Fair Inference for Discrete Latent Variable Models: An Intersectional Approach

Rashidul Islam*
raislam@visa.com
Visa Research, Visa Inc.
Atlanta, Georgia, USA

Shimei Pan
shimei@umbc.edu
Department of IS, UMBC
Baltimore, Maryland, USA

James R. Foulds
jfoulds@umbc.edu
Department of IS, UMBC
Baltimore, Maryland, USA

ABSTRACT

It is now widely acknowledged that machine learning models, trained on data without due care, often exhibit discriminatory behavior. Traditional fairness research has mainly focused on supervised learning tasks, particularly classification. While fairness in unsupervised learning has received some attention, the literature has primarily addressed fair representation learning of continuous embeddings. This paper, however, takes a different approach by investigating fairness in unsupervised learning using graphical models with discrete latent variables. We develop a fair stochastic variational inference method for discrete latent variables. Our approach uses a fairness penalty on the variational distribution that reflects the principles of intersectionality, a comprehensive perspective on fairness from the fields of law, social sciences, and humanities. Intersectional fairness brings the challenge of data sparsity in minibatches, which we address via a stochastic approximation approach. We first show the utility of our method in improving equity and fairness for clustering using naïve Bayes and Gaussian mixture models on benchmark datasets. To demonstrate the generality of our approach and its potential for real-world impact, we then develop a specialized graphical model for criminal justice risk assessments, and use our fairness approach to prevent the inferences from encoding unfair societal biases.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Unsupervised learning**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

fairness in AI, intersectionality, probabilistic graphical models, stochastic variational inference

ACM Reference Format:
Rashidul Islam, Shimei Pan, and James R. Foulds. 2024. Fair Inference for Discrete Latent Variable Models: An Intersectional Approach. In *International Conference on Information Technology for Social Good (GoodIT '24), September 04–06, 2024, Bremen, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677525.3678660

*The work was done while at Department of Information Systems, UMBC, USA.

GoodIT '24, September 04–06, 2024, Bremen, Germany

ACM ISBN 979-8-4007-1094-0/24/09
https://doi.org/10.1145/3677525.3678660

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have become ubiquitous. Increasingly, the automated decisions made by these systems have important real-life consequences, from credit scoring to the prediction of re-offending [46, 48]. However, implicit societal stereotypes in the data can often undermine the integrity of these decisions, leading to unfair discrimination against certain groups of people [2, 3, 7, 10, 47]. With the rising awareness and regulations, the AI community has devoted much effort to the development and enforcement of numerous quantifiable notions of fairness for AI/ML models [18, 21, 25, 35, 41]. The main paradigm for fair algorithms is to posit mathematical criteria of fairness across protected demographic groups, (e.g. by gender, and race) or similar individuals (e.g. persons with similar merits and risks) [6, 22]. The paradigm then enforces these criteria, when optimizing objective functions, by penalizing violations [5, 21, 30, 31] or by finding a transformation of data that provides fair latent representations [59].

From a fairness perspective, representation learning is appealing because deep learning-based vector representations often generalize to tasks that are unspecified at training time, implying that a properly designed fair network might operate as a kind of "parity bottleneck," reducing discrimination in unknown downstream tasks. Particularly, the goal of fair representation learning is to transform the data into continuous latent spaces that are invariant to protected attributes and useful to mitigate societal bias in tasks, e.g., classification. Most of the recent frameworks [19, 43, 43, 45, 60, 61] are built upon the variational autoencoder (VAE) [40], which can perform effective stochastic variational inference (SVI) and learning in continuous latent variables using backpropagation.

While the benefits of fair representations in continuous latent space in downstream tasks are clear, we are conversely interested in extending the success of variational techniques to fair inference in graphical models with discrete latent variables. As societal prejudices, societal disadvantages, underrepresentation of minorities, and intentional prejudices are inherent in historical data [3], inferences can naturally encode these harmful biases in the latent variables which should be prevented to mitigate discriminatory decisions. It is pertinent to acknowledge the existing work on fairness in causal latent variable models [41]. These works focus on fair causal inference in supervised settings, such as with class labels. They take into account the fairness-aware estimation of causal effects, attempt to handle confounding factors, and offer an important perspective on the fairness problem. Our work, however, explores a different yet complementary space. We focus on unsupervised settings with discrete latent variables, where the causal mechanisms are not explicitly modeled, but fairness considerations remain crucial.

Graphical models with discrete latent variables are used in numerous AI/ML methods including semi-supervised learning [38], and topic modeling [56]. Furthermore, discrete latent variables are a natural fit for complex reasoning, planning and predictive learning. E.g., an AI agent may learn a pattern that often occurs in the data, "*if you study hard, you will be successful,*" via a latent variable, *study hard*. This can be achieved by structuring the model such that the latent variable is associated with certain observable variables, like hours spent studying, grades obtained, etc. However, inference on discrete latent variables with backpropagation-based variational methods is difficult due to the inability to re-parameterize gradients. To address this, continuous relaxations in the VAE framework have been achieved via the Gumbel-Softmax re-parameterization trick [32] which defines a temperature-based continuous distribution, and converges to a discrete distribution in the zero-temperature limit. [12] addressed fairness in clustering problems for both k-center and k-median objectives, but not for general graphical models with discrete latents.

In this paper, we develop a *practical* framework for fair SVI on arbitrary graphical models with discrete latent variables to improve their equity and fairness. Our method is general and could be incorporated into probabilistic programming systems. Given a probabilistic graphical model, e.g. a custom model defined for a particular task, our goal is to perform inference such that the results are fair, e.g., they do not reflect negative stereotypes. For example, multiple studies demonstrated that police stop people from racial minority groups more frequently than whites [11, 23, 26, 27, 57]. In a traffic model, we may wish to prevent the inference that individuals of one demographic drive more aggressively than other demographics. Such an inference may in fact be warranted by the data, but it may be known –due to knowledge not encoded in the data, or known causes of the issue such as less data for minorities– that this inference cannot be correct, or it may be desirable to prevent it in order to avoid harm due to the use of the model, pursuant to Title VII [3]. Furthermore, our fairness intervention technique enforces an intersectional fairness notion [21] that guarantees fairness protections for all subsets of the protected attributes (e.g., black women, women, and black) which is consistent with the ethical principles of intersectionality theory [13, 15].

Our key contributions can be summarized as follows:

- We introduce the novel problem of fair inference in discrete latent variable models and propose its first solution. We formulate fair inference in unsupervised discrete latent variable models as a discrete VAE, and we design a practical stochastic variational inference algorithm, leveraging reparameterization strategies for discrete latents.
- We address the practical challenges that arise in this setting such as data sparsity in minibatches, via stochastic approximation, and we develop practical methods for addressing posterior collapse.
- We demonstrate the utility of our method for clustering analysis using naïve Bayes and Gaussian mixture models, and apply it to fair inference in a new special-purpose criminal justice decision model.

2 BACKGROUND

2.1 Variational Autoencoder (VAE)

Variational inference [29, 33] is an optimization approach to solve inference problems for latent variable models. We typically assume a generative model $p(x, z) = p(x|z)p(z)$ that produces a dataset $\mathbf{x} = \{x_j\}_{j=1}^n$ consisting of n i.i.d. individuals, generated using a set of K-dimensional discrete latent variables $\mathbf{z} = \{z_j\}_{j=1}^n$. Furthermore, each z_j is assumed to be generated from some prior distribution $p(z)$. We aim to compute the posterior distribution $p(z|x)$, which is assumed intractable over latent variables, and so approximations must be used. The key idea is to approximate $p(z|x)$ with a more tractable distribution $q(z)$, referred to as a variational distribution, and minimize the KL-divergence D_{KL} between them. The variational autoencoder (VAE) [40] performs variational inference in a latent Gaussian model where the variational posterior and model likelihood are parameterized by neural nets ϕ and θ, respectively. The VAE is generally implemented with a Gaussian prior $p_\theta(z) = \mathcal{N}(0, 1)$. The objective to maximize the evidence lower-bound (ELBO) is made differentiable by reparameterizing $z \sim q_\phi(z|x)$ with $z = \mu + \sigma \odot \mathcal{E}$, where $\mathcal{E} \sim \mathcal{N}(0, 1)$ as:

$$\mathcal{L}_{\text{VAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x)||p_\theta(z)]\,. \quad (1)$$

2.2 Gumbel-Softmax Trick

The Gumbel-Softmax trick is a reparameterization technique for training VAEs with discrete latent variables using the Gumbel-Softmax distribution [32, 44], leveraging the Gumbel-Max distribution, an efficient way to draw samples z from a k-dimensional categorical distribution with probabilities π as $z = \text{one\_hot}(\arg\max_i[g_i + \log \pi_i])$, where $g_i \sim \text{Gumbel}(0, 1)$ are i.i.d. samples. As "arg max" is not differentiable, in Gumbel-Softmax the softmax function is used as a continuous approximation:

$$z_i = \frac{\exp((g_i + \log \pi_i)/\tau)}{\sum_{k=1}^K \exp((g_k + \log \pi_k)/\tau)} \qquad \text{for } i = 1, \ldots, K\,, \quad (2)$$

where the temperature τ, which is is annealed towards 0, controls how closely samples from the Gumbel-Softmax distribution approximates the target distribution π. Reparameterizing via Gumbel-Softmax facilitates backpropagation.

3 METHOD: FAIR VARIATIONAL INFERENCE

3.1 Problem Formulation

Consider a generative probabilistic model $p_\theta(x, z)$ with discrete latent variables z. Let A be protected attributes, e.g. the individuals' gender and race, which may or may not be included in the typical (or non-sensitive) D-dimensional attribute vector x. We aim to perform fair inference on $p_\theta(x, z)$ with respect to A.

We propose to model this scenario as a VAE via Equation 1, leveraging the Gumbel-Softmax trick (Equation 2). Our proposed stochastic variational inference algorithm to compute the posterior distribution $p_\theta(z|x)$ achieves two properties: 1) it allows scalable inference that suits big data, and 2) it incorporates a fairness penalty function, which provides a simple and effective fairness intervention via backpropagation.

3.2 Inference Network

We use a neural network-based inference network $q_\phi(z|x)$ for the variational approximation to the intractable posterior $p_\theta(z|x)$, where weights and biases are variational parameters ϕ. Following [32], let the prior $p_\theta(z)$ be a uniform discrete distribution with probability $1/K$, although this can easily be generalized as shown in a later section for specialized modeling. The first step of our method is to reparameterize the variational distribution for latent variable sampling so that discrete distributions are re-represented as unconstrained distributions, in order to facilitate backpropagation. The two main options are the logistic-normal representation [56], and the Gumbel-softmax representation. If we use the logistic-normal representation, we reparameterize $q_\phi(z|x)$ using a mean μ and covariance matrix Σ via a logistic function. We focus on the Gumbel-softmax approach, since it performed better in preliminary experiments. To encode discrete z, the inference network basically outputs unnormalized log probabilities $\log \pi$ for the latent classes which are then used to reparameterize $q_\phi(z|x)$ using the Gumbel-Softmax trick in Equation 2.

3.3 Generative Network

To learn the generative model's parameters θ, unlike for the VAE, no neural network is used. It is generally impossible to accurately approximate the true joint distribution over observed and latent variables, including the true prior and posterior distributions over latent variables using the VAE framework due to the unidentifiability of the model [36], meaning that there are many different possible configurations of the latents that would generate the same observed variables. Khemakhem et al. [36] provided a solution that requires a factorized prior over the latent variables given an auxiliary observed variable, usually class labels.

3.3.1 Addressing Unidentifiability. As we desire to perform unsupervised learning, where no class labels are available, we present a different approach that produces identifiable and meaningful latent variables. We randomly initialize θ, while simply considering some informative hyper-priors $p_\alpha(\theta)$ on θ, where α are the hyper-parameters and fixed. In the context of latent variables, $p_\alpha(\theta)$ plays an important role in achieving identifiability and meaningfulness, meaning that the θ can be uniquely estimated from the observed data. Furthermore, by incorporating prior knowledge or beliefs about the latent variables, $p_\alpha(\theta)$ can be used to break ties in the ordering of latent variables. This is because $p_\alpha(\theta)$ provides prior information that constrains the parameter space of the model, reducing the ambiguity in the estimated parameters θ. For example, consider a model that involves latent variables such as "hard-working" and "not hard-working," and the observed variables such as "high income" and "low income." The use of informative priors on θ can allow us to determine which latent variable is more likely to be associated with the observed variable of "high income." For instance, the prior may be such that the "hard-working" latent variable is more likely to be associated with "high income" than the "not hard-working" latent variable. This effectively breaks any ties in the ordering of the latent variables, leading to more meaningful and interpretable results. In *vanilla model* with no fairness intervention, we then jointly optimize θ and ϕ via the ELBO:

$$\begin{aligned}\mathcal{L}(\theta,\phi) &= \mathbb{E}_{q_\phi}[\log p_\theta(x,z)] - \mathbb{E}_{q_\phi}[\log q_\phi(z|x)] \\ &= \mathbb{E}_{q_\phi}[\log p_\theta(x|z) + \log p_\theta(z) + p_\alpha(\theta)] - \mathbb{E}_{q_\phi}[\log q_\phi(z|x)] \,.\end{aligned} \tag{3}$$

3.4 Fair Inference Technique

We enforce fairness by adding a penalty to the ELBO objective. Our inference and learning objective is:

$$\min_{\theta,\phi} \; -\sum_{j=1}^{n} \mathcal{L}_j(\theta,\phi) + \lambda \mathcal{F}(\phi, A) \,, \tag{4}$$

where, $\mathcal{L}(\theta,\phi)$ is the ELBO of the vanilla model in Equation 3, $\mathcal{F}$ is a fairness penalty, and λ is a hyper-parameter that trades between the ELBO and fairness. In this work, our fair inference technique uses a fairness criterion that is motivated from an intersectionality perspective as a penalty term to measure violations, with regard to parity in the inferred discrete latent variables for intersecting protected groups. To design the fairness penalty on ELBO, we adapt the differential fairness (DF) metric [21], which was originally proposed for classification. DF extends the 80% rule to multiple protected attributes and outcomes, and provides an *intersectionality property*:

Let $s_1, \ldots, s_p$ be discrete-valued protected attributes, $A = s_1 \times s_2 \times \ldots \times s_p$. An inference mechanism $q_\phi(z|x)$ satisfies ϵ-DF with respect to A if for all x, and $z \in K$,

$$e^{-\epsilon} \leq \frac{p(q_\phi(z|x) = z|s_i)}{p(q_\phi(z|x) = z|s_j)} \leq e^{\epsilon} \,, \tag{5}$$

for all $(s_i, s_j) \in A \times A$ where $P(s_i) > 0$, $P(s_j) > 0$ (Proof given in [21]). Smaller ϵ is better, and $\epsilon = 0$ is ideal. In principle, we can measure ϵ-DF using the empirical data distribution. Let $N_{z,s} = \sum_{x \in \mathbf{x}: A=s} \mathbf{z}$ and N_s be the empirically estimated expected counts for latent assignments per group and for total population per group, respectively. Then *ϵ-DF* can be estimated via the posterior predictive distribution of a Dirichlet-multinomial, where scalar α is a Dirichlet prior with concentration parameter $K\alpha$, as:

$$e^{-\epsilon} \leq \frac{N_{z,s_i} + \alpha}{N_{s_i} + K\alpha} \frac{N_{s_j} + K\alpha}{N_{z,s_j} + \alpha} \leq e^{\epsilon} \,. \tag{6}$$

3.4.1 Intersectional Fair Variational Inference. The above equation and its gradients are sufficient for batch training. However, the reliable estimation of *ϵ-DF* on the inference mechanism in terms of A, denoted by $\epsilon(q_\phi(z|x), A)$, for a minibatch becomes statistically challenging due to data sparsity of intersectional groups [20]. For example, one or more missing intersectional groups for a minibatch is a typical scenario in the stochastic setting that can lead to inaccurate estimation of the fairness, an obstruction to scaling up the inference using SVI. To address data sparsity in $\epsilon(q_\phi(z|x), A)$, inspired by the noisy update technique in several SVI algorithms [29], we develop a stochastic approximation-based approach [52] that updates count parameters for each minibatch m as follows:

$$\begin{aligned} N_{z,s} &\coloneqq (1-\rho_t) N_{z,s} + \rho_t \frac{n}{m} \hat{N}_{z,s} \,, \\ N_s &\coloneqq (1-\rho_t) N_s + \rho_t \frac{n}{m} \hat{N}_s \,, \end{aligned} \tag{7}$$

where, $\hat{N}_{z,s}$ and $\hat{N}_s$ are empirically estimated noisy expected counts per group for a minibatch, and ρ_t is a step size schedule, typically annealed towards zero. These updates correspond to the updates

Algorithm 1 Intersectional Fair Stochastic Variational Inference

Require: Train data $\mathbf{x} = \{x_j\}_{j=1}^{n}$
Require: Trade-off parameter $\lambda > 0$
Require: Desired fairness ϵ_0
Require: Constant step-size for expected counts ρ_t
Require: Constant step-size for optimization algorithm ρ_o
Require: Randomly initialized generative model's parameters: θ, i.e., μ and σ
Require: Randomly initialized inference network's parameters: ϕ, i.e., MLP's weights and biases
Require: Fixed hyper-priors $p_\alpha(\theta)$
Require: Fixed prior $p_\theta(z)$
Output: Likelihood $p_\theta(x|z)$
Output: Variational Posterior $q_\phi(z|x)$

- For each epoch:
  - For each minibatch m:
    - Empirically estimate $\hat{N}_{z,s} = \sum_{x \in \mathbf{x}_m : A=s} \mathbf{z}_m$ and $\hat{N}_s$
    - Apply update: $N_{z,s} := (1-\rho_t) N_{z,s} + \rho_t \frac{n}{m} \hat{N}_{z,s}$
    - Apply update: $N_s := (1-\rho_t) N_s + \rho_t \frac{n}{m} \hat{N}_s$
    - Estimate $\epsilon(q_\phi(\mathbf{z}|\mathbf{x}), A)$:
      $$e^{-\epsilon} \leq \frac{N_{z,s_i}+\alpha}{N_{s_i}+K\alpha} \frac{N_{s_j}+K\alpha}{N_{z,s_j}+\alpha} \leq e^{\epsilon}$$
    - Compute fairness penalty:
      $$\mathcal{F}(\phi, A) = \max(0, \epsilon(q_\phi(\mathbf{z}|\mathbf{x}), A) - \epsilon_0)$$
    - Apply update using stochastic gradient descent with ρ_o via Equation 3 and 4:
      $$\min_{\theta,\phi} \; -\frac{1}{m} \sum_{j=1}^{m} \mathcal{L}_j(\theta, \phi) + \lambda \mathcal{F}(\phi, A)$$
      //in practice, *Adam* via backpropagation and autodiff

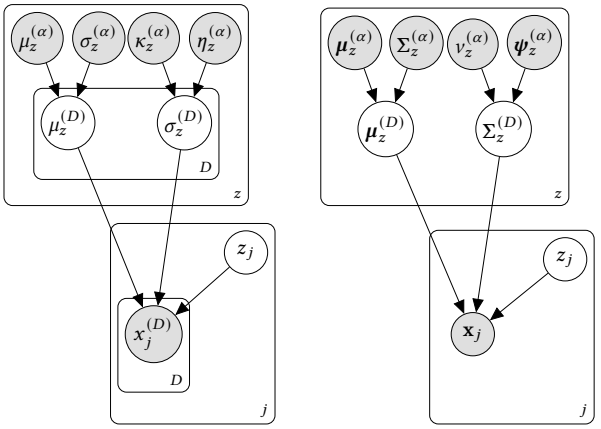
of the stochastic approximation algorithm, which is shown to converge to the counts from the full dataset, under mild conditions and for an appropriate sequence of step sizes [52]. In practice, we found that fixed ρ_t, selected as a hyper-parameter, is enough for successfully estimate $\epsilon(q_\phi(z|x), A)$ using the global counts $N_{z,s}$ and N_s, via Equation 6. The fairness penalty term is a hinge loss:

$$\mathcal{F}(\phi, A) = \max(0, \epsilon(q_\phi(z|x), A) - \epsilon_0) \,, \tag{8}$$

where ϵ_0 is the desired fairness, usually set to 0 to encourage perfect fairness. Finally, $\mathcal{F}(\phi, A)$ is plugged in Equation 4 to jointly optimize θ and ϕ in our fairness-preserving model, which we call the *DF-model*, using the Adam optimization algorithm on the objective via backpropagation and automatic differentiation. The pseudo-code for our fair inference method is provided in Algorithm 1.

3.5 Example: Naïve Bayes Model

We next discuss example models. First, consider the unsupervised naïve Bayes (NB) model, where we ensure fairness for cluster assignments z. Here, x is assumed to contain only categorical observed variables that are conditionally independent given z. The graphical model is provided in Figure 1 (a). Let θ be the mean $\mu_z^{(D)}$ and the standard deviation s.d. $\sigma_z^{(D)}$ for a logistic normal, that are generated from priors $\mathcal{N}(\mu_z^{(\alpha)}, \sigma_z^{(\alpha)})$ and $\Gamma(\kappa_z^{(\alpha)}, \eta_z^{(\alpha)})$, respectively. For simplicity, we choose to use the logistic normal here to reparameterize the model for sampling categorical datapoints, as it does not require annealing, unlike the Gumbel-Softmax. We can generate samples from the logistic normal as $y_z^{(D)} = \mathcal{S}(\mu_z^{(D)} + \sigma_z^{(D)} \mathcal{E})$, where $\mathcal{S}$ is the softmax function and $\mathcal{E} \sim \mathcal{N}(0, 1)$. Plugging $\log p_\theta(x|z) = \sum_z \sum_D z[x^{(D)} \log y_z^{(D)}]$ into Equations 3 and 4 provide the *Vanilla-NB* and *DF-NB* models, respectively.



(a) **Naïve Bayes model** (b) **Gaussian mixture model**

Figure 1: Examples of our settings using directed acyclic graphs: (a) Naïve Bayes (NB), and (b) Gaussian mixture (GMM) models for j individuals with D attributes. Here, z encodes *cluster assignments*. Our fair inference prevents z from reflecting negative stereotypes.

3.6 Example: Gaussian Mixture Model

For the Gaussian mixture model (GMM), continuous observed variables $\mathbf{x}$ are assumed to be generated from multivariate Gaussian distributions with mean vectors $\boldsymbol{\mu}_z^{(D)}$ and covariance matrices $\boldsymbol{\Sigma}_z^{(D)}$, depending on the cluster assignment z (Figure 1 (b)). Let the priors on these parameters be multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_z^{(\alpha)}, \boldsymbol{\Sigma}_z^{(\alpha)})$ and inverse Wishart $\mathcal{W}^{-1}(\nu_z^{(\alpha)}, \boldsymbol{\psi}_z^{(\alpha)})$, respectively. However, $\boldsymbol{\Sigma}_z^{(D)}$ is a positive semi-definite matrix which is generally infeasible to directly maintain in backpropagation-based gradient methods. To address this, we instead learn the real-valued factor $\boldsymbol{C}_z^{(D)}$ of the covariance matrix which is then used to form $\boldsymbol{\Sigma}_z^{(D)} = \boldsymbol{C}_z^{(D)} \boldsymbol{C}_z^{(D)^\intercal} + \mathbf{I}$. We then plug $\log p_\theta(x|z) = \sum_z z[\log \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_z^{(D)}, \boldsymbol{\Sigma}_z^{(D)})]$ into Equation 3 and 4 to achieve *Vanilla-GMM* and *DF-GMM*, respectively.

4 SPECIAL PURPOSE (SP) MODEL

Next, we present a special purpose (SP) graphical model for mitigating societal biases in risk assessments in the criminal justice system. Our methods and results should be taken as an illustration of how our approach is operationally effective rather than as an endorsement of the deployment of this approach to criminal justice systems, which would require further domain research, examination, stakeholder engagement, and vetting.

4.1 Motivation and Objective

AI and ML systems are increasingly integrated into criminal justice for tasks such as risk assessment, owing to their predictive capabilities [4, 17, 51]. However, these systems often perpetuate societal biases present in training data, posing ethical challenges. A pivotal 2016 ProPublica report [2] drew attention to significant biases in

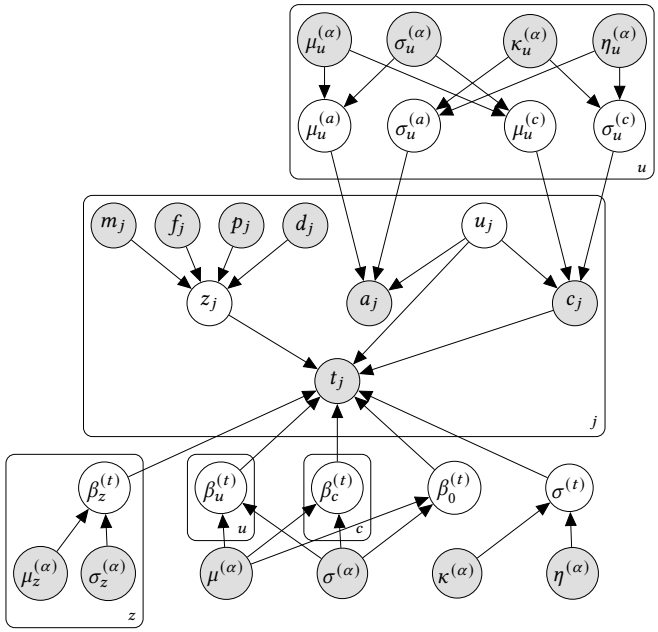


Figure 2: Graph for the special purpose (SP) model. The SP model provides investigative assistance to criminal justice professionals by encoding latent *risk of crime* (z) and latent *systems of oppression* (u). To prevent unfair societal biases, stakeholders can perform inference on the SP model with our fair inference method on any of these latent variables, depending on the context.

the widely-used COMPAS risk assessment tool. For instance, it was more likely to label black individuals as high-risk compared to their white counterparts, despite evidence to the contrary. This paper aims to address these concerns by introducing a probabilistic graphical model that works in conjunction with the COMPAS system. It incorporates COMPAS' scores and other variables to generate an alternative, fairness-focused risk assessment. Our goal is to help judges make more equitable decisions in bail and sentencing in courts already using COMPAS.

4.2 Model Development

We provide the directed acyclic graph for our SP model in Figure 2. In the SP model, we assume that the outcome of the risk assessment mechanism is jail time (t) for each offender j which is potentially influenced by some latent variables (z and u) and the observed degree of charges (c). To design the graphical model, we look into the existing literature for fairness from diverse fields including AI, humanity, law, and social sciences [10]. Although much of the literature in risk assessment views differences in the distributions of risk between protected groups as legitimate phenomena to be accounted for when determining the fairness of a system [54], the intersectionality framework aims for a counterpoint [21]. According to intersectionality theory, the distributions of risk are often influenced by unfair societal processes due to systemic structural disadvantages such as racism, sexism, inter-generational poverty, the school-to-prison pipeline, and the prison-industrial complex [15]. These unfair processes are termed *systems of oppression*. Inspired by intersectionality framework, we desire to encode a fair and equitable estimate of an individual's *risk of crime* (i.e. low, medium, or high) via z and *systems of oppression* via u (here encoded as a binary variable representing the level of impact), which along with the degrees of charges (c), are assumed to affect the jail time (t) outcome. Furthermore, the *risk of crime* (z) is considered to be influenced by the offender's historical record, including juvenile misdemeanors (m), juvenile felony charges (f), previous crime counts (p), and the COMPAS system's predicted decile scores (d). In contrast, *systems of oppression* (u) can lead the structural disadvantages toward the offenders in terms of their age (a), degrees of charges (c), and jail time (t). To reflect these in the graph, we formulate *risk of crime* (z) and *systems of oppression* (u) as downstream and upstream of the corresponding observed variables, respectively. Since jail time (t) is a real-valued observed variable, we formulate it using a regression model with corresponding coefficients $\boldsymbol{\beta}$ for *risk of crime* (z), *systems of oppression* (u), *degree of charges* (c), and an intercept term. We further posit informed hyper-priors on these coefficients to infer identifiable and meaningful latent variables.

Let $x^{(z)} = \{m, f, p, d\}$ and $x^{(u)} = \{a, c, t\}$ be observed variables directly associated with the latent variables z and u. Consider the prior $p_\theta(z|x^{(z)})$ over z be the Gumbel-Softmax whose distribution parameters $\log \pi$ are implemented as neural network outputs, and let prior $p_\theta(u)$ over u be the discrete uniform. Note that model parameters θ are weights and biases of $p_\theta(z|x^{(z)})$ network, all latent means (μ) and standard deviations (σ), and $\boldsymbol{\beta}$. From the DAG in Figure 2, the final objective for the *Vanilla-SP* model is:

$$
\begin{aligned}
&\mathcal{L}_{\text{SP}}(\theta, \phi) \\
&\quad = \mathbb{E}_{q_\phi}[\log p_\theta(x^{(z)}, x^{(u)}, t, z, u)] - \mathbb{E}_{q_\phi}[\log q_\phi(z, u|x^{(z)}, x^{(u)}, t)] \\
&\quad = \mathbb{E}_{q_\phi}[\log p_\theta(t|z, u, x^{(u)}) + \log p_\theta(x^{(u)}|u) + \log p_\theta(z|x^{(z)}) \\
&\qquad\qquad + \log p_\theta(u) + p_\alpha(\theta)] - \mathbb{E}_{q_\phi}[\log q_\phi(z|x^{(z)}, t)] \\
&\qquad\qquad - \mathbb{E}_{q_\phi}[\log q_\phi(u|x^{(u)}, t)]\,,
\end{aligned}
\tag{9}
$$

where $\log p_\theta(t|z, u, x^{(u)}) = \sum_z [\log \mathcal{N}(t; \beta_0^{(t)} + z\beta_z^{(t)} + \beta_u^{(t)} + \beta_c^{(t)}, \sigma^{(t)})]$ and $\log p_\theta(x^{(u)}|u)$ is implemented as log-likelihood of NB model in the previous section. Note that ϕ represents weights and biases of the inference networks $q_\phi(z|x^{(z)}, t)$ and $q_\phi(u|x^{(u)}, t)$.

Our SP model provides flexibility so that stakeholders can perform inference on the model with our fairness intervention approach on either of these latent variables, *risk of crime* (z) or *systems of oppression* (u), depending on the context. In this work, we train the *DF-SP* model using Equation 4 via Equation 9, where $\mathcal{F}$ is implemented in terms of $q_\phi(z|x^{(z)}, t)$, to prevent societal biases in inference on *risk of crime*. The fairness intervention on the *risk of crime* (z) ensures similar distributions of the low, medium, and high risks between intersecting groups. This upholds the philosophy, concordant with intersectionality theory, that different protected groups are not inherently differently prone to criminal behavior, but rather that observed disparities are primarily due to unfair systems of oppression [3].

We presented the SP model here in order to demonstrate our *fair latent variable modeling and inference methodology*. We emphasize

that further investigation and analysis from experts in criminal justice, law and social science would be necessary before considering the deployment of such a model in real systems. After performing such analysis, the eventual goal of this model is that the fairly inferred *risk of crime*, *systems of oppression* and predicted jail time will allow criminal justice professionals to better maintain the right balance between justice, fairness, and public safety. As such, our model represents a step toward a criminal justice system that is more equitable and fair.

5 PRACTICAL CONSIDERATIONS

Discrete latent variable models are known to be prone to the issue of posterior collapse [8, 39, 49, 53], a particular type of local optimum very close to the prior over latents, e.g., all individuals are assigned to same latent class. We found that better initialization of the parameters and smoothing out the functional space help to resolve this issue, which we implemented by using random restarts during the hyperparameter grid search on the development (dev) set and by using batch normalization on the output layer of inference networks, respectively.

However, the above tricks do not resolve the issue in training our SP models effectively. As we optimize a prior network along with model parameters and multiple inference networks, we found that SP models are more prone to the collapsing. Existing methods to avoid local optima such as annealing-based approaches to down-weight the KL term [1, 8] in early iterations of the training did not help. Finally, we were able to address the problem by using a *warm start* initialization procedure for $p_\theta(z|x^{(z)})$ as follows: 1) first pre-train by only maximizing the likelihood $p_\theta(t|z)p_\theta(z|x^{(z)})$, and 2) then fine tune the prior network by optimizing the full $\mathcal{L}_{\text{SP}}$.

6 EXPERIMENTS

We performed all experiments on the COMPAS dataset[1] (protected attributes: *race* and *gender*), the Adult 1994 U.S. census income data[2] (protected attributes: *race*, *gender*, USA vs non-USA *nationality*), and the Heritage Health Prize (HHP) dataset[3] (protected attributes: *age* and *gender*). The COMPAS, Adult and HHP datasets contain data instances for a total of 6.91K, 48.84K and 170.07K individuals, respectively. Our source code is provided in the GitHub.[4]

6.1 Experimental Settings

We validate and compare our models with two baseline models. For a typical baseline model that doesn't take fairness into account, we consider the Gumbel-Softmax (GS) reparameterization-based VAE model for discrete latent variables (GS-VAE) [32]. As there is no existing work that enforces fairness in completely unsupervised setting for discrete latent variables, the work from [43] is presumably the most relevant. They proposed an unsupervised fair VAE model $p_\theta(x|z, A)$ to factor out undesired information from the continuous latent variables z by the marginally independent protected attributes A. We extend this model for discrete latent variables by GS reparameterization and use it as a fair baseline model (GS-VFAE) for our experiments. Note that models referred to as "*vanilla*" throughout this section represent standard models, adhering to our latent variable modeling approach outlined in Equation 3, but they exclude our fairness intervention described in Equation 4.

We split the COMPAS into 60% train, 20% dev, and 20% test sets. For Adult dataset, we used the pre-specified train (32.56K) and test set (16.28K), and held-out 30% from the training data as the development ("dev") set. Finally, we held-out 10% from our larger data HHP as the test set, using the remainder for training, and further held-out 10% from the training data as the dev set. All the models were trained via the Adam optimizer using PyTorch on COMPAS, Adult and HHP datasets for a total of 50, 10 and 5 epochs, respectively. Finally, we performed grid search on the dev set to choose hyper-parameters, e.g., minibatch size, #neurons/hidden layer, learning rate, dropout, activation and random seed, from the same set of hyper-parameter values for all models.

6.2 Evaluation Protocols

To assess model fit, we calculate the average log-likelihood (LL) and average mutual information (MI) of the latent variable z across all observed variables. For clustering analysis, we measure commonly used Calinski-Harabasz (CH) score [9], where a higher score represents a model with better defined clusters, and Davies-Bouldin (DB) score [16], where a lower score represents a model with better separation between the clusters. In our SP model for criminal justice, we evaluate the predictive performance using LL, mean absolute error (MAE), mean squared error (MSE) and regression score (R^2) based on observed "jail time".

We employ several metrics to assess fairness, including ϵ-DF [21] for equitable treatment across intersecting groups and demographic parity (δ-DP) [18] for equal outcomes among protected groups. We also use the p%-Rule [58], which generalizes the 80% rule, and subgroup fairness (γ-SF) [34], which aims to prevent subset targeting, to further ensure fairness. These metrics are adapted for multi-dimensional latent variables, with the worst-case scenario reported as the final metric. For model selection, we start by identifying the best vanilla model based on LL. We then tune our fairness-focused models to optimize both LL and fairness metrics. *Final selections allow a minor LL degradation for improved fairness.*

6.3 Performance for Clustering

We evaluated the models on held-out test data for clustering analysis. For Adult data, models were trained on categorical observed variables like work classes, education levels, occupation types and income≥50K or not, where we aimed to infer z that represents whether an individual is "hard-working" or not. For our NB models, with the prior knowledge on the PDF of a logistic normal distribution, we set priors $\mathcal{N}(2, 1)$ and $\mathcal{N}(-2, 1)$ on μ_z to encode "hard-working" and not "hard-working," respectively, and the same prior $\Gamma(1, 2)$ on σ_z for both cases. Table 1 shows that the Vanilla-NB outperformed all models in clustering performance metrics like MI, CH and DB, while our DF-NB is the fairest model based on ϵ-DF, as well as several other fairness metrics (δ-DP (race) and p%-Rule (race)), with a small cost in performance.

In the HHP dataset, all models were trained on real-valued observations for hospitalized patients such as the estimation of mortality,

[1] https://tinyurl.com/2p8tbda2.
[2] https://archive.ics.uci.edu/ml/datasets/adult.
[3] www.kaggle.com/c/hhp.
[4] https://github.com/rashid-islam/fair_inference.

Table 1: Performance for clustering analysis on categorical variables in Adult dataset. Our DF-NB was the fairest model w.r.t. ϵ-DF along with several other fairness metrics, while the Vanilla-NB performed with highest clustering performances. Higher is better for measures with ↑, while lower is better for measures with ↓.

Models	MI ↑	CH ↑	DB ↓	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (race)	δ-DP ↓ (gender)	δ-DP ↓ (nation)	p%-Rule ↑ (race)	p%-Rule ↑ (gender)	p%-Rule ↑ (nation)
GS-VAE	0.132	1391.860	3.349	0.372	0.005	0.041	**0.005**	0.011	91.452	**98.916**	97.659
GS-VFAE	0.087	850.065	4.286	0.329	**0.003**	0.021	0.007	0.011	95.144	98.378	97.330
Vanilla-NB	**0.137**	**1510.138**	**2.776**	0.739	0.024	0.059	0.100	**0.004**	76.080	63.130	**98.307**
DF-NB	0.097	948.817	3.825	**0.221**	0.005	**0.011**	0.020	0.008	**96.543**	93.600	97.350

Table 2: Performance for clustering analysis on continuous variables in HHP dataset. Our DF-GMM was the fairest model w.r.t. ϵ-DF along with most of the other fairness metrics, while the Vanilla-GMM performed with highest clustering performances. Higher is better for measures with ↑, while lower is better for measures with ↓.

Models	MI ↑	CH ↑	DB ↓	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (age)	δ-DP ↓ (gender)	p%-Rule ↑ (age)	p%-Rule ↑ (gender)
GS-VAE	0.069	913.774	4.221	1.385	0.045	0.335	0.048	32.311	83.524
GS-VFAE	0.061	793.340	4.523	1.123	0.042	0.320	**0.040**	39.108	84.329
Vanilla-GMM	**0.106**	**1445.014**	**2.996**	2.269	0.064	0.416	0.104	16.060	77.261
DF-GMM	0.044	597.302	4.183	**0.275**	**0.021**	**0.078**	0.049	**84.338**	**89.964**

Table 3: Performance for our special purpose model on COMPAS dataset for criminal justice risk assessment. Predictive performances were measured w.r.t. observed "jail time," while fairness were measured w.r.t. z that encodes *risk of crime*. Higher is better for measures with ↑, while lower is better for measures with ↓.

Models	Measured in terms of observed "jail time"				Measured in terms of latent z					
	LL ↑	MAE ↓	MSE ↓	R^2 ↑	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (race)	δ-DP (gender) ↓	p%-Rule ↑ (race)	p%-Rule ↑ (gender)
Vanilla-SP	**-1.301**	**0.578**	**0.757**	**0.274**	1.744	0.035	0.143	0.093	40.844	43.929
DF-SP	-1.358	0.662	0.926	0.112	**1.304**	**0.022**	**0.074**	**0.048**	**41.145**	**51.587**

drug counts, lab counts and so on, where we aimed to group the patients into 3 clusters that may represent short, medium and long lengths of stay in hospital so that we can help stakeholders to properly allocate healthcare resources. In our GMM models, we set informed priors on $\boldsymbol{\mu}_z$ using cluster centers from a k-means clustering method on train data and same prior $\mathcal{W}^{-1}(D+2, \mathbf{I})$ on Σ_z for all clusters. Table 2 shows that our DF-GMM is the fairest model based on almost all fairness metrics (5 out of 6) with a loss in clustering, while the Vanilla-GMM performed as best and worst model with respect to clustering metrics and fairness metrics, respectively.

6.4 Performance for Criminal Risk Assessment

We investigate the performance of our SP models on the COMPAS system for criminal justice. In Table 3, we show detailed results for Vanilla-SP and DF-SP models. Note that we excluded the VAE-based baselines from this experiment since there is no straightforward way to extend them for the SP framework. According to the DAG presented in Figure 1 (c), we can evaluate the overall predictive performance in terms of "jail time." We measured fairness in terms of latent *risk of crime* z since fairness intervention was applied to z. As expected, we found that Vanilla-SP performed better w.r.t predictive performance metrics, but worse w.r.t all fairness metrics. DF-SP mitigated these biases with a little sacrifice in predictive performances. Table 4 shows fairness measures on latent *systems of oppression* built into our society, where Vanilla-SP outperforms DF-SP model. This result is intuitive from the DAG. Since both *risk of crime* and *systems of oppression* can alter the "jail time," improving fairness for one of them can increase disparity in the other.

We also looked into MI for COMPAS's score (MI = 0.079), inferred *risk of crime* by Vanilla-SP (MI = 0.072) and by DF-SP (MI = 0.048) with actually-occurred recidivism over a two-year period. Since COMPAS is a supervised learning-based system, it is expected that COMPAS shows a higher MI with the actual label, while our unsupervised Vanilla-SP and DF-SP performed with the comparable MI metric. Finally, in Figure 3, we visualized generated average "jail time" from our models in terms of all intersecting groups. We observe that Vanilla-SP reflected discrimination by predicting more "jail time" against a particular group, while DF-SP distributes similar "jail time" on average for all groups.

In order to achieve fairness in the risk assessments for criminal justice, the SP model uses an intersectional approach, taking into

Table 4: Fairness metrics measured on latent u which encodes *systems of oppression* against individuals.

Models	ϵ-DF ↓	γ-SF ↓	δ-DP ↓ (race)	δ-DP ↓ (gender)	p%-Rule ↑ (race)	p%-Rule ↑ (gender)
Vanilla-SP	**0.085**	**0.005**	0.027	**0.003**	94.133	**97.364**
DF-SP	0.100	0.006	**0.015**	0.019	**96.873**	96.063

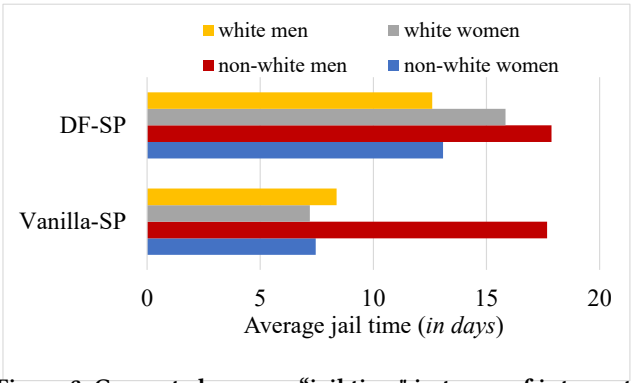


Figure 3: Generated average "jail time" in terms of intersecting protected groups for COMPAS data.

account not just individual characteristics but also the *systems of oppression* that may contribute to disparities in criminal behavior. The goal of this model is to strike a balance between justice, fairness, and public safety, by providing a fair inferred *risk of crime* and predicted "jail time." However, our method of achieving parity on "jail time" may have unintended consequences. In particular, increasing jail time for less disadvantaged groups may not be ideal from the perspective of criminal justice reform. This trade-off highlights the complexity of balancing fairness and justice in the criminal justice system and underscores the need for ongoing monitoring and evaluation of these models.

7 RELATED WORK

Our fairness intervention technique in this work is inspired by intersectionality, the core theoretical framework underlying the third-wave feminist movement [13, 15]. [21] proposed differential fairness which implements the principles of intersectionality with additional beneficial properties from a societal perspective regarding the law, privacy, and economics. While most of the fairness notions are defined for binary outcome and binary protected attribute, differential fairness conversely handles multiple outcomes and multiple protected attributes, simultaneously.

Much of the prior work that enforces fairness in variational inference [14, 42, 43, 55] using unsupervised probabilistic graphical models, e.g., VAE [40, 50], β-VAE [28], and FactorVAE [37], aim to learn fair representations of data using continuous latent variables for downstream classification tasks. [43] also proposed a semi-supervised VAE model that encourages statistical independence between continuous latent variables and protected attributes using a maximum mean discrepancy (MMD) [24] penalty. Through the lens of representation learning, there are other recent advances in building fair classifiers using fair representations. [59] proposed a neural network based supervised clustering model for learning fair representations that maps each data instance to a cluster, while the model ensures that each cluster gets assigned approximately equal proportions of data from each protected group. While this approach cannot leverage the representational power of a distributed representation, other work [19, 45, 60, 61] addressed this by developing joint framework using an autoencoder network to learn distributed representations along with an adversary network to penalize when protected attributes are predictable from representations and a classifier network to preserve utility-related information in the representations.

8 CONCLUSION AND FUTURE WORK

We have proposed a fair stochastic inference technique for unsupervised learning using probabilistic graphical models with discrete latent variables. Our method incorporates the principles of intersectionality, a comprehensive perspective on fairness, into the variational distribution through a fairness penalty and a stochastic approximation approach. We have also presented a special-purpose model for mitigating societal biases from risk assessments in criminal justice. Our empirical results show the benefits of our approach in sensitive tasks such as inferring merits or risks for individuals. Before deployment of our special-purpose model could be contemplated, feedback from criminologists and criminal justice stakeholders must be included in its design, which we plan to investigate in future.

ACKNOWLEDGMENTS

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant No.'s IIS1927486; IIS1850023; IIS2046381. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. Saurous, and K. Murphy. 2018. Fixing a broken ELBO. In *ICML*. PMLR, 159–168.
[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* 23 (2016).
[3] S. Barocas and A. Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[4] R. Berk and D. Berk. 2019. *Machine learning risk assessments in criminal justice settings.* Springer.
[5] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).
[6] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* (2018).
[7] T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, and A. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS* 29 (2016).
[8] S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio. 2016. Generating Sentences from a Continuous Space. In *CoNLL*. 10–21.
[9] T. Caliński and J. Harabasz. 1974. A dendrite method for cluster analysis. *Commun. Stat.* 3, 1 (1974), 1–27.
[10] A. Campolo, M. Sanfilippo, M. Whittaker, and K. Crawford. 2017. AI Now 2017 Report. *AI Now* (2017).

[11] L. Carroll and M. Gonzalez. 2014. Out of place: Racial stereotypes and the ecology of frisks and searches following traffic stops. *J. Res. Crime Delinq.* 51, 5 (2014), 559–584.
[12] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. 2017. Fair clustering through fairlets. *NeurIPS* 30 (2017).
[13] P. Collins. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* Routledge.
[14] E. Creager, D. Madras, J. Jacobsen, et al. 2019. Flexibly fair representation learning by disentanglement. In *ICML*. PMLR, 1436–1445.
[15] K. Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.* (1989), 139–167.
[16] D. Davies and D. Bouldin. 1979. A cluster separation measure. *TPAMI* (1979), 224–227.
[17] B. Dupont, Y. Stevens, H. Westermann, and M. Joyce. 2018. Artificial Intelligence in the Context of Crime and Criminal justice. *SSRN 3857367* (2018).
[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through awareness. In *Theo. Comp. Sci.* ACM, 214–226.
[19] H. Edwards and A. Storkey. 2016. Censoring representations with an adversary. In *ICML*.
[20] J. Foulds, R. Islam, K. Keya, and S. Pan. 2020. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *SDM*. SIAM, 424–432.
[21] J. Foulds, R. Islam, K. Keya, and S. Pan. 2020. An intersectional definition of fairness. In *ICDE*. IEEE, 1918–1921.
[22] J. Foulds and S. Pan. 2020. Are Parity-Based Notions of AI Fairness Desirable? *TCDE* 43, 4 (2020), 51–73.
[23] A. Gelman, J. Fagan, and A. Kiss. 2007. An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *J. Am. Stat. Assoc.* 102, 479 (2007), 813–823.
[24] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. 2008. A Kernel Method for the Two-Sample Problem. *JMLR* 1 (2008), 1–10.
[25] M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. *NeurIPS* 29 (2016).
[26] D. Harris. 1996. Driving while black and all other traffic offenses: The Supreme Court and pretextual traffic stops. *J. Crim. Law Criminol.* 87 (1996), 544.
[27] D. Harris. 1999. The stories, the statistics, and the law: Why driving while black matters. *Minn. Law Rev.* 84 (1999), 265.
[28] I. Higgins, L. Matthey, A. Pal, et al. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.
[29] M. Hoffman, D. Blei, C. Wang, and J. Paisley. 2013. Stochastic variational inference. *JMLR* (2013).
[30] R. Islam, H. Chen, and Y. Cai. 2024. Fairness without Demographics through Shared Latent Space-Based Debiasing. In *AAAI*, Vol. 38. 12717–12725.
[31] R. Islam, K. Keya, Z. Zeng, S. Pan, and J. Foulds. 2021. Debiasing career recommendations with neural fair collaborative filtering. In *WWW*. 3779–3790.
[32] E. Jang, S. Gu, and B. Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
[33] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. 1999. An introduction to variational methods for graphical models. *ML* 37, 2 (1999), 183–233.
[34] M. Kearns, S. Neel, A. Roth, and Z. Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*. PMLR, 2564–2572.
[35] K.r Keya, R. Islam, S. Pan, I. Stockwell, and J. Foulds. 2021. Equitable allocation of healthcare resources with fair survival models. In *SDM*. SIAM, 190–198.
[36] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*. PMLR, 2207–2217.
[37] H. Kim and A. Mnih. 2018. Disentangling by factorising. In *ICML*. PMLR, 2649–2658.
[38] D. Kingma, S. Mohamed, D. Jimenez, and M. Welling. 2014. Semi-supervised learning with deep generative models. *NeurIPS* 27 (2014).
[39] D. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. 2016. Improved variational inference with inverse autoregressive flow. *NeurIPS* 29 (2016).
[40] D. Kingma and M. Welling. 2014. Auto-Encoding Variational Bayes. In *ICML*.
[41] M. Kusner, J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual fairness. In *NeurIPS*. 4069–4079.
[42] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem. 2019. On the Fairness of Disentangled Representations. In *NeurIPS*, Vol. 32. 14611–14624.
[43] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. 2016. The Variational Fair Autoencoder. In *ICLR*.
[44] C. Maddison, A. Mnih, and Y. Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).
[45] D. Madras, E. Creager, T. Pitassi, and R. Zemel. 2018. Learning adversarially fair and transferable representations. In *ICML*. PMLR, 3384–3393.
[46] C. Munoz, M. Smith, and D. Patil. 2016. Big data: A report on algorithmic systems. *EOP* (2016).
[47] S. Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. NYU.
[48] C. O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.
[49] T. Pelsmaeker and W. Aziz. 2020. Effective Estimation of Deep Generative Language Models. In *ACL*. 7220–7236.
[50] D. Rezende, S. Mohamed, and D. Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*. PMLR, 1278–1286.
[51] C. Rigano. 2019. Using artificial intelligence to address criminal justice needs. *NIJ* 280 (2019), 1–10.
[52] H. Robbins and S. Monro. 1951. A stochastic approximation method. *Ann. Math. Stat.* (1951), 400–407.
[53] S. Semeniuta, A. Severyn, and E. Barth. 2017. A Hybrid Convolutional Variational Autoencoder for Text Generation. In *EMNLP*. 627–637.
[54] C. Simoiu, S. Corbett-Davies, S. Goel, et al. 2017. The problem of infra-marginality in outcome tests for discrimination. *AOAS* 11, 3 (2017), 1193–1216.
[55] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon. 2019. Learning controllable fair representations. In *AISTATS*. PMLR, 2164–2173.
[56] A. Srivastava and C. Sutton. 2017. Autoencoding Variational Inference for Topic Models. In *ICLR*.
[57] P. Warren, D. Tomaskovic-Devey, W. Smith, M. Zingraff, and M. Mason. 2006. Driving while black: Bias processes and racial disparity in police stops. *Criminology* 44, 3 (2006), 709–738.
[58] M. Zafar, I. Valera, M. Gomez, and K. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*. 1171–1180.
[59] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning fair representations. In *ICML*. PMLR, 325–333.
[60] H. Zhao, A. Coston, T. Adel, and G. Gordon. 2020. Conditional learning of fair representations. In *ICLR*.
[61] H. Zhao and G. Gordon. 2019. Inherent tradeoffs in learning fair representations. *NeurIPS* 32 (2019).

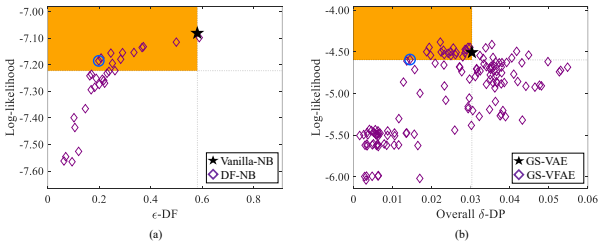


Figure 4: Selection strategy for fair models (*purple diamonds*): (a) DF-NB and (b) GS-VFAE. The best typical model *black asterisk*: (a) Vanilla-NB and (b) GS-VAE, based on log-likelihood. The best fair model (*blue circle*) is selected from the *orange area* that satisfies our pre-defined rule, according to the respective fairness metric.

Our fair models consider both ELBO and fairness, potentially affecting predictive performance. As illustrated in Figure 4 (a), we selected the best typical models (GS-VAE and Vanilla-NB) based on log-likelihood (LL) using grid search over hyper-parameters (black asterisk). The fair model, DF-NB, employed the same hyper-parameters as the best Vanilla-NB, with grid search conducted only for the fairness trade-off parameter λ (purple diamonds). For GS-VFAE, which lacks an explicit trade-off parameter, a full grid search was conducted (purple diamonds). Fair models were chosen based on the best fairness metrics, allowing a 2% LL degradation from the best typical model (orange area). We used an overall δ-DP metric for GS-VFAE, averaging δ-DP for each protected attribute. When deploying these methods in practice, the slack tolerance can be adjusted based on stakeholders' preferences.