entropy

MDPI

Article

Constrained Reweighting of Distributions: An Optimal Transport Approach

Abhisek Chakraborty *, Anirban Bhattacharya and Debdeep Pati

Department of Statistics, Texas A&M University, College Station, TX 77843, USA; anirbanb@stat.tamu.edu (A.B.); debdeep@stat.tamu.edu (D.P.)

* Correspondence: abhisek_chakraborty@tamu.edu

Abstract: We commonly encounter the problem of identifying an optimally weight-adjusted version of the empirical distribution of observed data, adhering to predefined constraints on the weights. Such constraints often manifest as restrictions on the moments, tail behavior, shapes, number of modes, etc., of the resulting weight-adjusted empirical distribution. In this article, we substantially enhance the flexibility of such a methodology by introducing a nonparametrically imbued distributional constraint on the weights and developing a general framework leveraging the maximum entropy principle and tools from optimal transport. The key idea is to ensure that the maximum entropy weight-adjusted empirical distribution of the observed data is close to a pre-specified probability distribution in terms of the optimal transport metric, while allowing for subtle departures. The proposed scheme for the re-weighting of observations subject to constraints is reminiscent of the empirical likelihood and related ideas, but offers greater flexibility in applications where parametric distribution-guided constraints arise naturally. The versatility of the proposed framework is demonstrated in the context of three disparate applications where data re-weighting is warranted to satisfy side constraints on the optimization problem at the heart of the statistical task—namely, portfolio allocation, semi-parametric inference for complex surveys, and ensuring algorithmic fairness in machine learning algorithms.

Keywords: complex surveys; demographic parity; entropy; optimal transport; portfolio allocation

Citation: Chakraborty, A.; Bhattacharya, A.; Pati, D. Constrained Reweighting of Distributions: An Optimal Transport Approach. *Entropy* **2024**, *26*, 249. https://doi.org/10.3390/e26030249

Academic Editor: Dawn E. Holmes

Received: 15 January 2024
Revised: 6 March 2024
Accepted: 7 March 2024
Published: 11 March 2024

1. Introduction

The maximum entropy principle [1,2] states that in situations characterized by uncertainty and limited prior knowledge-guided constraints, the optimal choice among all feasible probability distributions is the probability distribution that is the least informative or most uniformly spread. This idea is at the heart of numerous statistical tasks and has permeated into every corner of modern machine learning research. Prominent instances of such constrained entropy maximization include applications in image reconstruction [3], ill-posed inverse problems [4], portfolio optimization [5], generalized methods of moment models [6], natural language processing [7], network analysis [8], and reinforcement learning [9], to name a few. We refer the readers to Cover and Thomas [10], Kardar [11] for book-length reviews.

For maximum entropy inference, the specified constraints imposed on the probability distributions frequently manifest as constraints pertaining to moments [6], tail characteristics [12], distributional shapes [13], modal counts, and similar properties. In many cases, however, constructing constraints with the desired level of flexibility is challenging, if not unfeasible—refer to Sections 3 and 5 for specific examples in the context of inference in complex surveys and moment conditions based on portfolio optimization, respectively. On a related note, a recent article [14] introduced a flexible framework for the introduction of more elaborate constraints on probability distributions in the context of conducting robust Bayesian inference.

In this article, we offer a novel solution to this problem via introducing a probability distribution-guided constrained entropy maximization framework that not only offers

versatility but also enhances the interpretability of the inferential output. The main concept revolves around ensuring that a weight-adjusted empirical distribution of the observed data closely aligns with a predetermined family of probability distributions, measured through a statistical distance [15]. Importantly, the family of probability distributions is potentially continuous, but any weighted-adjusted empirical distribution of the observed data is discrete. This eliminates the possibility of adopting many common statistical discrepancies, e.g., Kullback–Leibler, total variation, or Hellinger's distance, to place the probability distribution-guided constraints. In practice, we need to exercise extreme care to ensure that our choice is tailored to the application of interest. For homogeneity of exposition across all scenarios in this article, we consider the Wasserstein metric [16,17].

The idea of data re-weighting is, of course, not new. Ref. [18] suggested elevating the likelihood of individual observations using data-driven weights, to conduct robust inference under mild model misspecification. Ref. [19] proposed a data re-weighting scheme to align the data with a different target distribution, enabling inference under covariate shifts. Other compelling ideas involving re-weighting can be found in fair learning [20], natural language processing [21], variational tempering [22], etc. Complementing the existing literature, we propose a versatile data re-weighting framework, borrowing from the maximum entropy principle and optimal transport, that is useful in a multitude of statistical tasks.

The rest of the paper is organized as follows. The general framework of the proposed probability distribution-guided constrained entropy maximization is introduced in Section 2. Sections 3–5 present applications of our methodology in the context of semi-parametric inference in complex surveys, in ensuring demographic parity in machine learning algorithms, and entropy-based portfolio optimization, respectively. Finally, we conclude with a discussion.

2. General Framework

Let $[a]$ denote the set of integers $\{1, \ldots, a\}$. Let Ω denote the set of all possible discrete distributions ω with atoms $\mathbf{s} = (s_1, \ldots, s_m)^{\mathrm{T}}$. The entropy of the discrete probability distribution $\sum_{i=1}^{m} w_i \delta_{s_i}(\cdot)$ is defined by

$$\mathrm{H}_m(\mathbf{w}) = -\sum_{i=1}^{m} w_i \log w_i,$$

where δ is Dirac's delta function. The entropy $\mathrm{H}_m(\mathbf{w})$ is a measure of randomness, which is maximized at the discrete uniform distribution with $w_i = 1/m$ for all i. In many statistical tasks, the core challenge constitutes optimizing a functional $\mathcal{F} : \Omega \to \Omega'$ with respect to ω subject to a constraint $\omega \in \Omega_0 (\subset \Omega)$. A simple example is when $\mathbf{s} = (s_1, \ldots, s_m)^{\mathrm{T}}$ is the observed sample itself. Then, the set Ω is simply characterized by the class of weighted empirical distributions of the observed data,

$$\Omega = \Big\{ \omega = \sum_{j=1}^{m} w_j \, \delta_{s_j}(\cdot) : \sum_{j=1}^{m} w_j = 1, \; w_j \geq 0, \; j \in [m] \Big\}.$$

In the following, we shall provide more general examples where the constraint set Ω_0 can be identified with a subset of an $(m-1)$-dimensional probability simplex $\mathcal{S}_{m-1} = \{\mathbf{w} : \sum_{i=1}^{m} w_i = 1, \; w_i > 0, \; i \in [m]\}$, for some $m \in \{1, 2, \ldots\}$.

Given $\mathbf{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$, parametric inference constitutes approximating the empirical distribution $(1/n) \sum_{i=1}^{n} \delta_{s_i}(\cdot)$ via a parametric family of distributions $\{f_\theta : \theta \in \Theta\}$, and learning the parameter θ from data. Such procedures often fall prey to model misspecification [23], leading to untrustworthy inferences. To avoid complete model specification, a popular class of semi-parametric approaches [24] operate under a milder assumption that the weight-adjusted empirical distribution $\sum_{i=1}^{n} w_i \delta_{s_i}(\cdot)$ satisfies moment restrictions of the form $\sum_{i=1}^{n} w_i \, g(s_i, \theta) = 0$, where g is a vector of known functions on $\mathbf{R}^d \times \Theta$. In numerous

instances, achieving such moment-based constraints with the intended degree of flexibility proves to be arduous, if not practically impossible—we elaborate on this more in the sequel. To that end, in this article, we offer a middle ground between the fully parametric and semi-parametric moment condition models that allows for flexible modeling assumptions while enjoying coherent interpretability similar to parametric inference. We propose to operate under a restriction of the form $\mathrm{D}(\sum_{i=1}^{n} w_i \delta_{s_i}(\cdot), f_\theta) \leq \varepsilon$, where D is a statistical distance and ε is a user-defined hyperparameter. Our goal is to infer θ while allowing for mild deviations from the parametric model f_θ, and ε measures the maximum allowable discrepancy.

Inference under moment condition models often involves computing the maximum entropy weight-adjusted empirical distribution of $(s_1, \ldots, s_n)^{\mathrm{T}}$ that satisfies some pre-specified moment conditions [6,25]. That is, for every $\theta \in \Theta$, we calculate $\sum_{i=1}^{n} w_i^\star(\theta)\, \delta_{s_i}(\cdot)$, where $\mathbf{w}^\star(\theta) = \arg\max_{\mathbf{w} \in \mathcal{S}_{n-1}} \mathrm{H}_n(\mathbf{w})$, subject to $\sum_{i=1}^{n} w_i g(s_i, \theta) = 0$. Under the proposed framework, we too appeal to the maximum entropy principle and compute the maximum entropy weight-adjusted empirical distribution of $\mathbf{s}$ that satisfies the parametric distribution-guided constraint. That is, for every $\theta \in \Theta$, we calculate $\sum_{i=1}^{n} w_i^\star(\theta)\, \delta_{s_i}(\cdot)$, where

$$\mathbf{w}^\star(\theta) = \underset{\mathbf{w} \in \mathcal{S}_{n-1}}{\arg\max}\, \mathrm{H}_n(\mathbf{w}) \text{ subject to } \mathrm{D}\left(f_\theta,\ \sum_{i=1}^{n} w_i\, \delta_{s_i}(\cdot) \right) \leq \varepsilon, \tag{1}$$

where D is a statistical distance, and ε is a user-defined parameter. In the ensuing applications in this article, we often solve the dual optimization problem for operational ease. In this case, for each $\theta \in \Theta$ and $\lambda \geq 0$, we calculate $\sum_{i=1}^{n} w_i^\star(\theta)\, \delta_{s_i}(\cdot)$ such that

$$\mathbf{w}^\star(\theta) = \underset{\mathbf{w} \in \mathcal{S}_{n-1}}{\arg\max} \left[\mathrm{H}_n(\mathbf{w}) - \lambda\, \mathrm{D}\left(f_\theta,\ \sum_{i=1}^{n} w_i\, \delta_{s_i}(\cdot) \right) \right]. \tag{2}$$

The parameter λ controls the extent of departure from the guiding parametric distribution.

One pivotal aspect yet to be addressed within the proposed framework is that a weight-adjusted empirical distribution is discrete, but, in the context of a specific problem, the guiding distribution f_θ is potentially continuous. For instance, in Section 5, in the context of entropy-based portfolio allocation, f_θ takes the form of a skew normal distribution [26]. This precludes the utilization of several standard statistical distances, such as total variation, Hellinger's distance, χ^2 distance, etc., for the implementation of the distance-based constraint. In this article, due to its versatility, we employ the Wasserstein metric [17] with the L_2 cost as the distance measure D. To that end, we briefly recall some relevant facts about the 2-Wasserstein metric. The Wasserstein space $\mathbf{P}_2(\mathbf{R}^d)$ is defined as the set of probability measures μ with finite moment of order 2, i.e., $\{\mu : \int_{\mathbf{R}^d} \|x\|^2 d\mu(x) < \infty\}$, where $\|\cdot\|$ is the Euclidean norm on $\mathbf{R}^d$.

Definition 1. *For $p_0, p_1 \in \mathbf{P}_2(\mathbf{R}^d)$, let $\pi(p_0, p_1) \subset \mathbf{P}_2(\mathbf{R}^d \times \mathbf{R}^d)$ denote the subset of joint probability measures (or* couplings*) ν on $\mathbf{R}^d \times \mathbf{R}^d$ with marginal distributions p_0 and p_1, respectively. Then, the 2-Wasserstein distance W_2 between p_0 and p_1 is defined as $W_2^2(p_0, p_1) = \inf_{\nu \in \pi(p_0, p_1)} \int_{\mathbf{R}^d \times \mathbf{R}^d} \|y_0 - y_1\|^2\, d\nu(y_0, y_1)$.*

Importantly, if both $p_0, p_1 \in \mathbf{P}_2(\mathbf{R})$ with quantile functions F_0^{-1}, F_1^{-1}, we have a highly tractable expression [27]: $W_2^2(p_0, p_1) = \int_{[0,1]} \left[F_0^{-1}(q) - F_0^{-1}(q) \right]^2 dq$. This is heavily utilized in the subsequent sections. With this, we have all the essential components to delve into the specific applications of interest.

The central idea of re-weighting observations subject to constraints is reminiscent of the empirical likelihood framework (EL; [28–31]) for conducting non-parametric inference. An EL approximates the underlying distribution with a discrete distribution supported at the observed data points and obtains the induced maximum likelihood of the parameter of interest defined through constraints, by effectively profiling out the nuisance parameters. Qin and Lawless [32] hugely expanded the scope of EL by integrating it with estimating equa-

tions. EL has also been adapted to the dependent data setup; see Nordman and Lahiri [33] for a detailed review in the context of time series data. Operationally, EL-based methods enjoy great computational simplicity. The computation of the EL at $\theta \in \Theta$ amounts to solving the convex optimization problem $\{\max_w \sum_{i=1}^n \log w_i \mid w_i > 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i g(s_i, \theta) = 0\}$. On the theoretical side, [34] demonstrated that the EL estimator exhibits desirable higher-order asymptotic properties in a well-specified setup. Newey and Smith [34] also showed that the EL estimator may not be $\sqrt{n}$-convergent when $g(\cdot)$ is unbounded. On the other hand, Schennach [35] showed that the exponentially tilted empirical likelihood (ETEL) attains the same asymptotic bias and variance as EL, as well as retaining $\sqrt{n}$-convergence under model misspecification. ETEL minimizes the Kullback–Leibler divergence of this discrete distribution with the empirical distribution of the observed data subject to satisfying the estimating equation. Operationally, ETEL-based methods continue to enjoy the same computational simplicity as EL, since its computation at $\theta \in \Theta$ amounts to solving the convex optimization problem $\{\max_w \sum_{i=1}^n -w_i \log w_i \mid w_i > 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i g(s_i, \theta) = 0\}$.

The direct application of the ETEL optimization routine to our setup is challenging as the moment conditions describing the parameters of general parametric models, especially those beyond exponential families, can be quite cumbersome or even unavailable. Instead, our approach proceeds by constraining a weighted empirical distribution of the observed data $\sum_{i=1}^n w_i \delta_{s_i}$ to be close to the parametric model F_θ with respect to a statistical metric. The key advantage of the proposed framework lies in the greater flexibility that it offers compared to existing EL or ETEL procedures based on moment conditions. Let us explain this with the application of the proposed framework to the portfolio optimization problem. Volatility Feedback Theory [36] posits that market volatility can influence subsequent returns. Specifically, it suggests that periods of high volatility can lead to skewed returns, where extreme price movements are more likely to occur. This theory underscores the idea that heightened volatility tends to be associated with increased risk and uncertainty in financial markets, potentially resulting in non-normal return distributions characterized by asymmetry and fat tails. Skewed returns indicate that the probability of extreme events, either positive and negative, is higher than what would be expected under a normal distribution. Consequently, skew normal distributions are routinely utilized to model observed returns [37,38]. This is an example where a distributional constraint arises naturally, explaining the added utility of the proposed framework.

The added flexibility potentially comes at a cost. The constraint of the form $\mathrm{D}(f_\theta, \sum_{i=1}^n w_i \, \delta_{s_i}(\cdot)) \leq \varepsilon$ is non-linear in the weights, posing a greater computational challenge in computing the ETEL. However, we find that augmented Lagrangian methods [39,40] and conic solvers [41] via the R interface [42] of constrained non-linear optimization solvers (for example, NLopt (Johnson [43]) and CVX (Grant and Boyd [44])) serve our purpose in the class of problems that we consider.

On the theoretical front, it is indeed interesting to study the asymptotic properties of the ETEL estimator under our distributional constraint and compare it to the asymptotic properties of the ETEL estimators with moment constraints that were eluded to earlier. Such study, however, will come with unique challenges due to the potentially non-convex nature of the set $\{(w_1, \ldots, w_m)^{\mathrm{T}} : \mathrm{D}(f_\theta, \sum_{i=1}^n w_i \, \delta_{s_i}(\cdot)) \leq \varepsilon, \sum_{j=1}^m w_j = 1, \, w_j \geq 0, \, j \in [m]\}$. In particular, the asymptotic bias and higher-order variance of the ETEL estimator subject to moment-based constraints are obtained via obtaining the stochastic expansions of the ETEL estimator of the parameter of interest and the Lagrange multiplier [34]. These stochastic expansions are in turn obtained by first studying the consistency and asymptotic normality of the estimator as prerequisites. To develop similar large-sample properties for the ETEL estimator subject to distributional constraints, the existing techniques are not directly useful. We need to devise appropriate regularity conditions to develop the required stochastic expansions from scratch. Thus, it is well beyond the scope of the current article and presents an opportunity for future investigations.

An instance of application of the proposed framework emerges within the realm of semi-parametric inference in complex survey data [45,46]. In survey sampling, we wish to infer about a collection of features of a finite population $\mathcal{P} := \{X_i,\ i \in [N]\}$. We are provided with a non-representative sample $(x_1, \ldots, x_n)$ obtained from $\mathcal{P}$ via a complex survey, and the corresponding survey weights $\pi = (\pi_1, \ldots, \pi_n)$, $0 < \pi_i < \infty$. In the general framework, this task involves finding the optimal $\omega \in \Omega_0 \subset \Omega$ such that

$$\Omega = \Big\{\omega = \sum_{j=1}^{n} w_j \delta_{s_i}(\cdot) : \sum_{j=1}^{n} w_j = 1, w_j \geq 0,\ j \in [n]\Big\},$$

where $(s_1, \ldots, s_n) = (x_1^\star, \ldots, x_n^\star)$ is an i.i.d. pseudo-sample of size n obtained from the complex survey sample $(x_1, \ldots, x_n)$, via weighted finite population Bayesian bootstrap [47–49] to adjust for the survey weights, and $m = n$. The restriction Ω_0 is dictated by the parametric model that the analyst posits on finite population $\mathcal{P}$ to infer about the features of interest in the finite population.

The next application in this article deals with the issue of ensuring demographic parity [50,51] in machine learning algorithms. Suppose that we have data $(x_i, y_i, a_i) \in \mathcal{X} \times \mathcal{Y} \times \{S, T\}$ for n individuals on covariate $x \in \mathbf{R}^p$, continuous response $y \in \mathbf{R}$, and *protected/sensitive* attribute A with labels $\{S, T\}$. For the sake of simplicity in exposition, we further assume that $a_i = S$, $i \in [n_S]$, $a_i = T$, $i \in [n] \setminus [n_T]$ and $n = n_S + n_T$. The goal is to learn a predictive rule, $h : \mathcal{X} \times \{S, T\} \to \mathcal{Y}$, that satisfies the specific notion of demographic parity. Refer to Section 4 for details. We shall see that this task involves finding the optimal $\omega \in \Omega_0 \subset \Omega$ such that

$$\Omega = \Big\{\omega = \sum_{j=1}^{n_T} w_{n_S+j}\, \delta_{s_j}(\cdot) : \sum_{j=1}^{n_T} w_{n_S+j} = 1, w_{n_S+j} \geq 0,\ j \in [n_T]\Big\},$$

where $s_j = -L(\theta_{(T)} \mid x_j)$, $j \in [n] \setminus [n_S]$ is the negative of the loss function utilized to learn the predictive rule h for individuals with $a = T$, and $\theta_{(T)}$ are the associated parameters. The optimality of ω and restriction Ω_0 are determined by the notion of demographic parity utilized.

An application of a slightly modified version of the general framework is identified in portfolio allocation problems [5,52,53], where the goal is to identify the optimal atoms of the discrete distributions, rather than the weights assigned to the atoms. This task translates to finding the optimal $\omega \in \Omega_0 \subset \Omega$ such that

$$\Omega = \Big\{\omega = \frac{1}{n}\sum_{i=1}^{n} \delta_{s_i}(\cdot) : \sum_{j=1}^{d} w_j = 1, w_j \geq 0,\ j \in [d]\Big\},$$

where $s_i = \sum_{j=1}^{d} w_j r_{i,j}$, $i \in [n]$; refer to Section 5 for details. The optimality criterion and the restriction Ω_0 are driven by the fund manager's portfolio allocation objectives and the assumed model for the return distribution, respectively.

3. Semi-Parametric Inference in Complex Surveys

Survey data [45,46] commonly arise from complex sampling methods such as stratification and multistage sampling, wherein individuals in the finite population have unequal probabilities of inclusion in the sample. Prominent instances of extensive surveys implementing these methodologies include the National Health and Nutrition Examination Surveys (NHANES), the British Household Panel Survey (BHPS), the Household Income and Labour Dynamics in Australia (HILDA) survey, etc. In complex surveys, the survey sample lacks representativeness, since the individuals with varying demographic characteristics in the finite population of interest have varying probabilities of selection in the sample. Consequently, traditional methods of inference and estimation result in bias and the poor coverage of estimators.

A prevalent approach to addressing this challenge entails carefully exploiting the sampling weights available with complex survey datasets. These weights could be used to rectify the biases introduced by the unequal probability sampling and enable us to create pseudo-equal probability samples from the population. If a survey participant falls within a demographic group with a low probability of selection or response, their weight is increased accordingly. Commonly, the available information only includes the survey dataset and the associated sampling weights for each unit in the sample. That is, there is limited or no information available about the complex sampling methodology or the precise technique employed to derive these weights. This situation presents a compelling inferential challenge, which we shall delve into further in the following discussion.

Assume that we have a finite population $\mathcal{P} := \{X_i,\ i \in [N]\}$, and we wish to infer about a collection of features of $\mathcal{P}$. We are provided with a non-representative sample $x = (x_1, \ldots, x_n)$ obtained via a complex survey design, and the corresponding survey weights $\pi = (\pi_1, \ldots, \pi_n)$, $0 < \pi_i < \infty$. It is assumed that the weights have been designed so that π_i is inversely proportional to the likelihood that the survey design selects an observation with the same demographic characteristics as observation x_i. That is, observations with a lower probability of being selected than they would have under a simple random sampling approach are assigned greater weight than they would receive in a simple random sampling scenario. Conversely, observations with a higher probability of selection receive lower weights than they would in a simple random sampling setup. The π_is are scaled to ensure that $\sum_{i=1}^{n} \pi_i = n$.

3.1. Related Works

Pseudo-maximum likelihood (PMLE)-based approaches are very popular with regard to conducting parametric inference with complex survey data, where we posit a parametric model f_θ to model $\mathcal{P}$ and θ encodes the features of interest of $\mathcal{P}$. The pseudo-loglikelihood of θ takes the form $\mathcal{L}(\theta) = \sum_{i=1}^{n} \pi_i \log f_\theta(x_i)$ [45,54]. The pseudo-likelihood estimate of $\hat{\theta}_{\text{PMLE}}$ satisfies the first-order condition $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \sum_{i=1}^{n} \pi_i \frac{\partial}{\partial \theta} \log f_\theta(x_i)$. Under a certain regularity condition [23],

$$\sqrt{n}(\hat{\theta}_{\text{PMLE}} - \theta_0) \xrightarrow{\text{d}} \text{N}(0, H_\pi^{-1} V_\pi H_\pi^{-1}),$$

where θ_0 is the true value of θ, and H_π and V_π are estimated by

$$\hat{H}_\pi = \frac{1}{n} \sum_{i=1}^{n} \pi_i \frac{\partial^2}{\partial \theta \partial \theta^T} \log f_\theta(x_i)\big|_{\theta=\hat{\theta}_{\text{PMLE}}},$$
$$\hat{V}_\pi = \frac{1}{n} \sum_{i=1}^{n} \pi_i \frac{\partial \log f_\theta(x_i)}{\partial \theta} \frac{\partial \log f_\theta(x_i)}{\partial \theta^T}\Big|_{\theta=\hat{\theta}_{\text{PMLE}}}$$

respectively.

As an alternative, a semi-parametric inference framework can be developed where the feature of interest θ of the finite population $\mathcal{P} := \{X_i,\ i \in [N]\}$, instead of a parametric family of distributions, as earlier, is described by the set of estimating equations $\frac{1}{N} \sum_{i=1}^{N} g(X_i, \theta) = 0$, with a vector of known functions g. This approach avoids the complete parametric specification of the model, and it is widely utilized in statistics and econometrics [6,25]. Given a sample $x = (x_1, \ldots, x_n)^{\text{T}}$ and survey weights $x = (\pi_1, \ldots, \pi_n)^{\text{T}}$, the exponentially tilted empirical likelihood [55] is given by

$$L_{\text{MCM}}(\theta) = \Big\{ \prod_{i=1}^{n} w_i^\star : w^\star = \arg\max_{w} \text{H}_n(\mathbf{w}),\ \mathbf{w} \in \mathcal{S}_{n-1},\ \sum_{i=1}^{n} w_i [\pi_i g(x_i, \theta)] = 0 \Big\}.$$

Here, and elsewhere, we use MCM as an acronym for the *moment condition model*. When the convex hull of $\cup_{i=1}^{n} g(x_i, \theta)$ contains the origin, it leads to $L_{\text{MCM}}(\theta) = \prod_{i=1}^{n} w_i^{\star}(\theta)$, with

$$w_i^{\star}(\theta) = \frac{\exp[\pi_i \lambda(\theta)^{\mathrm{T}} g(x_i, \theta)]}{\sum_{j=1}^{n} \exp[\pi_j \lambda(\theta)^{\mathrm{T}} g(x_j, \theta)]}$$

and $\lambda(\theta) = \arg\min_{\eta} n^{-1} \sum_{i=1}^{n} \exp[\pi_i \eta^{\mathrm{T}} g(x_i, \theta)]$. When the convex hull condition is not satisfied, $L_{\text{MCM}}(\theta \mid x_1, \ldots, x_n)$ is set to zero.

3.2. Proposed Methodology

Importantly, it is often difficult, if not impossible, to place more flexible constraints on the parameter of interest via moment conditions. In this article, we intend to provide additional flexibility to the ETEL framework via providing the scope for statistical distance-based parametric distribution-guided constraints. However, it is not straightforward to accomplish this in the context of complex survey data, due to the presence of the survey weights. To carefully circumnavigate this issue, we first reconstruct M pseudo-true populations of size N from the observed complex survey sample of size n via weighted finite population Bayesian bootstraping [47–49] to adjust for the survey weights; next, we draw an i.i.d. pseudo-sample of size n from each of the pseudo-true populations, and we finally construct an ETEL based on each of the M pseudo-samples. Given the m-th i.i.d. pseudo-sample $(x_{m,1}^{\star}, \ldots, x_{m,n}^{\star})$, $m \in [M]$, the exponentially tilted empirical likelihood with a parametric distribution-guided constraint takes the form

$$\begin{aligned} L_{\text{BDCM}}(\theta) = \Bigg\{ & \prod_{i=1}^{n} w_i^{\star} : w^{\star} = \arg\max_{w} \mathrm{H}_n(\mathbf{w}),\ \mathbf{w} \in \mathcal{S}_{n-1}, \\ & \sum_{i=1}^{n} w_i g(x_{m,i}^{\star}, \theta) = 0,\ W_2^2 \Bigg[\sum_{i=1}^{n} w_i \delta_{x_i^{\star}}(\cdot),\ f_{\theta} \Bigg] \leq \varepsilon \Bigg\}, \end{aligned}$$

where δ is the indicator function, f_{θ} is the parametric distribution of choice, and ε is a user-defined parameter denoting the maximum extent of departure from the parametric distribution of choice. Here, and elsewhere, we use BDCM as an acronym for *bootstrapped distributionally constrained models*. Importantly, the inference on the M pseudo-true samples can be carried out in parallel. The final estimate of θ is obtained via combining the estimates obtained from the M i.i.d. pseudo-samples.

3.3. Experiments

Based on the numerical experiments in [45], we design simulation studies to compare the proposed distribution-guided entropy maximization approach with the popular pseudo-likelihood approach. Suppose that the random variables (X, Z) jointly follow a bivariate normal distribution with mean $(\mu_x, \mu_z)' = (0, 10)'$, marginal variances $(\sigma_x^2, \sigma_z^2)' = (4, 16)'$, and correlation $\rho \in \{0.1, 0.5, 0.8\}$. The variable X is the variable of interest; we aim to estimate its mean μ_x and variance σ_x^2. The variable Z is a selection variable, i.e., the Z-value of a population unit determines the probability of inclusion of the unit in the sample. Specifically, we posit that the inclusion probability of X_s in the sample is given by $\pi_s^{\star} = \Phi(\beta_0 + \beta_1 Z_s)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. When a population unit is included in the sample, we observe x_s and assign a survey weight π_s such that $\pi_s \propto 1/\pi_s^{\star}$. Importantly, we assume that we do not directly observe Z_s. The selected sample of size n is denoted as $(\mathbf{x}, \boldsymbol{\pi})'$. We scale the weights such that they sum up to 1, and we have $\pi_s = \frac{(1/\pi_s^{\star})}{\sum_{j=1}^{n}(1/\pi_j^{\star})}$, $\quad s \in [n]$. The objective is to utilize $(\mathbf{x}, \boldsymbol{\pi})'$ to estimate the population parameters of interest (μ_x, σ_x^2).

We generate $N = 100,000$ values of (X_s, Z_s) as a finite population. We set $\beta_0 = 0.1$, $\beta_1 = -1.8$ and draw samples of sizes $n \in \{500, 1000, 1500, 2000, 2500\}$ from the finite population. Under each data generating setup, we utilize 100 Monte Carlo simu-

lations. For the pseudo-maximum likelihood (PMLE) approach, we simply posit the model $f_\theta \equiv \text{Normal}(\mu_x, \sigma_x^2)$. For the proposed BCDM approach, we assume the moment constraint based on the function $g(x, \mu_x) = x - \mu_x$, and the Wasserstein distance constraint based on the parametric family of distributions $f_\theta \equiv \text{Normal}(\mu_x, \sigma_x^2)$. For each of the replicates, we choose $M = 500$; to ensure the comparability of PMLE and BDCM, we set $\varepsilon = W_2^2\left[\sum_{i=1}^n 1/n\delta_{x_i^*}(\cdot),\ f_{\hat{\theta}}\right]$, where $\hat{\theta}$ is the estimate of θ obtained via PMLE. The bias and the coverage of the pseudo-maximum likelihood and moment condition model-based estimators for varying data generating mechanisms are presented in Table 1. A case study with complex survey data from the National Health and Nutrition Examination Surveys (NHANES) is provided later. .

Table 1. Average bias (=$||(\mu, \sigma^2) - (\hat{\mu}, \hat{\sigma^2})||$) and coverage (within brackets) of the MLE, PMLE, BPPE [56], and BDCM estimators for varying data generating mechanisms.

n	ρ	0.1	0.5	0.8
500	MLE	0.19 (0.91)	0.68 (0.48)	1.11 (0.45)
	BPPE	0.67 (0.82)	0.65 (0.72)	0.71 (0.78)
	PMLE	0.16 (0.94)	0.16 (0.91)	0.16 (0.94)
	BDCM	0.16 (0.92)	0.16 (0.93)	0.16 (0.95)
1000	MLE	0.16(0.87)	0.69(0.48)	1.11(0.42)
	BPPE	0.15 (0.92)	0.18 (0.90)	0.18(0.92)
	PMLE	0.11 (0.94)	0.10 (0.94)	0.10 (0.92)
	BDCM	0.11 (0.93)	0.10 (0.94)	0.10 (0.96)
1500	MLE	0.15(0.84)	0.68(0.47)	1.11(0.42)
	BPPE	0.12 (0.94)	0.10 (0.89)	0.12 (0.90)
	PMLE	0.09(0.94)	0.08(0.94)	0.07 (0.93)
	BDCM	0.09(0.94)	0.08(0.93)	0.08(0.94)
2000	MLE	0.15 (0.81)	0.68 (0.48)	1.10 (0.40)
	BPPE	0.09 (0.92)	0.08 (0.92)	0.07(0.92)
	PMLE	0.07(0.95)	0.07(0.95)	0.06(0.92)
	BDCM	0.07 (0.95)	0.07 (0.97)	0.07 (0.97)
2500	MLE	0.15 (0.75)	0.68 (0.47)	1.10 (0.39)
	BPPE	0.06(0.94)	0.07(0.88)	0.06 (0.92)
	PMLE	0.06(0.96)	0.07 (0.94)	0.06 (0.92)
	BDCM	0.06(0.97)	0.06 (0.95)	0.06 (0.94)

3.4. National Health and Nutrition Examination Surveys (NHANES) Data Analysis

The NHANES is a series of surveys designed to assess the health and nutritional status of individuals in the United States. The data extracted are from the NHANES 2009–2010 [46], and they contain information on binary indicators of high cholesterol, race, age, etc., and survey weights for 8591 individuals. For this exercise, we assume that these 8591 individuals make up a finite population and obtain samples of size $n \in \{250, 500, 1000, 2000\}$ according to the survey weights. We fit a logistic regression to model the binary indicator of high cholesterol as a function of race and age. For each $n \in \{250, 500, 1000, 2000\}$, we utilize 100 Monte Carlo simulations. For the distribution-guided entropy maximization approach, we assume constraints on the score function of the logistic regression. The coverage of the moment condition model-based estimates of the regression coefficients is presented in Table 2.

Table 2. NHANES data. Bias $= ||\beta - \hat{\beta}||$ and coverage of the moment condition model-based estimates of the regression parameters for varying sample sizes.

n	250	500	1000	2000
Coverage	0.95	0.95	0.96	0.96
Bias	0.42	0.27	0.18	0.13

4. Demographic Parity

Discrimination pertains to the unfair treatment of individuals based on specific demographic characteristics known as protected attributes. The goal of demographic parity or statistical parity [50,51] in machine learning is to design algorithms that yield fair inferences devoid of discrimination due to membership in certain demographic groups determined by a protected attribute. First, we introduce the mathematical formalization of the notions of demographic parity. To that end, we assume that X denotes the feature vector used for predictions, A is the protected attribute with two levels $\{S, T\}$, and Y is the response. Parity constraints are phrased in terms of the distribution over (X, A, Y). Two definitions are in order.

Definition 2 (Demographic parity, [50])**.** *A predictor h satisfies demographic parity under the distribution over (X, A, Y) if $h(X)$ is independent of the protected attribute A, i.e., $\mathbf{P}[h(X) \geq z \mid A = S] = \mathbf{P}[h(X) \geq z \mid A = T] = \mathbf{P}[h(X) \geq z]$, for all z.*

Definition 3 (Demographic parity in expectation, [50])**.** *A predictor h satisfies demographic parity under the distribution over (X, A, Y) if $h(X)$ is independent of the protected attribute A, i.e., $\mathbf{E}[h(X) \mid A = S] = \mathbf{E}[h(X) \mid A = T] = \mathbf{E}[h(X)]$.*

4.1. Proposed Methodology

Although the notions of demographic parity in Definitions 2 and 3 coincide when we work with binary responses, the latter may be amenable to simple computational algorithms [57] compared to the general definition. However, the notion of demographic parity in expectation is somewhat prohibitive since one cannot control the predictor h over its entire domain. For example, depending on the application of interest, we may be solely interested in controlling the tails of the predictor [58]. Returning to our semi-parametric inference framework, we offer a flexible as well as a computationally feasible compromise between the notions in Definitions 2 and 3. To that end, we introduce the notion of demographic parity in the Wasserstein metric next.

Definition 4 (Demographic parity in Wasserstein metric)**.** *A predictor h achieves demographic parity in the Wasserstein metric with bias ε, under the distribution over (X, A, Y), if $W_2^2\left[F_{h_S}, F_{h_T}\right] \leq \varepsilon$, where F_{h_k} is the empirical distribution of h under sub-population k, i.e., $h(X) \mid A = k, k \in \{S, T\}$.*

Suppose that we have data $(x_i, y_i, a_i) \in \mathbf{R}^d \times \mathbf{R} \times \{S, T\}$ for n individuals on p-dimensional covariate x, univariate continuous response y, and the levels of the protected attribute $a \in \{S, T\}$. For the sake of simplicity in exposition, we also assume that $a_i = S$, $i \in [n_S]$ and $a_i = T$, $i \in [n] \setminus [n_S]$, where $n = n_S + n_T$. Next, we posit a predictive model $y_i = h(x_i, \theta_{(a_i)}) + e_i$, $e_i \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma^2)$, $i \in [n]$, where h is potentially non-linear and $(\theta_{(S)}, \theta_{(T)})$ is the model parameter of interest to be estimated under the demographic parity constraint $W_2^2\left[F_{h_S}, F_{h_T}\right] \leq \varepsilon$. In particular, we consider the empirical CDF of h under sub-population S, $F_{h_S} = 1/n_S \sum_{i=1}^{n_S} \delta_{h(x_{iS})}(\cdot)$, and a weighted empirical CDF of h under sub-population T, $F_{h_T} = \sum_{i=n_S+1}^{n} w_i \, \delta_{h(x_{iT})}(\cdot)$. Here, δ is the Dirac delta measure. The goal is to infer about $(\theta_{(S)}, \theta_{(T)}, w)$, ensuring that the demographic parity constraint, i.e., F_{h_S}, F_{h_T}, is close with respect to W_2^2, and, at the same time, the extent of re-weighting in F_{h_T} is minimal, i.e., the entropy $-\sum_{i=n_S+1}^{n} w_i \log w_i$ is close to the maximal entropy $\log n_T$. A related idea in Jiang et al. [59] deals with W_1 constrained fair classification problems, but our approach of additionally re-weighting the observations offers more flexibility, with possible ramifications in the study of fairness in misspecified models.

We achieve this through an *in-model* approach solving the optimization problem

$$\max_{w,\theta_{(S)},\theta_{(T)},\sigma^2} \Bigg[-\frac{1}{n_S}\sum_{i=1}^{n_S} l_i(\theta_{(S)} \mid x_i) - \sum_{i=n_s+1}^{n} w_i l_i(\theta_{(T)} \mid x_i) \\ -(1-\lambda^\star) W_2^2\left[F_{h_S}, F_{h_T}\right] - \lambda^\star \sum_{i=n_s+1}^{n} w_i \log w_i \Bigg] \tag{3}$$

where $\sum_{i=n_s+1}^{n} w_i = 1$ and $l_i(\theta_{(a_i)} \mid x_i) = (y_i - h(x_i, \theta_{(a_i)}))^2/2\sigma^2$, $i \in [n]$. For a resulting re-weighting vector $w^\star = (w^\star_{n_S+1}, \ldots, w^\star_n)'$, we can obtain a fair prediction at a new $x \in T$ via a weighted kernel density estimate at x. As a competitor to the *in-model* scheme, motivated by popular post-processing schemes to ensure fairness [60,61], we utilize a *two-step* procedure.
Step 1: We obtain model parameter estimates by $(\hat{\theta}_{(S)},\ \hat{\theta}_{(T)},\ \hat{\sigma}^2) =$

$$\underset{\theta_{(S)},\,\theta_{(T)},\,\sigma^2}{\arg\max} \Bigg[-\frac{1}{n_S}\sum_{i=1}^{n_S} l_i(\theta_{(S)} \mid x_i) - \frac{1}{n_T}\sum_{i=n_s+1}^{n} l_i(\theta_{(T)} \mid x_i) \Bigg] \tag{4}$$

followed by a post-processing step at $(\hat{\theta}_{(S)},\ \hat{\theta}_{(T)},\ \hat{\sigma}^2)$ to obtain $w^\star$
Step 2:

$$\arg\max_{w} \Bigg[-(1-\lambda^\star)\ W_2^2\left[F_{h_S}, F_{h_T}\right] - \lambda^\star \sum_{i=n_s+1}^{n} w_i \log w_i \Bigg]. \tag{5}$$

A case study on algorithmic mental health monitoring is provided next. An additional case study on algorithmic criminal risk assessment is also included.

4.2. Distress Analysis Interview Corpus (DAIC)

The Distress Analysis Interview Corpus (DAIC) [62] is a multi-modal clinical interview collection, accessible upon request via the DAIC-WOZ (https://dcapswoz.ict.usc.edu/ (accessed on 16 October 2023)) website. Computer agents based on such clinical interviews are deemed to be used to make mental health diagnoses in relation to certain employment decisions, and concerns about the fairness of such tools with respect to the biological gender of the individuals have been raised. Specifically, we focus on predicting the PHQ-8 score, which captures the individual's severity of depression, as a function of the individual's verbal signals during the clinical interviews, while biological gender serves as a protected attribute. In particular, the Fourier series analysis of the speech signals of the individuals yields verbal attributes of interest, which in turn could be potentially used in the diagnosis of the individual's severity of depression. Therefore, it is of interest to develop novel methods to produce predictions while avoiding disparate treatment on the basis of biological gender. More precisely, we wish to ensure that the demographic parity constraint is satisfied here, which, in this context, simply dictates that the weighted empirical CDFs of the biological gender-specific fitted PHQ-8 scores are *identical* or *similar*.

The PHQ-8 scores range from 0 to 27, with a score of 0–4 considered none or minimal, 5–9 mild, 10–14 moderate, 15–19 moderately severe, and 20–27 severe. In this application, we work with the PHQ-8 (continuous response), biological gender (binary protected attribute), and 17 derived audio/verbal features (continuous covariates) corresponding to the $n = 107$ subjects. The PHQ-8 scores for the two biological genders show a clear discrepancy. Therefore, we shall assess the relative performance of the *in-model* scheme (3) and the *two-step* scheme (4) and (5) in ensuring demographic parity with respect to biological gender (refer to Figures 1 and 2. As earlier, for the sake of simplicity of exposition, we use linear regression (i.e., h is linear in the covariates) as our predictive model of choice. When we fit the predictive model without any fairness constraint, the fitted empirical cumulative distribution functions corresponding to the two biological genders are widely different. Our *in-model* scheme, as well as the *two-step*, scheme significantly reduces the discrepancy owing to its in-built fairness-based regularization. As noted earlier, the *in-model* scheme provides lower bias since it performs the two-step optimization simultaneously.

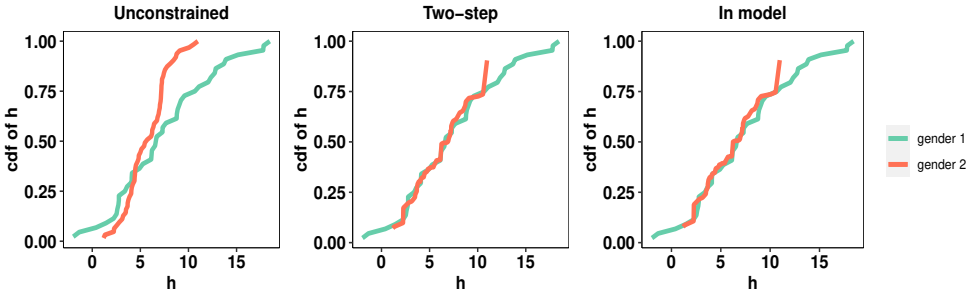

Figure 1. ***Distress Analysis Interview Corpus.*** Empirical CDFs of fitted h for the two groups, with no fairness constraint ($W_2 = 19.32$), fair post-processing ($W_2 = 2.24$), and fair model fitting with ($W_2 = 0.79$), respectively, at $\lambda^\star = 0$.

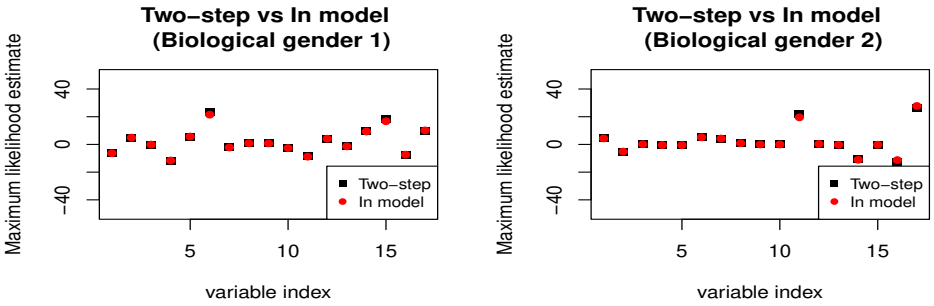

Figure 2. ***Distress Analysis Interview Corpus.*** Maximum likelihood estimates of the regression coefficients under both two-step and in-model schemes. In the in-model scheme, the estimates are slightly modified since the regression coefficients and the weights assigned to the data are learned simultaneously. For details on the in-model and two-step approaches, refer to Equations (3), (4), and (5), respectively.

4.3. COMPAS Recidivism Data Analysis

We consider a case study of algorithmic criminal risk assessment. We shall focus on the popular COMPAS dataset [63], which includes information on the criminal history of defendants in Broward County, Florida, available from the propublica website (https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis (accessed on 16 October 2023)). For each individual, several features of the criminal history are available, such as the number of past felonies, misdemeanors, and juvenile offenses; additional demographic information includes the sex, age, and ethnic group of each defendant. We focus on predicting the two-year recidivism score y (continuous) as a function of the defendant's demographic information, except for race and criminal history x, while race (categorical) serves as a protected attribute. Algorithms for the creation of such predictions are routinely used in courtrooms to advise judges, and concerns about the fairness of such tools with respect to the race of the defendants have been raised. Therefore, it is of interest to develop novel methods to produce predictions while avoiding disparate treatment on the basis of the protected attribute, race. More precisely, we wish to ensure that the demographic parity constraint is satisfied, which, in this context, simply dictates that the weighted empirical CDFs of the race-specific fitted recidivism scores are *identical* or *similar*.

For simplicity of exposition, we only consider two levels for the protected attribute of race, namely African-American and non-African-American, and consider a sub-sample of the entire dataset with 100 defendants corresponding to each level of the protected attribute. As a covariate, for each defendant, we consider the demographic information—sex (binary), age (continuous), and marital status (categorical)—and criminal status—legal status (categorical), supervision level (categorical), and custody status (categorical). We use linear regression (i.e., h is linear in the covariates) as our predictive model of choice; the methodology readily extends to more complicated models. The raw recidivism scores

for African-Americans versus non-African-Americans show a clear discrepancy). We shall assess the relative performance of the *in-model* scheme in (3) and the *two-step* scheme in (4) and (5) in ensuring demographic parity with respect to the protected attribute of race (refer to Figures 3 and 4). When we fit the predictive model without any fairness constraint, the fitted empirical cumulative distribution functions corresponding to the two sub-populations are widely different. Our *in-model* scheme, as well as the *two-step* scheme, significantly reduces the discrepancy owing to the in-built fairness-based regularization. As expected, the *in-model* scheme provides slightly lower bias since it performs the two-step optimization simultaneously.

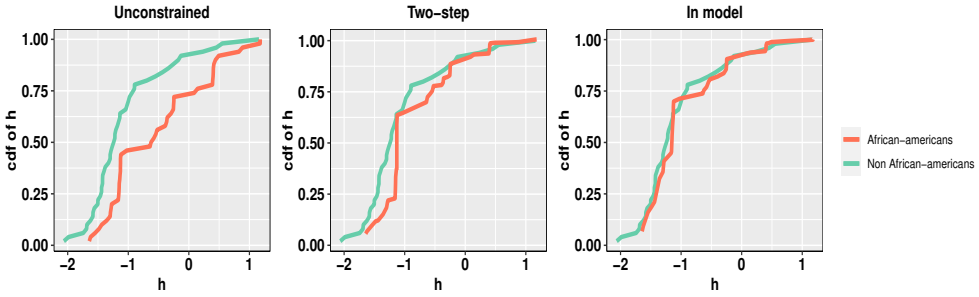

Figure 3. ***COMPAS dataset.*** Empirical CDFs of fitted h for the two groups, with no fairness constraint $(W_2 = 0.72)$, fair post-processing $(W_2 = 0.05)$, and fair model fitting with $(W_2 = 0.02)$, respectively, at $\lambda^\star = 0$.

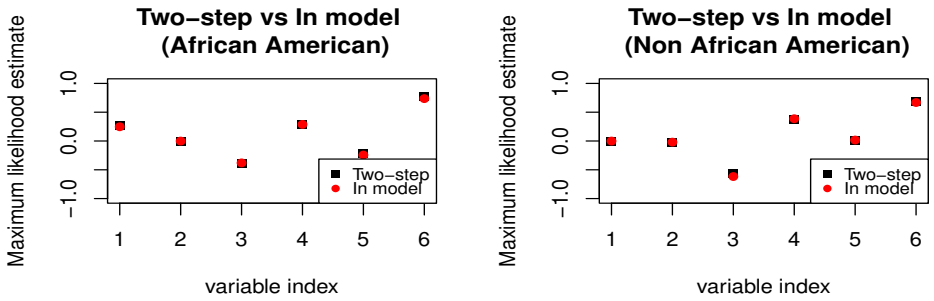

Figure 4. ***COMPAS dataset.*** Maximum likelihood estimates of the regression coefficients under both two-step and in-model schemes. In the in-model scheme, the estimates are slightly modified since the regression coefficients and the weights assigned to the data are learned simultaneously.

5. Entropy-Based Portfolio Allocation

We present an application of the proposed parametric distribution-guided entropy maximization framework to portfolio allocation problems [5,52,53]. Portfolio optimization is concerned with the allocation of an investor's wealth over several assets to optimize specific objective(s) based on historical data on asset returns. To elucidate the problem clearly, let $R_{(i)} = (R_{i,1}, R_{i,2}, \ldots, R_{i,d})'$ be the excess returns on d risky assets recorded over time $i \in [n]$. The portfolio $(w_1, \ldots, w_d)$ is a vector of weights that represents the investor's relative allocation of their wealth satisfying $\sum_{i=1}^{d} w_i = 1$ and $w_i \geq 0, i \in [d]$. The goal is to learn the $(w_1, \ldots, w_d)$ subject to specific constraints based on historical data.

5.1. Related Works

Markowitz's mean variance optimization [53] is widely recognized as one of the foundational formulations of the portfolio selection problem. The traditional mean variance (MV) optimal portfolio weights [53] are obtained via

$$\text{argmax}_w \left[w^{\mathrm{T}} \mu - \frac{\lambda}{2} w^{\mathrm{T}} \Sigma w \right],$$

such that $\sum_{i=1}^{d} w_i = 1$, where $\mu = (\mu_1, \ldots, \mu_d)^{\mathrm{T}} = (1/n)\sum_{i=1}^{n} R_{(i)}$ and $\Sigma = (1/n)\sum_{i=1}^{n}(R_{(i)} - \mu)(R_{(i)} - \mu)^{\mathrm{T}}$ are the mean and variance of the historical return, and $\lambda > 0$ is a risk aversion parameter. Given a specific mean and covariance matrix, the Markowitz paradigm offers an elegant approach to achieve an efficient allocation, where the pursuit of higher expected returns inevitably entails greater risk. However, in this framework, it is essential either for the asset returns to follow a normal distribution or for the utility to solely depend on the first two moments.

Real-world financial returns, as indicated by empirical evidence [64], diverge from normal distribution assumptions and commonly exhibit heavier tails and a lack of symmetry. For instance, Volatility Feedback Theory [36] posits that market volatility can influence subsequent returns. Specifically, it suggests that periods of high volatility can lead to skewed returns, where extreme price movements are more likely to occur. This theory underscores the idea that heightened volatility tends to be associated with increased risk and uncertainty in financial markets, potentially resulting in non-normal return distributions characterized by asymmetry and fat tails. Skewed returns indicate that the probability of extreme events, either positive and negative, is higher than what would be expected under a normal distribution. Consequently, skew normal distributions are routinely utilized to model observed returns [37,38].

Specifically, in the context of portfolio optimization, refs. [65,66] proposed to utilize higher-order moments in the portfolio allocation problem. However, portfolios created using sample moments of stock returns tend to be excessively concentrated in a small number of assets, which contradicts the fundamental principle of diversification. To that end, several approaches are proposed in the literature that ensure the shrinkage of the portfolio weights towards maximum diversification [5,67,68], i.e., they maximize the entropy of the portfolio weights. In particular, ref. [5] proposed to obtain the portfolio weights solving the optimization problem $\arg\max_w \mathrm{H}_d(\mathbf{w})$ subject to $\sum_{i=1}^{d} w_i\mu_i \geq \mu_0$, $w^{\mathrm{T}}\Sigma w \leq \sigma_0^2$, such that $\sum_{i=1}^{d} w_i = 1$ and (μ_0, σ_0^2) are the target mean and variance of the portfolio return. In essence, this approach constitutes obtaining the portfolio weight via entropy maximization subject to moment-based constraints.

5.2. Proposed Methodology

Importantly, empirical evidence suggests that there is merit in modeling asset returns via non-normal distributions [65,69], e.g., a skew normal distribution [26,70]. However, it is often unfeasible to place more flexible constraints on the portfolio weights in terms of moment conditions. In this section, we intend to provide additional flexibility to the entropy-based portfolio optimization framework via providing the scope for statistical distance-based parametric distribution-guided constraints. Our semi-parametric framework provides a formidable alternative to the existing literature, since (a) we can flexibly specify the distribution of the expected return and (b) the entropy provides direct handling of portfolio diversity. We achieve this by obtaining portfolio weights via the optimization problem $\arg\max_w \mathrm{H}_d(\mathbf{w})$ subject to $W_2^2\left[\frac{1}{n}\sum_{i=1}^{T}\delta_{w^T R_{(i)}}(\cdot), f_{\theta_0}\right] \leq \varepsilon$, such that $\sum_{i=1}^{d} w_i = 1$. Here, $\frac{1}{n}\sum_{i=1}^{n}\delta_{w^T R_{(i)}}(\cdot)$ is the empirical distribution of the portfolio return, f_θ is the centering parametric family of the distribution of choice, θ_0 is the fixed target value of θ, and ε is a user-defined parameter. For practical purposes, it is useful to express the optimization problem above as the following:

$$\arg\min_{w}\left[(1-\lambda^\star)W_2^2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w^T R_{(i)}}(\cdot),\ f_{\theta_0}\right) - \lambda^\star\, b_d \mathrm{H}_d(\mathbf{w})\right] \tag{6}$$

such that $\sum_{i=1}^{d} w_i = 1$ and $b_d = 1/\log d$. This choice of b_d is convenient since it ensures that $b_d\mathrm{H}_d(\mathbf{w}) \in [0,1]$. Further, the user-defined parameter $\lambda^\star \in [0,1]$ controls the balance between the portfolio diversity and the extent of deviation from the target distribution f_{θ_0}.

For exposition in this article, we choose f_{θ_0} to be a skew normal distribution [26] with parameters $\theta = (\omega, \zeta, \alpha)'$. The probability density function of a skew normal dis-

tribution $\mathrm{SN}(\zeta, \omega, \alpha)$ is given by $f(z) = \frac{2}{\omega}\phi\left(\frac{z-\zeta}{\omega}\right)\Phi\left(\alpha\left(\frac{z-\zeta}{\omega}\right)\right)$, $z \in \mathbf{R}$ where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the probability density function and cumulative density function of the standard normal distribution. For $\alpha = 0$, we can recover the normal distribution as the absolute value of skewness increases and the absolute value of α increases. For $\alpha > 0$, the distribution is left-skewed, and it is right-skewed for $\alpha < 0$. If $Z \sim \mathrm{SN}(\zeta, \omega, \alpha)$, then we have $\mu_0 = \mathbf{E}(Z), \sigma_0^2 = \mathrm{Var}(Z)$, $\gamma_0 = \text{Skewness}$. This allows us to set (ζ, ω, α) to achieve target $\theta_0 = (\mu_0, \sigma_0^2, \gamma_0)$ of the portfolio return distribution. This resulting skew normal density with fully specified parameters then serves as the target distribution to calculate the portfolio weights based on (6). The user can select any flexible probability distribution to model the portfolio return and follow the prescribed procedure to compute the target parameter values. The suggested methodology could potentially be utilized with more versatile target distributions, such as the generalized skew normal distribution [71]. Nevertheless, addressing this aspect is beyond the scope of the current paper, although it is a promising avenue for future investigation.

5.3. Historical Stock Returns Data Analysis

We consider the stock returns data of 5 companies (AMZN, AAPL, XOM, T, MS) for the period of January 2000 to December 2020, publicly available from Yahoo! Finance. The data are aggregated at a monthly level. The goal is to compare the mean variance optimal portfolio and the proposed parametric distribution-guided portfolio allocation framework. First, we compute the mean variance optimal portfolio for varying values of the risk aversion parameter $\lambda \in [0, 10]$. Figure 5 records the skewness, excess kurtosis, and number of zero portfolio weights for the mean variance optimal portfolio for varying λ. We focus on λ set at 1—a choice at which 3 out of 5 portfolio weights are 0, and the optimal portfolio return distribution is negatively skewed and leptokurtic. This exposes the fact that, once we have fixed the λ, the mean variance optimal portfolio optimization framework does not offer direct control over the portfolio diversity, and we will potentially obtain portfolio allocations concentrated on very few assets. Next, we fix the parameters of a skew normal density $\theta = (\omega, \zeta, \alpha)'$ such that its mean, variance, and skewness match the quantities of the mean variance optimal portfolio return at $\lambda = 1$. Finally, we compute the skew normal distribution-guided maximum entropy portfolio for varying values of the balance parameter $\lambda^\star \in [0, 1]$ in (6). Figure 6 presents the entropy of the portfolio weights and the departure of the portfolio return distribution from the guiding skew normal distribution as a function of $\lambda^\star \in [0, 1]$. This showcases that, contrary to the mean variance optimal portfolio allocation, here, the fund manager can choose a specific $\lambda^\star$ to ensure the desired level of portfolio diversity, while maintaining fidelity towards a pre-specified distribution of the portfolio return distribution.

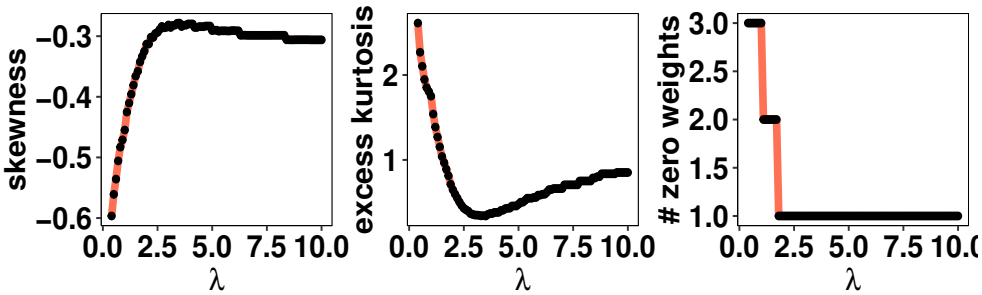

Figure 5. Limitations of mean variance optimal portfolio. (i) The skewness and excess kurtosis plots provide evidence that the normality assumption for expected returns does not hold. (ii) A small value of λ leads to zero weight to several assets.

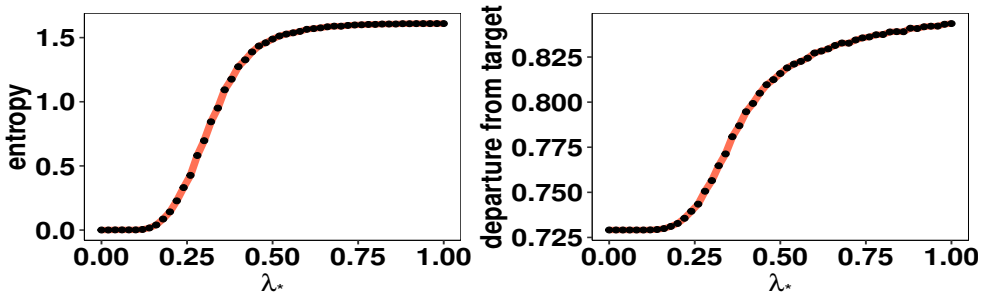

Figure 6. With a fixed target skew normal return, varying values of $\lambda^\star \in [0,1]$ provide different balances between diversity and departure from the target. The desired degree of diversification can be achieved $\lambda_\star$ via a simple grid search on $\lambda_\star \in [0,1]$.

6. Concluding Remarks

We introduce a nonparametrically oriented framework that aims to align the maximum entropy weight-adjusted empirical distribution of observed data closely with a predefined and potentially continuous probability distribution, while permitting mild deviations. The framework's versatility is showcased in three distinct applications. We anticipate the proposed methodology's utility in numerous other statistical tasks requiring data re-weighting, e.g., robustness [18], covariate shifts [18], ill-posed inverse problems [4], etc.

Author Contributions: Conceptualization, A.C., A.B. and D.P.; methodology, A.C., A.B. and D.P.; software, A.C.; validation, A.C.; formal analysis, A.C.; investigation, A.C.; resources, A.C.; data curation, A.C.; writing—original draft preparation, A.C., A.B. and D.P.; writing—review and editing, A.C., A.B. and D.P.; visualization, A.C.; supervision, A.B. and D.P.; project administration, A.C.; funding acquisition, A.C., A.B. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev. Ser.* **1957**, *106*, 620–630. [CrossRef]
2. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
3. Skilling, J.; Bryan, R.K. Maximum entropy image reconstruction: General algorithm. *Mon. Not. R. Astron. Soc.* **1984**, *211*, 111–124. [CrossRef]
4. Gamboa, F.; Gassiat, E. Bayesian methods and maximum entropy for ill-posed inverse problems. *Ann. Stat.* **1997**, *25*, 328–350. [CrossRef]
5. Bera, A.K.; Park, S.Y. Optimal portfolio diversification using the maximum entropy principle. *Econom. Rev.* **2008**, *27*, 484–512. [CrossRef]
6. Chib, S.; Shin, M.; Simoni, A. Bayesian estimation and comparison of moment condition models. *J. Am. Stat. Assoc.* **2018**, *113*, 1656–1668. [CrossRef]
7. Gudivada, V.N. Computational analysis and understanding of natural languages: Principles, methods and applications. In *Handbook of Statistics*; Elsevier: Amstardam, The Netherlands, 2018.
8. de Abril, I.M.; Yoshimoto, J.; Doya, K. Connectivity inference from neural recording data: Challenges, mathematical bases and research directions. *Neural Netw.* **2018**, *102*, 120–137.
9. Eysenbach, B.; Levine, S. Maximum entropy RL (provably) solves some robust RL problems. *CoRR* **2021**. Available online: https://arxiv.org/abs/2103.06257 (accessed on 16 October 2023).
10. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2012.
11. Kardar, M. *Statistical Physics of Particles*; Cambridge University Press: Cambridge, UK, 2007.
12. Einmahl, J.H.J.; Krajina, A.; Segers, J. A method of moments estimator of tail dependence. *Bernoulli* **2008**, *14*, 1003–1026. [CrossRef]
13. Chernozhukov, V.; Newey, W.K.; Santos, A. Constrained conditional moment restriction models. *Econometrica* **2023**, *91*, 709–736. [CrossRef]

14. Chakraborty, A.; Bhattacharya, A.; Pati, D. Robust Probabilistic Inference via a Constrained Transport Metric. 2023. Available online: https://arxiv.org/abs/2303.10085 (accessed on 16 October 2023).
15. Rachev, S.T.; Stoyanov, S.; Fabozzi, F.J. *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimization: The Ideal Risk, Uncertainty, and Performance Measures*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
16. Santambrogio, F. Optimal Transport for Applied Mathematicians. Calculus of Variations, Pdes and Modeling. 2015. Available online: https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf (accessed on 16 October 2023).
17. Villani, C. *Topics in Optimal Transportation*; American Mathematical Society. 2003. Available online: https://www.math.ucla.edu/~wgangbo/Cedric-Villani.pdf (accessed on 16 October 2023).
18. Wang, Y.; Kucukelbir, A.; Blei, D.M. Robust probabilistic modeling with bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning—Volume 70*; JMLR. org: Sydney, Australia, 2017; pp. 3646–3655.
19. Wen, J.; Yu, C.-N.J.; Greiner, R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In Proceedings of the International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014.
20. Yan, B.; Seto, S.; Apostoloff, N. Forml: Learning to Reweight Data for Fairness. 2022. Available online: https://arxiv.org/abs/2202.01719 (accessed on 16 October 2023).
21. Ramas, J.G.; Le, T.; Chen, B.; Kumar, M.; Kay Rottmann, K. Unsupervised training data reweighting for natural language understanding with local distribution approximation. In Proceedings of the EMNLP 2022, Abu Dhabi, UAE, 7–11 December 2022. Available online: https://www.amazon.science/publications/unsupervised-training-data-reweighting-for-natural-language-understanding-with-local-distribution-approximation (accessed on 16 October 2023).
22. Mandt, S.; McInerney, J.; Abrol, F.; Ranganath, R.; Blei, D. Variational tempering. In *Artificial Intelligence and Statistics*; Cadiz, Spain, 2016; pp. 704–712.
23. White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. Available online: http://www.jstor.org/stable/1912526 (accessed on 16 October 2023). [CrossRef]
24. Hall, A. *Generalized Method of Moments*; Oxford University Press: Oxford, UK, 2005.
25. Chib, S.; Shin, M.; Simoni, A. Bayesian estimation and comparison of conditional moment models. *arXiv* **2021**, arXiv:2110.13531.
26. Azzalini, A.; DALLA Valle, A. The multivariate skew-normal distribution. *Biometrika* **1996**, *83*, 715–726. [CrossRef]
27. Panaretos, V.M.; Zemel, Y. Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 405–431. [CrossRef]
28. Owen, A.B. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **1988**, *75*, 237–249. [CrossRef]
29. Owen, A.B. Empirical likelihood confidence regions. *Ann. Stat.* **1990**, *18*, 90–120. [CrossRef]
30. Owen, A.B. Empirical likelihood for linear models. *Ann. Stat.* **1991**, *19*, 1725–1747. [CrossRef]
31. Owen, A.B. *Empirical Likelihood*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2001.
32. Qin, J.; Lawless, J.L. Empirical likelihood and general estimating equations. *Ann. Stat.* **1994**, *22*, 300–325. [CrossRef]
33. Nordman, D.J.; Lahiri, S.N. A review of empirical likelihood methods for time series. *J. Stat. Plan. Inference* **2014**, *155*, 1–18. [CrossRef]
34. Newey, W.; Smith, R.J. Higher-order properties of gmm and generalized empirical likelihood estimators. *Econometrica* **2004**, *72*, 219–255. [CrossRef]
35. Schennach, S.M. Point estimation with exponentially tilted empirical likelihood. *Ann. Stat.* **2007**, *35*, 634–672. [CrossRef]
36. Brown, K.C.; Harlow, W.V.; Tinic, S.M. Risk aversion, uncertain information, and market efficiency. *J. Financ. Econ.* **1988**, *22*, 355–385. [CrossRef]
37. De Luca, G.; Loperfido, N. A skew-in-mean garch model for financial returns. In *Skew-Elliptical Distributions and Their Applications: A Journey beyond Normality*; CRC/Chapman & Hall: Boca Raton, FL, USA, 2004; pp. 205–222.
38. Peiro, A. Skewness in financial returns. *J. Bank. Financ.* **1999**, *23*, 847–862. [CrossRef]
39. Birgin, E.G.; Martínez, J.M. Improving ultimate convergence of an augmented lagrangian method. *Optim. Methods Softw.* **2008**, *23*, 177–195. [CrossRef]
40. Conn, A.R.; Gould, N.I.M.; Toint, P. A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. *Siam J. Numer. Anal.* **1991**, *28*, 545–572. [CrossRef]
41. Becker, S.R.; Candès, E.J.; Grant, M. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.* **2011**, *3*, 165–218. [CrossRef]
42. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022. Available online: https://www.R-project.org/ (accessed on 16 October 2023).
43. Johnson, S.G. *The Nlopt Nonlinear-Optimization Package*; The Comprehensive R Archive Network; R Foundation for Statistical Computing: Vienna, Austria, 2022.
44. Grant, M.; Boyd, S. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences*; Blondel, V., Boyd, S., Kimura, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 95–110. Available online: http://stanford.edu/~boyd/graph_dcp.html (accessed on 16 October 2023).
45. Gunawan, D.; Panagiotelis, A.; Griffiths, W.; Chotikapanich, D. Bayesian weighted inference from surveys. *Aust. N. Z. J. Stat.* **2020**, *62*, 71–94. [CrossRef]
46. Lumley, T. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R*; John Wiley and Sons: Hoboken, NJ, USA, 2010.

47. Cohen, M.P. *The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs*; Technical report; National Center for Education Statistics: New Jersey Avenue, NW, USA; Washington, DC, USA, 1997.
48. Dong, Q.; Elliott, M.R.; Raghunathan, T.E. A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **2014**, *40*, 29–46.
49. Lo, A.Y. A bayesian method for weighted sampling. *Ann. Stat.* **1993**, *21*, 2138–2148. [CrossRef]
50. Agarwal, A.; Dudík, M.; Wu, Z.S. Fair regression: Quantitative definitions and reduction-based algorithms. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; International Machine Learning Society (IMLS): Bellevue, WA, USA, 2019; pp. 166–183.
51. Gajane, P.; Pechenizkiy, M. On Formalizing Fairness in Prediction with Machine Learning. 2018. Available online: https://www.fatml.org/media/documents/formalizing_fairness_in_prediction_with_ml.pdf (accessed on 16 October 2023).
52. Elton, E.J.; Gruber, M.J.; Brown, S.J.; Goetzmann, W.N. *Modern Portfolio Theory and Investment Analysis*; Wiley: Hoboken, NJ, USA, 2014; ISBN 978-1118469941.
53. Markowitz, H. Portfolio selection. *J. Financ.* **1952**, *7*, 77–91. Available online: http://www.jstor.org/stable/2975974 (accessed on 16 October 2023).
54. Wooldridge, J.M. Inverse probability weighted estimation for general missing data problems. *J. Econom.* **2007**, *141*, 1281–1301. Available online: https://www.sciencedirect.com/science/article/pii/S0304407607000437 (accessed on 16 October 2023). [CrossRef]
55. Schennach, S.M. Bayesian exponentially tilted empirical likelihood. *Biometrika* **2005**, *92*, 31–46 . [CrossRef]
56. León-Novelo, L.G.; Savitsky, T.D. Fully Bayesian estimation under informative sampling. *Electron. J. Statist.* **2019**, *13*, 1608–1645. [CrossRef]
57. Fitzsimons, J.; Al Ali, A.; Osborne, M.; Roberts, S. A general framework for fair regression. *Entropy* **2019**, *21*, 741. [CrossRef]
58. Yang, D.; Lafferty, J.; Pollard, D. Fair Quantile Regression. 2019. Available online: https://arxiv.org/abs/1907.08646 (accessed on 16 October 2023).
59. Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; Chiappa, S. Wasserstein fair classification. In *Machine Learning Research, Proceedings of the 35th Uncertainty in Artificial Intelligence Conference, Tel Aviv, Israel, 22–25 July 2019*; Adams. R.P., Gogate, V., Eds.; PMLR, Association for Uncertainty in Artificial, Intelligence: Pittsburg, PA, USA, 2020; Volume 115, pp. 862–872. Available online: https://proceedings.mlr.press/v115/jiang20a.html (accessed on 16 October 2023).
60. Nandy, P.; DiCiccio, C.; Venugopalan, D.; Logan, H.; Basu, K.; El Karoui, N. Achieving fairness via post-processing in web-scale recommender systems. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), Seoul, Republic of Korea, 21–24 June 2022; Association for Computing Machinery: New York, NY, USA, 2022: pp. 715–725. [CrossRef]
61. Xian, R.; Yin, L.; Zhao, H. Fair and optimal classification via post-processing. In *Machine Learning Research, Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023*; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; PMLR: Bellevue, WA, USA, 2023; Volume 202, pp. 37977–38012. Available online: https://proceedings.mlr.press/v202/xian23b.html (accessed on 16 October 2023).
62. Gratch, J.; Artstein, R.; Lucas, G.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. The distress analysis interview corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; European Language Resources Association (ELRA): Paris, France, 2014; pp. 3123–3128. Available online: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf (accessed on 16 October 2023).
63. Aliverti, E.; Lum, K.; Johndrow, J.E.; Dunson, D.B. Removing the influence of group variables in high-dimensional predictive modelling. *J. R. Stat. Soc. Ser. Stat. Soc.* **2021**, *184*, 791–811. Available online: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12613 (accessed on 16 October 2023). [CrossRef]
64. Mills, T.C. Modelling skewness and kurtosis in the london stock exchange ft-se index return distributions. *Statistician* **1995**, *44*, 323–332. [CrossRef]
65. Liechty, M.W.; Harvey, C.R.; Liechty, J.C.; Müller, P. Portfolio selection with higher moments. *Quant. Financ.* **2010**, *10*, 469–485. [CrossRef]
66. Mehlawat, M.K.; Gupta, P.; Khan, A.Z. Portfolio optimization using higher moments in an uncertain random environment. *Inf. Sci.* **2021**, *567*, 348–374. Available online: https://www.sciencedirect.com/science/article/pii/S0020025521002565 (accessed on 16 October 2023). [CrossRef]
67. Kang, Y.L.; Tian, J.-S.; Chen, C.; Zhao, G.-Y.; Li, Y.F.; Wei, Y. Entropy based robust portfolio. *Phys. Stat. Mech. Its Appl.* **2021**, *583*, 126260. Available online: https://www.sciencedirect.com/science/article/pii/S0378437121005331 (accessed on 16 October 2023). [CrossRef]
68. Zhou, R.; Yang, Z.; Yu, M.; Ralescu, D.A. A portfolio optimization model based on information entropy and fuzzy time series. *Fuzzy Optim. Decis. Mak.* **2015**, *14*, 381–397. [CrossRef]
69. Park, J. *Finding Bayesian Optimal Portfolios with Skew-Normal Returns*; Elsevier: Rochester, NY, USA, 2021; 48p. [CrossRef]

70. Roberts, C. A correlation model useful in the study of twins. *J. Am. Stat. Assoc.* **1966**, *61*, 1184–1190. Available online: http://www.jstor.org/stable/2283207 (accessed on 16 October 2023). [CrossRef]
71. Loperfido, N. Generalized skew-normal distributions. In *Skew-Elliptical Distributions and Their Applications: A Journey beyond Normality*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2004; pp. 65–80.