the overlap experiments in 4.2, the KDE failed to restore the smooth distribution in all three groups. Standard VAEs works well when efficient training data were provided for group 1 and 2, but did poorly for group 3 when only 10 points were used to training. The Contrastive-VAE was best able to reproduce the ground truth probability densities and their overlap.

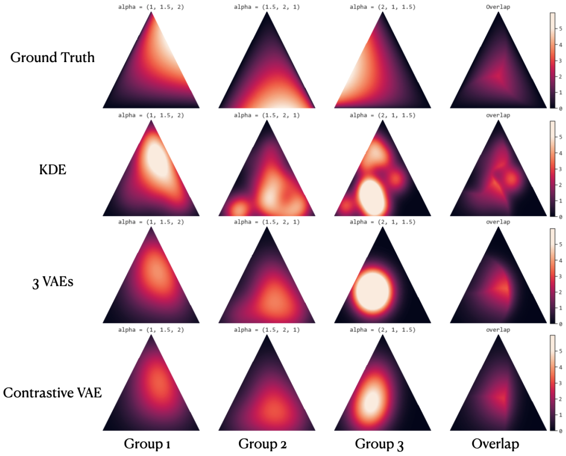

Figure 3: Values of probability densities plotted on the 2D simplex with color. Group 1-3 refers to the 3 Dirichlet distributions. Training set size for each of the groups are 1000, 50 and 10. Reconstructions of the probability densities and their overlap are compared using KDE, 3 independent VAEs and the Contrastive-VAE. Mean squared errors are 0.66 for KDE, 0.12 for 3 VAEs and 0.08 for Contrastive-VAE.

4.4. Results of Treatment effect estimation

Figure 4 (b) - (c) shows the mean squared error of the estimated treatment effect compared to simulated ground truth. To see the effect of the common support assumption, we altered the overlap threshold that decides which predictions are used to compute the treatment effect. With 0 overlap probability threshold, the common support assumption is completely ignored and the error in the estimated treatment effect is the largest. When non-zero overlap probability thresholds are used, the treatment effect estimation is only computed for patients who have an overlap probability that is larger than this threshold. The mean squared error decreases as the threshold increases, thereby demonstrating that accurate estimation of the treatment effect requires significant common support.

4.5. Simulation Results of Confidence Interval

Figure 5 shows the simulated results of three different feature distributions and the estimated ITE as well as its confidence interval. Figure 5 (a) - (c) show the situation where the two groups distribution are identical, where (a) gives the ground truth distribution of the feature in two treatment groups. The simulated outcome and training samples are shown in (b). (c) shows the estimated ITE and 95% confidence interval with regard to the overlap distribution. In this case, we see an extremely small interval in the region that contain training samples, as the common support assumption is fully satisfied. (d)-(f) show a partial overlapping situation. Here, the interval is dramatically smaller in the overlap region, compared to outside, which intuitively demonstrate the importance of common support for the ITE estimation. (g) - (i) show the other extreme case where there's hardly any over-

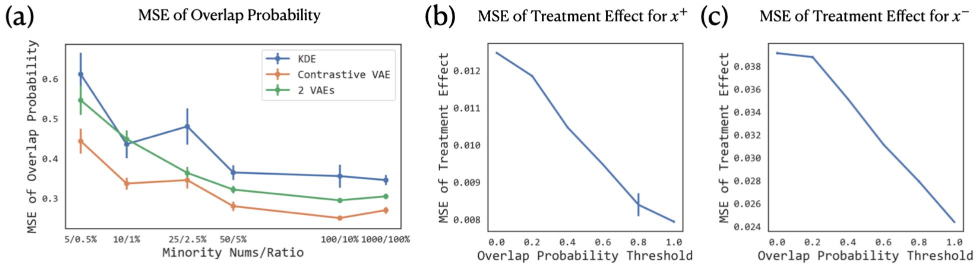

Figure 4: (a) Mean squared error of the overlap probability estimated by KDE, Contrastive-VAE and 2 standard VAE for 2 Dirichlet distributions with different levels of class imbalance. (b)-(c) Mean squared error of predicted treatment effect vs. overlap probability threshold for patients who received the treatment and who did not. All error bars represent the standard error over 10 bootstraps.

lap between the two distributions, where the extremely large interval indicates the ITE estimation is barely reliable when considering the common support assumption.

5. Experiments on Real Clinical Data

5.1. Experimental Design

We applied our method to a real world clinical data set, the Global Registry of Acute Coronary Events (GRACE). GRACE enrolled over 70,000 patients from 250 hospitals in 30 countries (Fox et al., 2014). Patients enrolled in the GRACE registry were diagnosed with an acute coronary syndrome (a constellation of signs and symptoms consistent with reduced blood flow to the heart). Patients were followed and their outcomes and therapeutic interventions were recorded.

We chose 2 major treatments, Percutaneous Coronary Intervention (PCI) and Coronary Artery Bypass Grafting (CABG), and trained a Contrastive-VAE with three separate private latent variables, representing each of the three groups - those who only receive a PCI, those who only had a CABG, and patients who received neither treatment as the control group.

In order to estimate the effect of the treatments, we consider patients that have common support in the treated and the control groups. For example, for the treatment PCI, we compare the distribution of the patients who received PCI to those who did not receive either PCI or CABG, and selected patients with confidence interval of ITE that below a threshold. Within the selected patients, we used the trained Contrastive-VAE to estimate the treatment effect of PCI, where the outcome of interest is death within 6 months of presentation. Those patients with a treatment effect greater than 6% were considered to be the effective group. A cutoff of 6% was used because this corresponds to the prevalence of death in the overall dataset. The non-effective group, on the other hand, corresponds to the patients whose estimated treatment effect is smaller than 6%. We compared patient characteristics between the effective group and the non-effective group for different thresholds of confidence interval to analyze the importance of the common support.

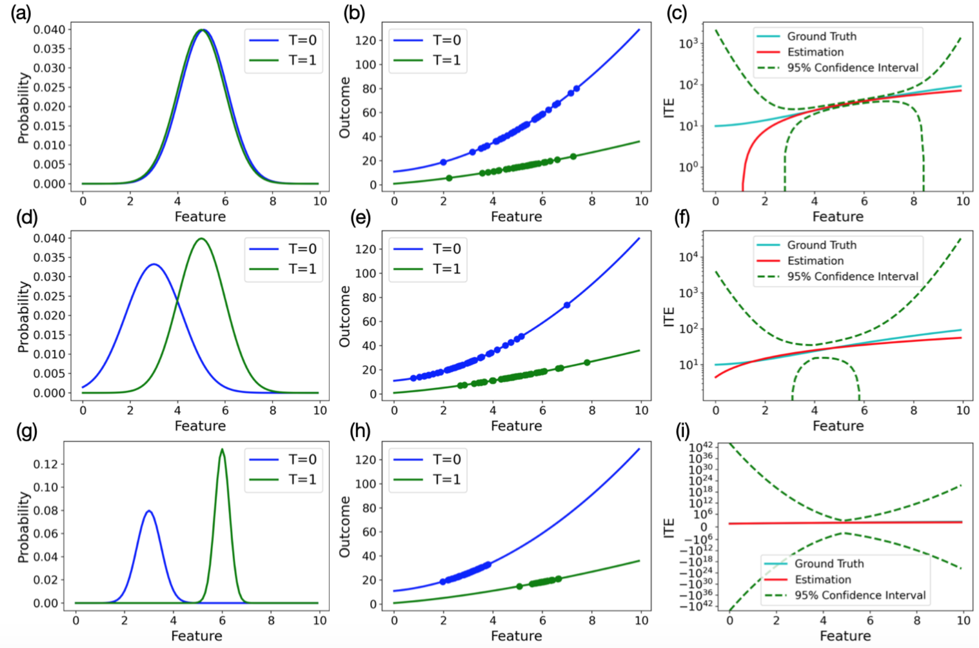

Figure 5: Confidence interval with regard to common support for different overlap levels. (a), (d) and (g) show the ground truth distribution of the two treatment groups. (b), (e) and (g) show the simulated outcome. (c), (f) and (i) show the predicted ITE and its confidence interval with regard to common support.

5.2. Treatment Effects and Common Support

Figure 6 (a) - (b) shows the average age and systolic blood pressure for patients in the CABG effective and non-effective group, and the corresponding p value that quantifies the statistical significance of this difference. Similarly, Figure 6 (c) shows the expected KILLIP class (a metric that quantifies the extent of heart failure on clinical exam at presentation) for patients in the PCI effective and non-effective group. As we can see from the figures, considering the confidence interval with regard to the distribution overlap can change the group characteristics significantly, and therefore leads to completely different clinical conclusions. For example, for the CABG treatment in (a) and (b), by considering only the patients with ITE within a small confidence interval, one can conclude that CABG is effective for patients with a younger age and lower systolic blood pressure ($p < 0.05$), however, a similar analysis that uses all the data (i.e., a large confidence interval) suggests that CABG does not derive a benefit irrespective of age ($p > 0.05$). Similarly, for the PCI treatment, conclusions that are made without considering the extent of the common support might be not valid when the overlap is considered, as shown in Figure 6 (c). When the threshold for the confidence interval is large, one can conclude the treatment is effective for patients with a larger KILLIP class, but the conclusion would be not valid if we restrict the patients to those with smaller confidence interval for their ITE.

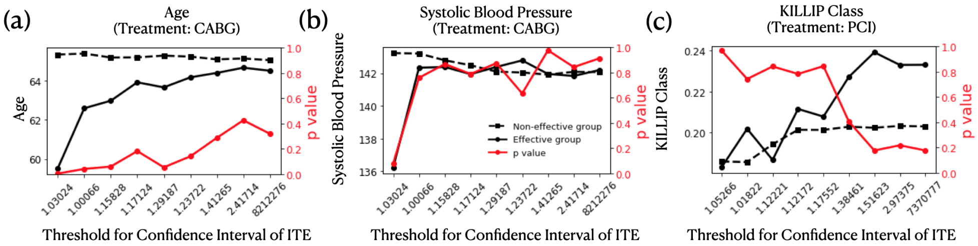

Figure 6: (a) - (b) Average age and systolic blood pressure compared between effective and non-effective groups for treatment CABG. (c) Similar result for KILLP class in threatment PCI.

6. Discussions

In this work, we demonstrate that both the treatment effect and the common support can be accurately estimated using a single Contrastive-VAE. The key point that makes our method different from traditional propensity matching approaches is that we approach the problem in a parametric way, where we model distributions explicitly for each of the treatment groups. The method allows us to model multiple treatment groups simultaneously and effectively deals with data scarcity - a common problem in real world datasets where patients can receive multiple different treatments. We demonstrate that a Contrastive-VAE can be used to discover meaningful clinical insights, even when data are highly imbalanced and sometimes scarce for certain treatment combinations.

A Contrastive-VAE is appropriate for this class of problems because it leverages the shared information between different treatment groups. Although patients in different groups may be treated differently, they often share the same diagnosis, and latent factors that lead to similar observed clinical or demo-graphical features. Having said this, it is important to stress that the shared information is an assumption and should be treated as a inductive-bias that arises from domain specific knowledge. If the two groups do not share any common information, a situation that is not typical of treatment groups in observational datasets, then the Contrastive-VAE may not yield suitable estimates of the common support.

In order to explain the estimated overlap probability and makes it easier for clinical applications, we proposed a effective confidence interval with regard to overlap. The number of sample size in the vanilla definition of confidence interval is replaced by a effective sample size $N \min(P^{+}(x), P^{-}(x))$. Here we use the probability, in stead of density, to make the weighting factor a number between 0 and 1. For continuous variables, this was achieved by discretizing the feature space and multiplying the density by a pre-chosen volume size. However, the volume size for different data sets and distributions might be chosen differently and therefore makes it difficult to compare the confidence interval between data sets. One possible solution is to first map the feature space to a latent space of fixed size and asses the common support assumption in the latent space. In this way, the volume size will be a fixed factor and comparison between different data sets will be unbiased.

References

Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

Richard Ambrosino, Bruce G Buchanan, Gregory F Cooper, and Michael J Fine. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 304. American Medical Informatics Association, 1995.

Mathias Carl Blom, Awais Ashfaq, Anita Sant'Anna, Philip D Anderson, and Markus Lingman. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: a retrospective, population-based registry study. *BMJ open*, 9(8): e028015, 2019.

Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

Wangzhi Dai, Kenney Ng, Kristen Severson, Wei Huang, Fred Anderson, and Collin Stultz. Generative oversampling with a contrastive variational autoencoder. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 101–109. IEEE, 2019.

Xavier de Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2(2):187–199, 2014.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Keith AA Fox, Gordon FitzGerald, Etienne Puymirat, Wei Huang, Kathryn Carruthers, Tabassome Simon, Pierre Coste, Jacques Monsegu, Philippe Gabriel Steg, Nicolas Danchin, et al. Should patients with acute coronary disease be stratified for management according to their risk? derivation, external validation and outcomes using the updated grace risk score. *BMJ open*, 4(2), 2014.

Melissa M Garrido, Amy S Kelley, Julia Paris, Katherine Roza, Diane E Meier, R Sean Morrison, and Melissa D Aldridge. Methods for constructing and assessing propensity scores. *Health services research*, 49(5):1701–1720, 2014.

Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the Amer-*

ican Statistical Association, 113(521):390–400, 2018.

Akram Mohammed, Pradeep SB Podila, Robert L Davis, Kenneth I Ataga, Jane S Hankins, and Rishikesan Kamaleswaran. Using machine learning to predict early onset acute organ failure in critically ill intensive care unit patients with sickle cell disease: Retrospective study. *Journal of Medical Internet Research*, 22(5):e14693, 2020.

Paul D Myers, Benjamin M Scirica, and Collin M Stultz. Machine learning improves risk stratification after acute coronary syndrome. *Scientific reports*, 7(1):1–12, 2017.

Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Paul R Rosenbaum et al. *Design of observational studies*, volume 10. Springer, 2010.

Donald B Rubin. *Matched sampling for causal effects.* Cambridge University Press, 2006.

Kristen A Severson, Soumya Ghosh, and Kenney Ng. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4862–4869, 2019.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016.