Counterfactual Inference of Second Opinions

Abstract

Automated decision support systems that are able to infer second opinions from experts can potentially facilitate a more efficient allocation of resources—they can help decide when and from whom to seek a second opinion. In this paper, we look at the design of this type of support systems from the perspective of counterfactual inference. We focus on a multiclass classification setting and first show that, if experts make predictions on their own, the underlying causal mechanism generating their predictions needs to satisfy a desirable set invariant property. Further, we show that, for any causal mechanism satisfying this property, there exists an equivalent mechanism where the predictions by each expert are generated by independent sub-mechanisms governed by a common noise. This motivates the design of a set invariant Gumbel-Max structural causal model where the structure of the noise governing the sub-mechanisms underpinning the model depends on an intuitive notion of similarity between experts which can be estimated from data. Experiments on both synthetic and real data show that our model can be used to infer second opinions more accurately than its non-causal counterpart.

1 INTRODUCTION

In decision making under uncertainty, seeking opinions from multiple human experts tends to improve the overall quality of the decisions. For example, in medicine, second opinions have been shown valuable for establishing diagnoses and initiating treatment [Burger et al., 2020] as well as reducing the number of unnecessary procedures [Leape, 1989, Althabe et al., 2004]. In machine learning, ground truth labels are determined by carefully aggregating multiple noisy labels provided by different experts [Zhang et al., 2016] and inconsistencies between these noisy labels help developing more robust models [Peterson et al., 2019]. Unfortunately, the timeliness and quality of the decisions is often compromised due to a shortage of experts, which prevents each decision to be informed by multiple experts' opinions.

In this context, we argue that the development of automated decision support systems that, given an expert's opinion on a decision instance and a set of features, are able to infer other experts' opinions will enable a more efficient allocation of resources. On the one hand, these systems could prevent (prioritize) seeking other experts' opinions when they are unlikely (likely) to bring new perspectives. On the other hand, these systems could also help identify those experts whose opinion is most likely to disagree with that of the expert sought first. Here, it is worth noting that several studies have also argued that decision support systems that identify disagreement between experts may help identify when a decision instance would benefit most from a second opinion [Raghu et al., 2019, Lim et al., 2021]. However, these studies do not focus on inferring other experts' opinions given an expert's opinion on a decision instance and a set of features, as we do in our work.

More specifically, we consider a multiclass classification setting where, for each instance, experts form their opinions on their own (*i.e.*, without communicating).[1] In this setting, each expert's opinion reduces to a label prediction. Then, our goal is to design decision support systems that, given an expert's prediction on an instance with a set of features, are able to infer other experts' predictions about the same instance, as illustrated in Figure 2 in Appendix A. To this end, one could resort to standard supervised learning. Under this perspective, for each instance, the given expert's prediction would be just an additional feature about the instance. Unfortunately, this would limit the applicability of the resulting supervised learning model to the unrealistic

[1] This setting fits a variety of real-world applications. For example, when a patient is diagnosed by multiple doctors, each doctor diagnoses the patient separately.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

scenario where, for each possible pair of experts, we observe a sizeable number of instances where both experts made a prediction. To circumvent this limitation, we look at the design of the above systems from the perspective of counterfactual inference.

Our contributions. We first show that, if experts form their opinions for each instance on their own, the underlying causal mechanism generating the experts' predictions needs to satisfy a certain set invariant property. Moreover, we further show that any structural causal model satisfying the above set invariant property (in short, any SI-SCM) also satisfies two additional desirable properties:

(i) there exists an equivalent SI-SCM where each expert's predictions are generated by independent submechanisms governed by a common (multidimensional) noise;
(ii) given an expert's prediction on an instance with a set of features, the conditional interventional distribution and the counterfactual distribution of another expert's predictions entailed by the SI-SCM are identical[2].

These properties suggest the following natural strategy to design and train SI-SCM based decision support systems. In a first step, we can use interventional data about each expert—her predictions on a set of instances—to determine the structure of each sub-mechanism separately. One can view this step as multiple independent supervised learning problems, one per expert. In a second step, we can use a small amount of interventional data about multiple experts making predictions about a joint set of instances to characterize similarity across experts and factorize the noise governing the sub-mechanisms into a set of noise components. In a way, in this second step, we are adding a wrapper to the supervised learning models characterizing each expert's sub-mechanism to be able to make counterfactual predictions about second opinions.

To implement the above strategy, we introduce a specific class of SI-SCMs based on the Gumbel-Max structural causal model [Oberst and Sontag, 2019] (in short, Gumbel-Max SI-SCM) and characterize similarity across pairs of experts using the concept of counterfactual stability[3]. In the Gumbel-Max SI-SCM, each expert's submechanism is governed by a Gumbel-Max noise variable and submechanisms of similar experts may be governed by the same noise variable. Further, we show that the problem of uniquely associating each of these noise variables with disjoint sets of mutually similar experts given data reduces to the clique partioning problem, a problem known to be NP-hard [Grötschel and Wakabayashi, 1989, 1990], and propose a simple randomized greedy algorithm with good performance.

Finally, we experiment with synthetic and real data comprising of $20{,}426$ expert predictions over $1{,}560$ natural images. The results on synthetic data show that our randomized greedy algorithm can successfully recover the disjoint sets of mutually similar experts underpinning a specific Gumbel-Max SI-SCM from data. The results on real data show that the (counterfactual) predictions provided by the Gumbel-Max SI-SCM are more accurate that those provided by its non-causal counterpart.

Further related work. Predictions by different experts have been typically studied separately, *i.e.*, without conditioning on an observed prediction by a given expert [Dawid and Skene, 1979, Welinder and Perona, 2010, Guan et al., 2018, Kerrigan et al., 2021, Straitouri et al., 2022]. One could think of the observed prediction just as an additional feature when inferring other experts' predictions, however, this would limit the applicability of existing inference methods to scenarios where, for each pair of experts, we observe a sizeable number of instances where both experts made a prediction, as discussed previously. More broadly, our work is not the first to use counterfactual reasoning in expert prediction [Bica et al., 2020]. However, previous work has used counterfactual reasoning to quantify an expert's preference over counterfactual outcomes rather than to infer other experts' predictions conditioning on a given expert's prediction.

Counterfactual inference has a long and rich history [Imbens and Rubin, 2015]. However, it has mostly focused on estimating quantities related to the interventional distribution of interest such as, *e.g.*, the conditional average treatment effect (CATE). A few notable exceptions are by Oberst and Sontag [2019] and Tsirtsis et al. [2021], which use the Gumbel-Max SCM to reason about counterfactual distributions in Markov decision processes (MDPs), and by Lorberbom et al. [2021], which introduces a parameterized family of causal mechanisms that generalize the Gumbel-Max SCM and are specifically-tuned to a distribution of observations and interventions of interest. However, the Gumbel-Max structural causal model has not been used previously to reason about counterfactual expert predictions.

2 PRELIMINARIES

Given a set of random variables[4] $\mathbf{X} = \{X_1, \dots X_n\}$, a structural causal model (SCM) $\mathcal{M}$ defines a complete data-generating process via a collection of assignments

$$X_i = f_i(\mathbf{PA}_i, U_i),$$

[2]Under the conditional interventional distribution, both experts have made a prediction but we only observe one of them. Under the counterfactual distribution, only one expert has made a prediction, which we observe.

[3]Counterfactual stability is, in general, an axiomatic requirement imposed to counterfactual distributions [Oberst and Sontag, 2019]. However, in SI-SCMs, it is verifiable from interventional data due to (ii), as shown in Theorem 7.

[4]We denote random variables with capital letters and realizations of random variables with lower case letters.

where $\mathbf{PA}_i \subseteq \mathbf{X} \setminus X_i$ are the direct causes of X_i, $\mathbf{F} = \{f_1, \ldots, f_n\}$ are deterministic causal mechanisms, $\mathbf{U} = \{U_1, \ldots, U_n\}$ are jointly independent noise random variables, and $P(\mathbf{U})$ denotes the (prior) distribution of the noise variables. Here, note that the noise variables $\mathbf{U}$ are the only source of stochasticity and, given an observational distribution $P(\mathbf{X})$, there always exists a distribution $P(\mathbf{U})$ and mechanisms $\mathbf{F}$ so that $P = P^{\mathcal{M}}$, where $P^{\mathcal{M}}$ is the distribution entailed by $\mathcal{M}$.

Two SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ over variables $\mathbf{X}$ and $\mathbf{U}$, with noise distribution $P(\mathbf{U})$ and mechanisms $\mathbf{F}$ and $\tilde{\mathbf{F}}$ respectively, are equivalent if, for all $i \in [n]$, it holds that

$$x_i = f_i(\mathbf{pa}_i, u_i) \iff x_i = \tilde{f}_i(\mathbf{pa}_i, u_i).$$

for any realization $\mathbf{PA}_i = \mathbf{pa}_i$ and $P(\mathbf{U})$-almost every u_i.

Given a SCM $\mathcal{M}$, an atomic intervention $\mathcal{I}$ corresponds to assigning a variable a fixed value, *e.g.*, $\mathcal{I} = \text{do}[X_i = x]$. The intervened SCM is denoted as $\mathcal{M}^{\mathcal{I}}$ and the interventional distribution entailed by the intervened SCM as $P^{\mathcal{M};\mathcal{I}}$. Moreover, given the (possibly partial) observation $\mathbf{X} = \mathbf{x}$, we can define a counterfactual SCM $\mathcal{M}_{\boldsymbol{X}=\mathbf{x}}$ where the noise variables $\boldsymbol{U}$ are distributed according to the posterior distribution $P(\boldsymbol{U} \mid \boldsymbol{X} = \mathbf{x})$. Then, we can view a counterfactual statement as an intervention $\mathcal{I}$ in the SCM $\mathcal{M}_{\boldsymbol{X}=\mathbf{x}}$ and denote the counterfactual distribution entailed by the intervened counterfactual SCM $\mathcal{M}^{\mathcal{I}}_{\boldsymbol{X}=\mathbf{x}}$ as $P^{\mathcal{M} \mid \mathbf{X}=\mathbf{x};\mathcal{I}}$.

The Gumbel-Max SCM is a specific class of SCM in which the causal mechanism for a random variable Y is defined as

$$f_y(\mathbf{PA}, \mathbf{U}) := \underset{j}{\operatorname{argmax}}\{\log P(Y = j \mid \mathbf{PA}) + U_j\} \quad (1)$$

and each noise variable $U_j \sim \text{Gumbel}(0, 1)$. Here, note that the interventional distribution $P^{\mathcal{M} \mid \mathbf{PA}=\mathbf{pa}}(Y)$ entailed by a Gumbel-Max SCM $\mathcal{M}$ is exactly $P(Y \mid \mathbf{PA} = \mathbf{pa})$.

3 COUNTERFACTUAL INFERENCE OF SECOND OPINIONS

We consider a multi-class classification task where, for each instance, a human expert $h \subseteq \mathcal{H}$ makes a label prediction $y_h \in \mathcal{Y} = \{1, \ldots, k\}$ based on multiple sources of information, which are (imperfectly) summarized by a feature vector $x \in \mathcal{X}$. Here, we assume that experts make predictions on their own (*i.e.*, without communicating with each other) and the assignment of experts to instances is independent of the identity of the instances and their feature vectors. Then, our goal is to design an automated decision support system that, given a prediction y_h from an expert h about an instance summarized by a feature vector x, is able to infer what prediction $y_{h'}$ another expert $h' \neq h$ would have made about the *same* instance if she had been asked. Here, note that two different instances may be (imperfectly) summarized by the same feature vector x, however, we are interested in a counterfactual prediction about the *same* instance.

Our starting point is to view the above counterfactual statement as an intervention in a particular counterfactual SCM. More specifically, let $\mathcal{M}$ be a SCM defined by the assignments

$$\mathbf{Y} = f_{\mathbf{Y}}(X, Z, U), \quad Z = f_Z(V), \quad \text{and} \quad X = f_X(W) \quad (2)$$

where U, V and W are (multidimensional) independent noise variables, $f_{\mathbf{Y}}$, f_Z and f_X are given deterministic causal mechanisms (or functions), and $\boldsymbol{Y} = (Y_h)_{h \in Z}$ are the predictions by a set of human experts $Z \subseteq \mathcal{H}$. Then, we can express the above counterfactual statement as an intervention $\mathcal{I} = do[Z = \{h'\}]$ in the counterfactual SCM $\mathcal{M}_{X=x, Z=\{h\}, \boldsymbol{Y}=y_h}$ and, to infer the label prediction $y_{h'}$, we just need to resort to the counterfactual distribution $P^{\mathcal{M} \mid X=x, Z=\{h\}, \boldsymbol{Y}=y_h;\, do[Z=\{h'\}]}(\boldsymbol{Y})$.

At this point, one may argue that, even if we find a noise distribution $P(U)$ and a function $f_{\boldsymbol{Y}}$ under which the conditional distribution $P^{\mathcal{M}}(\boldsymbol{Y} \mid X)$ is a good fit for observed historical predictions by experts, we would be unable to validate how accurate our counterfactual label predictions are using data. In general, this is true since counterfactual reasoning lies within level three in the "ladder of causation" [Pearl, 2009]. In this context, previous work resorts instead to axiomatic assumptions about the causal mechanism of the world [Oberst and Sontag, 2019, Tsirtsis et al., 2021, Noorbakhsh and Rodriguez, 2021]. In our setting, this would reduce to specifying how differences across experts may have lead to a different prediction while holding "everything else" fixed. However, in what follows, we will show that, if experts do not communicate with each other, the above SCM satisfies a set invariance property that surprisingly implies that the above counterfactual distribution coincides with an interventional conditional distribution. This enables a data-driven design and validation of our SCM based decision support system.

4 RELATING THE COUNTERFACTUAL AND INTERVENTIONAL WORLDS

To build some intuition on the reasons why, if experts do not communicate, certain type of counterfactual and interventional distributions are identical, we start with a simple example. Let $h, h' \in \mathcal{H}$ be two different experts and consider the following two questions:

1. Both experts have made a label prediction about an instance (*i.e.*, $Z = \{h, h'\}$) but we only observe the prediction $Y_h = c$ made by h, what is the prediction made by h'?

2. One of the experts has made a label prediction $Y_h = c$ about an instance (*i.e.*, $Z = \{h\}$) and we observe it, what would the prediction made by h' be if she had made a prediction?

The first question is of conditional nature while the second is a counterfactual one. In general, the answer to both questions may differ, for example, if experts communicate with each other before making their predictions in the first case. However, if experts do not communicate, the answer to both questions should be identical. More formally, the following conditional interventional distribution and counterfactual distribution of the expert should be equal:

$$P^{\mathcal{M}\,;\,do[Z=\{h,h'\}]}(Y_{h'} \mid X = x, Y_h = c) \\ = P^{\mathcal{M}\,|\,X=x,Z=\{h\},Y_h=c\,;\,do[Z=\{h'\}]}(Y_{h'}). \quad (3)$$

More generally, we will now show that, if experts do not communicate with each other, the above equality is a direct consequence of a set invariance property satisfied by the SCM defined in Eq. (2).

Set Invariant SCMs (SI-SCMs). If experts do not communicate with each other before making a prediction, the mechanism $f_{\mathbf{Y}}$ has a set invariant value over expansions (supersets) of Z. For example, consider one single expert h has made a prediction $f_{\mathbf{Y}}(x, \{h\}, u) = c$ about a specific instance. Then, one can conclude that, if instead of a single expert, a set of experts $\zeta \subseteq \mathcal{H}$ such that $h \in \zeta$ would have made predictions about the same instance (*i.e.*, x and u does not change), expert h would have made the same prediction, *i.e.*, $(f_{\mathbf{Y}}(x, \zeta, u))_h = c$. More formally, we define the set invariance property as follows:

Definition 1 (Set Invariance)**.** *A mechanism $f_{\mathbf{Y}}$ for variable $\mathbf{Y}$ is set invariant with respect to Z if, for any two realizations $Z = \zeta$ and $Z = \zeta'$ such that $\zeta \subseteq \zeta'$, it holds that*

$$f_{\mathbf{Y}}(x, \zeta, u) = (f_{\mathbf{Y}}(x, \zeta', u))_\zeta \quad \text{for all } x \in \mathcal{X}, u \in \mathcal{U}\ .$$

A SCM $\mathcal{M}$ with such a mechanism is set invariant for $\mathbf{Y}$.

In this context, we can resort to the following theorems to conclude that, within the class of set invariant SCMs (SI-SCMs), we just need to focus on SCMs whose causal mechanism $f_{\mathbf{Y}}$ can be expressed by submechanisms f_{Y_h} governed by a common noise variable:[5]

Theorem 2. *Any SCM $\mathcal{M}$ with mechanism $f_{\mathbf{Y}}$ of the form $f_{\mathbf{Y}}(X, Z, U) = (f_{Y_h}(X, U))_{h \in Z}$, where $f_{Y_h} \colon \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$ are arbitrary functions, is set invariant for $\mathbf{Y}$.*

Theorem 3. *For any SI-SCM $\mathcal{M}$, there exists an equivalent SI-SCM $\mathcal{M}'$ with causal mechanism*

$$f'_{\mathbf{Y}}(X, Z, U) = (f'_{Y_h}(X, U))_{h \in Z} := (f_{\mathbf{Y}}(X, \{h\}, U))_{h \in Z}.$$

Moreover, the above theorems also make clear that any correlation between experts' predictions is caused by the common noise and features but not the causal mechanism.

[5] All proofs can be found in Appendix B

Equality between the counterfactual distribution and the conditional interventional distribution. Returning to our simple motivational example, note that, if a SCM is set invariant, the answers to the counterfactual and the conditional questions 1 and 2 are the same as long as the noise $u \sim P(U \mid X = x, Y_h = c)$ is the same. In particular, for question 1, the answer is $Y_{h'} = (f(x, \{h, h'\}, u))_{h'}$, for question 2, the answer is $Y_{h'} = f(x, \{h'\}, u)$, and since $f_{\mathbf{Y}}$ is set invariant, both answers are equal. More generally, for arbitrary sets of experts, we can easily conclude that equality holds if and only if $f_{\mathbf{Y}}$ is set invariant.

Next, to show that, if a SCM is set invariant, then the equality of distributions in Eq. (3) holds, we first present a more general theorem that states that, if we expand the set of experts who make predictions, the corresponding interventional distribution of $\mathbf{Y}$ does not change:

Theorem 4. *Let SCM $\mathcal{M}$ be set invariant for $\mathbf{Y}$. Then, for any $\zeta, \zeta' \in \mathcal{H}$ such that $\zeta \subseteq \zeta'$, it holds that*

$$P^{\mathcal{M}\,;\,do[Z=\zeta]}(\mathbf{Y} = \mathbf{y} \mid X) \\ = P^{\mathcal{M}\,;\,do[Z=\zeta']}((\mathbf{Y})_\zeta = \mathbf{y} \mid X)$$

for any $\mathbf{y} \in \mathcal{Y}^{|\zeta|}$ where $(\mathbf{Y})_\zeta$ denotes the predictions by the experts in the subset $\zeta \subseteq \zeta'$.

The above theorem is straight forward to show using that, due to the set invariance property, the prediction values of mechanism $f_{\mathbf{Y}}$ for (x, ζ, u) are equal to the values for (x, ζ', u) for experts in ζ and, due to the independence between the noise and the intervention, the noise distribution does not change. A direct conclusion is that, no matter how many experts make predictions, the conditional interventional distribution of a single expert's prediction does not change, as formalized by the following corollary:

Corollary 1. *Let SCM $\mathcal{M}$ be set invariant for $\mathbf{Y}$. Then, for any $h \in \mathcal{H}$ and $\zeta \subseteq \mathcal{H}$ such that $h \in \zeta$, it holds that*

$$P^{M\,;\,do[Z=\{h\}]}(Y_h \mid X) = P^{M\,;\,do[Z=\zeta]}(Y_h \mid X).$$

Similarly, we can derive the desired equality between the counterfactual distribution and the conditional interventional distribution by using the set invariance of mechanism $f_{\mathbf{Y}}$ and the fact that the noise distribution changes equally in both scenarios. More formally, we have the following corollary:

Corollary 2. *Let SCM $\mathcal{M}$ be set invariant for $\mathbf{Y}$. Then, for any $h, h' \in \mathcal{H}$ and $\zeta \subseteq \mathcal{H}$ such that $h, h' \in \zeta$, it holds that*

$$P^{\mathcal{M}\,|\,X=x,Z=\{h\},\mathbf{Y}=c\,;\,do[Z=\{h'\}]}(\mathbf{Y}) \\ = P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_{h'} \mid X = x, Y_h = c)\ .$$

for any $x \in \mathcal{X}$ and $c \in \mathcal{Y}$.

5 CHARACTERIZING MUTUALLY SIMILAR EXPERTS

Given a SI-SCM model $\mathcal{M}$ where each expert's predictions Y_h are generated by a sub-mechanism f_{Y_h}, our goal in this section is to characterize mutually similar experts. Later on, this will help us factorize the noise U governing the sub-mechanisms f_{Y_h} underpinning the model into a set of independent noise components and uniquely associate each of these noise components with disjoint sets of mutually similar experts given data.

To this end, we first start by characterizing similarity between a pair of experts $h, h' \in \mathcal{H}$. To this end, we resort to the recently introduced notion of counterfactual stability [Oberst and Sontag, 2019]. More specifically, we argue that two experts h and h' are *similar* if $\mathcal{M}$ satisfies counterfactual stability for h, h' with respect to $\mathbf{Y}$.

Definition 5 (Counterfactual stability)**.** *A SCM $\mathcal{M}$ satisfies counterfactual stability for h, h' with respect to $\mathbf{Y}$ if, for all $\zeta, \zeta' \subseteq \mathcal{H}$ such that $h \in \zeta$ and $h' \in \zeta'$ and for all $c' \neq c$, the condition*

$$\frac{P^{\mathcal{M}\,;\,do[Z=\zeta']}(Y_{h'} = c \mid X)}{P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_h = c \mid X)} \geq \frac{P^{\mathcal{M}\,;\,do[Z=\zeta']}(Y_{h'} = c' \mid X)}{P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_h = c' \mid X)}$$

implies that $P^{\mathcal{M}\,|\,X, Z=\zeta, Y_h=c\,;\,do[Z=\zeta']}(Y_{h'} = c') = 0$, where $Y_h = c$ is the observed outcome under $do[Z = \zeta]$.

For example, consider a scenario where a doctor needs to decide what treatment option—surgery ($Y = 0$), radiation ($Y = 1$) or chemotherapy ($Y = 2$)—will be more beneficial for a patient with a tumor, imperfectly summarized by a feature vector x. Assume doctor h decides the most beneficial option is surgery, *i.e.*, $Y_h = 0$, and we know that, for patients with similar x, doctor h' is generally more likely to operate and less likely to resort to therapy than doctor h. Then, if doctors h and h' are similar, as defined in Eq. (5), we expect doctor h' would have also decided the most beneficial option is surgery for the given patient, if consulted, *i.e.*, $Y_{h'} = 0$. Here, whenever two doctors h and h' are *not* similar, one could argue that it is because they weigh any (hidden) factor specific to the patient at hand differently[6].

Unfortunately, in general, we cannot use data to verify if two experts h and h' are similar. This is because our notion of similarity relies on a counterfactual distribution, $P^{\mathcal{M}\,|\,Y_h=i\,;\,\mathrm{do}[Z=\zeta']}$, and counterfactual reasoning lies within level three in the "ladder of causation" [Pearl, 2009]. However, we will now define a notion of conditional stability that is verifiable using interventional data and, in the case of SI-SCMs, is both a sufficient and necessary condition for counterfactual stability—if conditional stability holds, we can conclude that two experts are similar.

Definition 6 (Conditional stability)**.** *A SCM $\mathcal{M}$ satisfies conditional stability for two experts $h, h' \in \mathcal{H}$ with respect to $\mathbf{Y}$ if, for all $\zeta \subseteq \mathcal{H}$ such that $h, h' \in \zeta$ and for all $c' \neq c$, the condition*

$$\frac{P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_{h'} = c \mid X)}{P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_h = c \mid X)} \geq \frac{P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_{h'} = c' \mid X)}{P^{\mathcal{M}\,;\,do[Z=\zeta]}(Y_h = c' \mid X)} \tag{4}$$

implies that $P^{\mathcal{M};do[Z=\zeta]}(Y_{h'} = c' \mid X, Y_h = c) = 0$.

Here, note that, for SI-SCMs, we only need to verify the condition in Eq. (4) for the sets $\zeta = \{h\}$ and $\zeta = \{h'\}$ because no matter how many experts make predictions, the conditional interventional distributions in Eq. (4) do not change, as shown in Corollary 1. Then, the following Theorem formalizes the equivalence between conditional and counterfactual stability:

Theorem 7. *Let SCM $\mathcal{M}$ be set invariant for $\mathbf{Y}$. Then, $\mathcal{M}$ satisfies counterfactual stability for $h, h' \in \mathcal{H}$ with respect to $\mathbf{Y}$ iff it satisfies conditional stability.*

Once we have a notion of similarity between pairs of experts that we can verify from data, we can characterize groups of mutually similar experts. In this context, it will be useful to introduce the following notion of pairwise counterfactual stability (in short, PCS), which extends counterfactual stability to groups of experts $\zeta \subseteq \mathcal{H}$ of arbitrary size.

Definition 8 (Pairwise Counterfactual Stability)**.** *A SCM $\mathcal{M}$ satisfies pairwise counterfactual stability for a group of experts $\zeta \subseteq \mathcal{H}$ with respect to $\mathbf{Y}$ if it satisfies counterfactual stability for any $h, h \in \zeta$.*

Similarly as in the case with a pair of experts, one can also define pairwise conditional stability and it immediately follows from Theorem 7 that, for SI-SCM, pairwise conditional and counterfactual stability are equivalent, as formalized by the following Corollary.

Corollary 3. *Let SCM $\mathcal{M}$ be set invariant for $\mathbf{Y}$. Then, $\mathcal{M}$ satisfies pairwise counterfactual stability for $\zeta \in \mathcal{H}$ with respect to $\mathbf{Y}$ iff it satisfies pairwise conditional stability.*

6 GUMBEL-MAX SI-SCM

In this section, we build upon our theoretical results to develop the Gumbel-Max SI-SCM, a new class of SI-SCM based on the Gumbel-Max SCM.

Given a set of experts $\mathcal{H}$, the Gumbel-Max SI-SCM partitions $\mathcal{H}$ into disjoint sets of experts $\Psi = \{\psi\}_{\psi \in \Psi}$, as defined in Section 5, and associate all experts within each set to the same multidimensional noise variable. More formally, the Gumbel-Max SI-SCM is defined as follows:

[6] In general, note that similarity between experts does not always deterministically enforce the observed expert's prediction on the counterfactual prediction. In the example above, this happens because Eq. (5) holds for the two remaining label values. Rather, it allows us to identify experts with different decision making criteria.

Definition 9 (Gumbel-Max SI-SCM)**.** *The Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ is a specific class of SCM in which the causal mechanism for $\mathbf{Y}$ is defined as*

$$f_{\mathbf{Y}}(X, Z, U) = (f_h(X, U))_{h \in Z},$$

with

$$f_h(X, U_{\psi(h)}) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\{\log P(Y_h = c \mid X) + U_{\psi(h),c}\},$$

where $\psi(h) \in \Psi$ denotes the subgroup expert h belongs to and each noise variable $U_{\psi(h),c} \sim$ Gumbel$(0, 1)$.

By definition, the Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ is set invariant for $\mathcal{Y}$ and, for any $\zeta \subseteq \mathcal{H}$ and $h \in \zeta$, it holds that $P^{\mathcal{M}(\Psi);\text{do}[Z=\zeta]}(Y_h \mid X) = P(Y_h \mid X)$. Moreover, all experts within each group $\psi \in \Psi$ are mutually similar, as formalized by the following Theorem:

Theorem 10. *The Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ satisfies pairwise counterfactual stability (PCS) for each group $\psi \in \Psi$ with respect to $\mathbf{Y}$.*

Finally, note that, for $\Psi = \mathcal{H}$, the Gumbel-Max SI-SCM reduces to the original Gumbel-Max SCM defined in Eq. 1. Therefore, one can view the Gumbel-Max SI-SCM as a generalization of the original Gumbel-Max SCM where, instead of a single multidimensional noise variable U for all $h \in \mathcal{H}$, one has several noise variables U_ψ, one per group.

Estimating counterfactual distributions. Given a prediction $Y_h = c$ by an expert h, we can compute an unbiased finite sample Monte-Carlo estimator of the counterfactual distribution $P^{\mathcal{M}(\Psi) \mid X=x, Z=\{h\}, \boldsymbol{Y}=y_h \,;\, do[Z=\{h'\}]}(\boldsymbol{Y})$ for the prediction $Y_{h'}$ of another expert $h' \neq h$ as follows:

$$P^{\mathcal{M}(\Psi) \mid X=x, Z=\{h\}, \boldsymbol{Y}=y_h \,;\, do[Z=\{h'\}]}(\boldsymbol{Y}) \approx \frac{1}{T} \sum_{t \in T} \mathbb{1}[c = f_{h'}(x, \mathbf{u}_t)] \quad (5)$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_T$ are samples from the posterior distribution $P^{\mathcal{M}(\Psi) \mid X=x, Z=\{h\}, \boldsymbol{Y}=y_h \,;\, do[Z=\{h'\}]}(U_{\psi(h')})$ of the noise variable $U_{\psi(h')}$. Here, we can use an efficient procedure to sample from the above noise posterior distribution, described elsewhere [Oberst and Sontag, 2019, Maddison et al., 2015]. Moreover, note that, if $h \notin \psi(h')$, the posterior distribution coincides with the prior $P^{\mathcal{M}(\Psi)}(U_{\psi(h')})$.

Partitioning experts into mutually similar groups. In the Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$, for each expert $h \in \mathcal{H}$, we can estimate the conditional distribution $P(Y_h \mid X)$ using any machine learning model trained using historical predictions made by the expert h. However, to fully define $\mathcal{M}(\Psi)$, we need to partition the set of experts $\mathcal{H}$ into disjoint sets of experts Ψ given a small amount of historical data about multiple experts making predictions about a joint set of instances. To this end, we proceed as follows.

First, we look for violations of the conditional stability condition throughout the historical data. Whenever there exists a sample for which the predictions by two different experts h and h' violate the conditional stability[7], we conclude that h and h' cannot belong to the same group ψ. Further, we also conclude that any pair of experts whose predictions did not violate conditional stability and were at least once observed for the same sample *can* be similar. However, since conditional stability is not a transitive property, there may be multiple valid partitions $\mathcal{P} = \{\Psi\}$ of the experts into disjoint sets that are consistent with the above conclusions. To decide among them, we would like to pick the partition $\Psi \in \mathcal{P}$ under which the counterfactual distributions $P^{\mathcal{M}(\Psi) \mid X=x, Z=\{h\}, \boldsymbol{Y}=y_{h'} \,;\, do[Z=\{h\}]}(\boldsymbol{Y})$ provide the best goodness of fit. More formally, we would like to solve the following maximization problem:

$$\underset{\Psi \in \mathcal{P}}{\text{maximize}} \quad \sum_{h, h' \in \mathcal{H}} \mathcal{L}(\mathcal{M}(\mathcal{H}), h', h) - \mathcal{L}(\mathcal{M}(\Psi), h', h), \quad (6)$$

where $\mathcal{L}(\mathcal{M}(\cdot), h', h)$ denotes an average (empirical) loss whenever we observe Y_h and infer the label prediction $Y_{h'}$ using the counterfactual distribution $P^{\mathcal{M}(\cdot) \mid X=x, Z=\{h\}, \boldsymbol{Y}=y_h \,;\, do[Z=\{h'\}]}(\boldsymbol{Y})$. Here, we measure goodness of fit in terms of average loss reduction with respect to the counterfactual distributions entailed by the causal model $\mathcal{M}(\mathcal{H})$ because this will allow us to reduce the number of pairs (h, h') we need to consider.

Next, we reduce the above problem to the clique partitioning problem [Grötschel and Wakabayashi, 1989, 1990]. More specifically, let $\mathcal{G} = (\mathcal{H}, \mathcal{E})$ be an undirected graph where, if $(h, h') \in \mathcal{E}$, then h and h' *can* be similar, as concluded from the data. Then, it readily follows that finding a partition Ψ of $\mathcal{H}$ is equivalent to finding a clique cover for $\mathcal{G}$[8]. Now, let the weight $w(h, h')$ of each edge $(h, h') \in \mathcal{E}$ be given by:

$$\begin{aligned} w(h, h') = \mathcal{L}(\mathcal{M}(\mathcal{H}), h, h') - \mathcal{L}(\mathcal{M}(\Psi), h', h) \\ + \mathcal{L}(\mathcal{M}(\mathcal{H}), h', h) - \mathcal{L}(\mathcal{M}(\Psi), h, h') \end{aligned}$$

Then, we can rewrite the maximization problem defined in Eq. 6 as a clique partitioning problem[9]:

$$\begin{aligned} \underset{\Psi}{\text{maximize}} \quad & \sum_{\psi \in \Psi} \sum_{h, h' \in \psi} w(h, h') \\ \text{subject to} \quad & \Psi \text{ is a clique cover for } \mathcal{G}, \end{aligned} \quad (7)$$

where note that we only need to consider pairs of experts $h, h' \in \psi$ because, otherwise, $w(h, h') = 0$ since the corresponding counterfactual distributions entailed by $\mathcal{M}(\Psi)$

[7] A violation occurs whenever Eq. (4) holds but we observe $Y_h = c$ and $Y_{h'} = c'$.

[8] Ψ is a clique cover for $\mathcal{G}$ iff Ψ is a partition of $\mathcal{H}$, *i.e.*, $\bigcup_{\psi \in \Psi} \psi = \mathcal{H}$ and $\psi \cap \psi' = \emptyset$ for all $\psi, \psi' \in \Psi$, and vertices in $\psi \in \Psi$ form a clique in $\mathcal{G}$.

[9] In most of the literature, the problem is defined for complete graphs. However, for arbitrary graphs, one can simply include the missing edges and assign positive infinite weights so that they are not included in a solution [Brimberg et al., 2017]

and $\mathcal{M}(\mathcal{H})$ coincide. The clique partitioning problem is a NP-Hard problem [Grötschel and Wakabayashi, 1989, 1990], however, we found that a simple randomized greedy algorithm works well in our setting, as shown in Figure 3 in Appendix D. Refer to Appendix C for more details about the algorithm.

7 EXPERIMENTS ON REAL DATA

In this section, we compare the performance of the proposed Gumbel-Max SI-SCM at inferring second opinions against several competitive baselines using a dataset with real expert predictions over natural images. Appendix D contains additional experiments on synthetic data where we assess the performance of Algorithm 1 at recovering group of mutually similar experts on synthetic data.[10]

Data description and experimental setup. We experiment with the dataset CIFAR-10H [Peterson et al., 2019], which contains $10{,}000$ images taken from the test set of the standard dataset CIFAR-10 [Krizhevsky et al., 2009]. Each of these images belongs to $n = 10$ classes and contains label predictions from approximately 50 human annotators. In total, the images are annotated by $2{,}571$ different human annotators (from now on, experts).[11] Since the classification task is relatively easy for humans, there are many images ($\sim$35%) in which there is full agreement between experts—all experts make the same label prediction. Here, motivated by the empirical observation that, in medical diagnosis, there is typically a 20% per-instance disagreement among experts [Van Such et al., 2017, Elmore et al., 2015], we filter out the above mentioned images in which there is full agreement. Moreover, we split the remaining images into two disjoint sets at random—a training set and a test set—and filter out data from any expert who made less than 130 and 20 predictions in training and test set, respectively, and whose predicted labels in the training data do not cover all class labels. After these preprocessing steps, the resulting training and test sets contain $1{,}257$ and 303 images, respectively, annotated by $|\mathcal{H}| = 114$ experts, where each image in the training (test) set is annotated by at least one (two) expert(s).

To find the groups of mutually similar experts underpinning our Gumbel-Max SI-SCM, we run Algorithm 1 on the training set. Within the Gumbel-Max SI-SCM, we estimate the conditional distribution $P^{\mathcal{M}\,;\,\text{do}[Z=\{h\}]}(Y_h \mid X)$ for each expert h using a Gaussian Naive Bayes model (GNB) trained using also the training set (one GNB per expert)[12]. Each GNB model uses 20 dimensional feature vectors computed by running PCA on a 512 dimensional normalized feature vector extracted using VGG19 [Simonyan and Zisserman, 2014]. Both during training and test, given an observed label prediction Y_h by an expert h, we infer the prediction $Y_{h'}$ by another expert h' using the most likely label under (an estimate of) the corresponding counterfactual distribution. To estimate each counterfactual distribution, we use $T = 1{,}000$ samples from the noise posterior distribution.

Table 1: Overall test accuracy

Model	$h, h' \in \mathcal{H}$	$h, h' \in \psi$	$h \in \psi, h \in \psi'$
Gumbel-Max SI-SCM	66.8%	79.9%	45.1%
GNB	48.9%	51.3%	45.1%
GNB + CNB	62.0%	66.0%	55.2%

Baselines and evaluation metrics. We compare the performance of our trained Gumbel-Max SI-SCM with two baselines:

— The "*GNB*" baseline uses only the same Gaussian Naive Bayes models (GNB), one per expert, used by our trained Gumbel-Max SI-SCM. More specifically, given an observed label prediction Y_h by an expert h, it infers the prediction $Y_{h'}$ by another expert h' using the most likely label under the estimate of the conditional distribution $P^{\mathcal{M}\,;\,\text{do}[Z=\{h'\}]}(Y_{h'} \mid X)$ given by the corresponding GNB.

— The "*GNB + CNB*" baseline uses the same Gaussian Naive Bayes models (GNB), one per expert, used by our trained Gumbel-Max SI-SCM and a Categorical Naive Bayes (CNB) model, one per expert, that estimates $P^{\mathcal{M}\,|\,\text{do}[Z=\{h,h'\}]}(Y_{h'} \mid Y_h)$.[13] More specifically, given an observed label prediction Y_h by an expert h, it infers the prediction $Y_{h'}$ by another expert h' using the most likely label under the product of distributions $P^{\mathcal{M}\,;\,\text{do}[Z=\{h'\}]}(Y_{h'} \mid X) \times P^{\mathcal{M}\,;\,\text{do}[Z=\{h,h'\}]}(Y_{h'} \mid Y_h)$, as estimated by the corresponding GNB (first term) and CNB (second term).

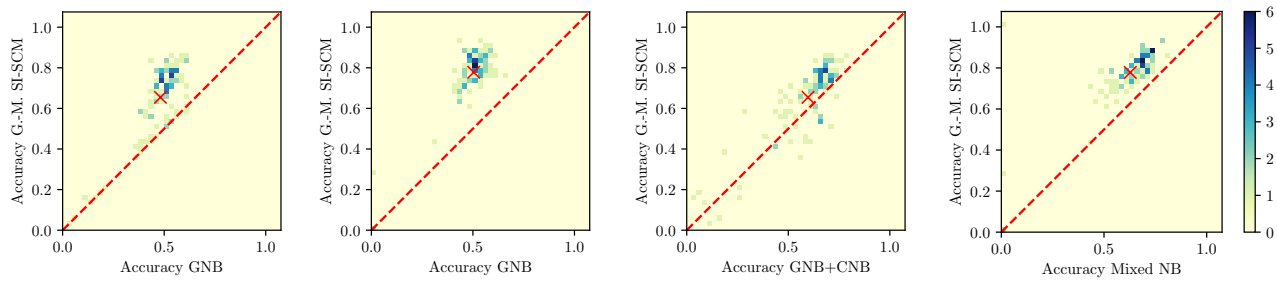
To compare the performance of our trained Gumbel-Max SI-SCM and both baselines, for each sample in the test set, we pick each of the corresponding expert label predictions Y_h as the observed prediction in turn and infer the value of the other predictions $Y_{h'}$. Here, we compute the overall

[10] To facilitate research in this area, we will release the code for the empirical evaluation with the final version of the paper. We have submitted an anonymous version of the code with the supplementary material.

[11] The dataset CIFAR-10H is one of the only publicly available datasets containing a relatively large number of samples with multiple label predictions by different experts per sample, necessary to train the proposed Gumbel-Max SI-SCM. However, since our methodology and theoretical results are rather general, our model may also be useful in other applications.

[12] In the CIFAR-10H dataset, experts are assigned to images (presumably) at random. Therefore, it holds that $P^{\mathcal{M}\,;\,\text{do}[Z=\{h\}]}(Y_h \mid X) = P(Y_h \mid X, Z = \{h\})$ and we can use observational data to estimate the interventional conditional distribution $P^{\mathcal{M}\,;\,\text{do}[Z=\{h\}]}(Y_h \mid X)$.

[13] The CNB uses a "one-hot" encoding of the observed prediction Y_h as a single $|\mathcal{H}|$-dimensional feature where, for each dimension, it uses an additional label value to denote than an expert's label prediction has not been observed.



$h, h' \in \mathcal{H}$ $h, h' \in \psi$ $h, h' \in \mathcal{H}$ $h, h' \in \psi$

(a) Gumbel-Max SI-SCM vs. GNB (b) Gumbel-Max SI-SCM vs. GNB+CNB

Figure 1: Per-expert test accuracy achieved by our model and both baselines on the CIFAR-10H dataset. In each panel, the y-axis always measures the per-expert test accuracy achieved by our method and the x-axis measures the per-expert accuracy achieved by one of the baselines. For each cell, the darkness is proportional to the number of experts with the corresponding test accuracies.

accuracy as well as the per-expert accuracy and distinguish among three scenarios: (i) $h, h' \in \mathcal{H}$; (ii) $h, h' \in \psi$; and, (iii) $h \in \psi, h' \in \psi', \psi \neq \psi'$.

Results. We start by reporting that, during the training of our Gumbel-Max SI-SCM, Algorithm 1 found 352 violations of the conditional stability condition between pairs of experts and partitioned the experts into fifteen disjoint groups of mutually similar experts, where seven of these groups were singletons. Refer to Appendix E for more details regarding the groups identified by Algorithm 1.

Next, we report the overall accuracy achieved by our model and the baselines in Table 1. We find that, in general ($h, h' \in \mathcal{H}$), our model infers the expert predictions more accurately than both baselines and this competitive advantage comes from instances in which the observed prediction is by an expert h who belongs to the same group of mutually similar experts as the expert h' whose prediction we infer ($h, h' \in \psi$). In fact, the GNB+CNB baseline is more accurate whenever both experts h and h' do not belong to the same group ($h \in \psi$, $h' \in \psi'$, $\psi \neq \psi'$). Moreover, we also find that the GNB+CNB baseline infers the expert predictions more accurately whenever both experts belong to the same group of mutually similar experts identified by Algorithm 1. In Appendix E, we report the confusion matrix of the above counterfactual predictions.

Finally, we report the per-expert h' accuracy achieved by our model and both baselines in Figure 1.[14] The results show that, in general ($h, h' \in \mathcal{H}$), our model infers the expert predictions more accurately than the baselines for a majority of the experts (103 and 89, out of 114, compared to GNB and GNB+CNB, respectively). Moreover, if we restrict our attention to observed label predictions by experts h belonging to the same group of mutually similar experts as the expert h' whose prediction we infer ($h, h' \in \psi$), our model infers the expert prediction more accurately for almost all experts h' (100 and 101, out of 103, compared to GNB and GNB+CNB, respectively). Additionally, Figure 6 in Appendix E shows that, for most experts (87 out of 103), the GNB+CNB baseline infers the expert predictions $Y_{h'}$ more accurately if the observed prediction Y_h is by an expert h belonging to the same group of mutually similar experts as the expert h' ($h, h' \in \psi$) than if it is by an expert h belonging to a different group ($h \in \psi$, $h' \in \psi'$, $\psi \neq \psi'$).

8 CONCLUSION

In this work, we have addressed the problem of inferring second opinions by experts from the perspective of counterfactual inference. We have focused on a multiclass classification setting and showed that, if experts make predictions on their own, the underlying causal mechanism generating their predictions needs to satisfy a desirable set invariant property. Moreover, we have introduced the set invariant Gumbel-Max structural causal model, a new class of structural causal model whose structure and counterfactual predictions about second opinions by experts can be validated using interventional data.

Our work opens up many interesting avenues for future work. For example, we assume experts do not communicate before forming their opinion—they make predictions on their own. Although this assumption may be satisfied in some real-world applications, it would be interesting to relax it. Moreover, we have validated our model using a single real dataset. It would be valuable to validate the counterfactual predictions about second opinions provided by our model using additional datasets from other applications. Finally, it would be important to carry out user studies in which the inferred second opinions provided by our model are shared with domain experts (*e.g.*, medical doctors).

[14] Here, we would like to acknowledge that, for the scenario where $h, h' \in \psi$, we could not compute the per-expert accuracy for eleven experts—seven of these experts belong to singleton groups and the remaining four do not predict any of the same samples predicted by other experts in their mutually similar groups.

References

Fernando Althabe, José M Belizán, José Villar, Sophie Alexander, Eduardo Bergel, Silvina Ramos, Mariana Romero, Allan Donner, Gunilla Lindmark, Ana Langer, et al. Mandatory second opinion to reduce rates of unnecessary caesarean sections in latin america: a cluster randomised controlled trial. *The Lancet*, 363(9425):1934–1940, 2004.

Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning" what-if" explanations for sequential decision-making. *arXiv preprint arXiv:2007.13531*, 2020.

Jack Brimberg, Stefana Janićijević, Nenad Mladenović, and Dragan Urošević. Solving the clique partitioning problem as a maximally diverse grouping problem. *Optimization Letters*, 11(6):1123–1135, 2017.

Pascal M Burger, Jan Westerink, and Bram EL Vrijsen. Outcomes of second opinions in general internal medicine. *PloS one*, 15(7):e0236048, 2020.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Ulrich Dorndorf and Erwin Pesch. Fast clustering algorithms. *ORSA Journal on Computing*, 6(2):141–153, 1994.

Joann G Elmore, Gary M Longton, Patricia A Carney, Berta M Geller, Tracy Onega, Anna NA Tosteson, Heidi D Nelson, Margaret S Pepe, Kimberly H Allison, Stuart J Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313 (11):1122–1132, 2015.

Martin Grötschel and Yoshiko Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming*, 45(1):59–96, 1989.

Martin Grötschel and Yoshiko Wakabayashi. Facets of the clique partitioning polytope. *Mathematical Programming*, 47(1):367–387, 1990.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Lucian L Leape. Unnecessary surgery. *Health Services Research*, 24(3):351, 1989.

Justin Lim, Christina Ji, Michael Oberst, Saul Blecker, Leora Horwitz, and David Sontag. Finding regions of heterogeneity in decision-making via expected conditional covariance. *Advances in Neural Information Processing Systems*, 34, 2021.

Guy Lorberbom, Daniel Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhi Lu, Yi Zhou, and Jin-Kao Hao. A hybrid evolutionary algorithm for the clique partitioning problem. *IEEE Transactions on Cybernetics*, pages 1–13, 2021. doi: 10.1109/TCYB.2021.3051243.

Chris J. Maddison, Daniel Tarlow, and Tom Minka. A* sampling, 2015.

Kimia Noorbakhsh and Manuel Gomez Rodriguez. Counterfactual temporal point processes. *arXiv preprint arXiv:2111.07603*, 2021.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

Gintaras Palubeckis, Armantas Ostreika, and Arūnas Tomkevičius. An iterated tabu search approach for the clique partitioning problem. *The Scientific World Journal*, 2014, 2014.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290. PMLR, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Provably improving expert predictions with conformal prediction. *arXiv preprint arXiv:2201.12006*, 2022.

Noriyoshi Sukegawa, Yoshitsugu Yamamoto, and Liyuan Zhang. Lagrangian relaxation and pegging test for the clique partitioning problem. *Advances in Data Analysis and Classification*, 7(4):363–391, 2013.

Stratis Tsirtsis, Abir De, and Manuel Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems*, 34, 2021.

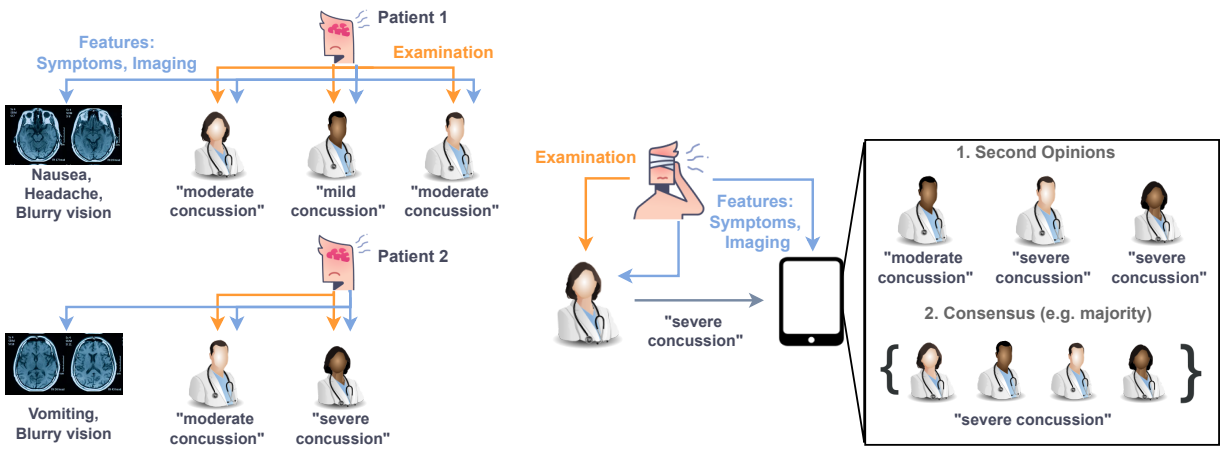
Monica Van Such, Robert Lohr, Thomas Beckman, and James M Naessens. Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice*, 23(4):870–874, 2017.

Peter Welinder and Pietro Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 25–32. IEEE, 2010.

Jing Zhang, Xindong Wu, and Victor S Sheng. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4):543–576, 2016.

Yi Zhou, Jin-Kao Hao, and Adrien Goëffon. A three-phased local search approach for the clique partitioning problem. *Journal of Combinatorial Optimization*, 32(2):469–491, 2016.

A EXAMPLE OF DATASET AND USE CASE FOR OUR DECISION SUPPORT SYSTEM



(a) Training dataset

(b) Use case

Figure 2: An example of a training dataset and use case of our decision support system on a medical application. In panel (a), for each patient, multiple doctors assess the severity of a concussion on the basis of a set of features. In panel (b), given a doctor's assessment of the severity of a concussion and a set of features, our decision support system infers other fellow doctors' assessment of the severity of the concussion.

B PROOFS

Proof of Theorem 2. Let $\zeta \subseteq \zeta' \subseteq \mathcal{H}$ and both non-empty. Then, for any $x \in \mathcal{X}, u \in \mathcal{U}$, we have that

$$(f(x, \zeta', u))_\zeta = ((f_h(x, u))_{h \in \zeta'})_\zeta = (f_h(x, u))_{h \in \zeta} = f(x, \zeta, u).$$

Proof of Theorem 3. Let $\mathcal{M}'$ be constructed from $\mathcal{M}$ by changing causal mechanism f with f'. To prove equivalence of $\mathcal{M}$ and $\mathcal{M}'$, we only need to show that, for any $u \in \mathcal{U}, x \in \mathcal{X}, \zeta \in \mathcal{P}(\mathcal{H}) \setminus \emptyset$, it holds that

$$\mathbf{y} = f(x, \zeta, u) \iff \mathbf{y} = f'(x, \zeta, u), \tag{8}$$

as only the causal mechanism was altered in the construction of $\mathcal{M}'$. Let $h \in \zeta$ be an arbitrary expert, then

$$(f'(x, \zeta, u))_h \overset{def.}{=} f(x, \{h\}, u) = (f(x, \zeta, u))_h,$$

where the last equality holds because f is set invariant. Thus, $f(x, \zeta, u) = f'(x, \zeta, u)$ for all $x \in \mathcal{X}, \zeta \in \mathcal{H}, u \in \mathcal{U}$.

Proof of Theorem 4. For clarity, we explicitly write $\mathbf{Y}_{Z=\{h'\}}$ and $\mathbf{Y}_{Z=\zeta}$ to better distinguish the two interventional outcomes. For discrete probability distribution $P(U)$ the right probability is given by

$$P^{M;do[Z:=\zeta']}(\mathbf{Y}_\zeta = \mathbf{y} \mid X = x) = \sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta', u))_\zeta = \mathbf{y})],$$

whereas the left is given by

$$P^{M;do[Z:=\zeta]}(\mathbf{Y} = \mathbf{y} \mid X = x) = \sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta, u) = \mathbf{y}].$$

Because f is a set invariant mechanism over Z, $(f(x, \zeta', u))_\zeta = f(x, \zeta, u)$, thus,

$$\sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta, u) = \mathbf{y}] = \sum_{u \in \mathcal{U}} P(U = u) \cdot \mathbb{1}[f(x, \zeta', u))_\zeta = \mathbf{y}].$$

The proof is analogous for the continuous probability distributions $P(U)$.

Proof of Corollary 1. Choose $\zeta = \{h\}$ in Theorem 4 and note that abusing notation $\mathbf{Y}_{Z=\zeta}$ is in this case equivalent to Y_h.

Proof of Corollary 2. Using the definition of counterfactual distributions, the proof is analogous to the proof of Theorem 4 but using the posterior distribution $P(U \mid X = x, Z = \{h\}, Y_h = c)$. Let $\zeta \subseteq H$ be so that $h, h' \in \zeta$. For all $c \in \mathcal{Y}$, by definition, we have that

$$P^{M|X=x,Z=\{h\},\mathbf{Y}=c;do[Z:=\{h'\}]}(\mathbf{Y} = c') = \sum_{u \in \mathcal{U}} P(U = u \mid X, (\mathbf{Y}_{Z=\{h\}})_h = c) \cdot \mathbb{1}[f(x, \{h'\}, u) = c']. \tag{9}$$

Using that $\mathcal{M}$ is set invariant we get that

$$(\mathbf{Y}_{Z=\{h\}})_h = f(x, \{h\}, U) = (f(x, \zeta, U))_h = (\mathbf{Y}_{Z=\zeta})_h \quad \text{and} \quad \mathbb{1}[f(x, \{h'\}, u)) = c'] = \mathbb{1}[(f(x, \zeta, u))_{h'} = c'].$$

Thus, Eq. (9) is equal to

$$\sum_{u \in \mathcal{U}} P(U = u \mid X, (\mathbf{Y}_{Z=\zeta})_h = c) \cdot \mathbb{1}[(f(x, \zeta, u))_{h'} = c'] \stackrel{def.}{=} P^{M;do[Z:=\zeta]}(Y_{h'} = c' \mid X = x, Y_h = c).$$

Proof of Theorem 7. Note that for all $\zeta_1, \zeta_2, \zeta_3 \subseteq H$ so that $h \in \zeta_1, h' \in \zeta_2, h, h' \in \zeta_3$ holds that

$$P^{\mathcal{M}|X=x,Z=\zeta_1,Y_h=c;do[Z=\zeta_2]}(Y_{h'} = c') = 0 \iff P^{\mathcal{M}|X=x,Z=\{h\},Y_h=c;do[Z=\{h'\}]}(Y_{h'} = c') = 0, \tag{10}$$

$$P^{\mathcal{M}|X=x,Z=\{h\},Y_h=c;do[Z=\{h'\}]}(Y_{h'} = c') = 0 \iff P^{\mathcal{M};do[Z=\zeta_3]}(Y_{h'} = c' \mid X = x, Z = \zeta_3, Y_h = c) = 0, \tag{11}$$

where Eq. (10) follows from the definition of set invariance and Eq. (11) follows from Corollary 2. Recall that $p_\zeta(h, c) := P^{\mathcal{M};do[Z=\zeta]}(Y_h = c \mid X)$ and $p_h(c) := P^{\mathcal{M};do[Z=\{h\}]}(Y_h = c \mid X)$.

It follows from Corollary 1 that, for all $h \in H$, $c \in \mathcal{Y}$ and ζ, ζ' so that $h \in \zeta$ and $h \in \zeta'$, we have that

$$p_\zeta(h, c) = p_h(h, c) = p_{\zeta'}(h, c).$$

Thus, following implications hold

$$\frac{p_{\zeta_2}(h', c)}{p_{\zeta_1}(h, c)} \geq \frac{p_{\zeta_2}(h', c')}{p_{\zeta_1}(h, c')} \iff \frac{p_{h'}(h', c)}{p_h(h, c)} \geq \frac{p_{h'}(h', c')}{p_h(h, c')} \iff \frac{p_{\zeta_3}(h', c)}{p_{\zeta_3}(h, c)} \geq \frac{p_{\zeta_3}(h', c')}{p_{\zeta_3}(h, c')}.$$

With these set of implications, it is straight forward to imply one statement from the other.

Proof of Theorem 10. Let ψ be a subgroup in Ψ so that $|\psi| \geq 2$. Let h and h' denote two arbitrary experts in subgroup psi. As the Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ is set invariant, it is enough to show that pairwise conditional stability condition is satisfied for pair h and h'. Analogously to Oberst and Sontag [2019], we proceed by proving the contrapositive, that for all sets ζ, so that $h, h' \in \zeta$, and $c \neq c'$

$$P^{\mathcal{M}(\Psi);do[Z=\zeta]}(Y_{h'} = c' \mid X, Y_h = c) \neq 0 \implies \frac{p_\zeta(h', c)}{p_\zeta(h, c)} < \frac{p_\zeta(h, c')}{p_\zeta(h, c')}.$$

If the conditional probability is positive, almost surely there must exist Gumbel noise variables $g_{\psi,c}$ and $g_{\psi,c'}$ such that

$$\log P(Y_h = c \mid X) + g_{\psi,c} > \log P(Y_h = c' \mid X) + g_{\psi,c'}$$
$$\log P(Y_{h'} = c \mid X) + g_{\psi,c} < \log P(Y_{h'} = c' \mid X) + g_{\psi,c'},$$

as the submechanisms f_h and $f_{h'}$ of each expert share the noise vector of the subgroup ψ.

Recall that, by set invariance, $p_\zeta(h, c) = P(Y_h = c \mid X)$ for all ζ, h, c. Hence, we can substitute the probabilities in both inequalities. Then, we further subtract the first inequality from the second which cancels out the Gumbel noises. Finally, using the properties of the logarithm function, the inequality is rearranged deriving the implication.

$$\log p_\zeta(h', c) - \log p_\zeta(h, c) < \log p_\zeta(h', c') - \log p_\zeta(h, c'),$$
$$\frac{p_\zeta(h', c)}{p_\zeta(h, c)} < \frac{p_\zeta(h, c')}{p_\zeta(h, c')}.$$

This proves that the Gumbel-Max SI-SCM $\mathcal{M}(\Psi)$ satisfies the pairwise conditional stability condition

$$\frac{p_\zeta(h', c)}{p_\zeta(h, c)} \geq \frac{p_\zeta(h', c')}{p_\zeta(h, c')} \implies P^{\mathcal{M}(\Psi);do[Z=\zeta]}(Y_{h'} = c' \mid X, Y_h = c) = 0,$$

for any two experts in the same subgroup in Ψ.

C RANDOMIZED GREEDY ALGORITHM FOR THE CLIQUE PARTITIONING PROBLEM

The idea behind the simple greedy randomized Algorithm 1 is to sequentially grow a clique starting from a random vertex in G' until no vertices can be added, remove this clique from the graph and repeat this process on the remaining graph until no vertices are left. For the current clique ψ the set of vertices that can be added, called candidate set, consists of vertices that have edges to all the vertices in ψ so that the sum these edge weights is positive. The expert with maximum sum of is added next to the clique, and the candidate set is updated.

The updated candidate set is a subset of the previous set, thus, the sum of edge weights of a vertex connected to the updated clique can be computed in constant time by considering the previous value and the weight of the edge to the newly added vertex. If no edge exists, the vertex can be removed from the candidate set. Algorithm 1 can thus be implemented in $O(|\mathcal{E}|)$.

Algorithm 1 Greedy Algorithm for the Clique Partitioning Problem, $N(\psi)$ denotes the set of vertices not in ψ with edges to all vertices in ψ

Input: weighted, undirected graph $G = (H, E, w)$
while G not empty **do**
  pick random vertex h
  $\psi \leftarrow h$
  while $N(\psi)$ not empty **do**
    $h^* \leftarrow \arg\max_{h' \in N(\psi)} \sum_{h \in \psi} w(\{h', h\})$
    if $\sum_{h \in \psi} w(\{h^*, h\}) > 0$ **then**
      $\psi \leftarrow h^*$
    else
      break
    end if
  end while
  $\Psi \leftarrow \psi$
  delete ψ from G
  $\psi = \emptyset$
end while
return Ψ

We note that, since the algorithm maximizes the sum of weights for each clique sequentially, it's performance in recovering a partition maximizing the overall sum depends on the sequence of sampled vertices which we start each clique from. To stabelize the algorithm's performance, one can rerun the algorithm a few times and choose among the returned partitions the one maximizing the overall sum of edge weights between vertices in the same set (see objective function in optimazation problem 7). In the experiments on real (synthetic) data, we run the algorithm 10 (5) times.

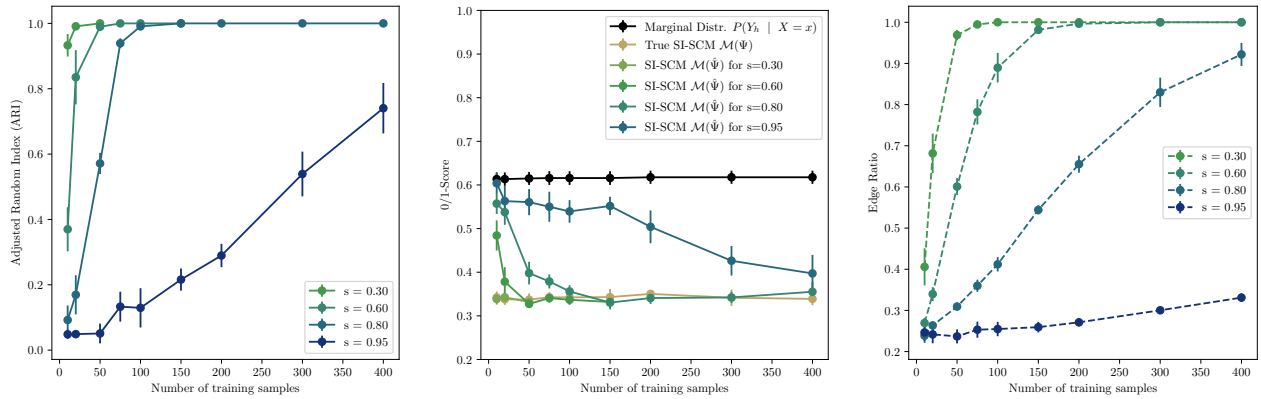
D EXPERIMENTS ON SYNTHETIC DATA

In this section, we assess the performance of Algorithm 1 at recovering the groups of mutually similar experts underpinning our Gumbel-Max SI-SCM using synthetic data.

Experimental setup. We consider a synthetic prediction task with $k = 5$ labels and 20 features per sample, whose values we sample uniformly at random from the interval $[0, 1]$, and a set $\mathcal{H}$ of 48 synthetic experts. These synthetic experts make label predictions according to a Gumbel-Max SI-SCM with five disjoint groups of mutually similar experts Ψ, *i.e.*, each expert within a group $\psi \in \Psi$ use the same Gumbel noise within the model[15]. Moreover, for each expert, the probability $P(Y_h = c \mid X = x)$ is given by a multinomial logit model with random weight coefficients $w = (w_1, \ldots, w_5)$, which we also sample uniformly at random from the interval $[0, 1]$ independently for each expert, *i.e.*,

$$P(Y_h = c \mid X = x) = \frac{\exp(w_c \cdot x)}{\sum_{j \in \mathcal{Y}} \exp(w_j \cdot x)}. \tag{12}$$

[15]The groups in the partition Ψ contain 6, 7, 11, 11 and 13 experts.



(a) Adjusted random index (ARI) (b) Average 0/1-loss on a held-out set (c) Edge ratio

Figure 3: Performance of Algorithm 1 at recovering the partition Ψ given the true probabilities $P(Y_h = c \mid X = x)$ under different amounts of data m and sparsity levels $s \in (0, 1)$. To compute the mean and the variance in panels (a) and (b), we run each experiment five times.

We measure the performance of Algorithm 1 at recovering the partition Ψ given the true probabilities $P(Y_h = c \mid X = x)$ under different amounts of training data m and sparsity levels $s \in (0, 1)$. The sparsity level s controls the average number of observed expert predictions per sample, *i.e.*, for each sample, all experts make a prediction but we only observe $\max\{2, (1 - s)|\mathcal{H}|\}$, picked at random. Here, note that, as the sparsity level s decreases (increases) and the amount of training data increases (decreases), it is easier (harder) to recover the partition Ψ.

As a measure of the difficulty of each inference problem, we will use the edge ratio r, defined as the fraction of pair of experts who belong to the same group $\psi \in \Psi$, among all pairs whose predictions did not violate conditional stability and were at least once observed for the same sample. As performance metrics, we will use:

- The adjusted random index (ARI), which measures similarity between the partition $\hat{\Psi}$ returned by Algorithm 1 and the true partition Ψ. Its value lies in the interval $[0, 1]$ where 1.0 means full recovery and 0.0 means a completely random partition was recovered with no similarity to the true one.
- The average $0/1$-loss on a held-out set (with 100 samples) of a predictor that, given an observed label Y_h, returns the most likely label $Y_{h'}$ under the inferred counterfactual distribution $P^{\mathcal{M}(\hat{\Psi}) \mid X=x, Z=\{h'\}\mathbf{Y}=y_{h'}; \text{do}(Z\{h\})(\mathbf{Y})}$. Here, to estimate the inferred counterfactual distributions, we use 500 samples.

As a point of comparison, for the second performance metric, we also compute the average $0/1$-loss over the same held-out set of two other predictors that, given an observed label Y_h, return the most likely label $Y_{h'}$ under the true counterfactual distributions $P^{\mathcal{M}(\Psi) \mid X=x, Z=\{h'\}\mathbf{Y}=y_{h'}; \text{do}(Z\{h\})(\mathbf{Y})}$ and the counterfactual distribution $P^{\mathcal{M}(\mathcal{H}) \mid X=x, Z=\{h'\}\mathbf{Y}=y_{h'}; \text{do}(Z\{h\})(\mathbf{Y})}$, respectively.

Results. Figure 3 summarizes the results, which show that, as long as the edge ratio $r > 0.3$, the inferred partition $\hat{\Psi}$ is very similar to the true partition Ψ (*i.e.*, the value of ARI is very close to 1) and the $0/1$-losses of the predictors that use $\mathcal{M}(\hat{\Psi})$ and $\mathcal{M}(\Psi)$ respectively are very similar. Here, note however that, even the predictor that uses the true model $\mathcal{M}(\Psi)$ has a non zero $0/1$-loss is not error free because, given an observed expert prediction Y_h and feature vector x, the expert prediction $Y_{h'}$ is not deterministic.

E ADDITIONAL FIGURES FOR EXPERIMENTS ON REAL DATA

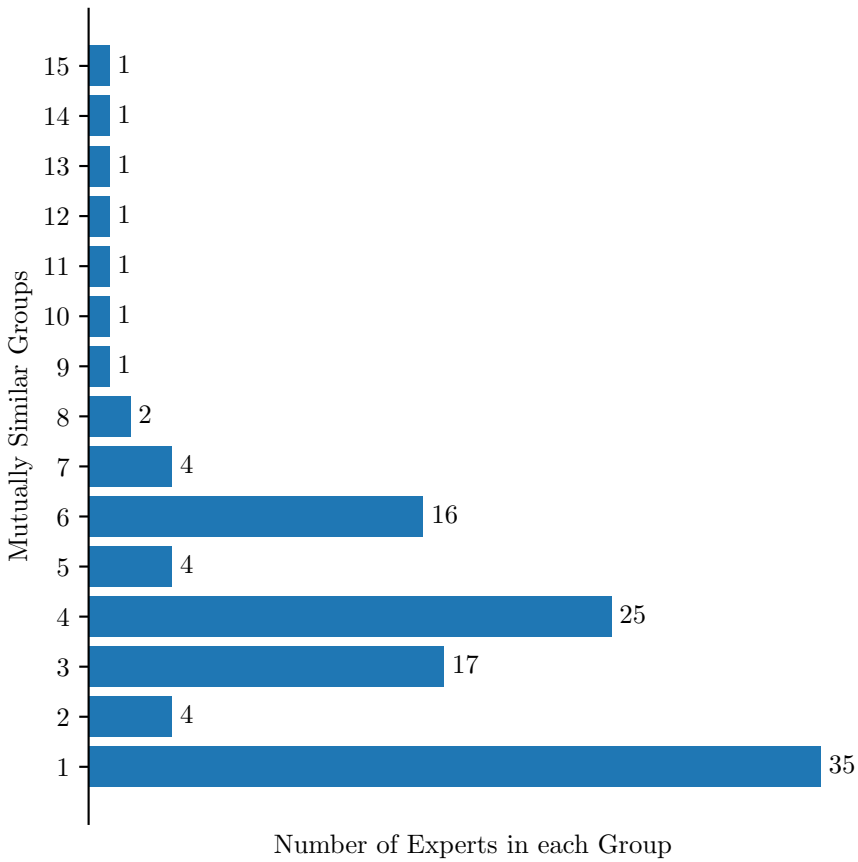
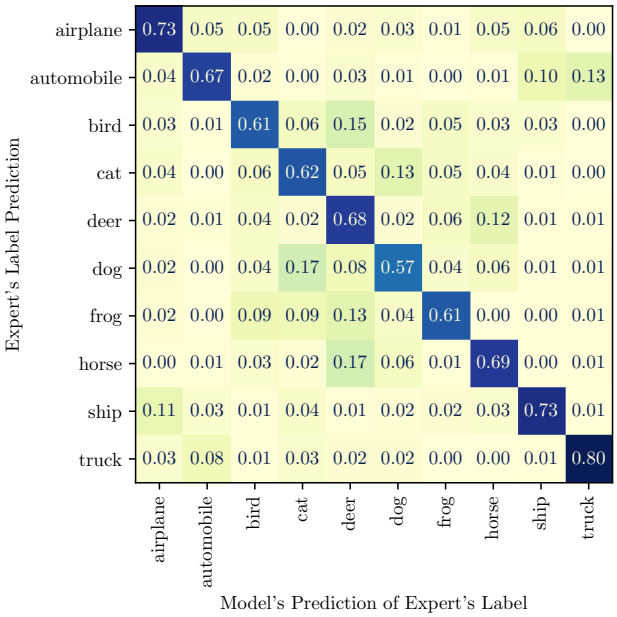
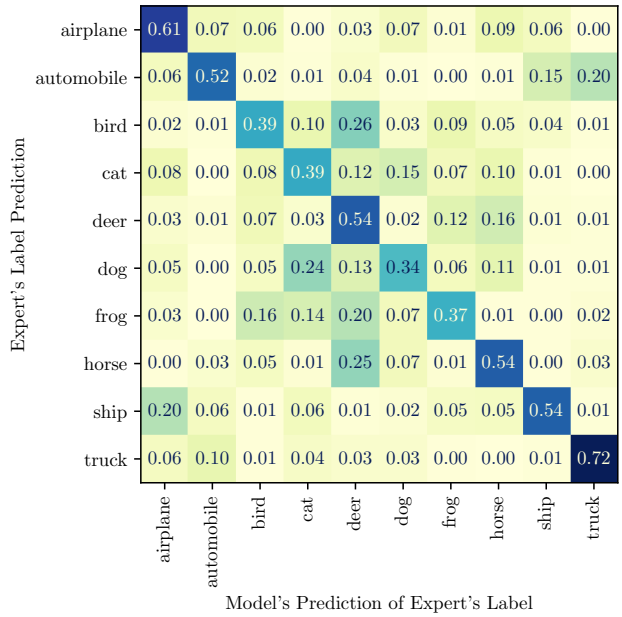


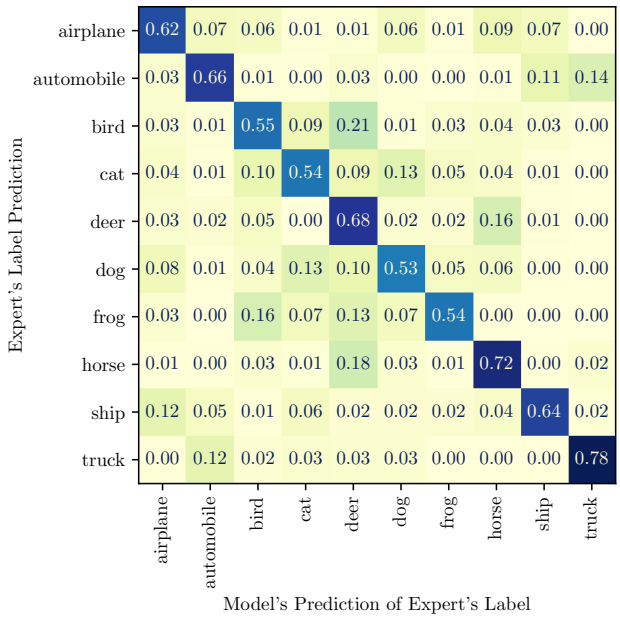
Figure 4: Size of the mutually similar expert groups returned by Algorithm 1 for the preprocessed CIFAR-10H dataset.



(a) Gumbel-Max SI-SCM



(b) GNB



(c) GNB+CNB

Figure 5: Confusion matrices of the counterfactual predictions of our model and the predictions of the two baselines for expert's labels on the test dataset.

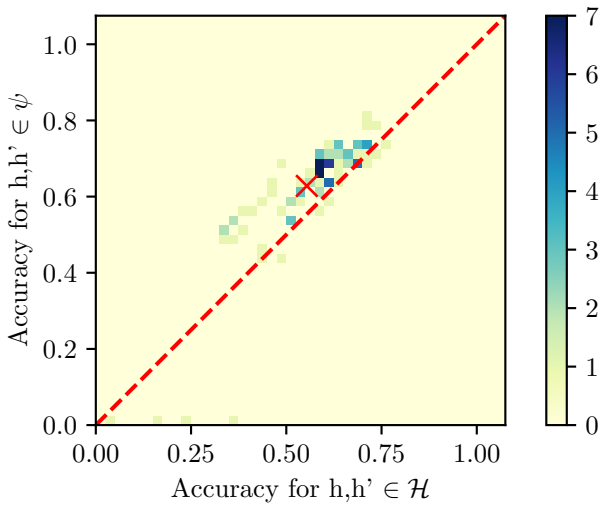


Figure 6: Per-expert test accuracy achieved by the baseline GNB+CNB on the preprocessed CIFAR-10H dataset. For each expert h', the y-axis measures the test accuracy whenever the observed expert h belongs to the same group of mutually similar experts as h' and the x-axis measures the test accuracy whenever h does not belong to the same group. For each cell, the darkness is proportional to the number of experts with the corresponding test accuracies.