Bayesian Hierarchical Models for Counterfactual Estimation

Natraj Raman
J.P.Morgan AI Research

Daniele Magazzeni
J.P.Morgan AI Research

Sameena Shah
J.P.Morgan AI Research

Abstract

Counterfactual explanations utilize feature perturbations to analyze the outcome of an original decision and recommend an actionable recourse. We argue that it is beneficial to provide several alternative explanations rather than a single point solution and propose a probabilistic paradigm to estimate a diverse set of counterfactuals. Specifically, we treat the perturbations as random variables endowed with prior distribution functions. This allows sampling multiple counterfactuals from the posterior density, with the added benefit of incorporating inductive biases, preserving domain specific constraints and quantifying uncertainty in estimates. More importantly, we leverage Bayesian hierarchical modeling to share information across different subgroups of a population, which can both improve robustness and measure fairness. A gradient based sampler with superior convergence characteristics efficiently computes the posterior samples. Experiments across several datasets demonstrate that the counterfactuals estimated using our approach are valid, sparse, diverse and feasible.

1 INTRODUCTION

Large-scale adoption of decision critical AI solutions requires explaining the rationale for a particular prediction. Counterfactual explanations (Wachter et al., 2017) provide an intuitive mechanism to reason even over complex models by defining a set of feature perturbations that would change the outcome to a favourable decision, thereby explaining the factors that led to the original decision.

Many counterfactuals are possible since there are several paths for an instance to achieve the desired outcome. After all, a loan applicant could qualify for a mortgage with different combinations of increasing the collateral amount, obtaining an educational degree or working longer hours. The focus traditionally (Dhurandhar et al., 2018; Pawelczyk et al., 2020; Looveren and Klaise, 2021) though has been on finding a unique path that is defined by the smallest possible perturbation of an instance. We argue that a theoretically ideal single explanation is overly restrictive and propose a mechanism to offer multiple alternative counterfactual explanations in a Bayesian framework.

While there has been some investigations before (Rodríguez et al., 2021; Mothilal et al., 2020; Russell, 2019) on generating diverse counterfactuals, they have largely been cast in the frequentist optimization setting producing point-estimates. In contrast, we follow a Bayesian approach that models the perturbations in a probabilistic paradigm and regard them as random variables with distribution functions.

This probabilistic treatment offers a variety of benefits over traditional models as summarized in Figure 1. For example, the posterior distribution of the perturbations can be directly used to sample a diverse set of counterfactuals in a single shot without resorting to the use of ensemble models or exotic post-hoc selection constraints. A distribution oriented approach (Gelman et al., 1995) is asymptotically unbiased, can natively handle multimodal parameters and quantify the uncertainty in predictions using variances, interquartile ranges, credible intervals and entropy. Such variability measures can help in producing robust counterfactual estimates Dutta et al. (2022), thereby improving the overall reliability.

Another appealing aspect of our approach is the ability to customize the counterfactual generation (Watson, 2022) through appropriate prior distributions. Consider a user who is less likely to take a *Sales* job when compared with other job types. We can encode this belief through an asymmetric categorical prior that places reduced probability mass on the *Sales* type, and vary this mass depending on the relative extent of the user's preference. The incorporation of such specific knowledge through informative priors enables greater flexibility and personalization in counterfactual generation.

The generated counterfactuals must respect the inherent dependencies between the features in order to produce ex-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

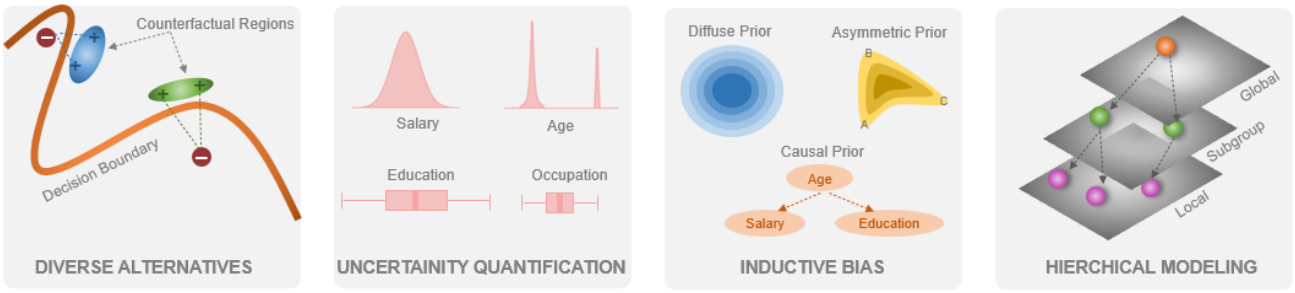

Figure 1: Advantages of Bayesian approach to counterfactual estimation. Several alternative explanations can be sampled from the posterior distribution, the uncertainty in predicted variables can be summarized through credible intervals, user specific customization can be formalized using informative priors and the hierarchical structure can capture relative effects.

amples that are feasible (Mahajan et al., 2019; Poyiadzi et al., 2020) in the real world. For example, we may want to preserve the correlation between *Age* and *Salary* variables in the observed data. Our solution allows the introduction of conditional distributions that can capture the causal structure relating the variables, thereby modeling the feature interactions explicitly.

When presenting a counterfactual, it is useful to compare the magnitude of change in the feature value of an instance with population level counterfactuals. For example, in the spirit of fairness (Wu et al., 2019), it may be of interest to determine whether the education level suggested by the counterfactual to get a loan is discriminative or not relative to other typical instances. These comparisons could also span across multiple hierarchical levels considering demographic subgroups within the population such as *Male* and *Female* or *White*, *Black* and *Asian*. A natural way to model these dependencies across instances is by using a Bayesian hierarchical structure that can share information across data groups and yet capture variations at different levels of the data, thereby improving the quality of the estimates and enabling fairness evaluation.

We propose here a three-level hierarchical counterfactual model in which the perturbations of a local instance depends on its corresponding subgroup, which in turn depends on the population. We account both for continuous and categorical variables, treating the latter as first-class citizens. Due to the intractable nature of the posterior distribution, we utilize Hamiltonian Monte Carlo (HMC) (Betancourt and Girolami, 2015) to derive the posterior samples. HMC makes use of the geometric information estimated via first-order derivatives of the target distribution and this enables us to efficiently explore the parameter space.

We use three public tabular datasets namely Adult Income (Kohavi and Becker, 1996), German Credit (Hofmann, 1994) and HELOC (FICO, 2018) for demonstrating the efficacy of our approach. In particular, we show that the Bayesian model compares well with point-estimate based solutions in measures such as validity, proximity and sparsity, with the added benefit of its support for diversity, robustness, causality, personalization and borrowing strength.

To summarize, our main contributions are: (a) A distribution oriented approach to counterfactual generation that can produce diverse alternatives and quantify uncertainty, (b) Incorporation of inductive biases through informative priors, and (c) A hierarchical formulation that preserves feature dependencies, promotes information sharing and enables subgroup analysis.

2 RELATED WORK

Counterfactual explanations has several research themes (Verma et al., 2020) and our primary focus is on the diversity angle. While most methods produce a single counterfactual (Dhurandhar et al., 2018; Pawelczyk et al., 2020; Looveren and Klaise, 2021), the generation of multiple counterfactuals had been considered before. The preferred route to support alternate examples is to incorporate an explicit term in the optimization objective function or perform an efficient search. For example in Mothilal et al. (2020), a diversity constraint is included in the loss function by building on detrimental point processes, while in Rodríguez et al. (2021), a collection of latent perturbations are searched to identify the attributes that will change the decision. Dandl et al. (2020) formalize the search as a multi-objective optimization problem that returns a Pareto set of counterfactuals. Differently, Russell (2019) employs a mixed-integer programming solver that is integrated with a set of constraints to generate diverse explanations. Smyth and Keane (2022); Albini et al. (2022) follow a nearest neighbour search to generate multiple counterfactuals.

The primary difference with the above methods is that we cast counterfactual generation as sampling from posterior distributions. Even methods that claim to be distribution-aware deal with the empirical data distribution (Kanamori et al., 2020) or the distribution of classifier model parameters (Bui et al., 2021), unlike our focus on the counterfactual

generation process.

A distribution-centric framework such as ours produces uncertainty estimates that can be used to measure the stability, reliability and consistency of the generated counterfactuals. Few recent works such as the extensions to LIME and SHAP in Slack et al. (2021b), the latent variable deep generative model in Ley et al. (2022) and the optimal transport for Gaussian mixtures in Nguyen et al. (2022) do consider uncertainty estimates. However, their focus is either on generating local explanations of a classifier prediction or on the predictive uncertainty of the classification model.

Recommending a set of actions that can achieve favourable outcomes in practice often requires treating algorithmic recourse as a causal problem (Barocas et al., 2017). Karimi et al. (2022) perform causal reasoning for recourse under imperfect knowledge by employing Gaussian processes and conditional variational autoencoders while Dominguez-Olmedo et al. (2022) study adversarial robustness of recourse from the lens of causality. The causal relationship between features are used to capture downstream effects of recourse actions and enforce fairness in von Kügelgen et al. (2022). While causal recourse is not our primary focus, causal dependencies can be incorporated in our model through conditional distributions.

A distinguishing aspect of our work is the treatment of counterfactuals in a hierarchical setting. Existing methods that can provide explanations at local and global levels such as Plumb et al. (2020); Becker et al. (2021) lack a principled framework that can inherently support multiple levels as ours does. Works that can potentially support several levels, as in Rawal and Lakkaraju (2020); Kanamori et al. (2022), differ from ours in their objective and technique.

3 MODEL

Let $\mathbf{x} = (x_1, ..., x_{d_{cont}}, x_{d_{cont}+1}, ..., x_{d_{cont}+d_{cat}})$ be an observed instance that contains d_{cont} number of continuous values and d_{cat} number of categorical values with $|\mathbf{x}| = d_{cont} + d_{cat} = d$. Each instance is associated with one of K different groups and let k be the subgroup corresponding to $\mathbf{x}$. Let $f : x \rightarrow [0, 1]$ be a binary classifier function that is differentiable.

If $f(\mathbf{x}) <= 0.5$, we would like to generate a counterfactual explanation $\mathbf{x}^*$ for $\mathbf{x}$ such that $f(\mathbf{x}^*) > 0.5$. We write

$$\begin{aligned} \mathbf{x}^* &= \mathbf{z} \odot \mathbf{x} + \Delta, \\ \mathbf{\Delta} &= (\delta_1, ..., \delta_{d_{cont}}, \eta_{d_{cont}+1}, ..., \eta_d), \end{aligned} \tag{1}$$

where Δ is a set of parameters that models the perturbations, with δ being the change in value of continuous variables and η the modified value of a categorical variable. Here $\mathbf{z} \in \{0, 1\}^d$ is a vector of indicators that is set to 1 for continuous variables and $\odot$ denotes element-wise multiplication.

In traditional counterfactual frameworks, Δ is a set of fixed but unknown values that must be determined based on some optimization function. Instead, in our Bayesian paradigm, Δ is modeled as a random variable having a probability distribution. This allows the incorporation of problem specific knowledge (or the lack of it) on these parameters through appropriate prior distributions. Furthermore, we can sample multiple values corresponding to different modes of the distribution, thereby generating a diverse set of explanations.

3.1 Hierarchical Bayes

The modular nature of a Bayesian framework allows simplifying complex setups necessitated by counterfactual generation requirements. For example, we may wish to measure how a counterfactual generated for a particular instance differs from other instances in its subgroup or across the groups. This requirement to capture variations at different levels of the data can be formulated naturally as a hierarchical model, where we have a population level set of parameters that are shared with the parameters at a group level which in turn are shared at the individual instance level. Such a hierarchical structure also enables borrowing of information from other related data, through which the robustness of parameter estimates can be improved.

We describe a three level hierarchical counterfactual model with the population level being referred as $L1$, the group level as $L2$ and the local instance level as $L3$. Each continuous variable δ is bestowed a Gaussian prior $\mathcal{N}(\mu, \sigma)$, with the mean μ being shared across the hierarchical levels while the variance σ remains independent. Intuitively, the parameter value of a subgroup is centered around the population level value but can still deviate from it. Similarly, the parameters of all the local instances can take distinct values, yet are dependent on its parent and remains within a range of its corresponding subgroup. A level can optionally be dropped, in which case the dependency shifts upwards. Formally, the generative steps are defined as:

$$\mu_i^{L1} \sim \mathcal{N}(\mu_i^0, \sigma_i^{L1}) \quad \forall i = 1...d_{cont} \tag{2}$$

$$\mu_{k,i}^{L2} \sim \mathcal{N}(\mu_i^{L1}, \sigma_{k,i}^{L2}) \quad \forall i = 1...d_{cont}, k = 1...K \tag{3}$$

$$\delta_i \sim \mathcal{N}(\mu_{k,i}^{L2}, \sigma_i^{L3}) \quad \forall i = 1...d_{cont}. \tag{4}$$

The relationship between the variables are modeled using conditional distributions. Given a causal model where variable i is the parent of j, the dependency structure is defined using a linear approximation with parameters m and c as

$$\delta_j | \delta_i \sim \mathcal{N}(m\delta_i + c, \sigma_j). \tag{5}$$

A categorical variable η is endowed with a scaled Dirichlet prior $Dir(\alpha\beta)$, where $\beta \in \mathbb{R}_+^L$ is a vector whose length depends on the number of categories L and α is a scale parameter. The categorical values themselves are sampled

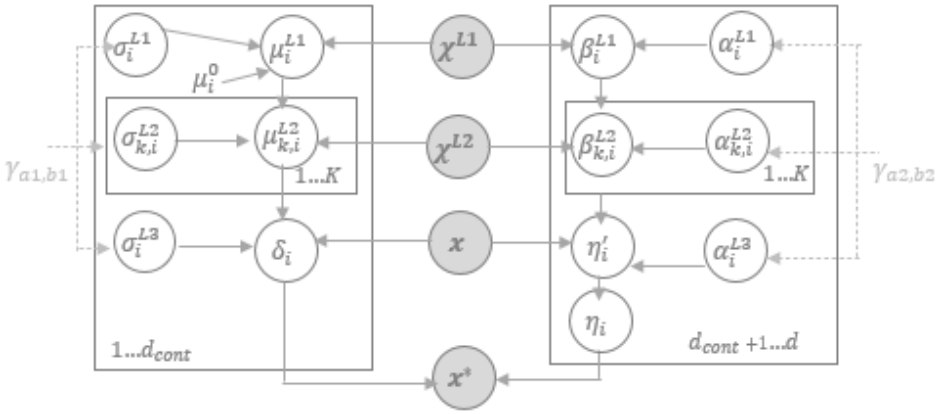

Figure 2: Hierarchical counterfactual model in plate notation. The left side of the figure corresponds to the dependency graph of continuous variables while the right side reflects the categorical variables.

from the probabilities estimated from the Dirichlet prior using a Multinomial distribution. Similar to the continuous variables, each parameter depends on the parameter at its corresponding parent level while still retaining the flexibility to diverge. The generative steps here are formalized as:

$$\beta_i^{L1} \sim Dir(\alpha_i^{L1}) \quad \forall i = d_{cont}{+}1...d \tag{6}$$

$$\beta_{k,i}^{L2} \sim Dir(\alpha_{k,i}^{L2}\beta_i^{L1}) \quad \forall i = d_{cont}{+}1...d, k = 1...K \tag{7}$$

$$\eta_i^{'} \sim Dir(\alpha_i^{L3}\beta_{k,i}^{L2}) \quad \forall i = d_{cont}{+}1...d \tag{8}$$

$$\eta_i \sim Mult(\eta_i^{'}) \quad \forall i = d_{cont}{+}1...d. \tag{9}$$

The variance parameters σ are further assigned independent inverse gamma priors with hyper-parameters $(\gamma_{a1}, \gamma_{b1})$. Similarly, α follows a gamma prior with hyper-parameters $(\gamma_{a2}, \gamma_{b2})$. The conditional dependency graph of the model is illustrated in Figure 2.

3.2 Posterior Inference

Let $\Theta = \{\Delta, \mu, \sigma, \alpha, \beta\}$ be the set of all parameters. The central inference problem is to estimate these parameters given an observed instance $\mathbf{x}$ and a training dataset

$$\mathcal{D} = (\chi_1^{L2+}, ..., \chi_K^{L2+}, \chi_1^{L2-}, ..., \chi_K^{L2-}, \chi^{L1+}, \chi^{L1-}), \tag{10}$$

where χ_k^{L2+} represents a set of training instances corresponding to subgroup k such that $f(.) > 0.5$ while χ^{L1+} is the set of instances pooled across all the subgroups. Similarly χ_k^{L2-} and χ^{L1-} denote the negative instances. The posterior density of the parameters is given as

$$p(\Theta|\mathbf{x}, \mathcal{D}) = \int f(\mathbf{x}^*|\Theta)p(\mathbf{x}|\mathbf{x}^*, \Theta) \prod_k p(\chi_k^{L2}|\Theta)p(\chi^{L1}|\Theta)p(\Theta)d\Theta. \tag{11}$$

Here $\mathbf{x}^*$ is constructed as in equation (1) and $p(\Theta)$ is the prior distribution for the parameters as defined in equations (2) to (9). The first likelihood term in the above posterior captures the probability of a counterfactual to be valid while the second term encourages close proximity between an original instance and its counterfactual. Specifically,

$$p(\mathbf{x}|\mathbf{x}^*, \Theta) \propto e^{\frac{-\|\mathbf{x}^* - \mathbf{x}\|_2}{2\lambda}}, \tag{12}$$

where λ is a bandwidth hyper-parameter.

The likelihood of the training instances at the population level is written as

$$p(\chi^{L1}|\Theta) \propto \prod_{\mathbf{y} \in \chi^{L1-}} f(\mathbf{y}^*|\Theta)e^{\frac{-\|\mathbf{y}^* - \bar{\chi}^{L1+}\|_2}{2\lambda}} \tag{13}$$

where $\mathbf{y}^*$ is the counterfactual constructed from a negative training instance based on the parameters at $L1$ and $\bar{\chi}^{L1+}$ is the expected value of the positive instances in the feature space. Intuitively, we wish to find the counterfactuals of negative training instances at the population level in such a way that they are in close proximity to the positive instances of the training data. A similar construct follows for $p(\chi_k^{L2}|\Theta)$.

3.3 Sampling Mechanism

An approximation to equation (11) must be developed since exact posterior inference is not feasible. Simulation techniques such as Markov Chain Monte Carlo (MCMC) (Neal, 1993) methods allow drawing a sequence of correlated samples that can be used to estimate the intractable integrals. However, considering the large number of parameters and the complex nature of the posterior distribution induced by the presence of the classifier function, traditional MCMC methods such as Metropolis and Gibbs sampling will struggle to converge to the target distribution.

Hamiltonian Monte Carlo (HMC) (Betancourt and Girolami, 2015) sampling methods enable efficient exploration of such complex parameter spaces by incorporating the

gradient of the log posterior. The geometric information provided by these gradients can guide the chain towards regions of high posterior density, thereby reducing the number of samples required for convergence. Exploiting the fact that $f(.)$ is differentiable, we use the No-U-Turn-Sampler (NUTS) (Hoffman et al., 2014) variant of HMC for computing the posterior samples.

The $(\Delta_1, ...\Delta_N)$ samples produced by this sampling mechanism are used to derive N different $\mathbf{x}^*$ that can serve as a diverse set of counterfactuals for a given $\mathbf{x}$. The uncertainty in the generated samples can be quantified using measures such as variances, interquartile ranges and credible intervals, while a point-estimate if required can be obtained through summaries or ranking the samples by cost metrics.

4 EXPERIMENTS

We discuss here the experiment setup, present counterfactual evaluations and assess convergence properties. More details can be found in the supplementary material.

4.1 Evaluation Setup

Datasets: To evaluate our approach, we consider the following datasets: (a) Adult Income (Kohavi and Becker, 1996) - a dataset containing the income factors of various individuals such as *Gender*, *Race*, *Marital Status*, *Education*, *Workclass*, *Occupation*, *Age* and *Hours*. Except for *Age* and *Hours*, the rest of the variables are categorical and the classification objective is to predict whether an individuals' income exceeds \$50K. (b) German Credit (Hofmann, 1994) - a dataset that includes 20 different attributes of persons who takes a credit in a bank, with the vast majority of these attributes being categorical and a binary label indicating if an individual is a credit risk or not. and (c) HELOC (FICO, 2018) - a dataset with information on customers who received a home equity line of credit. It has over 20 features that are predominantly continuous and a binary label as to whether a customer paid back the loan or not.

Classification Model: We train a non-linear neural-network model with 2 layers and 200 hidden neurons as the classifier. The categorical values are converted to a smoothed one-hot encoded vector while the continuous values are normalized to be between 0 and 1 in the feature space. We obtain an accuracy of 80% for Adult Income, 75% for German Credit and 72% for HELOC. We use only the instances that are correctly classified by the model in ground-truth when evaluating the counterfactuals.

Settings: Throughout the experiments, we used a burn-in of 5000 samples, and a target sample size of 1000, which is sufficiently large. The standard normal distribution and a symmetric Dirichlet were used. The gamma hyper-parameters were set to a unit value and a value of 0.7 was used for the bandwidth.

4.2 Flattened Bayes Evaluation

We first focus on the assessment of a flattened model where only the local instances are considered. This would allow comparing the benefits of using a Bayesian model over a traditional point-estimate based counterfactual model.

Qualitative: Table 1 shows a few examples of counterfactual samples that were generated for an instance in Adult Income dataset, where a realistic setting is followed in which variables such as *Age*, *Gender*, *Race* and *Marital Status* are frozen while the values of other variables are allowed to change. The generated samples provide a wider range of options to the target user and highlights the various possibilities to have an income above \$50K. For example, the first sample indicates an option where the user can continue to retain the current values of *Workclass*, *Occupation* and *Hours*, while the second sample allows to retain only *Workclass* and *Education*. The samples also help address questions such as *"What if I have to work fewer hours than I currently do"*. Contrast this with a point-estimate model where the counterfactual must be re-generated for each *what-if* question and the potential set of options for the user may not be evident.

Quantitative: We also perform a quantitative comparison of the Bayesian method with the traditional point-estimate (Wachter et al., 2017) technique over various measures such as validity, sparsity and proximity for all the three datasets in Table 2. All the features are assumed to be mutable here and the ADAM optimizer (Kingma and Ba, 2015) is used to obtain the point estimates. The validity metric captures the percentage of negative instances in the test set for which a valid counterfactual was generated, and the coverage appears nearly identical between the two methods. The sparsity metric highlights the percentage of features being used, while proximity is a measure of the closeness between the counterfactual and the original instance in the feature space. The point-estimates seem to perform slightly better for these two attributes. This is unsurprising because the variety in samples is an important consideration of the Bayesian method and this requires an increase both in the number and change in magnitude of the features.

Diversity: A key benefit of the Bayesian counterfactual method is its ability to generate multiple options for the user. It is desirable for the counterfactual samples not to be trivial modifications and rather be a diverse choice. We compare the diversity of the counterfactuals produced by the Bayesian model with a point-estimate model that is initialized with different random seeds similar to Mothilal et al. (2020). The diversity metric is computed by calculating the distance between a counterfactual and all its peer samples and averaging over them. For a continuous variable the distance is based on the $l1$ norm scaled by median absolute deviation in the training set, while for a categorical variable it is based on whether the value has changed or not. Figure 3 illustrates this comparison by plotting the

Table 1: Examples of generated counterfactual samples for a negative label instance in Adult Income dataset.

	Immutable				Mutable				
	Age	Gender	Race	Marital	Workclass	Education	Occupation	Hours	Label
Original	69	Male	White	Married	Self-emp	HS-grad	Sales	30	<=50K
Counterfactual Samples	–				Self-emp	Prof-School	Sales	30	>50K
					Self-emp	HS-grad	Blue-Collar	43	>50K
					Gov	HS-grad	Professional	42	>50K
					Private	Masters	White-Collar	25	>50K

Table 2: Comparison of generated counterfactuals between Bayesian Posteriors and Point Estimates.

	Bayesian Posterior			Point Estimates		
Dataset	Validity % ↑	Sparsity % ↑	Proximity ↓	Validity % ↑	Sparsity % ↑	Proximity ↓
Adult Income	100	50	1.8	100	38	1.2
German Credit	97	50	2.8	100	35	2.5
HELOC	94	48	3.3	91	36	2.3

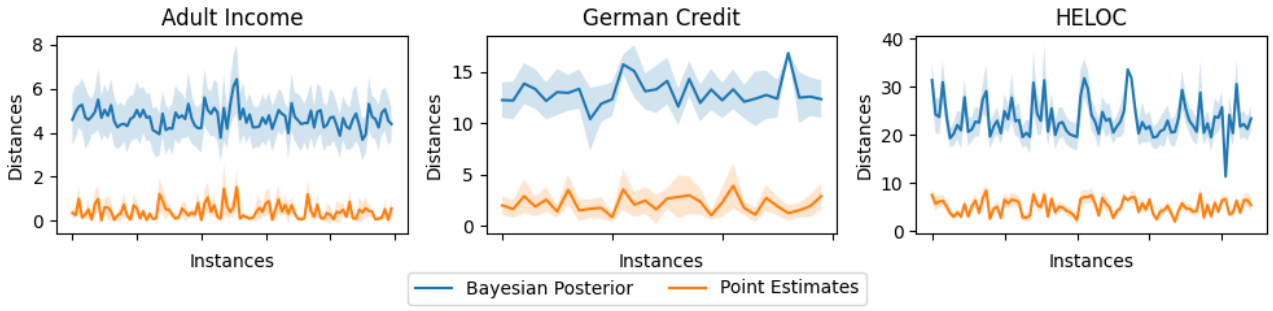

Figure 3: Diversity comparison between Bayesian Posteriors and randomly initialized Point-Estimates for different datasets. Larger distances imply greater diversity between the counterfactual samples of an instance.

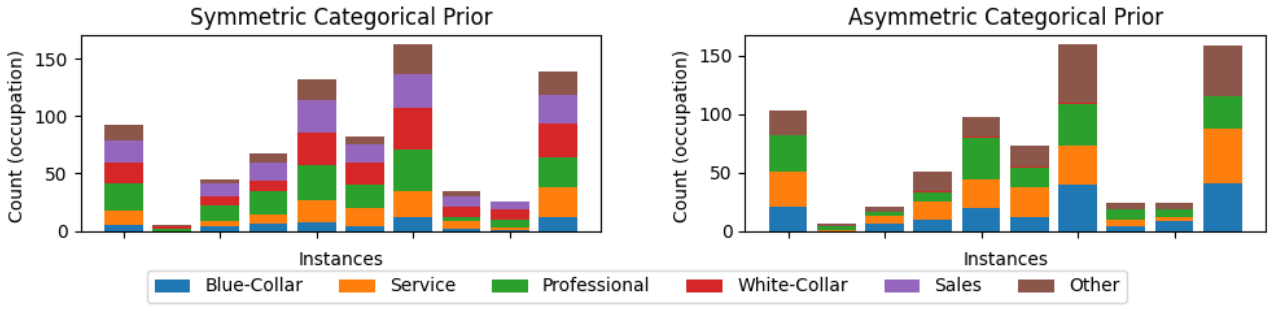

Figure 4: Inductive bias for categorical values in the generated counterfactuals of Adult Income dataset. *left*: Posterior samples from a symmetric Dirichlet prior on the *Occupation* categorical variable. *right*: Asymmetric prior with negligible mass on *Sales* and *White-Collar* values.

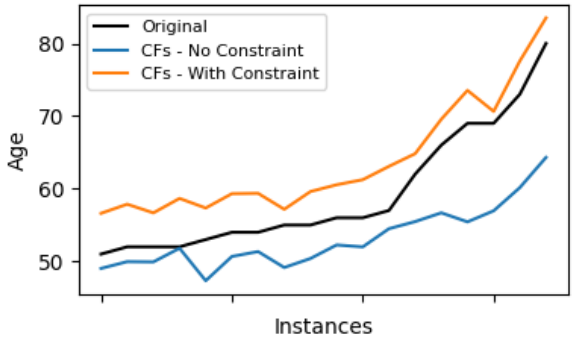

Figure 5: Unary constraint that dissuades a counterfactual *Age* value to be lower than its original in the AdultIncome dataset. The model with a truncated prior produces feasible counterfactuals.

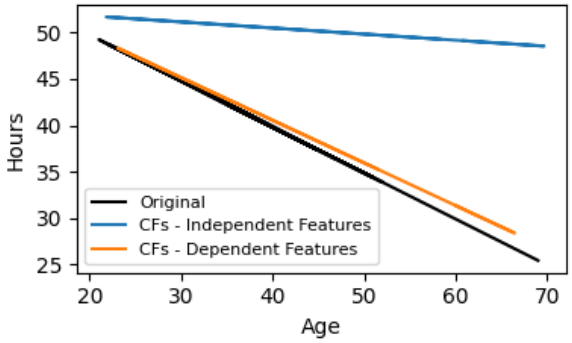

Figure 6: Binary constraint with a monotonic trend between *Hours* and *Age* variable pairs in the AdultIncome dataset. The model with a conditional prior preserves the feature relationship in generated counterfactuals.

mean distance and their variance. The optimization function used in point-estimate models seem to converge towards a narrow region despite random initialization. In contrast, it is evident that across datasets the Bayesian methods owing to their distribution oriented approach produce considerably diverse samples.

Inductive Bias: When generating the counterfactuals for an instance, we may want to incorporate customizations that are driven by apriori beliefs. For instance, a user may be less inclined to change their education levels beyond a Masters degree or would prefer not working longer hours. Such beliefs can be naturally inducted through asymmetric and truncated Bayesian priors. Figure 4 presents the samples generated from two different priors for the categorical variable *Occupation* in Adult Income dataset. The left side figure uses a symmetric Dirichlet prior with all the categories being equally likely. The right side figure uses an asymmetric prior that reduces the mass on *White-Collar* category (i.e. it is less likely than others) while the mass on *Sales* variable is set to a negligible value. Consequently, the figure on the right side doesn't produce any samples for *Sales* category and only a few samples for *White-Collar* category. Such targeted customizations are difficult to achieve in traditional models.

Feature Constraints: We evaluate feasibility by validating whether the counterfactuals satisfy constraints entailed by a given causal model. Similar to Mahajan et al. (2019), we consider an unary constraint where it is infeasible for the *Age* variable to decrease in the generated counterfactual and a binary constraint where there is a monotonic trend between the *Age* and *Hours* variables. The former is modeled using a truncated prior while the latter uses the linear approximation in (5) and the counterfactuals are generated for a subset of data points. Figure 5 highlights that the counterfactual *Age* values are consistently greater than the original when incorporating the constraint through a domain specific prior. Similarly, Figure 6 shows that when *Hours* and *Age* are negatively correlated, modeling their feature dependency explicitly generates counterfactuals that preserve their relationship.

4.3 Multi-level Bayes Evaluation

Hierarchical models provide the advantage of allowing relative comparisons of counterfactuals generated across different levels in the hierarchy and thus help assess counterfactual fairness. We first consider a two-level hierarchy, and in Figure 7 compare the values of counterfactual samples generated at the local instance level to the global population level. The left figure uses a box plot to display the median and whiskers of samples from the ordinal *Education* variable in Adult Income dataset. The middle figure plots the mode and dispersion of the nominal *Workclass* variable while the right figure displays the median and credible intervals for the continuous variable *Hours*. The population level aggregated counterfactual values for these variables are shown in a red dashed line. These illustrations allow drawing inferences such as *the education levels required for a local counterfactual seems lower than the global average*, or that *the work category required for local instances is compatible with global standards* and so on.

For the three-level hierarchy, we focus on the *Hours* variable and consider two different groups corresponding to the *Gender* and *Race* categorical variables. The top part of Figure 8 plots the median and credible intervals for $Hours$, along with the values at population level (red dashed line) and subgroups level (green dashed line). This enables the simultaneous visualization of how the values at local levels compare against the global values and the *Male* and *Female* subgroups. The bottom part of Figure 8 contrasts the values for *Hours* against *Black*, *White* and *Other* race categories. See supp. for more results.

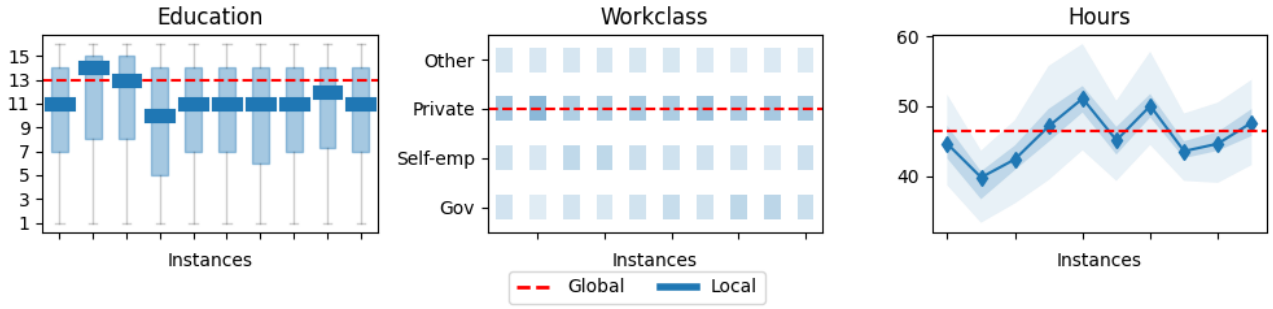

Figure 7: Global vs Local Counterfactuals in Adult Income. *left*: Box plot of ordinal variable *Education* with the median value of samples for different instances mostly below the global value 13. *middle*: The samples mode coincides with the global for nominal variable *Workclass*. *right*: The credible interval region for continuous variable *Hours* covers the global value.

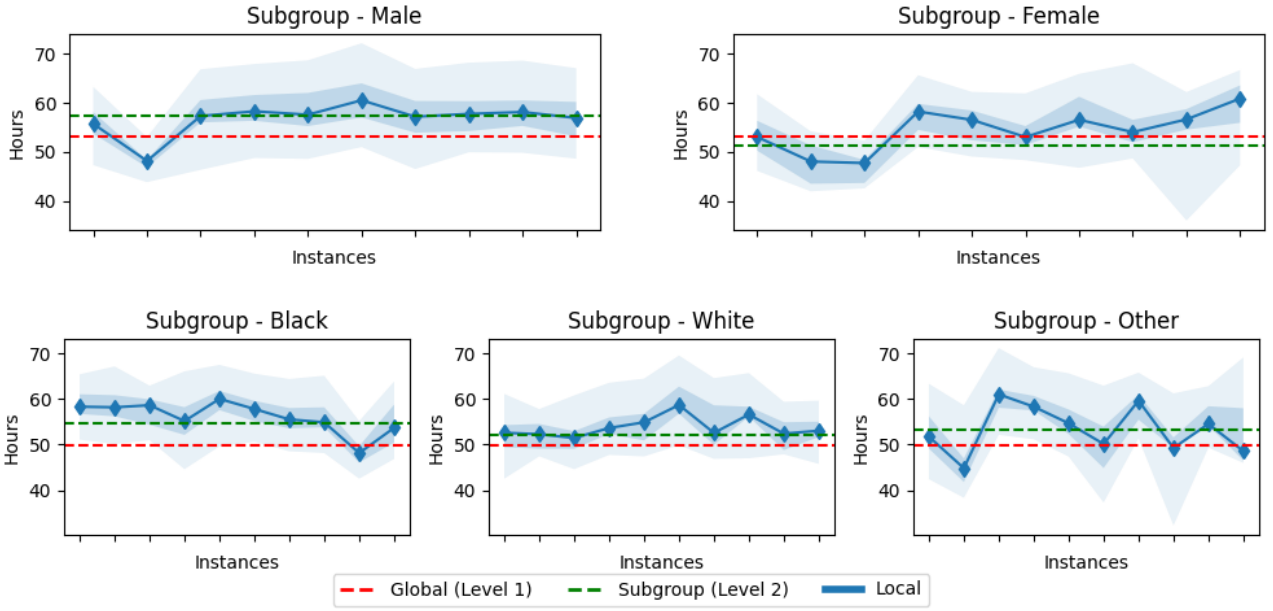

Figure 8: Multi-level hierarchical model showing whether the counterfactual of a local instance coincides with or deviates from subgroups within the data (Level 2) and the entire dataset (Level 1) for *Hours* variable in Adult Income. *top*: Subgroups *Male* and *Female* of *Gender* variable. *bottom*: Subgroups *Black*, *White* and *Other* of *Race* variable.

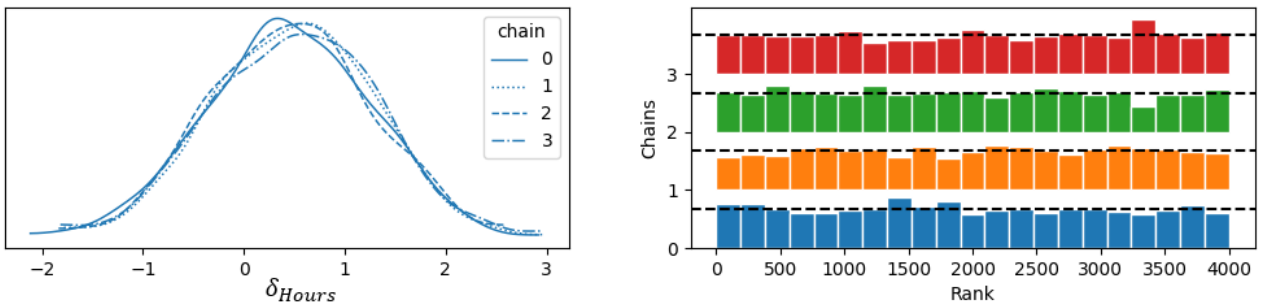

Figure 9: Convergence of samples for *Hours* variable in Adult Income dataset. *left*: Density estimate of the posterior samples for four different chains. *right*: Rank plots of posterior draws showing no substantial difference across the chains.

4.4 Convergence Diagnostics

It is critical to monitor whether the samples produced using an MCMC method converges (Roy, 2020) to the target posterior distribution. The generated samples maybe degenerate if they are strongly correlated with each other and are not effectively independent. To assess convergence, we run 4 different chains initialized at various starting points and check if the obtained distribution is similar across the chains. The most widely used convergence diagnostic is the scale reduction factor $\hat{R}$, which compares the variance of all the chains mixed together with the variance of individual chains. The observed $\hat{R}$ was well under 1.1, thus not detecting any convergence problems. We also inspected the effective sample size and found it to be large, confirming the absence of auto-correlations in the chains. Figure 9 displays two other convergence metrics. In the left side, the kernel density estimates of the posterior draws are shown for the *Hours* variable and it can be seen that the distributions appear similar across the chains. Additionally, we visualize the histograms of ranked posterior draws for each chain in the right figure, and as advocated in Vehtari et al. (2021) observe that the rank plots of all chains appear similar thereby indicating a good mixing of the chains.

5 LIMITATIONS

A valid criticism of the Bayesian methods is the inordinate number of computational steps required for convergence. For collecting N samples of V variables, HMC has a complexity of $\mathcal{O}(NV^{5/4})$. Even though the cost can be amortized for higher levels in the hierarchy and efficient parallelization can scale out the computations, the inference duration may still be intractable for real-time performance. The use of HMC also implies that the classifier function must be differentiable. Consequently, our solution is not purely model agnostic. While alternate sampling methods that can handle black box models exist, such solutions may pose challenges in convergence. The advent of AutoGrad and the prevalence of deep learning though makes our differentiable assumption reasonable. Finally, in order to model feature dependencies, we assume that the structural causal model is known apriori. However, in practice this information may not be available and learning them automatically is preferable. It must also be noted that we do not support complex causal relationships over multiple features.

6 SOCIETAL IMPACT

Recent studies (Kasirzadeh and Smart, 2021; Slack et al., 2021a) have highlighted the importance of understanding the vulnerabilities and potential for misuse of counterfactuals. In particular, sufficient attention must be paid to ensure that the generated counterfactuals provide actionable recommendations (Karimi et al., 2021). Our work contributes positively towards allowing people to act, by providing a diverse choice to the users instead of patronizing them with a single ideal option, and personalizing the generation process with informative priors that can offer tailor made solutions. It is also important in a social context to verify whether the interventions vary based on protected attributes (Coston et al., 2020). The multi-level setup described here opens a pathway to perform such assessments by comparing the extent of deviations across different subgroups.

7 CONCLUSION

We presented a mechanism to generate multiple alternative counterfactuals in a probabilistic setting and characterized the perturbations at several levels of abstraction. Our formulation can support prior beliefs, handle multimodal parameters, furnish uncertainty metrics and compare across population levels, all while producing a diverse set of choices. The experiment results confirm the benefits offered by the proposed Bayesian framework. In future, we wish to include complex causal relationships between the features, enforce fair recourse and extend to black box classification models.

Acknowledgments

This paper was prepared for information purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Albini, E., Long, J., Dervovic, D., and Magazzeni, D. (2022). Counterfactual shapley additive explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1054–1070.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

Becker, M., Burkart, N., Birnstill, P., and Beyerer, J. (2021). A step towards global counterfactual explanations: Approximating the feature space through hierarchical division and graph search. *Adv. Artif. Intell. Mach. Learn.*, 1(2):90–110.

Betancourt, M. and Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4.

Bui, N., Nguyen, D., and Nguyen, V. A. (2021). Counterfactual plans under distributional ambiguity. In *International Conference on Learning Representations*.

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 582–593.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. (2020). Multi-objective counterfactual explanations. In *Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I*, pages 448–469. Springer.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31.

Dominguez-Olmedo, R., Karimi, A. H., and Schölkopf, B. (2022). On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pages 5324–5342. PMLR.

Dutta, S., Long, J., Mishra, S., Tilli, C., and Magazzeni, D. (2022). Robust counterfactual explanations for tree-based ensembles. In *International Conference on Machine Learning*, pages 5742–5756. PMLR.

FICO (2018). Fico xml challenge.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Hofmann, H. (1994). UCI machine learning repository.

Kanamori, K., Takagi, T., Kobayashi, K., and Arimura, H. (2020). Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *IJCAI*, pages 2855–2862.

Kanamori, K., Takagi, T., Kobayashi, K., and Ike, Y. (2022). Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1870. PMLR.

Karimi, A.-H., Schölkopf, B., and Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. (2022). Towards causal algorithmic recourse. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 139–166. Springer.

Kasirzadeh, A. and Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kohavi, R. and Becker, B. (1996). UCI machine learning repository.

Ley, D., Bhatt, U., and Weller, A. (2022). Diverse, global and amortised counterfactual explanations for uncertainty estimates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7390–7398.

Looveren, A. V. and Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer.

Mahajan, D., Tan, C., and Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada.

Nguyen, T.-D. H., Bui, N., Nguyen, D., Yue, M.-C., and Nguyen, V. A. (2022). Robust bayesian recourse. In *Uncertainty in Artificial Intelligence*, pages 1498–1508. PMLR.

Pawelczyk, M., Broelemann, K., and Kasneci, G. (2020). Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132.

Plumb, G., Terhorst, J., Sankararaman, S., and Talwalkar, A. (2020). Explaining groups of points in low-dimensional representations. In *International Conference on Machine Learning*, pages 7762–7771. PMLR.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.

Rawal, K. and Lakkaraju, H. (2020). Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198.

Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., and Vazquez, D. (2021). Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1056–1065.

Roy, V. (2020). Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 7:387–412.

Russell, C. (2019). Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28.

Slack, D., Hilgard, A., Lakkaraju, H., and Singh, S. (2021a). Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, 34:62–75.

Slack, D., Hilgard, A., Singh, S., and Lakkaraju, H. (2021b). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404.

Smyth, B. and Keane, M. T. (2022). A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*, pages 18–32. Springer.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: an improved r for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718.

Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., and Schölkopf, B. (2022). On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Watson, D. (2022). Rational shapley values. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1083–1094.

Wu, Y., Zhang, L., and Wu, X. (2019). Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Supplementary Material: Bayesian Hierarchical Models for Counterfactual Estimation

A EVALUATION METRICS

We formally define the quantitative metrics validity, proximity, sparsity and diversity that were used in the experiments here for completeness. Given a test set of negative outcomes χ^-, these metrics are computed as

$$validity = \frac{|\{\mathbf{x} \in \chi^- \mid \exists \mathbf{x}^* : f(\mathbf{x}^*) > 0.5\}|}{|\chi^-|}$$

$$proximity = \frac{1}{|\chi^-|} \sum_{\mathbf{x} \in \chi^-} \min_{\mathbf{x}^* \in \mathbf{x}^{cfs}} \|\mathbf{x}^* - \mathbf{x}\|_2$$

$$sparsity = \frac{1}{|\chi^-|} \sum_{\mathbf{x} \in \chi^-} \frac{1}{|\mathbf{x}^{cfs}|} \sum_{\mathbf{x}^* \in \mathbf{x}^{cfs}} \frac{\sum_{l=1}^{d_{cont}} \mathbb{I}(Q(x_l^*), Q(x_l)) + \sum_{l=d_{cont}+1}^{d} \mathbb{I}(x_l^*, x_l)}{d}$$

$$diversity(\mathbf{x}^{cfs}) = \frac{1}{|\mathbf{x}^{cfs}|^2} \sum_{i=1}^{|\mathbf{x}^{cfs}|-1} \sum_{j=i+1}^{|\mathbf{x}^{cfs}|} \Big[\sum_{l=1}^{d_{cont}} \frac{|x_l^i - x_l^j|}{MAD_l} + \sum_{l=d_{cont}+1}^{d} \mathbb{I}(x_l^i, x_l^j) \Big]$$

where $\mathbf{x}^{cfs}$ is the set of counterfactual samples of $\mathbf{x}$, $\mathbb{I}$ is an indicator function that evaluates to 1 if both arguments are equal, Q is a quantization function that coarsely bins a continuous feature into 10 discrete intervals and MAD is the median absolute deviation computed from 80% of the dataset that was marked as training data.

B ROBUSTNESS EVALUATION

We hypothesize that the multi-level Bayes structure enables information sharing through which the quality of parameter estimates can be improved. While there are several themes for quality such as robustness of the estimates w.r.t input perturbations or classification model changes, we focus here on whether using the hierarchical structure enables the counterfactuals to lie in dense regions of the data manifold. Data supported counterfactuals tend to be model invariant and result in realistic recourses, and hence data density is an important metric for counterfactual robustness.

We consider two measures namely the distance of a counterfactual to its k nearest neighbors and the local outlier factor, which computes the deviation of the local density of a counterfactual with respect to its neighbors. The data points within the positive class in the training dataset is used as the neighborhood. The distance between any two data points i and j uses the same metric defined above for diversity (the term within the square brackets).

Figure 10 plots the neighborhood distance and local outlier factor for different values of k, from the HELOC dataset. We partition the dataset into different numbers of clusters using the standard k-Means algorithm, and evaluate the generated counterfactual in the neighborhood of its corresponding cluster in the ground-truth. We can see in the top figure that the neighborhood distance is consistently smaller when using a hierarchical model. Similarly, the bottom figure shows that the outlier percentage is reduced when a multi-level Bayes model is utilized. This confirms that using a hierarchical model encourages the desired behavior of a counterfactual to reside in the neighborhood of its subgroup.

C FAIRNESS EVALUATION

Our solution computes the perturbations at different levels of abstraction and it can be exploited to compare the counterfactual of an instance directly with group level counterfactuals. When the groups correspond to protected attributes such as *Gender* or *Race*, it provides a mechanism to verify whether the distributions are identical for each protected group and thereby assess counterfactual fairness.

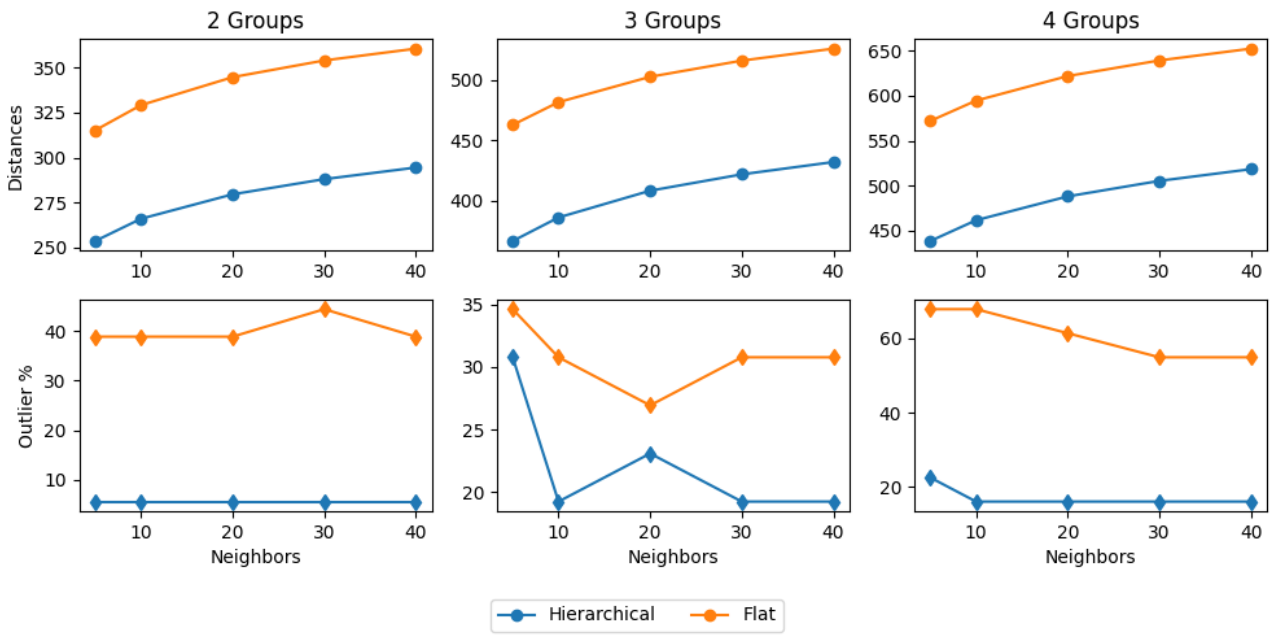

Figure 10: Data support for generated counterfactuals in HELOC dataset. *top*: Neighborhood distances. *bottom*: Percentage of outliers. In both cases, the hierarchical model exhibits the desired behavior of a smaller value.

Table 3: Recourse cost at the subgroup level (*left*) and at the instances level (*right*) in AdultIncome dataset. The difference in cost can be used to determine recourse fairness.

Group	Sub Group	Cost
Gender	Male	5.42 ± 1.05
Gender	Female	5.09 ± 1.01
Race	White	5.32 ± 1.02
Race	Black	5.44 ± 0.99
Race	Other	5.53 ± 1.20

Instance	Gender	Race	Age	Occupation	Workclass	Cost
1	Male	White	69	Sales	Self-Emp	6.74
2	Male	White	28	Blue-Collar	Private	6.54
3	Female	White	52	White-Collar	Gov	5.96
4	Female	Black	44	Sales	Private	4.93
5	Male	Other	23	White-Collar	Private	7.41

In Table 3 we present the cost of recourse both at the subgroup level (left side) and at the instance level (right side) for the AdultIncome dataset. The costs are aggregated using the mean function to handle multiple counterfactual samples. While there maybe many problem specific definitions of recourse cost, we restrict ourselves to the distance function as a proxy for the cost. As before, the distance between an original data point i and its counterfactual j is computed from the $l1$ norm scaled by medium absolute deviation for continuous features and the change in value for categorical features.

The recourse fairness at a group level is the difference in cost between the subgroups, while at an instance level it is the difference between the cost for a particular instance and the protected group for which we wish to compare against. For example, we can see that the *Race - Other* subgroup has a cost greater than its peers. Similarly, the cost for *Instance 4* is less than both the *Female* and *Black* subgroups the instance belongs to. Given a problem defined scalar threshold for the cost difference, we can now calculate the demographic parity and thus measure fairness.

D PARAMETER ANALYSIS

Besides convergence, the number of posterior samples is also relevant for the diversity in generated counterfactuals. We plot the diversity measure against different numbers of samples in Figure 11. The initial burn-in samples that were used for tuning is ignored and only the samples in equilibrium are considered. We observe that the counterfactual diversity in general increases with the number of samples used across the three datasets. However, they do plateau indicating that only a selected number of truly divergent recourse options is available. In practice, a top-k ranking of these samples based on a domain specific metric maybe necessary before presenting the recommendations to the user.

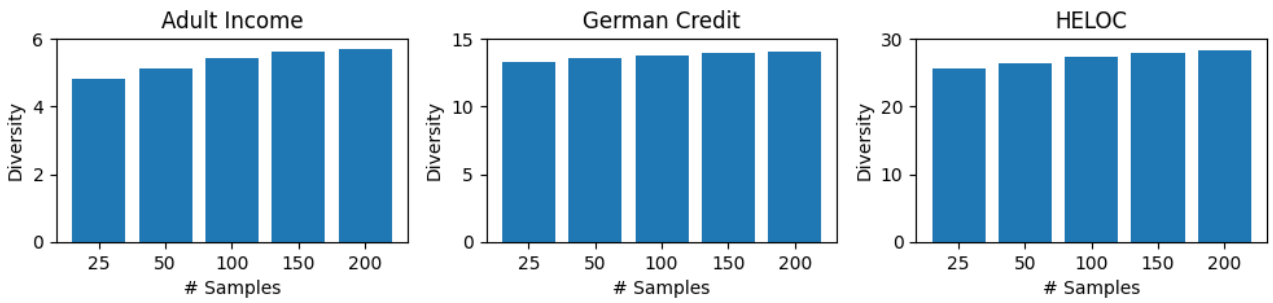

Figure 11: Sensitivity analysis on the number of posterior samples. The counterfactual diversity increases with the sample size.

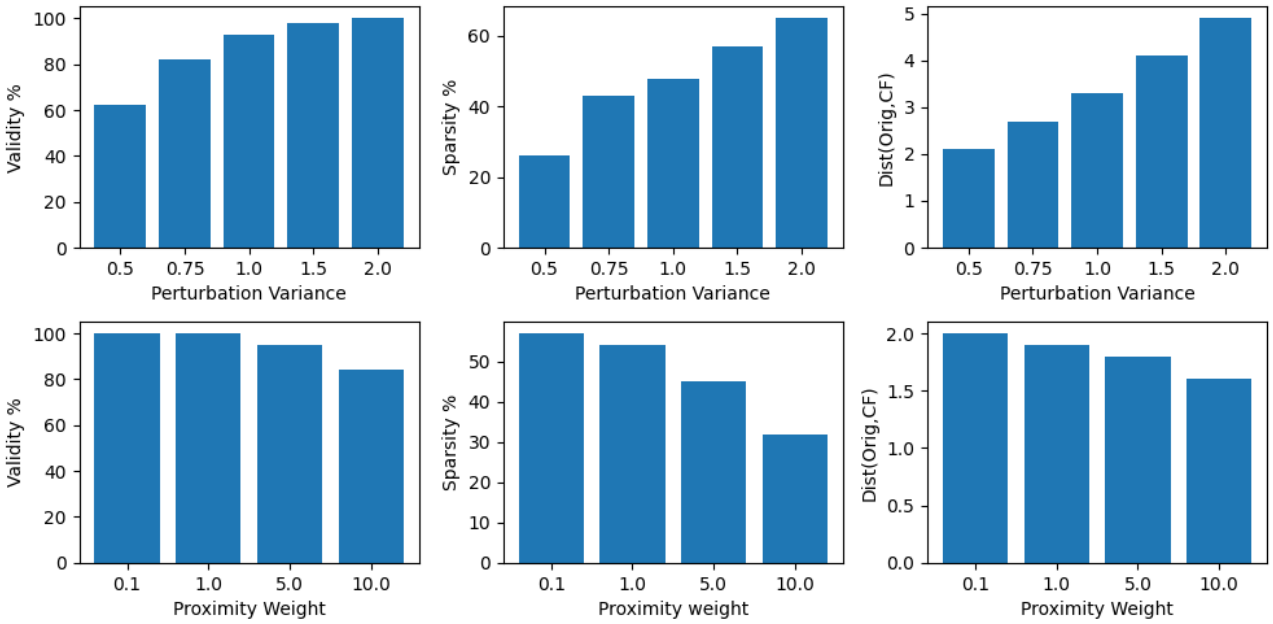

Figure 12: Control over the generated counterfactual properties. *top*: Change in the perturbation variance parameter for the HELOC dataset. *bottom*: Modifications to the importance for proximity in Adult Income dataset.

The extent of perturbations can be controlled using the prior values. In particular, by adjusting the variance parameter σ of a continuous feature's normal distribution, we can influence the generated counterfactuals. The top part of Figure 12 shows how validity, sparsity and proximity changes with σ for the HELOC dataset. When the perturbation variance is high, there are more options for constructing the counterfactuals and consequently the number of instances for which a counterfactual can be generated increases (*top left*). However, it also implies that the percentage of features used (*top center*) and the distance between an original and counterfactual point (*top right*) also increases.

Modifying the prior parameters for perturbations maybe inconvenient, especially when there is a mixture of categorical and continuous features. An alternate mechanism to vary the generation process is to scale the second term that encourages proximity in equation (11). The bottom part of Figure 12 plots the change in validity, sparsity and proximity for different weights assigned to this term in the Adult Income dataset. As the importance assigned to the proximity between an original and its counterfactual increases, this constrained setting results in the reduction of validity (*bottom left*) and the number of features used (*bottom center*). However, the counterfactual point is now more closer to the original data point (*bottom right*). Both the variance and proximity weight parameters provide effective control to tailor the outcomes based on a problem specific scenario.