Multiple Causal Inference with Latent Confounding

Rajesh Ranganath
Courant Institute of Mathematical Sciences
New York University
rajeshr@cims.nyu.edu

Adler Perotte
Department of Biomedical Informatics
Columbia University
adler.perotte@columbia.edu

Abstract

Causal inference from observational data requires assumptions. These assumptions range from measuring confounders to identifying instruments. Traditionally, these assumptions have focused on estimation in a single causal problem. In this work, we develop techniques for causal estimation in causal problems with multiple treatments. We develop two assumptions based on shared confounding between treatments and independence of treatments given the confounder. Together these assumptions lead to a confounder estimator regularized by mutual information. For this estimator, we develop a tractable lower bound. To fit the outcome model, we use the residual information in the treatments given the confounder. We validate on simulations and an example from clinical medicine.

1 Introduction

Causal inference aims to estimate the effect one variable has on another. Such causal inferences form the heart of inquiry in many domains, including estimating the value of giving a medication to a patient, understanding the influence of genetic variations on phenotypes, and measuring the impact of job training programs on income.

Assumption-free causal inferences rely on randomized experimentation [7, 27]. Randomized experiments break the relationship between the intervention variable (the treatment) and variables that could alter both the treatment and the outcome—confounders. Though powerful, randomized experimentation fails to make use of large collections of non-randomized observational data (like electronic health records in medicine) and is inapplicable where broad experimentation is infeasible (like in human genetics). The counterpart to experimentation is causal inference from observational data. Causal inference from observational data requires assumptions. These assumptions include measurement of all confounders [32], the presence of external randomness that partially controls treatment [3], and structural assumptions on the randomness [14].

Though causal inference from observational data has been used in many domains, the assumptions the underlie these inferences focus on estimation in a single causal problem. But many real-world applications consist of multiple causal problems. For example, the effects of genetic variation on various phenotypes [6] or the effects of medications from order sets in clinical medicine [25] consist of multiple causal problems rather than a single causal problem. Considering multiple causal problems make new kinds of assumptions possible.

We formalize *multiple causal inference*, a collection of causal inference problems with multiple treatments and a single outcome. We develop a set of assumptions under which confounders can be estimated when unmeasured. Two assumptions form the starting point: that the treatments share confounders, and that given the shared confounder, all of the treatments are independent. This kind of shared confounding structure can be found in many domains such as genetics. In genetics, the multiple causal problems are the effects of different genetic variations on different phenotypes. The shared unobserved confounding structure across problems comes from population structure or local correlation structure (such as linkage disequilibrium).

These two assumptions of shared confounding and independent treatments given the confounder imply a probabilistic model for the treatments, confounder, and outcome. This model is akin to generalized factor analysis. Yet shared confounding and independent treatments still leave ambiguity in how to estimate the confounder. This ambiguity lies in how much information the confounder contains about a single treatment given the rest of the treatments.

To respect shared confounding, the information between the confounder and a treatment given the rest of the treatments should be minimal. However, forcing this information to zero makes the confounder independent of the treatments. This can violate the assumption of independence given the shared confounder. To resolve the tension between our two assumptions, we develop the principle of minimal information. This minimal information principle says that the true confounder has minimal information with each treatment, given the rest of the treatments—subject to the constraint that the confounder should predict each treatment best, given the rest of the treatments. In other words, we try find the confounder that has lowest information with the treatments, while still rendering them independent.

We develop an algorithm for estimating the confounder based on the minimal information principle. The algorithm relies on independently reconstructing each treatment given a stochastic confounder estimated from all of the treatments, while regularizing the mutual information each treatment has with the confounder, given the rest of the treatments. The mutual information is intractable, so we build a lower bound called the multiple causal lower bound (MCLBO).

The last step in building a causal estimator is to build the outcome model. Traditional outcome models regress the confounders and treatments to the outcome [24]. However, since the confounder estimate is a stochastic function of the treatments, it contains no new information about the response over the treatments—a regression on both the estimated confounder and treatments can ignore the estimated confounder. Instead, we build regression models using the *residual information* in the treatments and develop an estimator to compute these residuals. The assumptions and algorithms we develop strengthen the theoretical basis for existing causal estimation with unobserved confounders such as causal estimation with linear mixed models (LMMs) [17, 21].

We demonstrate our approach on a large simulation study. Though traditional methods like principal component analysis (PCA) adjustment [42] closely approximate the family of techniques we describe, we find that our approach more accurately estimates the causal effects, especially when the confounder dimensionality is misspecified. Finally, we apply the MCLBO to control for confounders in a medical prediction problem on health records from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC III) clinical database [16]. We show the effects we recover match those in the literature.

Related Work. Causal inference has a long history in many disciplines including statistics, computer science, and econometrics. A full review is outside of the scope of this article, however, we highlight some of recent advances in building flexible causal models. Wager and Athey [38] develop random forests to capture variability in treatment effects [38]. Hill [13] uses Bayesian nonparametric methods to model the outcome response. Louizos et al. [22] build flexible latent variables to correct for confounding when proxies of confounders are measured, rather than the confounders themselves. Johansson et al. [15], Shalit et al. [34] develop estimators with theoretical guarantees by building representations that penalize differences in confounder distributions between the treated and untreated.

The above approaches focus on building estimators for single treatments, where either the confounder or a non-treatment proxy is measured. In contrast, computational genetics has developed a variety of methods to control for unmeasured confounding in genome-wide association studies (GWAS). Genome-wide association studies have multiple treatments in the form of genetic variations across multiple sites. Yu et al. [42], Kang et al. [17], Lippert et al. [21] estimate kinship matrices between individuals using a subset of the genetic variations, then fit a LMM where the kinship provides the covariance for random effects. Song et al. [35] adjust for confounding via factor analysis on discrete variables and use inverse regression to estimate individual treatment effects. Tran and Blei [37] build implicit models for genome-wide association studies and describe general implicit causal models in the same vein as Kocaoglu et al. [19]. A formulation of multiple

causal inference was also proposed by [40]; they take a model-based approach in the potential outcomes framework that leverages predictive checks.

Our theoretical grounding for multiple causal inference complements this earlier work. We develop the two assumptions of shared confounding and of independence given shared confounders. We describe the information ambiguity inherent in estimating confounders in multiple causal inference and show that a probabilistic model can be insufficient to resolve the ambiguity. We develop a confounder estimator to resolve this by regularization, while minimizing the leave-one-treatment-out prediction error. Earlier work estimates confounders by choosing their dimensionality to not overfit. This matches the flavor of the estimator we develop.

2 Multiple Causal Inference

The trouble with causal inference from observational data lies in confounders, variables that affect both treatments and outcome. The problem is that the observed statistical relationship between the treatment and outcome may be partially or completely due to the confounder. Randomizing the treatment breaks the relationship between the treatment and confounder, rendering the observed statistical relationship causal. But the lack of randomized data necessitates assumptions to control for potential confounders. These assumptions have focused on causal estimation with a single treatment and a single outcome. In real-world settings such as in genetics and medicine, there are multiple treatments. We now define the multiple causal inference problem, detail assumptions for multiple causal inference, and develop new estimators for the causal effects given these assumptions.

Multiple causal inference consists of a collection of causal inference problems. Consider a set of T treatments t_i indexed by i and single outcome y. For example, t_i could be the ith medication for a disease given to a patient and y could be the severity of that disease. The patient's unmeasured traits induce a relationship between the treatments and the outcome. The goal of multiple causal inference is to simultaneously estimate the causal effects for all T treatments. We develop two assumptions under which these causal effects can be estimated in the presence of unobserved confounders.

Shared Confounding. The first assumption we make to identify multiple causal effects is that of *shared confounders*. The shared confounder assumption posits that confounders are shared across all of the treatments. Under this assumption, each treatment provides a view on the shared confounder. This provides the means to control for confounding on other treatments. This assumption is natural in many problems. For example, in GWAS, treatments are genetic variations and the outcome is a phenotype. Due to correlations in genetic variations caused by ancestry, the treatments share confounding structure.

Independence Given Unobserved Confounders. The shared confounding assumption does not identify the causal effects, since there can be links from treatment t_i to treatment t_j. In the presence of these links, we cannot get a clear view of the shared confounder. The is because the dependence between t_i and t_j may be due either to confounding or to the direct link between the pair of treatments. To address this issue, we assume that treatments are independent given unobserved confounders. In the discussion, we explore strategies to loosen this assumption.

Implied Model. We developed two assumptions: shared confounding and independence given the confounder. Together, these assumptions imply the existence of an unmeasured variable $\mathbf{z}$ with some unknown distribution such that when conditioned on, the treatments become independent:

Proposition 1 *Let $\boldsymbol{\epsilon}$ be independent noise terms, and f, g, h be functions. Then, shared confounding and independence given unobserved confounding imply a generative process for the data that is*

$$\mathbf{z} = f(\boldsymbol{\epsilon}_{\mathbf{z}}), \quad t_i = h_i(\boldsymbol{\epsilon}_i, \mathbf{z}), \quad y = g(\boldsymbol{\epsilon}_y, \mathbf{z}, t_1, ...t_T). \tag{1}$$

We require that any given value of the treatments, t_i, be expressible as a function of the treatment noise, $\boldsymbol{\epsilon}_i$, given any value of the confounder, $\mathbf{z}$. Also, we require that the outcome, y, be a

non-degenerate function of the treatment noise, $\boldsymbol{\epsilon}_i$, via the treatments, t_i. In other words, the treatment noise, $\boldsymbol{\epsilon}_i$, must influence the outcome.

If this model were explicitly given, posterior inference would reveal the causal effects. However, an issue remains. Since the information to infer $\mathbf{z}$ comes from the treatments and the outcome, flexible models lead to a problematic ambiguity. Specifically, it is unclear how much information the outcome and each treatment reveal about the confounder. These assumptions lead to the *principle of minimal information*. This principle says that the confounder retains the minimum amount of information about each treatment needed to reconstruct the other treatments. We formalize the notion of information and develop an estimator that controls the information in the next section.

3 Confounder Estimation in Multiple Causal Inference

We develop an estimator for the confounder in multiple causal inference without directly specifying a generative model. This estimator finds the confounder with minimum information that reconstructs each treatment given the other treatments. The estimator works via information-based regularization and cross-validation in a way that is agnostic to the particular functional form of the estimator. We present a tractable lower bound to estimate the confounder.

Information Ambiguity. We formalize the notion of information using mutual information [8]. Let $\mathbb{I}(a, b)$ denote the mutual information. The mutual information is nonnegative and is zero when a and b are independent. To understand the ambiguity in building stochastic confounder estimators, consider the information between the estimated confounder and treatment i given the remaining treatments $\mathbf{t}_{-i}$: $\mathbb{I}(t_i, \mathbf{z} \mid \mathbf{t}_{-i})$. We call this the additional mutual information (AMI). It is the additional information a treatment can provide to the confounder, over what the rest of the treatments provide. The additional mutual information takes values between zero and some nonnegative number. The maximum indicates that $\mathbf{z}$ and $\mathbf{t}_{-i}$ perfectly predict t_i.[1] This range parameterizes the ambiguity of how much information the confounder encodes about treatment i, over the information present in the remaining treatments.[2]

At first glance, letting $\mathbb{I}(t_i, \mathbf{z} \mid \mathbf{t}_{-i}) > 0$ seems to violate shared confounding because the confounder $\mathbf{z}$ has information about a treatment that is not in the other treatments. But setting $\mathbb{I}(t_i, \mathbf{z} \mid \mathbf{t}_{-i}) = 0$ forces the confounder to be independent of all of the treatments. This is in tension with the assumption of the independence of treatments given the confounder. Since if the confounder-estimated entropy $\mathbb{H}(t_i \mid \mathbf{t}_{-i})$ is bigger than the true entropy under the population sampling distribution F, $\mathbb{H}_F(t_i \mid \mathbf{t}_{-i})$, the treatments cannot be independent given the confounder.

The minimal information principle can be seen as a balancing of independence given a shared confounder and having a purely shared confounder. This principle can formally be stated as trying to minimize the AMI $\mathbb{I}(t_i, \mathbf{z} \mid \mathbf{t}_{-i})$ subject to good predictions via the confounder by minimizing predictive entropy $\mathbb{H}(t_i \mid \mathbf{t}_{-i})$.

Confounder Estimation. The most general form of a confounder estimator is a function that takes the following as input: noise $\boldsymbol{\epsilon}$, parameters $\boldsymbol{\theta}$, treatments $\mathbf{t}_j$, and outcome y_j. Using the outcome without extra assumptions is inherently ambiguous. The ambiguity lies in how much of y_j is retained in $\mathbf{z}_j$. The only statistics we observe about y come from y or its cross statistics with $\mathbf{t}$. From eq. (1), we know that the cross statistics provide a way to estimate $\mathbf{z}$ the confounder. However, since the outcome depends on the treatments, cross statistics between the treatment and outcome could either be from the confounder or from the direct relationship between the treatments and outcome. This ambiguity cannot be resolved without further assumptions like assuming a model. Therefore we focus on estimating the confounder without using the outcome, where the outcome has been marginalized out.

A generic stochastic confounder estimator with marginalized outcome is a stochastic function of the treatments and noise with parameters $\boldsymbol{\theta}$. The posterior of the latent confounder in a model is an example of such a stochastic estimator. To respect the assumptions, we want to find a $\boldsymbol{\theta}$

[1] When all variables are discrete, the upper bound is the entropy $\mathbb{H}(t_i \mid \mathbf{t}_{-i})$.

[2] In the appendix, we provide an explicit construction of equivalent models that have different information.

such that conditional on the confounder, the treatments are independent. With the estimator we can directly control the AMI $\mathbb{I}(t_i, \mathbf{z} \mid \mathbf{t}_{-i})$ via regularization. This contrasts classical latent variable models, where parameters like the dimensionality, flexibility of the likelihood, and number of layers in a neural network control the additional mutual information.

For compactness, we drop the confounder's functional dependence on $\boldsymbol{\epsilon}$ and write $\mathbf{z} \sim p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t})$. Let $p(\mathbf{t})$ be the empirical distribution over the observed treatments, let $\boldsymbol{\beta}$ be a parameter, and let α be a regularization parameter. Then we can define an objective that tries to reconstruct each treatment independently given $\mathbf{z}$, while controlling the additional mutual information:

$$\max_{\boldsymbol{\theta}, \boldsymbol{\beta}} \quad \mathbb{E}_{\mathbf{t} \sim p(\mathbf{t})} \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t})} \left[\sum_{i=1}^{T} \log p_{\boldsymbol{\beta}}(t_i \mid \mathbf{z}) \right] - \alpha \sum_{i=1}^{T} \mathbb{I}_{\boldsymbol{\theta}}(t_i, \mathbf{z} \mid \mathbf{t}_{-i}). \tag{2}$$

We will suppress the index i in $p_{\boldsymbol{\beta}}$ when clear. The distributions $p_{\boldsymbol{\beta}}$ and $p_{\boldsymbol{\theta}}$ can be from any class; in practice we use conditional distributions built from neural networks, e.g., $\mathbf{z} \sim \text{Normal}(\mu_{\boldsymbol{\theta}}(\mathbf{t}), \sigma_{\boldsymbol{\theta}}(\mathbf{t}))$, where μ and σ are neural networks. This objective finds the $\mathbf{z}$ that can reconstruct $\mathbf{t}$ most accurately, assuming the treatments are conditionally independent given $\mathbf{z}$. Equation (2) can be viewed as an autoencoder where the code is regularized to limit the additional mutual information, thereby preferring to keep information shared between treatments.

The information regularizer is similar to regularizers in supervised learning. Consider how well the confounder predicts a treatment when estimated conditional on the rest of the treatments. When α is too small for a flexible model, the confounder memorizes the treatment so the prediction error $\mathbb{H}(t_i \mid \mathbf{t}_{-i})$ is big. When α is too large, $\mathbf{z}$ is independent of the treatments so again the prediction error is big. This mirrors choosing the regularization coefficient in linear regression. When the regularization is too large, the regression coefficients ignore the data, and when it is too small, the regression coefficients memorize the data. As in regression, α can be found by cross-validation. Minimizing the conditional entropy directly rather than by cross-validation leads to the degenerate solution of $\mathbf{z}$ having all the information in $\mathbf{t}$.

Since we do not have access to $p(t_i, \mathbf{z} \mid \mathbf{t}_{-i})$, the objective contains an intractable mutual information term. We develop a tractable objective based on the conditional entropy decomposition of conditional mutual information.

Multiple Causal Lower Bound. The conditional mutual information can be written in terms of conditional entropies as

$$\mathbb{I}_{\boldsymbol{\theta}}(t_i, \mathbf{z} \mid \mathbf{t}_{-i}) = \mathbb{H}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i}) - \mathbb{H}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i}, \mathbf{t}_i) = \mathbb{H}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i}) - \mathbb{H}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}).$$

The second term comes from the entropy of $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t})$ and is tractable. But the first term requires marginalizing out the treatment t_i. This conditional entropy with marginalized treatment is not tractable, so we develop a lower bound. Let $p(t_i)$ be the marginal distribution of treatment i; expanding the integral gives

$$-\mathbb{H}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i}) \geq \int p(\mathbf{z} \mid \mathbf{t}) p(\mathbf{t}) \mathbb{E}_{\hat{t}_i^k \sim p(t_i)} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k) \right] d\mathbf{z} d\mathbf{t} := \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i}).$$

The lower bound follows from Jensen's inequality. Unbiased estimates of the lower bound can be computed via Monte Carlo. The bound becomes tight as K goes to the number of observations. Substituting this back, the information-regularized confounder estimator objective gives

$$\mathcal{L} = \mathbb{E}_{\mathbf{t} \sim p(\mathbf{t})} \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t})} \left[\sum_{i=1}^{T} \log p_{\boldsymbol{\beta}}(t_i \mid \mathbf{z}) \right] + \alpha \sum_{i=1}^{T} (\mathbb{G}(\mathbf{z} \mid \mathbf{t}_{-i}) + \mathbb{H}(\mathbf{z} \mid \mathbf{t})). \tag{3}$$

This gives a tractable lower bound to the information-regularized objective called the multiple causal lower bound (MCLBO).

Algorithm. To optimize the MCLBO, we use stochastic gradients by passing the derivative inside expectations [41]. These techniques underly black box variational inference algorithms [29, 18, 31]. We derive the full gradients for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in the appendix. With these gradients, the algorithm can be summarized as follows. We choose a range of α values and fit the confounder

estimator using the MCLBO. We then select the α that minimizes the entropy $\sum_i \mathbb{H}(\mathbf{t}_i \mid \mathbf{t}_{-i})$ on held-out treatments. In practice, we allow a small relative tolerance for larger α's over the best held-out prediction to account for finite sample estimation error. The algorithm can be seen as learning an autoencoder. The code of the this autoencoder minimizes the information retained about each treatment subject to the code predicting each t_i best when the code is computed only from $\mathbf{t}_{-i}$, the remaining treatments.

Necessity of Minimal Information. Even with an explicit probabilistic model, the minimal information principle is necessary. We provide an example model and population treatment distribution that demonstrates this. Suppose the true treatments $\mathbf{t}_j$ come from an unconfounded model, where all of the $\mathbf{t}_j$ are independent. Consider a latent variable model where each observation has a latent variable $\mathbf{z}$ and treatment vector $\mathbf{t}$. Let W be a matrix of parameters, let κ and σ be hyperparameters, and let j index observations. Then the model is

$$\mathbf{z}_j \sim \mathcal{N}(0, \sigma), \quad \mathbf{t}_j \sim \mathcal{N}(W\mathbf{z}_j, \kappa). \tag{4}$$

The maximum likelihood estimate for the model in eq. (4) with latent size equal to data size given this true model is $W^* = I(1-\kappa)$, up to rotations. The posterior distribution is

$$p(\mathbf{z} \mid \mathbf{t}) = \mathcal{N}\left(\frac{\sigma}{\sigma+\kappa}\mathbf{t}, \sigma^2\left(1 - \frac{\sigma}{\sigma+\kappa}\right)\right).$$

From the posterior, we see that small κ leads to the posterior $\mathbf{z}$ memorizing the treatments $\mathbf{t}$—the model learns that all of $\mathbf{t}$ is confounded by $\mathbf{z}$, while the true treatments are unconfounded. This problem occurs because there are a class of models indexed by κ that model the observed $\mathbf{t}$ and have the same conditional entropy and predictive likelihoods. All additional mutual information regularization values α lead to the same prediction. Satisfying the minimal information principle forces identification, in this case preserving the true unconfounded treatments.

4 Estimating the Outcome Model

In traditional observational causal inference, the possible outcomes are independent of the treatments given the confounders, so predictions given confounders are causal estimates. With the do notation that removes any influence from confounding variables, we have

$$\mathbb{E}[y \mid do(\mathbf{t} = \mathbf{t}^*)] = \mathbb{E}_{p(\mathbf{z})}\mathbb{E}[y \mid do(\mathbf{t} = \mathbf{t}^*), \mathbf{z}] = \mathbb{E}_{p(\mathbf{z})}\mathbb{E}[y \mid \mathbf{t}^*, \mathbf{z}].$$

So to estimate the causal effect, it suffices to build a consistent regression model. However with estimated confounders that are stochastic functions of the treatment, this relationship breaks: $\mathbb{I}(\mathbf{z}, y) \leq \mathbb{I}(\mathbf{t}, y)$ and $\mathbb{I}(y, \mathbf{z} \mid \mathbf{t}) = 0$. The confounder has less information about the outcomes than the treatments themselves. Given the treatments, the confounders provide no information about the outcome. The lack of added information means that if we were to simply regress $\mathbf{t}$ and $\mathbf{z}$ to y, the regression could completely ignore the confounder. Building outcome models conditional on $\mathbf{z}$ addresses this issue, but this requires doing regression for every value of $\mathbf{z}$.

A regression conditional on the confounder can only use the part of the treatment that is independent of the confounder. Recovering these independent components allows for the use of regression. Formally, let $\boldsymbol{\epsilon}_i$ be the independent component of the ith treatment, then we would like to find a distribution $p(\boldsymbol{\epsilon}_i \mid \mathbf{z}, \mathbf{t})$ that maximizes

$$\mathbb{E}_{p(\mathbf{t})p_\theta(\mathbf{z}\mid\mathbf{t})\prod_i p(\boldsymbol{\epsilon}_i \mid \mathbf{z}, t_i)}\left[\sum_{i=1}^{T} \log p(t_i \mid \mathbf{z}, \boldsymbol{\epsilon}_i)\right], \text{such that } \mathbb{I}(\boldsymbol{\epsilon}_i, \mathbf{z}) = 0. \tag{5}$$

Optimizing this objective over $p(\boldsymbol{\epsilon}_i \mid \mathbf{z}, \mathbf{t})$ and $p(t_i \mid \mathbf{z}, \boldsymbol{\epsilon}_i)$ provides stochastic estimates of the part of t_i independent of $\mathbf{z}$. We call this leftover part $\boldsymbol{\epsilon}_i$ the residuals. These residuals are independent of the confounders and can be used in conjunction with them to estimate the outcome. The residuals mirror instrumental variables; they are independent and affect the outcome only through the treatments. Optimizing eq. (5) can be a challenge both due to the intractable mutual information constraint and that $p(t_i \mid \mathbf{z}, \boldsymbol{\epsilon}_i)$ may have degenerate density.

In the appendix, we provide a general estimation technique for the residuals. Here we focus on the degenerate case. Suppose $t_i \sim p_{\boldsymbol{\beta}}(t_i \mid \mathbf{z})$ in eq. (2) can be for, some function d, written

as $t_i = d(\mathbf{z}, \boldsymbol{\epsilon}_i)$ for $\boldsymbol{\epsilon}_i$ drawn independently. Then if d is invertible for every fixed value of $\mathbf{z}$, the residuals $\boldsymbol{\epsilon}_i$ that satisfy eq. (5) can be found via inversion. That is, eq. (5) is optimal if $\boldsymbol{\epsilon}_i = d^{-1}(\mathbf{z}, t_i)$, since $\boldsymbol{\epsilon}_i$ is independent of $\mathbf{z}$ by construction and in conjunction with $\mathbf{z}$ perfectly reconstructs $\mathbf{t}_i$. This means when the reconstruction in eq. (2) is invertible, independent residuals are easy to compute.

To learn the outcome model, we regress with the residuals and confounder by maximizing

$$\mathbb{E}_{p(y,\mathbf{t})p_\theta(\mathbf{z}\mid\mathbf{t})p(\boldsymbol{\epsilon}_i\mid\mathbf{z},\mathbf{t})}[\log p(y\mid\mathbf{z},\boldsymbol{\epsilon})]. \tag{6}$$

Since $\boldsymbol{\epsilon}$ and $\mathbf{z}$ are independent, they provide different information to y. To compute the desired causal estimate, $p(y\mid\mathbf{z}, do(\mathbf{t}))$ given the learned $p(y\mid\mathbf{z},\boldsymbol{\epsilon})$, we can substitute:

$$p(y\mid\mathbf{z}, do(\mathbf{t}=\mathbf{t}^*)) = p(y\mid\mathbf{z},\mathbf{t}^*) = p(y\mid\mathbf{z},\boldsymbol{\epsilon}=d^{-1}(\mathbf{t}^*,\mathbf{z})). \tag{7}$$

The right hand side is in terms of known quantities: the outcome model from eq. (6) and the $\boldsymbol{\epsilon}$ from the confounder estimation in eq. (3). The causal estimate of $do(\mathbf{t}=\mathbf{t}^*)$ can be computed by averaging over $p(\mathbf{z})$. We call the process of confounder estimation with the MCLBO followed by outcome estimation with residuals the multiple causal estimation via information (MCEI). We formalize causal recovery with MCEI in a simple setting.

Proposition 2 *If the confounder is finite dimensional and the treatments are i.i.d. given the confounder, then the multiple causal estimator in eq. (2) combined with eq. (7) recovers the correct causal estimate as $T \to \infty$, $N \to \infty$.*

The intuition is that as $T \to \infty$, we get perfect estimates of $\mathbf{z}$. Moreover, the amount of information about each treatment in the confounder goes to zero, so the information ambiguity gets resolved and we return to the classical causal inference setup. Asymptotics require constraints for the outcome model to be well defined. For example, with normally distributed outcomes the variance needs to be finite [10]. This proposition can be generalized to non-identically distributed treatments where posterior concentration occurs.

5 Experiments

We demonstrate our approach on a challenging simulation where the noise also grows with the amount of confounding. We study multiple variants of this simulation across many replications. In total, we estimate over a hundred different models. We also study a real-world example from medicine. Here, we look at the effects of various lab values prior to entering the intensive care unit on the length of stay in the hospital.

Simulation. Consider a model with real-valued treatments. Let n index the observations and i the treatments. Let W be a parameter matrix, σ be the simulation standard deviation, and D be the dimensionality of $\mathbf{z}$. The treatments are drawn conditional on an unobserved $\mathbf{z}_n$ as

$$\mathbf{z}_n \sim \text{Normal}(0,\gamma), \quad \boldsymbol{\epsilon}_n \sim \text{Normal}(0, 1-\gamma), \quad t_{i,n} \sim \text{Normal}(W\mathbf{z}+\boldsymbol{\epsilon}_n, \sigma), \tag{8}$$

where γ scales the influence of $\mathbf{z}_n$ on each of the treatments. The outcomes are drawn in a way so that the effects of $\boldsymbol{\epsilon}_n$ and $\mathbf{z}_n$ cancel each other out. Let b be a vector of weights and σ_y be the outcome standard deviation. Then the outcomes are

$$y_n \sim \text{Normal}((1-\gamma)b^\top\boldsymbol{\epsilon} - \gamma b^\top W\mathbf{z}_i, \sigma_y).$$

The amount of confounding grows with γ. Moreover since the weights b are shared and opposite in sign between the treatment part $b^\top\boldsymbol{\epsilon}$, and confounding part $b^\top(W\mathbf{z}_i)$, the effects when simply considering the observed confounded treatment $\mathbf{t}$ can cancel each other out. This means to have any hope of recovering the true b requires $\mathbf{t}_i$ be split apart into the confounded and unconfounded parts. This, combined with the fact the noise grows relative to the unconfounded portion as γ increases, makes this simulation a challenge.

We compare our approach to the PCA correction [42] and the LMM [17, 21]. These approaches should perform well since the process in eq. (8) matches the assumptions behind probabilistic PCA [5]. Both these methods fall into the theoretical framework we developed. For confounder estimation by the MCLBO, we limit our approach to have a similar number of parameters. Details are in the appendix.

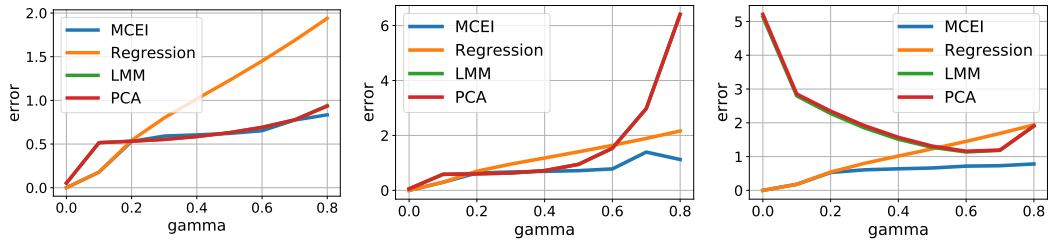

Figure 1: Simulation results for the correctly specified (left), underspecified (center), overspecified (right) confounder dimensionality. MCEI performs similar or better than PCA and the mixed model and better tolerates misspecification. The LMM performs the same as PCA.

We study two cases. First, we correctly give all methods the right dimensionality $D = 2$. Second, we misspecify: all methods use a smaller dimension 2, while the true $D = 4$, and the reverse setting where the dimensionality (40) is much bigger than the true $D = 2$. We measure MSE to the true parameters scaled by the true norm. We simulate 10,000 observations with 50 treatments over 5 redraws. We describe the remaining simulation parameters in the appendix.

Figure 1 shows the results. The left panel plots the error for varying levels of confounding when the confounder dimension is correctly specified. We find that confounder estimation with MCEI performs similar to or better than PCA and the LMM. This difference is larger when the confounding grows. This is mostly due to the better variance control from fitting all treatments at once (PCA with all treatments has degenerate design). The middle and right panels show MCEI tolerates misspecification better than PCA and the linear mixed model.

Clinical Experiment. Length of stay (LOS) is defined as the duration of a hospital visit. This measure is often used as an intermediate outcome in studies due to the associated adverse primary outcomes. Patient flow is important medically because unnecessarily prolonged hospitalization places patients at risk for hospital acquired infections (among other adverse outcomes). These can be difficult to treat and are associated with significant morbidity, mortality, and cost. Studies have found a 1.37% increase in infection risk and 6% increase in any adverse event risk for each excess LOS day [12, 2]. Also, it is of operational concern for hospitals because reimbursement for medical care is increasingly tied to visit episodes rather than to discrete products or services provided [28].

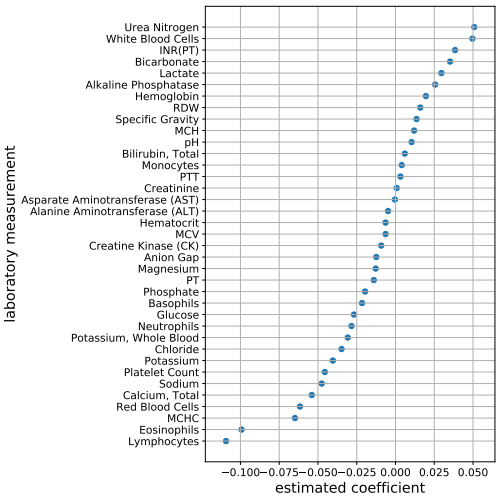

Figure 2: Causal estimates for effects of laboratory values on ICU length of stay.

The dataset studied in this experiment is comprised of 25753 ICU visits and 37 laboratory tests from the MIMIC III clinical database [16]. We applied our MCEI approach to laboratory tests measured in the emergency department prior to admission as treatments, and a binarized LOS based on the average value as outcome. Laboratory test values were shifted, log-transformed, and then standardized with missing values imputed by randomly sampling the empirical distribution of the laboratory test type.

The results are shown in fig. 2 and correlate well with findings in the literature regarding factors influencing LOS. For example, elevated blood urea nitrogen is associated with states of hypovolemia and hypercatabolism and has been linked to increased LOS in pancreatitis and stroke patients [11, 20]. Elevated white blood cells, or leukocytosis, is one of the main markers for infection and, as expected, infection has been associated with increased LOS, particularly when systemic [4, 36]. Other findings, such as an inverse relationship to potassium (hypokalemia) and platelets (thrombocytopenia) are also supported by the literature [26, 9].

6 Discussion

We formalized two assumptions needed for multiple causal inference, namely shared confounding and independence between treatments given the shared confounders. Together, these assumptions imply a minimal information principle that tries to find the confounder that has minimal information with the treatments while rendering the treatments conditionally independent given the confounder. We developed a tractable algorithm to estimate the confounder based on the minimal information principle. We showed how stochastic residuals can be used to estimate the outcome model, and we demonstrated our approach in simulations and on ICU data.

Many future directions remain. First, the assumptions we made are likely not tight. For example, the independence between treatments given the shared confounder could be relaxed to allow a finite number of dependencies between observations. The intuition is that if there is a limited amount of dependence between treatments, the confounder can be estimated from the other treatments. Next, in the algorithm to estimate the information, the lower bound can be replaced by likelihood ratio estimation. This has the added benefit of removing slack from the bound, while also improving numerical stability by avoiding differences of numbers of arbitrary magnitude. In this work, we focused on estimation with a single outcome. With multiple outcomes, new kinds of estimators that are simpler can be developed.

Acknowledgments

We would like to acknowledge Jaan Altosaar, Fredrik Johansson, Rahul Krishnan, Bharat Srikishan, and Alexander D'Amour for helpful discussion and comments.

References

[1] Agakov, F. V. and Barber, D. (2004). An auxiliary variational method. In *Neural Information Processing*, pages 561–566.

[2] Andrews, L. B., Stocking, C., Krizek, T., Gottlieb, L., Krizek, C., Vargish, T., and Siegler, M. (1997). An alternative strategy for studying adverse events in medical care. *The Lancet*, 349(9048):309–313.

[3] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

[4] Beyersmann, J., Kneib, T., Schumacher, M., and Gastmeier, P. (2009). Nosocomial infection, length of stay, and time-dependent bias. *Infection Control & Hospital Epidemiology*, 30(3):273–276.

[5] Bishop, C. (2016). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

[6] Consortium, W. T. C. C. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661.

[7] Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

[8] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[9] Crowther, M. A., Cook, D. J., Meade, M. O., Griffith, L. E., Guyatt, G. H., Arnold, D. M., Rabbat, C. G., Geerts, W. H., and Warkentin, T. E. (2005). Thrombocytopenia in medical-surgical critically ill patients: prevalence, incidence, and risk factors. *Journal of critical care*, 20(4):348–353.

[10] D'Amour, A. (2018). (Non-)identification in latent confounder models. *http://www.alexdamour.com/blog/public/2018/05/18/non-identification-in-latent-confounder-models/*.

[11] Faisst, M., Wellner, U. F., Utzolino, S., Hopt, U. T., and Keck, T. (2010). Elevated blood urea nitrogen is an independent risk factor of prolonged intensive care unit stay due to acute necrotizing pancreatitis. *Journal of critical care*, 25(1):105–111.

[12] Hassan, M., Tuckman, H. P., Patrick, R. H., Kountz, D. S., and Kohn, J. L. (2010). Hospital length of stay and probability of acquiring infection. *International Journal of pharmaceutical and healthcare marketing*, 4(4):324–338.

[13] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[14] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

[15] Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029.

[16] Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

[17] Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348.

[18] Kingma, D. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR-14)*.

[19] Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2017). Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*.

[20] Lin, W.-C., Shih, H.-M., and Lin, L.-C. (2015). Preliminary prospective study to assess the effect of early blood urea nitrogen/creatinine ratio-based hydration therapy on poststroke infection rate and length of stay in acute ischemic stroke. *Journal of Stroke and Cerebrovascular Diseases*, 24(12):2720–2727.

[21] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833.

[22] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6449–6459.

[23] Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*.

[24] Morgan, S. L. and Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.

[25] O'connor, C., Adhikari, N. K., DeCaire, K., and Friedrich, J. O. (2009). Medical admission order sets to improve deep vein thrombosis prophylaxis rates and other outcomes. *Journal of hospital medicine*, 4(2):81–89.

[26] Paltiel, O., Salakhov, E., Ronen, I., Berg, D., and Israeli, A. (2001). Management of severe hypokalemia in hospitalized patients: a study of quality of care based on computerized databases. *Archives of internal medicine*, 161(8):1089–1095.

[27] Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

[28] Press, M. J., Rajkumar, R., and Conway, P. H. (2016). Medicare's new bundled payments: design, strategy, and evolution. *Jama*, 315(2):131–132.

[29] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

[30] Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333.

[31] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

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

[33] Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226.

[34] Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085.

[35] Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature genetics*, 47(5):550.

[36] Talmor, M., Hydo, L., and Barie, P. S. (1999). Relationship of systemic inflammatory response syndrome to organ dysfunction, length of stay, and mortality in critical surgical illness: effect of intensive care unit resuscitation. *Archives of surgery*, 134(1):81–87.

[37] Tran, D. and Blei, D. M. (2017). Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*.

[38] Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

[39] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM.

[40] Wang, Y. and Blei, D. M. (2018). The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*.

[41] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer.

[42] Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203.

A Appendix

Explicit Ambiguity. We make the idea of explicit ambiguity precise by constructing two models. Both models have the same distribution of treatments and outcomes and have treatments that are independent of the outcome. Take the model

$$\begin{aligned}\mathbf{z} &= f(\boldsymbol{\epsilon}_{\mathbf{z}})\\ t_i &= h_i(\boldsymbol{\epsilon}_i, \mathbf{z})\\ y &= g(\boldsymbol{\epsilon}_y, \mathbf{z}, t_1, ...t_T),\end{aligned}$$

and the model

$$\begin{aligned}\mathbf{z} &= f(\boldsymbol{\epsilon}_{\mathbf{z}}), \boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_T\\ t_i &= h_i(\mathbf{z})\\ y &= g(\boldsymbol{\epsilon}_y, \mathbf{z}, h_1(\mathbf{z}), ..., h_T(\mathbf{z})).\end{aligned}$$

Both of these models satisfy the independence of treatments given the shared confounder and have the same joint distribution on $\mathbf{t}, y$. But the second model differs in key way. It assumes all of the treatments are due to confounding. Since the two models have the same observed distribution, we need assumptions to choose between them.

We could start by arguing that this model does not satisfy shared confounding because parts of $\mathbf{z}$ only relate to a single treatment. Howeverm limiting $\mathbf{z}$ by requiring it to contain only shared information $\mathbb{I}_{\boldsymbol{\theta}}(t_i, \mathbf{z} \mid \mathbf{t}_{-i}) = 0$ requires the confounder be independent of the treatments. In practice it can be hard to prevent a confounder from memorizing the $\boldsymbol{\epsilon}_i$, as even a single dimensional $\mathbf{z}$ with flexible h_i can memorize the treatments. This is why we need the *minimal information principle* along with the regularizer it induces.

Negative Entropy Lower Bound

$$\begin{aligned}-\mathbb{H}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i}) &= \int p(\mathbf{z}, \mathbf{t}_{-i}) \log p(\mathbf{z} \mid \mathbf{t}_{-i}) d\mathbf{z} d\mathbf{t}_{-i}\\ &= \int \int p(\mathbf{z}, \mathbf{t}_{-i} \mid t_i) p(t_i) dt_i \log p(\mathbf{z} \mid \mathbf{t}_{-i}) d\mathbf{z} d\mathbf{t}_{-i}\\ &= \int p(\mathbf{z}, \mathbf{t}_{-i} \mid t_i) p(t_i) \log p(\mathbf{z} \mid \mathbf{t}_{-i}) d\mathbf{z} d\mathbf{t}\\ &= \int p(\mathbf{z} \mid \mathbf{t}_{-i}, t_i) p(\mathbf{t}_{-i} \mid t_i) p(t_i) \log p(\mathbf{z} \mid \mathbf{t}_{-i}) d\mathbf{z} d\mathbf{t}\\ &= \int p(\mathbf{z} \mid \mathbf{t}) p(\mathbf{t}) \log p(\mathbf{z} \mid \mathbf{t}_{-i}) d\mathbf{z} d\mathbf{t}\\ &= \int p(\mathbf{z} \mid \mathbf{t}) p(\mathbf{t}) \log \left[\int p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i) p(t_i = \hat{t}_i) d\hat{t}_i \right] d\mathbf{z} d\mathbf{t}\\ &\geq \int p(\mathbf{z} \mid \mathbf{t}) p(\mathbf{t}) \mathbb{E}_{\hat{t}_i^k \sim p(t_i)} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k) \right] d\mathbf{z} d\mathbf{t}\\ &:= \mathbb{G}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{t}_{-i})\end{aligned}$$

The lower bound follows via the relationship from Jensen's inequality,

$$\begin{aligned}\log \left[\int p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i) p(t_i = \hat{t}_i) d\hat{t}_i \right] &= \log \frac{1}{K} \sum_{k=1}^{K} \left[\int p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k) p(t_i = \hat{t}_i^k) d\hat{t}_i^k \right]\\ &= \log \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\hat{t}_i^k \sim p(t_i)} \left[p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k) \right]\\ &\geq \mathbb{E}_{\hat{t}_i^k \sim p(t_i)} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k) \right]. \qquad (9)\end{aligned}$$

Algorithm 1: Confounder estimation via lower bound for fixed α

Input : Reconstruction $p_{\beta}(\mathbf{t}_i \mid \mathbf{z})$,
Stochastic confounder estimate $p_{\theta}(\mathbf{z} \mid \mathbf{t})$
Output : Confounder estimate parameters θ
Initialize β and θ randomly.
while *not converged* **do**
  Compute unbiased estimate of $\nabla_{\theta}\mathcal{L}$. (eq. (11))
  Compute unbiased estimate of $\nabla_{\beta}\mathcal{L}$. (eq. (12))
  Update θ and β using stochastic gradient ascent.
end

Proposition 1. Independence given the confounder means that t_i is independent of t_j given the unobserved confounder. Shared confounding means there is only a single confounder $\mathbf{z}$. Since the form of f is arbitrary, the distribution on $\mathbf{z}$ is arbitrary. Also, since h_i is arbitrary the distribution of t_j given $\mathbf{z}$ is arbitrary. Thus the generative process in Equation (1) constructs treatments that are conditionally independent given the confounder. It can represent any distribution for each treatment given the confounder. The confounder can also can take any distribution. This means that Equation (1) can represent any distribution of treatments that satisfy both assumptions, of shared confounding and of independence given confounding. The outcome function g is arbitrary and so can be chosen to match any true outcome model.

Gradients of the MCLBO. The first term in the MCLBO denoted $\mathbb{G}$ and the conditional entropy are all integrals with respect to the distribution $p_{\theta}(\mathbf{z} \mid \mathbf{t})$. To compute stochastic gradients, we differentiate under the integral sign as in variational inference. For simplicity, we assume that a sample from $p_{\theta}(\mathbf{z} \mid \mathbf{t})$ can be generated by transforming parameter-free noise $\epsilon \sim s$ through a function $\mathbf{z} = \mathbf{z}(\epsilon, \theta, \mathbf{t})$. This assumption leads to simpler gradient computation [18, 31].

Define

$$\tilde{p}_{\theta}(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k) = \frac{\tilde{p}_{\theta}(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k)}{\sum_{j=1}^{K} \tilde{p}_{\theta}(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^j)}. \tag{10}$$

Then the gradient with respect to θ can be written as

$$\begin{aligned}\nabla_{\theta}\mathcal{L} =& \mathbb{E}_{p(\mathbf{t})}\mathbb{E}_{s(\epsilon)}\left[\nabla_{\theta}\mathbf{z}(\epsilon, \theta, \mathbf{t})\nabla_{\mathbf{z}}\sum_{i=1}^{T}\log p_{\beta}(t_i \mid \mathbf{z})\right] \\ &+ \alpha\sum_{i=1}^{T}\mathbb{E}_{p(\mathbf{t})}\mathbb{E}_{s(\epsilon)}\mathbb{E}_{\hat{t}_i^k \sim p(t_i)}\left[\sum_{k=1}^{K}\tilde{p}_{\theta}(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k)\nabla_{\theta}\log p_{\theta}(\mathbf{z} \mid \mathbf{t}_{-i}, \hat{t}_i^k)\right] \\ &- \alpha\mathbb{E}_{p(\mathbf{t})}\mathbb{E}_{s(\epsilon)}\left[\nabla_{\theta}\mathbf{z}(\epsilon, \theta, \mathbf{t})T\nabla_{\mathbf{z}}\log p_{\theta}(\mathbf{z} \mid \mathbf{t})\right].\end{aligned} \tag{11}$$

Sampling from the various expectations gives a noisy unbiased estimate of the gradient. The gradient for β is much simpler, as the sampled distributions do not depend on β:

$$\nabla_{\beta}\mathcal{L} = \mathbb{E}_{p(\mathbf{t})}\mathbb{E}_{p_{\theta}(\mathbf{z} \mid \mathbf{t})}\left[\sum_{i=1}^{T}\nabla_{\beta}\log p_{\beta}(\mathbf{t} \mid \mathbf{z})\right]. \tag{12}$$

Sampling from the observed data then sampling the confounder estimate gives an unbiased estimate of this gradient. The confounder estimation for a fixed value of α is summarized in Algorithm 1.

Equivalent Confounders. Invertible transformations of a random variable preserve the information in that random variable. Take two distributions for computing the stochastic confounder $\mathbf{z}_1 \sim p_1(\cdot \mid \mathbf{t})$ and $\mathbf{z}_2 \sim p_2(\cdot \mid \mathbf{t})$ where $\mathbf{z}_2$ can be written as an invertible function of $\mathbf{z}_1$. These two distributions have equivalent information for downstream tasks, such as building the outcome model or conditioning on the confounder. This equivalence means we have choice on which member in the equivalence class we choose. One way to narrow the choice is to enforce that the dimensions of $\mathbf{z}$ are independent by minimizing total correlation.

Connection to Factor Analysis. Factor analysis methods work by specifying a generative model for observations that independently generate each dimension of each observation. In its most general form this model is

$$\mathbf{z}_i = f(\boldsymbol{\epsilon}_z),$$
$$\mathbf{t}_{i,j} = h_j(\boldsymbol{\epsilon}_y, \mathbf{z}_i).$$

Inference in this model matches the reconstruction term inside our confounder estimator with a KL-divergence regularizer. If we allow for the parameters of the prior on $\mathbf{z}$ to be learned to maximize the overall likelihood, and if $\mathbf{z}$'s dimensions are independent, then inference corresponds to minimizing the reconstruction eq. (2) with a total correlation style penalty.

There are many ways to choose the complexity of the factor model. One choice is to find the smallest complexity model that still gives good predictions of $\mathbf{t}_i$ given $\mathbf{t}_{-i}$ (like document completion evaluation in topic models [39]). Here complexity is measure in terms of the dimensionality of $\mathbf{z}$ and the complexity of h_j and f. This choice tries to minimize the amount of information retained in $\mathbf{z}$, while still reconstructing the treatments well. This way to select the factor analysis model's complexity meets the condition of the minimum information principle. However, selecting discrete parameters like dimensionality give less fine-grained control over the information rates.

Proposition 2. If the data are conditionally i.i.d., then in the true model $\mathbf{z}$ concentrates as the number of treatments goes to infinity. In this setting, we can learn the model from Proposition 1 using the MCLBO. This follows because the information each treatment provides goes to zero as $T \to \infty$ since they are conditionally i.i.d., thus the true confounder (and posterior), up to information equivalences, is simply a point that maximizes the reconstruction term in the MCLBO subject to asymptotically zero AMI. This shows outcome estimation corresponds to simple regression with treatments and confounder (up to an information equivalence), which correctly estimates the causal effects as $N \to \infty$.

Estimating $\boldsymbol{\epsilon}_i$. The $\boldsymbol{\epsilon}_i$ estimation requires finding parameters $\boldsymbol{\lambda}$ and ξ that maximize

$$\mathbb{E}_{p(\mathbf{t})p_\theta(\mathbf{z}|\mathbf{t})\prod_i p_\lambda(\boldsymbol{\epsilon}_i|\mathbf{z},\mathbf{t}_i)}\left[\sum_{i=1}^{T}\log p_\xi(t_i \mid \mathbf{z}, \boldsymbol{\epsilon}_i)\right], \text{such that } \mathbb{I}[\boldsymbol{\epsilon}_i, \mathbf{z}] = 0.$$

The constraint can be baked into a Lagrangian with parameter κ,

$$\mathbb{E}_{p(\mathbf{t})p_\theta(\mathbf{z}|\mathbf{t})\prod_i p_\lambda(\boldsymbol{\epsilon}_i|\mathbf{z},\mathbf{t}_i)}\left[\sum_{i=1}^{T}\log p_\xi(t_i \mid \mathbf{z}, \boldsymbol{\epsilon}_i)\right] - \kappa\mathbb{I}[\boldsymbol{\epsilon}_i, \mathbf{z}].$$

The mutual information can be split into entropy terms:

$$\mathbb{I}[\boldsymbol{\epsilon}_i, \mathbf{z}] = \mathbb{H}(\boldsymbol{\epsilon}_i) - \mathbb{H}(\boldsymbol{\epsilon}_i \mid \mathbf{z}).$$

The first term can be bounded as before using multiple samples. The second term requires an alternative approach. We can use the entropy bounds with auxiliary distributions on the conditioning set, as used in variational inference [1, 33, 30, 23]. These bounds work with a distribution over the reverse conditioning set in this case $r(\mathbf{t}|\mathbf{z}, \boldsymbol{\epsilon}_i)$. For this, we can use the reconstruction distribution $p_\xi(t_i \mid \mathbf{z}, \boldsymbol{\epsilon}_i)$ and the fact that $p(\mathbf{z})$ and $p(\mathbf{t})$ do not depend on the parameters ξ and $\boldsymbol{\theta}$.

Confounder Parameterization and Simulation Hyperparameters. We limit the confounder to have similar complexity as PCA. We do this by using a confounder distribution with normal noise, where we restrict the mean of the confounder estimate to be a linear function of the treatments $\mathbf{t}$. The variance is independent and controlled by a two-layer (for second moments) neural network. We similarly limit the likelihoods and outcome model to have linear means and fixed variance.

For the remaining simulation hyperparameters, we set W and b to be the absolute value of draws from the standard normal. We fix the simulation standard deviation to 0.02 and fix outcome standard deviation to 0.1.