Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1–33

Robust Probabilistic Inference via a Constrained Transport Metric

Abhisek Chakraborty*, Anirban Bhattacharya*, and Debdeep Pati*

Abstract. Flexible Bayesian models are typically constructed using limits of large parametric models with a multitude of parameters that are often difficult to interpret. In this article, we offer a novel alternative by constructing an exponentially tilted empirical likelihood carefully designed to concentrate near a parametric family of distributions of choice with respect to a novel variant of the Wasserstein metric, which is then combined with a prior distribution on model parameters to obtain a robustified posterior. The proposed approach finds applications in a wide variety of robust inference problems, where we intend to perform inference on the parameters associated with the centering distribution in the presence of outliers. Our proposed transport metric enjoys great computational simplicity and is inherently parallelizable, exploiting the Sinkhorn regularization for discrete optimal transport problems. We demonstrate superior performance of our methodology when compared against state-of-the-art robust Bayesian inference methods. We also demonstrate the equivalence of our approach with a non-parametric Bayesian formulation under a suitable asymptotic framework, thereby testifying to its flexibility.

Keywords: empirical likelihood, entropy, non-parametric Bayes, robust inference, Wasserstein metric.

1 Introduction

In most modeling exercises, our objective is limited to approximating a few key features of the true data-generating mechanism to ensure interpretable inference. It is often futile, if not misleading, to try to model small-scale and complicated underlying contaminating effects. Thus, the interplay between model adequacy and robustness is a fundamental consideration in model-based inference. Consequently, robust inferential methods (Huber, 2011) possess an influential body of literature, that has permeated many modern areas of research including differential privacy (Dwork and Lei, 2009; Avella-Medina, 2021; Liu et al., 2021), algorithmic fairness (Wang et al., 2020a; Du and Wu, 2021), noise-robust training of deep neural nets (Han et al., 2018; Wang et al., 2020b), sequential decision making (Xu and Mannor, 2010; Chen et al., 2019), transfer learning (Shafahi et al., 2020), quantification learning (Fiksel et al., 2021), to name a few. Bayesian procedures, being almost exclusively model-based, inevitably fall prey to model mis-specification and/or perturbation of the data-generating mechanism – an issue that exacerbates as sample size increases (Miller and Dunson, 2019). Credible intervals obtained from such parametric Bayesian models under model mis-specification

*Department of Statistics, Texas A&M University, College Station, TX, USA, abhisek_chakraborty@tamu.edu

 https://doi.org/10.1214/25-BA1535

may not have the desired asymptotic coverage (Kleijn and van der Vaart, 2012). Non-parametric Bayes methods are routinely used to guard against such mis-specification, either by enlarging the parameter space to impart flexibility or by taking their limit to construct infinite-dimensional prior distributions (Müller and Quintana, 2004; Kleijn and van der Vaart, 2006; De Blasi and Walker, 2013).

Despite the success of non-parametric Bayes methods over the last few decades; see Müller et al. (2015) for a comprehensive review; the presence of a large number of non-identifiable parameters can be contentious, particularly when the interest is solely on simpler features of the population. This has led to a proliferation of pseudo-likelihood-based approaches (Chernozhukov and Hong, 2003; Jiang and Tanner, 2008; Hooker and Vidyashankar, 2011; Minsker et al., 2017; Grünwald and van Ommen, 2017; Miller and Dunson, 2019) targeted towards specific parameters of interest. However, these approaches typically lack generative model interpretations, making the calibration of the associated dispersion or temperature parameters challenging (Holmes and Walker, 2017; Grünwald and van Ommen, 2017).

Empirical likelihood (EL; Owen (2001)), which approximates the underlying distribution with a discrete distribution supported at the observed data points, offers an attractive alternative. Exponentially tilted Empirical Likelihood (ETEL) is a variant of this idea that minimizes the Kullback–Leibler divergence of this discrete distribution with the empirical distribution of the observed data subject to satisfying the estimating equation. Both EL and ETEL obtain the induced maximum likelihood of the parameter of interest defined through estimating equations, by effectively profiling out the nuisance parameters. One can import such likelihoods in a Bayesian framework (Lazar, 2003; Schennach, 2005; Chib et al., 2018, 2021). In fact, posterior credible intervals obtained from ETEL based Bayesian procedures have the correct frequentist coverage (Chib et al., 2021), thereby effectively circumnavigating the longstanding criticism associated with Bayesian inference under model mis-specification.

In this article, our goal is to develop a flexible Bayesian semi-parametric procedure that centers around a postulated parametric family F_θ, without explicitly modeling aspects of the underlying data-generating mechanism that are irrelevant to us. Alternatively, one may consider a non-parametric Bayes procedure (Ferguson, 1973; Teh, 2010; Antoniak, 1974; Lavine, 1994; Verdinelli and Wasserman, 1998), where the parametric guess F_θ (with density f_θ) assumes the role of the base measure, with the precision parameter controlling the extent of concentration around F_θ. However, unlike these approaches, we desire our approach to be devoid of nuisance parameters, and that the inference is solely targeted to the parameter of interest while retaining the interpretation of a generative probability model. In a sense, these goals are similar to that of EL (or ETEL), where one can simply consider the estimating equation $\mathrm{E}[\partial \log f_\theta(X)/\partial\theta] = 0$ to infer about the parameter θ. However, such estimating equations simply enforce specific constraints on the moments of the distribution, and will not be able ensure the vicinity of the distribution near a parametric family F_θ of interest.

In pursuit of this, we propose a novel adaptation of ETEL by centering P around F_θ using a suitable distance metric D, that encapsulates a more holistic discrepancy between the two distributions. More specifically, we restrict P within the neighborhood

$\mathrm{D}[P, F_\theta] < \varepsilon$, for some radius $\varepsilon > 0$. The proposed methodology is termed as D-BETEL. In an inferential task, this framework provides a good balance between modeling flexibility by adaptively tuning ε and interpretability, since we have the provision to invoke a non-parametric likelihood that concentrates around an interpretable parametric guess, where the nuisance parameters are profiled out within the ETEL framework. Naturally, a key ingredient in our proposal is the choice of the metric D that yields a non-trivial distance between F_θ and the empirical distribution on the observed data, and at the same time enjoys computational simplicity and straightforward extension to multivariate cases. Because F_θ is potentially absolutely continuous with respect to the Lebesgue measure and the empirical distribution on the observed data is discrete, it rules out many standard distances, for example, Kullback–Leibler, Hellinger, total variation, χ^2 etc. The 2-Wasserstein metric (Villani, 2003; Panaretos and Zemel, 2019), despite some limitations discussed later, provides a viable choice.

In ensuing applications, other than the choice of the metric, an equally important aspect is to allow the user to select from relatively broader class of distributions F_θ. Although having a fully flexible F_θ defeats the purpose of constructing a procedure devoid of nuisance parameters, we choose to work with elliptical mixture models (EMMs), which offers the user a sufficiently large class to choose from. We also note that the 2-Wasserstein metric does not allow for a computationally efficient multivariate extension, even for EMMs. To this end, we propose an important special case of D-BETEL, with a novel adaptation of the 2-Wasserstein metric by a *restriction* and an *augmentation* scheme. In the *restriction* scheme, we assume F_θ to be an EMM and adapt D by further restricting the coupling measures to the class of EMMs, which considerably reduce the computational cost and yet encompasses a rich class of coupling measures. However, this renders the metric to depend only on the mean and variance-covariance matrix of F_θ, and ignores finer comparison in the tails. We address this in the *augmentation* scheme, where we augment the coupling measure with a product of univariate couplings. This is tantamount to adding a sum of univariate Wasserstein metrics to our adaptation, which effectively captures tail features. Further, the restriction scheme can exploit an entropic regularization of discrete optimal transport (Le et al., 2019; Cuturi, 2013), that remains expressive and computationally tenable even in multivariate cases. The resulting regularized optimal transport metric for EMMs is termed as AugmeNteD and REstricted Wasserstein metric or ANDREW, and is utilized as a convenient metric of choice for the application section.

The rest of the article is organized as follows. Section 2 introduces the proposed Bayesian exponentially tilted empirical likelihood framework with distributional constraints in complete generality, and presents a posterior computation scheme. Section 3 presents an important special case of the proposed D-BETEL methodology, incorporating the elliptical mixture model as the centering family of distribution F_θ and a principled and computationally efficient adaptation of optimal transport as the distance metric D. In Section 4, we investigate the population-level target of inference under the distributionally constrained exponentially tilted empirical likelihood, central to D-BETEL, and examine its robustness properties to model misspecification. In Section 5, we demonstrate that one may view our proposed methodology as a non-parametric Bayes approach based on asymptotic equivalence relationship between our

framework and a hierarchical setup similar to the mixture of finite mixture models. Section 6 presents two applications of the proposed methodology, in model based clustering and generalized linear models. Finally, we conclude with a discussion.

2 D-BETEL: Bayesian ETEL with distributional constraints

Let $\mathcal{S}^{n-1} := \{v \in \mathbf{R}^n : \ v_i \geq 0, \ i = 1, \ldots, n, \ \sum_{i=1}^n v_i = 1\}$ be the $(n-1)$-dimensional unit simplex, and define $\mathcal{H}(v) = -\sum_{i=1}^n v_i \log v_i$ to be the Shannon entropy of a probability vector $v \in \mathcal{S}^{n-1}$, with the standard convention $0 \log 0 = 0$. Let $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ denote the observed data, which are assumed to be i.i.d. samples from a probability distribution P on $\mathcal{X} \subseteq \mathbf{R}^m$. Let $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$ be a parametric family of distributions posited for the x_is. We develop a flexible Bayesian semi-parametric procedure that centers around this parametric family, while allowing for flexible departures from it. The proposed methodology can be extended beyond the i.i.d. setting; see Section 6.2 for an illustration in the regression context. Our approach draws inspiration from the Bayesian exponentially tilted empirical likelihood (Bayesian ETEL or BETEL; Schennach (2005)) for moment condition models.

Moment condition models are specified by a collection of moment conditions $\mathbf{E}_P\big[g(X, \psi)\big] = 0_r$, where $\psi := \psi(P) \in \Gamma$ is the parameter of interest, $g : \mathcal{X} \times \Gamma \to \mathbf{R}^r$ is a vector of known functions or *estimating equations* ($r \geq d$), the expectation $\mathbf{E}_P$ is taken with respect to the unknown generating distribution P, and 0_r refers to a vector of zeros in $\mathbf{R}^r$. Operating under a non-parametric Bayesian framework, Schennach (2005) proposed a flexible prior on P with an entropy-maximizing flavor. Under a specific asymptotic regime that allowed analytic marginalization of nuisance parameters describing the generative model, the corresponding *marginal* posterior distribution of ψ, given a random sample $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ from P, was shown to approach a limiting distribution. This is called the BETEL posterior, given by,

$$\pi_{\mathrm{MCM}}(\psi \mid x_1, \ldots, x_n) \ \propto \ \pi(\psi) \ L_{\mathrm{MCM}}(\psi), \tag{2.1}$$

where the 'likelihood' $L_{\mathrm{MCM}}(\cdot)$ is called the exponentially tilted empirical likelihood,

$$L_{\mathrm{MCM}}(\psi) = \prod_{i=1}^n w_i^\star(\psi), \quad w^\star(\psi) = \left\{ \arg\max_{w \in \mathcal{S}^{n-1}} \mathcal{H}(w) \ : \ \sum_{i=1}^n w_i g(x_i, \theta) = 0 \right\}, \tag{2.2}$$

and $\pi(\cdot)$ denotes a prior distribution on Γ. Here and elsewhere, we use MCM as an acronym for *moment condition model*.

The maximization problem in (2.2) admits a non-trivial closed-form solution when the convex hull of $\cup_{i=1}^n g(x_i, \psi)$ contains the origin, leading to $L_{\mathrm{MCM}}(\psi) = \prod_{i=1}^n w_i^\star(\psi)$, with

$$w_i^\star(\psi) = \frac{\exp[\lambda(\psi)^{\mathrm{T}} g(x_i, \psi)]}{\sum_{j=1}^n \exp[\lambda(\psi)^{\mathrm{T}} g(x_j, \psi)]}, \quad \lambda(\psi) = \arg\min_{\eta} n^{-1} \sum_{i=1}^n \exp[\eta^{\mathrm{T}} g(x_i, \psi)].$$

When the convex hull condition is not satisfied, $\pi_{\text{MCM}}(\psi \mid x_1, \ldots, x_n)$ is set to zero. The minimization problem defining $\lambda(\psi)$ is convex, which leads to efficient computation of the ETEL likelihood L_{MCM}, and the corresponding BETEL posterior π_{MCM} can be sampled using standard Markov Chain Monte Carlo (MCMC) procedures. Chib et al. (2018) significantly contributed towards the theoretical underpinning of BETEL for moment condition models, proving Bernstein–von Mises (BvM) theorems and model selection consistency results under model mis-specification. They also numerically displayed its utility in wide-ranging econometric and statistical applications.

The feature of BETEL most relevant to our purpose is that while motivated from a non-parametric Bayesian angle, it *operationally* avoids a complete probabilistic specification of the data-generating mechanism. That is, the user only needs to specify a prior distribution on the parameter of interest ψ. In a similar spirit, our goal is to avoid a full non-parametric modeling of the data-generating distribution, and only place a prior distribution on the (typically low-dimensional) parameter θ describing the centering model F_θ. A direct application of the ETEL to our setup is challenging as moment conditions describing parameters of general parametric models, especially those beyond exponential families, can be quite cumbersome or even unavailable in an analytically tractable form. Moreover, there is no unique way of describing a finite number of moment restrictions describing F_θ. Instead, our approach is to design a modified likelihood by constraining a weighted empirical distribution of the observed data

$$\nu_{w,x} := \sum_{i=1}^{n} w_i \delta_{x_i} \tag{2.3}$$

to be close to the parametric model F_θ with respect to a statistical metric. Specifically, we propose a likelihood function

$$L_{\text{DCM}}(\theta) := \prod_{i=1}^{n} w_i^\star(\theta), \quad w^\star(\theta) = \left\{ \underset{w \in \mathcal{S}^{n-1}}{\arg\max}\, \mathcal{H}(w) \,:\, \mathrm{D}[F_\theta, \nu_{w,x}] \leq \varepsilon \right\}, \tag{2.4}$$

where $\mathrm{D}[\cdot,\cdot]$ is an appropriate statistical distance, $\varepsilon > 0$ is a concentration parameter which controls fidelity to the centering model, and DCM is an acronym for *distributionally constrained model*. With this DCM likelihood, and a prior distribution on the parameter θ, the corresponding posterior distribution is

$$\pi(\theta \mid x_1, \ldots, x_n) \;\propto\; \pi(\theta)\; L_{\text{DCM}}(\theta). \tag{2.5}$$

We refer to our formulation in (2.4) – (2.5) as the Bayesian ETEL subject to distributional constraint (D-BETEL).

The idea of centering the distribution of the observed data around a pre-specified parametric model is not new. In fact, the Dirichlet process prior (Ferguson, 1973; Teh, 2010) in Bayesian non-parametric is exactly designed to achieve this. Other related approaches include Antoniak (1974); Lavine (1994); Verdinelli and Wasserman (1998). Notably, the traditional non-parametric priors are often accompanied by a large number of uninterpretable nuisance parameters that result in a computational overhead. On

the contrary, D-BETEL directly arrives at a marginal posterior of the parameter of interest in a principled manner. We show in Section 5 that the D-BETEL posterior arises organically from a non-parametric Bayes model by marginalization of the nuisance parameters specifying a mixing measure, which has a mixture of finite mixtures (MFM; Miller and Harrison (2018)) interpretation. Therefore, D-BETEL inherits the flexibility of Bayesian nonparametric models while operationally being devoid of high-dimensional nuisance parameters.

A key ingredient in our proposal is the metric D. A basic requirement is that D (i) returns a non-trivial distance between a discrete and a continuous distribution. For the ensuing applications we also require that D: (ii) allows a straightforward multivariate extension, (iii) is computationally feasible and efficient, and (iv) effectively captures the discrepancies at the tail of the distributions. The requirement (i) itself rules out the applicability of many popular statistical distances/divergences like the Kullback–Leibler divergence, Hellinger distance, total variation distance, χ^2 distance, etc. The Cramer–von Mises metric on $\mathbf{R}$ satisfies (i), (iv), but its multivariate extension is not immediate. The p-Wasserstein metric (Villani, 2003) satisfies (i), (ii), and (iv), and is an attractive candidate. To discuss this further, we recall some relevant facts about the p-Wasserstein metric first.

Definition 1. *For $p \geq 1$, the Wasserstein space $\mathbf{P}_p(\mathbf{R}^d)$ is defined as the set of probability measures μ with finite moment of order p, that is, $\{\mu : \int_{\mathbf{R}^d} \|x\|^p \, d\mu(x) < \infty\}$, where $\|\cdot\|$ is the Euclidean norm on $\mathbf{R}^d$.*

Definition 2. *For $p_0, p_1 \in \mathbf{P}_p(\mathbf{R}^d)$, let $\mathcal{C}(p_0, p_1) \subset \mathbf{P}_p(\mathbf{R}^d \times \mathbf{R}^d)$ 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 p-Wasserstein distance W_p between p_0 and p_1 is defined as $W_p^p(p_0, p_1) = \inf_{\nu \in \mathcal{C}(p_0, p_1)} \int_{\mathbf{R}^d \times \mathbf{R}^d} \|y_0 - y_1\|^p \; d\nu(y_0, y_1)$.*

If both $p_0 \equiv F_\theta$ and $p_1 \equiv \nu_{w,x}$ belong to $\mathbf{P}_p(\mathbf{R})$ with quantile functions F_0^{-1}, F_1^{-1} respectively, we have tractable expression (Panaretos and Zemel, 2019) $W_p^p(p_0, p_1) = \int_{[0,1]} \left[F_0^{-1}(q) - F_0^{-1}(q)\right]^p dq$. However, such closed-form expressions beyond one dimension are unavailable. Fortunately, numerical approximations for the case $p = 2$ (i.e., the W_2 metric) are ubiquitous (Taskesen et al., 2022; Cuturi, 2013; Delon and Desolneux, 2020), even beyond one dimension. In particular, semi-discrete optimal transport schemes (Taskesen et al., 2022), that compute the W_2 distance between a discrete and a potentially continuous probability measure, are broadly applicable to our problem, and approximate numerical algorithm are available in the literature (Mirebeau, 2015; Kitagawa et al., 2017; Gerber and Maggioni, 2017). One may ensure further computational efficiency of D-BETEL, while maintaining fidelity towards required statistical considerations, via principled adaptations of the W_2 metric applicable for judiciously chosen family of F_θ; refer to Section 3 for a special case.

Another key ingredient of the D-BETEL mechanism is the hyper-parameter ε which determines how tightly $\nu_{w,x}$ sits around F_θ with respect to D. Clearly, it bears similarity to the concentration parameter in a Dirichlet process (Ferguson, 1973; Teh, 2010) which dictates fidelity to the base measure. While there is substantial literature on

tuning the concentration parameter of the Dirichlet process mixture model (Escobar and West, 1995; Ishwaran and Zarepour, 2000; McAuliffe et al., 2006), it still remains a notoriously difficult task. Since the nuisance parameters are effectively marginalized out in D-BETEL, we adopt a simple predictive approach to devise a data-driven and principled tuning scheme in Section 5. We now outline a posterior computation scheme for sampling from the D-BETEL posterior.

Posterior computation

We sample from the D-BETEL posterior $\pi(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta)\, L_{\mathrm{DCM}}(\theta)$ via the Metropolis–Hastings (MH) algorithm in all our examples, with carefully-designed proposal schemes for the parameter of interest θ. Refer to Sections 6.1 and 6.2 for specific instances of the sampler in model-based clustering, and generalized linear regression exercises, respectively. Implementation of any MH algorithm requires evaluation of the 'likelihood' L_{DCM}, which we discuss below. Throughout the remainder of this section, we assume the availability of an efficient numerical implementation of D that facilitates the posterior sampling scheme proposed in the sequel.

We undertake a closer look at the constrained entropy maximization problem at the core of our likelihood formulation in (2.4), and note that $\log \prod_{i=1}^{n} w_i^{-w_i} = \log n - \sum_{i=1}^{n} w_i \log(w_i/(1/n))$. For fixed $\theta \in \Theta$, solving the above maximization problem in (2.4) is equivalent to finding the probability vector $(w_1, \ldots, w_n)$ that minimizes the Kullback–Leibler divergence between the probabilities $w_1, \ldots, w_n$ assigned to each sample and the empirical probabilities $1/n, \ldots, 1/n$, subject to the distance constraint $\mathrm{D}[F_\theta, \nu_{w,x}] < \varepsilon$. Unfortunately, unlike the case for L_{MCM} (2.2), the optimization problem for L_{DCM} (2.4) does not allow a closed form solution. However, we can access augmented Lagrangian methods (Conn et al., 1991; Birgin and Martínez, 2008) and conic solvers (Becker et al., 2011) via the R interface (R Core Team, 2022) of constrained non-linear optimization solvers (for example, NLopt, Johnson (2022) and CVX, Grant and Boyd (2008)). In particular, for fixed $\theta \in \Theta$ and $\varepsilon > 0$, we can express the non-linear programming problem in (2.4) in standard form as:

$$\min_{w \in \mathcal{S}^{n-1}} \; -\mathcal{H}(w), \quad \text{subject to} \quad \mathrm{D}[F_\theta, \nu_{w,x}] \leq \varepsilon.$$

The associated Lagrangian function $\mathcal{L} : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}$ is defined as

$$\mathcal{L}(w, \lambda^*) = -\mathcal{H}(w) + \lambda^\star \, \mathrm{D}[F_\theta, \nu_{w,x}], \tag{2.6}$$

where $\lambda^\star$ is the Lagrange multiplier, and the Lagrange dual function $v : \mathbf{R} \to \mathbf{R}$ takes the form

$$v(\lambda^\star) = \inf_{w \in \mathcal{S}} \mathcal{L}(w, \lambda^\star). \tag{2.7}$$

This dual formulation enables us to access off-the-shelf augmented Lagrangian based optimization algorithms (Conn et al., 1991; Birgin and Martínez, 2008) to compute L_{DCM}. This completes the description of our posterior computation scheme.

We conclude the section by noting that utilizing off-the-shelf algorithms for semi-discrete optimal transport towards calculating the D-ETEL likelihood can still be prohibitive, since solving semi-discrete optimal transport problems are at least $\#P$-hard (Taskesen et al., 2022) and each evaluation of D-BETEL likelihood involves repeated computation of D. This motivates the need for a specialized adaptation of the W_2 metric and judicious choice of the centering family of distributions F_θ. To that end, the next section is devoted to a important special case of our proposed D-BETEL methodology, that remains computationally feasible across the ensuing applications.

3 D-BETEL with an AugmeNteD and REstricted Wasserstein metric (ANDREW) for Elliptical Mixture Models (EMM)

In this section, we present an important special case of the proposed D-BETEL methodology, introducing a principled and computationally efficient adaptation of W_2 as the distance metric D for a specific choice of the centering parametric family F_θ. Specifically, we restrict the centering family F_θ to Elliptical Mixture Models (EMM), which form a flexible family of generative models. EMMs notably include Gaussian and t location-scale mixtures, which can approximate a broad variety of density shapes, including multi-modality and skewness.

Given a center $m \in \mathbf{R}^d$, a positive (semi-)definite scale matrix $\Sigma \in \mathbf{R}^{d\times d}$, and a generator function $h : [0,\infty) \to (0,\infty)$, the elliptical distribution $\mathrm{ED}_h(m,\Sigma)$ is defined to be the distribution with characteristic function

$$t \to \exp(it^{\mathrm{T}}m)\ h(t^{\mathrm{T}}\Sigma t), \quad t \in \mathbf{R}^d. \tag{3.1}$$

A multivariate Gaussian distribution $\mathrm{N}_d(m,\Sigma)$ has characteristic function of the form in (3.1) with $h(z) = \exp(-z/2)$ for $z > 0$. Elliptical distributions (Muirhead, 2005) correspond to general non-negative functions h, and include multivariate normal and t distributions as special examples. A discrete mixture of such elliptical distributions, $\sum_{k=1}^K s_k\,\mathrm{ED}_h(m_k,\Sigma_k)$ with $s = (s_1,\ldots,s_K) \in \Delta^{K-1}$, provides a flexible tool for statistical modeling (Cambanis et al., 1981; Holzmann et al., 2006) and probabilistic embedding of complex objects (Muzellec and Cuturi, 2018; Le et al., 2019). Consequently, such elliptical mixture models (EMM) serve as an attractive candidate for our parametric centering family F_θ, with $\theta = \left(s, \{m_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K\right)$ in the most general case. In practice, one may assume additional structure by fixing K or setting $\Sigma_k = \Sigma$ for all k, or posit additional structure on Σ_k (such as $\Sigma_k = \sigma_k^2 I_d$), etc.

To proceed further, we introduce some notations. Let $p_0, p_1 \in \mathbf{P}_p(\mathbf{R}^d)$ be two Gaussian Mixture Models (GMMs), denoted by

$$p_j = \sum_{k=1}^{K_j} s_{jk}\mathrm{N}(m_{jk},\Sigma_{jk}), \quad \mathbf{s}_j = (s_{j1},\ldots,s_{jK_j})^\top \in \mathcal{S}^{K_j-1}, \quad j = 0,1.$$

In the context of this article, p_0 corresponds to the centering distribution F_θ and p_1 corresponds to $\nu_{w,x}$, as defined in equation (2.3), viewed as a limiting case of a GMM

with $K_1 = n$. The W_2 distance between p_0 and p_1 does not admit a closed form expression. An interesting line of work (Delon and Desolneux, 2020; Bion-Nadal and Talay, 2019; Cuturi, 2013) has emerged that modifies the W_2 metric, defined in (2), via restricting the class of coupling measures $\mathcal{C}(p_0, p_1)$ to a carefully chosen sub-family, which considerably reduces the computational cost and yet encompasses a rich class of coupling measures. In particular, Delon and Desolneux (2020) defined a modified W_2 metric, denoted by MW_2, by considering a restricted class of coupling measures $\mathcal{C}_{\mathrm{GMM}} := \big\{\mathcal{C}(p_0, p_1) \cap \mathrm{GMM}_{2d,K_0,K_1}\big\}$, where

$$\mathrm{GMM}_{2d,K_0,K_1} = \Big\{ \sum_{k=1}^{K_0}\sum_{l=1}^{K_1} \pi_{kl}\ \mathrm{N}(b_{kl}, \Omega_{kl}) : \pi_{kl} \geq 0,\ \sum_{k=1}^{K_0}\sum_{l=1}^{K_1} \pi_{kl} = 1 \Big\},$$

denotes the collection of all $(K_0 \times K_1)$-component mixture of Gaussian distributions on $\mathbf{R}^{2d}$. This relaxation enables them to obtain a tractable expression for $\mathrm{MW}_2(p_0, p_1)$. Particularly, to the interest of the current article, we set $p_0 \equiv \sum_{k=1}^{K_0} s_{0k}\mathrm{N}(m_{0k}, \Sigma_{0k})$ and $p_1 \equiv \sum_{k=1}^{K_1} s_{1k}\delta_{m_{1k}}$. Suppose $M = ((||m_{0k} - m_{1l}||^2)) \in \mathbf{R}^{K_0 \times K_1}$ denotes the quadratic cost matrix and $\Pi = ((\pi_{kl})) \in \mathbf{R}^{K_0 \times K_1}$ denotes the weight matrix, then

$$\mathrm{MW}_2^2(p_0, p_1) : = \inf_{\nu \in \mathcal{C}_{\mathrm{GMM}}} \mathbf{E}_\nu ||X_0^\star - X_1^\star||^2 = \inf_{\Pi \in \mathcal{C}(s_0, s_1)} \big\langle \Pi, M \big\rangle + \sum_{k=1}^{K_0} s_{0k}\mathrm{tr}(\Sigma_{0k}).$$

One may equip the D-BETEL methodology with the $\mathrm{D} \equiv \mathrm{MW}_2^2$. However, MW_2^2 suffers from two key limitations.

First, the expression involves the discrete optimal transport problem $\inf_{\Pi \in \mathcal{C}(s_0,s_1)} \big\langle \Pi, M \big\rangle$, that requires a cubic time complexity when solved via traditional simplex or interior-point methods. Fortunately, one may circumnavigate the computational challenge by appealing to Cuturi (2013), that proposed a entropic restriction on $\mathcal{C}(s_0, s_1)$, that in turn introduces entropic regularization term to the discrete optimal transport objective. This entropic regularization term make the discrete optimal transport problem strictly convex, and ensures that one can access linear convergence via Sinkhorn's fixed point iterations. In a shared spirit, one may restrict $\mathrm{GMM}_{2d,K_0,K_1}$ further to $\mathrm{GMM}^{\alpha}_{2d,K_0,K_1}$ by imposing the entropic restriction on $\mathcal{C}(s_0, s_1)$ as follows:

$$\Big\{ \Pi = ((\pi_{kl})) \in \mathbf{R}^{K_0 \times K_1} : \Pi 1_{K_1} = s_0, \Pi^{\mathrm{T}} 1_{K_0} = s_1, D_{\mathrm{KL}}\big[\Pi \mid\mid s_0 s_1^{\mathrm{T}}\big] \leq \alpha \Big\},$$

and define a collection of couplings $\mathcal{C}_{\mathrm{GMM},\alpha} = \big\{\mathcal{C}(p_0, p_1) \cap \mathrm{GMM}^{\alpha}_{2d,K_0,K_1}\big\}$. Particularly, to the interest of the current article, if $p_0 \equiv \sum_{k=1}^{K_0} s_{0k}\mathrm{N}(m_{0k}, \Sigma_{0k})$, $p_1 \equiv \sum_{k=1}^{K_1} s_{1k}\delta_{m_{1k}}$, we get a computationally convenient adaptation of $\mathrm{MW}_2^2(p_0, p_1)$ as follows

$$\mathrm{MW}^2_{2,\alpha}(p_0, p_1) := \inf_{\nu \in \mathcal{C}_{\mathrm{GMM},\alpha}} \mathbf{E}_\nu ||X_0^\star - X_1^\star||^2 = \inf_{\Pi \in \mathcal{C}(s_0, s_1)} \Big[\big\langle \Pi, M \big\rangle - \frac{1}{\lambda_\alpha}\mathcal{H}(\Pi) \Big] + \sum_{k=1}^{K_0} s_{0k}\mathrm{tr}(\Sigma_{0k}),$$

where λ_α depends on α and $\mathcal{H}(\Pi) = -\sum_{k,l} \pi_{kl} \log \pi_{kl}$.

The second key limitation of $\mathrm{MW}_2(p_0, p_1)$ is that its expression relies solely on the first and second-order moments of F_θ, rendering it incapable of adequately capturing the tail behavior. Even if we replace the GMMs by a versatile class of EMMs and restrict the class of coupling measures $\mathcal{C}(p_0, p_1)$ to $\mathcal{C}_{\mathrm{EMM},\alpha} = \mathcal{C}(p_0, p_1) \cap \mathrm{EMM}^{\alpha}_{2d,K_0,K_1}$, then also

$$\begin{aligned}\mathrm{MW}^2_{2,\mathrm{EMM},\alpha}(p_0, p_1) :&= \inf_{\nu \in \{\mathcal{C}_{\mathrm{EMM},\alpha}\}} \mathbf{E}_\nu ||X_0^\star - X_1^\star||^2 \\ &= \inf_{\Pi \in \mathcal{C}(s_0, s_1)} \left[\langle \Pi, M \rangle - \frac{1}{\lambda_\alpha} \mathcal{H}(\Pi) \right] + \nu_h \sum_{k=1}^{K_0} s_{0k} \mathrm{tr}(\Sigma_{0k})\end{aligned}$$

only depends on first and second-order moments, and we fail to capture the tail behavior of F_θ. For example, in Figure 1, let

$$p_0 \equiv \sum_{k=1}^{K_0} s_{0k} t_\eta(m_{0k}, \Sigma_{0k}),\ p_0' \equiv \sum_{k=1}^{K_0} s_{0k} t_{\eta'}(m_{0k}, \Sigma'_{0k}),\ p_1 \equiv \sum_{k=1}^{K_1} s_{1k} \delta_{m_{1k}}$$

and set $\eta' = \eta/m,\ \Sigma'_{0k} = \frac{\eta - 2m}{\eta - 2} \Sigma_{0k}$ for some $m \in \mathbf{Z}^+$ such that the variances of the multivariate t-distributions $t_\eta(m_{0k}, \Sigma_{0k})$ and $t_{\eta'}(m_{0k}, \Sigma'_{0k})$ match for $k = 1, \ldots, K_0$. Then, $\mathrm{MW}^2_{2,\mathrm{EMM},\alpha}(p_0, p_1) = \mathrm{MW}^2_{2,\mathrm{EMM},\alpha}(p_0', p_1)$, despite the differences in the tail behaviors of p_0 and p_1. In Supplementary Section 5 (Chakraborty et al., 2025), we compare D-BETEL with $\mathrm{D} = \mathrm{MW}_2$ and $\mathrm{D} = W_{\mathrm{AR}}$, proposed in the sequel, on generalized linear regression task.

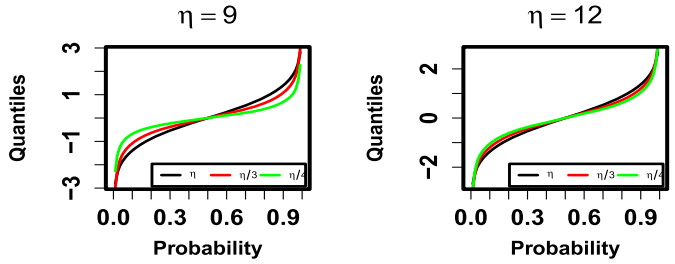

Figure 1: *The plots show that $MW_{2,\mathrm{EMM},\alpha}$ fails to capture the differences in the tail behavior of the probability distributions. Let $p_0 \equiv \sum_{k=1}^{K_0} s_{0k} t_\eta(m_{0k}, \Sigma_{0k})$, $p_0' \equiv \sum_{k=1}^{K_0} s_{0k} t_{\eta'}(m_{0k}, \Sigma'_{0k})$, $p_1 \equiv \sum_{k=1}^{K_1} s_{1k} \delta_{m_{1k}}$ and set $\eta' = \eta/m$, $\Sigma'_{0k} = \frac{\eta-2m}{\eta-2}\Sigma_{0k}$ for $m \in \{3, 4\}$, such that the variances of the multivariate t-distributions $t_\eta(m_{0k}, \Sigma_{0k})$ and $t_{\eta'}(m_{0k}, \Sigma'_{0k})$ match for $k = 1, \ldots, K_0$. Then, $MW^2_{2,\mathrm{EMM},\alpha}(p_0, p_1) = MW^2_{2,\mathrm{EMM},\alpha}(p_0', p_1)$, despite p_0 and p_1 being different probability distributions. In this plot, the two panels correspond to $\eta \in \{9, 12\}$, respectively.* Since the expression of W_{AR}, proposed in the sequel, additionally involves the marginal quantiles given by $\sum_{k=1}^{d} \int_0^1 \left(F_{0k}^{-1}(z) - F_{1k}^{-1}(z)\right)^2 dz$, it is capable of capturing the difference in the tail due to the different d.f. of the t.

Proposed metric

In view of the discussion so far in this section, we finally describe a new and carefully crafted strategy based on augmentation succeeded by restriction in the space of coupling measures that yield a transport metric that not only automatically inherits computational tractability of $\mathrm{MW}^2_{2,\alpha}(p_0, p_1)$, and is capable of accessing improved computational algorithms based on an entropic regularization of the discrete optimal transport (Cuturi, 2013), but also remains expressive. In essence, our novel strategy presents a general recipe for devising increasingly expressive transport metrics and describing a corresponding modified class of couplings, of which ANDREW introduced next is a specific example. To that end, we *augment* the class of coupling measures $\mathcal{C}(p_0, p_1) \subset \mathbf{P}_p(\mathbf{R}^d \times \mathbf{R}^d)$ into a class of coupling measures $\mathcal{C}(p_0^\star, p_1^\star) \subset \mathbf{P}_p(\mathbf{R}^{2d} \times \mathbf{R}^{2d})$, and then *restrict* $\mathcal{C}(p_0^\star, p_1^\star)$ to a carefully chosen sub-class of couplings. We describe the details below.

Definition 3. *Let $p_0, p_1 \in \mathbf{P}_p(\mathbf{R}^d)$ with $p_j = \sum_{k=1}^{K_j} s_{jk} ED_h(m_{jk}, \Sigma_{jk}), (j = 0, 1)$. Next, we shall consider an augmentation followed by a restriction scheme as follows:*
(a) Augmentation: Define probability distribution $p_0^\star \in \mathbf{P}_2(\mathbf{R}^{2d})$ as

$$p_0^\star := p_0 \otimes \widetilde{p}_0, \ \text{ with } \widetilde{p}_0 := p_{01} \otimes \ldots \otimes p_{0d}$$

and p_{0i} the i-th marginal of p_0. Clearly, if $X_0 \sim p_0$, and $\widetilde{X}_0$ independent of X_0 is distributed as $\widetilde{p}_0$, then $X_0^\star = (X_0, \widetilde{X}_0)^{\mathrm{T}} \sim p_0^\star$. Similarly, define $p_1^\star$. By construction we have

$$\mathcal{C}(p_0^\star, p_1^\star) = \mathcal{C}(p_0, p_1) \otimes \mathcal{C}(\widetilde{p}_0, \widetilde{p}_1) = \mathcal{C}(p_0, p_1) \otimes \big\{ \otimes_{i=1}^d \mathcal{C}(p_{0i}, p_{1i}) \big\}.$$

(b) Restriction: Suppose

$$\mathrm{EMM}_{2d,K_0,K_1} = \left\{ \sum_{k=1}^{K_0} \sum_{l=1}^{K_1} \pi_{kl} \ \mathrm{ED}_h(b_{kl}, \Omega_{kl}) : \pi_{kl} \geq 0, \ \sum_{k=1}^{K_0} \sum_{l=1}^{K_1} \pi_{kl} = 1 \right\}$$

denote the collection of all $(K_0 \times K_1)$-component mixture of identifiable elliptical distributions on $\mathbf{R}^{2d}$. Define a subset $\mathrm{EMM}^\alpha_{2d,K_0,K_1}$ of $\mathrm{EMM}_{2d,K_0,K_1}$ by imposing the entropic restriction

$$D_{\mathrm{KL}}\big[\Pi \ || \ s_0 s_1^{\mathrm{T}}\big] \leq \alpha, \ \text{ where } \Pi = ((\pi_{kl})) \in \mathbf{R}^{K_0 \times K_1}$$

is the joint probability matrix of the mixture weights, and s_0, s_1 are the respective marginals, that is, $\Pi 1_{K_1} = s_0, \Pi^{\mathrm{T}} 1_{K_0} = s_1$. Finally, define a collection of couplings $R^\alpha(p_0^\star, p_1^\star) \subset \mathcal{C}(p_0^\star, p_1^\star)$ as

$$R^\alpha(p_0^\star, p_1^\star) = \mathcal{C}_{\mathrm{EMM},\alpha} \otimes \big\{ \otimes_{i=1}^d \mathcal{C}(p_{0i}, p_{1i}) \big\}, \quad \mathcal{C}_{\mathrm{EMM},\alpha} = \mathcal{C}(p_0, p_1) \cap \mathrm{EMM}^\alpha_{2d,K_0,K_1}.$$

Refer to Figure 2 for a schematic representation of the proposed augmentation and restriction strategy. With these notations in place, we define ANDREW as

$$W^2_{\mathrm{AR}}(p_0, \ p_1) = \inf_{\nu \in R^\alpha(p_0^\star, p_1^\star)} \mathbf{E}_\nu ||X_0^\star - X_1^\star||^2. \tag{3.2}$$

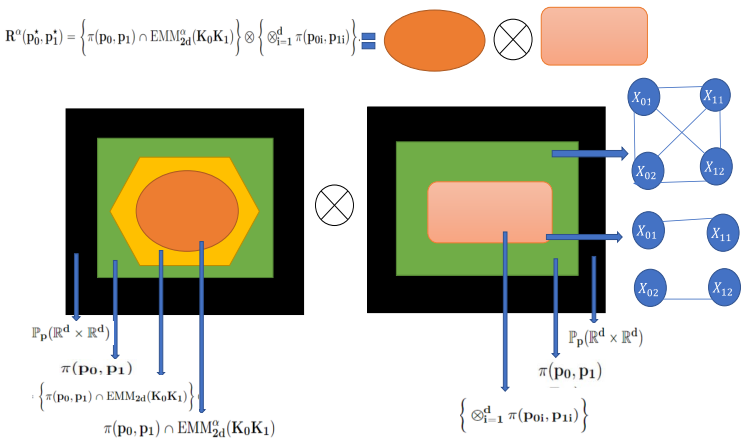

Figure 2: The augmentation and restriction scheme to construct $R^{\alpha}(p_0^{\star}, p_1^{\star})$, left: construction of $\left\{\mathcal{C}_{\mathrm{EMM},\alpha} = \mathcal{C}(p_0, p_1) \cap \mathrm{EMM}^{\alpha}_{2d,K_0,K_1}\right\}$, right: construction of $\left\{ \otimes_{i=1}^{d} \mathcal{C}(p_{0i}, p_{1i})\right\}$ with $d = 2$.

To cater to our original goal of centering the D-BETEL around an EMM, we now present a simplified form of W_{AR} for the case when one of p_0, p_1 is discrete.

Theorem 1. *Suppose* $p_0 \equiv \sum_{k=1}^{K_0} s_{0k}\mathrm{ED}_h(m_{0k}, \Sigma_{0k})$, $p_1 \equiv \sum_{k=1}^{K_1} s_{1k}\delta_{m_{1k}}$, *and* $M = ((||m_{0k} - m_{1l}||^2)) \in \mathbf{R}^{K_0 \times K_1}$ *be the quadratic cost matrix. Then, there exists* λ_{α} *depending on* α *such that,* $W^2_{\mathrm{AR}}(p_0, p_1) =$

$$\inf_{\Pi \in \mathcal{C}(s_0, s_1)} \left[\langle \Pi, M\rangle - \frac{1}{\lambda_{\alpha}} H(\Pi)\right] + \nu_h \sum_{k=1}^{K_0} s_{0k}\mathrm{tr}(\Sigma_{0k}) + \sum_{k=1}^{d} \int_0^1 \left(F_{0k}^{-1}(z) - F_{1k}^{-1}(z)\right)^2 dz,$$

where $\langle \Pi, M\rangle = \mathrm{tr}(\Pi^{\mathrm{T}} M)$, $H(\Pi) = -\sum_{k,l} \pi_{kl} \log \pi_{kl}$ *and* $F_{jk}^{-1}(\cdot)$ *is the quantile function of* X_{jk}.

We defer the proof and a cascade of required lemmas to Supplementary Section 1, and make some remarks about ANDREW here. We recall that, in the posterior computation scheme for D-BETEL, each evaluation of the likelihood involves several computations of the distance $\mathrm{D}[F_{\theta}, \nu_{w,x}] = W_{\mathrm{AR}}[F_{\theta}, \nu_{w,x}]$. Importantly, the expression above is completely tractable and computationally feasible. The entropic regularization term in W_{AR} makes the discrete optimal transport problem strictly convex, and consequently, it can access linear convergence via Sinkhorn's fixed point iterations (Cuturi, 2013). Secondly, since the expression of W_{AR} additionally involves the marginal quantiles, it is capable of capturing the difference in the tail. We believe the flexibility and the computational simplicity of our novel Wasserstein metric may render itself useful in many optimal transport-based machine learning applications, beyond what we discuss here, see the discussion section for some specific application domains.

4 Population level target of D-ETEL

In this section, we explore the population-level target of inference under the distributionally constrained exponentially tilted empirical likelihood (D-ETEL) mechanism in (2.4), which lies at the heart of D-BETEL, and the robustness it brings under model misspecification. We conduct this exploration via a combination of analytical calculations and numerical simulations. Precisely characterizing the target of estimation for D-ETEL in an analytical fashion is difficult and beyond the scope of this article. The difficulty arises from nonlinearities in the objective function as we discuss below. Even in the EL literature, where the objective function is comparatively simpler, such characterizations are not straightforward; see Section 2.3 of Schennach (2007) for a discussion. In this section, we fix D to be the Wasserstein metric with the squared Euclidean norm (Villani, 2003), denoted by W_2.

Recall that P denotes the unknown data generative model of interest. A proposed parametric class of models $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$ aims to adequately describe P. Suppose F_θ admits a probability density function f_θ. Under model misspecification, that is, when $P \notin \{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$, a statistical procedure targets the best approximation of P within the proposed parametric family $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$ in an appropriate sense. For example, under appropriate regularity conditions, the maximum likelihood estimator of θ, denoted by $\theta^\star$, converges to the KL projection of P within $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$, that is, to

$$\theta^\star = \arg\min_\theta \mathrm{KL}(P||F_\theta) = \arg\min_\theta E_{X\sim P}[-\log f_\theta(X)], \tag{4.1}$$

as the sample size $n \to \infty$ (White, 1982). Moreover, a usual Bayesian posterior also contracts around the same target parameter $\theta^\star$ under similar regularity conditions (Kleijn and van der Vaart, 2006). The proposed D-BETEL procedure replaces the usual parametric likelihood $L(\theta) = \prod_{i=1}^n f_\theta(X_i)$ with the exponentially tilted empirical likelihood $L_{\mathrm{DCM}}(\theta) = \prod_{i=1}^n w_i^\star(\theta)$ in (2.4), where the role of the parametric family F_θ is encapsulated inside the weights $\{w_i^\star(\theta)\}_{i=1}^n$. Therefore, it is natural to investigate which population summary the D-ETEL procedure targets. We offer a characterization below, and provide an empirical illustration of its robustness over the nearest KL point $\theta^\star$.

Inspecting the dual formulation in (2.6), it becomes apparent that the population-level target of D-ETEL can be obtained by the following two-step procedure:: (i) for a given $\theta \in \Theta$, obtain

$$\hat{Q}_\theta = \arg\min_Q \big[\mathrm{KL}(Q \,||\, P) + \lambda^\star \, \mathrm{D}(F_\theta, Q)\big], \tag{4.2}$$

where the minimization is over all probability measures Q. The class of densities $\hat{Q}_\theta$ can be interpreted as an enlargement of F_θ. Then, in step (ii) we set

$$\theta^\dagger = \arg\max_{\theta\in\Theta} E_{X\sim P}[\log \hat{Q}_\theta(X)]. \tag{4.3}$$

This follows since for large sample sizes, P is adequately described by the empirical distribution of the data $\nu_{(1/n,\ldots,1/n)^\mathrm{T},x}$, P_θ takes the form of a weighted empirical distribution $\nu_{w,x}$ as in (2.3) and $\mathrm{KL}(\nu_{w,x} \,||\, \nu_{(1/n,\ldots,1/n)^\mathrm{T},x}) = \log n - (-\sum_{i=1}^n w_i \log w_i)$.

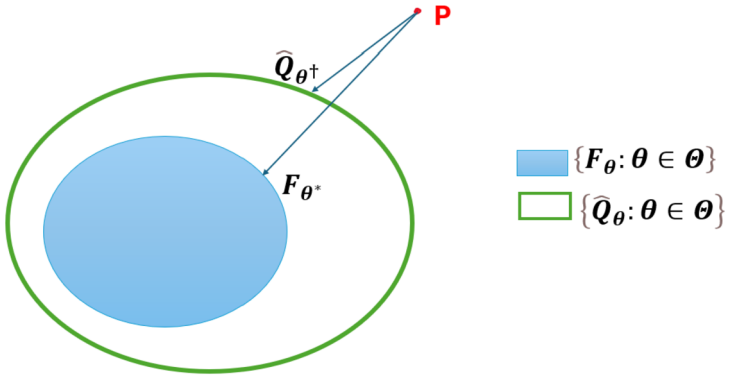

Figure 3: Here P denotes the unknown data generative model of interest. A proposed parametric class of models $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$ aims to adequately describe P. The $\{\hat{Q}_\theta, \theta \in \Theta\}$ should be interpreted as an enlargement of the class $\{F_\theta, \theta \in \Theta\}$, obtained via D-ETEL mechanism. This explains the robustness of the D-ETEL estimator $\theta^\dagger$, compared to the MLE $\theta^\star$, under model misspecification, i.e, when $P \notin \{F_\theta : \theta \in \Theta\}$.

In particular, when $\varepsilon \to 0$, or equivalently $\lambda^\star \to \infty$, we recover the parametric inference based on the family of distributions F_θ. When $\varepsilon > 0$, we can imagine $\{\hat{Q}_\theta, \theta \in \Theta\}$ to be an enlargement of $\{F_\theta, \theta \in \Theta\}$, refer to Figure 3 for a visual representation. In that case, assuming $P \notin \{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$, D-ETEL targets $(\theta^\dagger, P^\dagger := P_{\theta^\dagger})$ such that $\mathrm{KL}(P^\dagger \,||P)$ is minimum and $\theta^\dagger$ lies on the boundary of $\{\theta : \mathrm{D}(F_\theta, P^\dagger) \leq \epsilon\}$.

Note that, the optimization problem in (4.2)-(4.3) does not admit a closed-form solution even for the simplest choices of F_θ, for example, Gaussian. So, in the context of specific examples, we resort to solving the sample version of the optimization problem in (4.2)-(4.3) with an adequately large sample size. Then, we use the θ that maximizes L_{DCM} as a proxy for $\theta^\dagger$. In what follows, we consider an example, where the data is generated from a mildly skewed skew-normal distribution (Azzalini and Dalla Valle, 1996). This imitates a data generation scheme, where the underlying true distribution is univariate normal in the presence of mild contamination. The goal is to compute the target of estimation for the maximum likelihood procedure and for D-ETEL with F_θ as univariate normal distribution.

Example 1 (Skew-normal distribution). Motivated by the model based clustering example in Miller and Dunson (2019); see also Cai et al. (2020a); we generate a sample of size $n = 500$ from a univariate skew-normal distribution (Azzalini and Dalla Valle, 1996) with pdf $f(x) = 2\phi(x)\Phi(\alpha x)$, $x \in \mathbf{R}$ and the skewness parameter $\alpha \neq 0$. To model the generated data, we first consider the parametric class of models $F_{\mu,\sigma^2} = \{\mathrm{N}(\mu, \sigma^2), \mu \in \mathbf{R}, \sigma^2 > 0\}$ and compute the maximum likelihood estimate of (μ, σ^2). This involves calculating the best KL projection of the empirical distribution of the observed data within the parametric class F_{μ,σ^2}. As an alternative, we also consider computing the D-ETEL estimate of (μ, σ^2) with the centering family of distri-

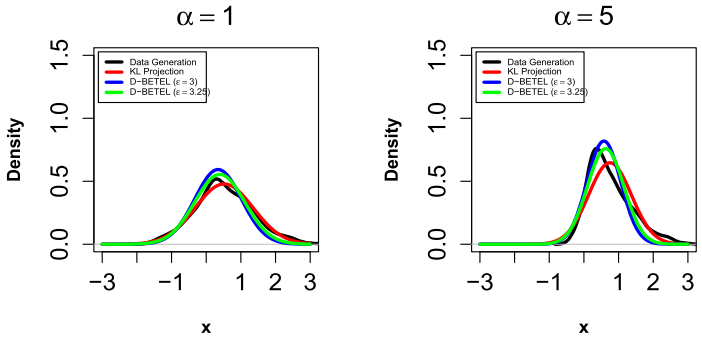

Figure 4: **Skew-normal distribution.** We plot the density estimates $f_{\theta^\star}$ corresponding to the MLE in (4.1) and $f_{\theta^\dagger}$ corresponding to the D-ETEL estimator in (4.3), where the data is generated from a univariate skew normal distribution (Azzalini and Dalla Valle, 1996) and the parametric family of interest is $F_{\mu,\sigma^2} = \{\mathrm{N}(\mu,\sigma^2), \mu \in \mathbf{R}, \sigma^2 > 0\}$. In particular, we generate a sample of size $n = 500$ from a univariate skew-normal distribution with pdf $f(x) = 2\phi(x)\Phi(\alpha x)$, $x \in \mathbf{R}$, with varying values of the skewness parameter $\alpha \in \{1, 5\}$. The plot consider D-ETEL estimates for varying value of $\varepsilon \in \{3, 3.5\}$. The specific values of ε is guided by the hyper-parameter tuning scheme, introduced in (5.4). In particular, the considered values of ε correspond to the two highest κ's in (5.4). For moderate skewness ($\alpha = 1$), the MLE of (μ, σ^2) or $\theta^\star$ is $(0.52, 0.70)$, and D-ETEL estimates with $\varepsilon = 3, 3.25$ are $(0.34, 0.45)$ and $(0.39, 0.51)$, respectively. For larger skewness ($\alpha = 5$), the MLE of (μ, σ^2) or $\theta^\star$ is $(0.75, 0.38)$, and D-ETEL estimates with $\varepsilon = 3, 3.25$ are $(0.58, 0.24)$ and $(0.63, 0.28)$, respectively.

butions F_{μ,σ^2}, based on the equation (2.4). This involves computing the best weighted empirical distribution of the observed data within the ε neighborhood of the parametric class F_{μ,σ^2} with respect to the metric D, as described in (4.2)-(4.3). Figure 4 plots the true skew normal density, and density estimates $f_{(\mu^\star,\sigma^{2\star})}$ corresponding to the MLE and $f_{(\mu^\dagger,\sigma^{2\dagger})}$ corresponding to the D-ETEL estimator for varying values of ε. For moderate skewness ($\alpha = 1$), the normal distributions with parameters estimated via D-ETEL provides a satisfactory description of the data generating mechanism, compared to the normal distributions with parameters estimated via maximum likelihood. For larger skewness ($\alpha = 5$), the normal distributions with parameters estimated via D-ETEL still provides a better visual description of the data generating mechanism, compared to the normal distributions with parameters estimated via maximum likelihood.

Having formally characterized the target of estimation of the D-ETEL method, it is instructive to examine the robustness properties of the proposed framework. For a heuristic justification of the robustness of D-ETEL, readers are referred to the Supplementary Section 2.

5 Non-parametric Bayes interpretation of D-BETEL

Before we move on to the specific applications of D-BETEL, we shall discuss a key feature of our proposal, that we briefly alluded to earlier, in concrete terms. In particular, we demonstrate that one may view our proposed methodology as a non-parametric Bayes approach based on centering mixture models around a specific parametric family by establishing an intriguing asymptotic equivalence relationship between our framework and a hierarchical setup similar to the mixture of finite mixture (MFM) models (Miller and Harrison, 2018). MFM and related non-parametric Bayesian priors (Ferguson, 1973; Antoniak, 1974; Pitman and Yor, 1997; Gnedin, 2009) can be recovered as variants of the popular Gibbs-type priors (Gnedin and Pitman, 2005). Such Gibbs-type priors are characterized by predictive distributions that are a linear combination of the prior guess and a weighted empirical measure. The asymptotic properties of the Gibbs-type priors are extensively studied in De Blasi et al. (2013). In the context of the current article, the asymptotic equivalence of the proposed framework and a hierarchical setup similar to the mixture of finite mixture (MFM) models enables us to formally identify D-BETEL as a generative model – a feature illusive to many existing pseudo-likelihood-based robust Bayesian methods. In particular, we offer a concrete probabilistic justification to D-BETEL by building a Bayesian hierarchical generative model centered around F_θ so that the marginal posterior of θ converges in distribution to the D-BETEL posterior under a limiting environment motivated by Schennach (2005). This result is established in Theorem 2. For notational convenience, we assume that the data dimension m and the parameter dimension d are identical; however, this assumption is not essential for the subsequent analysis.

In the following, we first briefly introduce a generative model for the data points $x_1, \ldots, x_n$ which closely mimics commonly used Bayesian nonparametric methods such as the mixture of finite mixture of Gaussian. The description proceeds via a probability model for the independent d-variate observations x_i conditional on its own set of parameters $\eta_i \in \mathbf{R}^d$, that is, $x_i \mid \eta_i \overset{ind.}{\sim} f(\cdot \mid \eta_i)$ for $i = 1, \ldots, n$. To impart flexibility, the random effects η_i are independently drawn from a common *mixing measure* $P^{(N)}$ defined on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$, where N is a positive integer involved in the description of $P^{(N)}$. This renders the marginal density of $x_i \mid P^{(N)}$ to be $\int f(x_i \mid \eta_i) P^{(N)}(d\eta_i)$, independently for $i = 1, \ldots, n$. The mixing distribution $P^{(N)}$ is parameterized through its associated nuisance parameters $\xi^\star = \big(k, b, \{\mu_h\}_{h=1}^k\big)$, where $k \in \mathbf{N}$, $b = (b_1, \ldots, b_k)$ such that each b_h is a positive integer subject to the constraint $\sum_{h=1}^k b_h = N$, and $\mu_h = (\mu_{h,1}, \ldots, \mu_{h,d})^{\mathrm{T}} \in \mathbf{R}^d$ for $h = 1, \ldots, k$. The details on the construction of the mixing measure $P^{(N)}$ is deferred to the next sub-section. Next, we induce a prior distribution on $P^{(N)}$ through a prior distribution on $\xi^\star$. To do so, we construct a joint prior on $(\xi^\star, \theta)$ hierarchically by first specifying the marginal prior on the parameter of interest θ, and then the conditional prior of $\xi^\star \mid \theta$ in terms of a θ dependent slice on the support of an unconditional distribution $\pi_{\infty,N}(\cdot)$ for $\xi^\star$. In essence, ξ^* act as a *bridge* between the data and the parameter of interest θ in the hierarchical formulation. This is where our modeling departs from a typical non-parametric Bayes model, where $P^{(N)}$ is the object of inference and θ is viewed as a derived quantity from $P^{(N)}$. Instead, in our framework, θ retains its own identity and $P^{(N)}$ is viewed as an infinite-dimensional nuisance parameter. In other words, $(P^{(N)}, \theta)$ describes a semi-parametric

object for inference, where $P^{(N)}$ is a flexible probability measure, and θ is the parameter of interest.

A constrained generative mechanism

We specify the details for each of the pieces of the generative model from top down in the sequel. First, the distribution f of the data given random effects is chosen to be an appropriate uniform distribution. Specifically, given $\tau > 0$, let

$$\begin{aligned} x_i \mid \eta_i, P^{(N)} &\overset{\text{ind.}}{\sim} \prod_{l=1}^{d} \text{Uniform}(\eta_{i,l} - \tau^{-1},\ \eta_{i,l} + \tau^{-1}),\ i = 1, \ldots, n, \\ \eta_i \mid P^{(N)} &\sim P^{(N)}, \end{aligned} \tag{5.1}$$

where $\eta_i = (\eta_{i,1}, \ldots, \eta_{i,d})^{\mathrm{T}}, i \in [N]$. The uniform kernel is chosen for analytic tractability in ensuing calculations. We expect the main results to hold for more general kernels, albeit with additional technical challenges.

Next, for any Borel set $A \in \mathcal{B}(\mathbf{R}^d)$, define $P^{(N)}(A)$ as $P^{(N)}(A) = \sum_{h=1}^{k} \pi_h \delta_{\mu_h}(A)$, where conditional on k, the mixture weights $(\pi_1, \ldots, \pi_k)$ are constructed via normalized counts, that is, $(\pi_1, \ldots, \pi_k) = (b_1/N, \ldots, b_k/N)$. We specify distributions on the pieces to define an *unconditional distribution* $\pi_{\infty,N}(\cdot)$ for $\xi^* = (k, \mu_1, \ldots, \mu_k, b_1, \ldots, b_k)^{\mathrm{T}}$,

$$\begin{aligned} (b_1, \ldots, b_k) \mid k &\sim \text{Multinomial}(N; 1/k, \ldots, 1/k) \\ \mu_h \mid k &\overset{\text{i.i.d.}}{\sim} \mathrm{H}^{(N)},\ h = 1, \ldots, k, \quad k \sim p(k) \equiv \text{Geometric}(p), \end{aligned} \tag{5.2}$$

where $\mathrm{H}^{(N)}$ is a suitably chosen d-dimensional "base" distribution, refer to assumptions in Supplementary Section for details. Given a draw of k, the k atoms $\{\mu_h\}_{h=1}^{k}$ are drawn independently from $\mathrm{H}^{(N)}$, and the count vector $(b_1, \ldots, b_k)$ that yields the mixture weights is drawn from Multinomial$(N; 1/k, \ldots, 1/k)$, instead of direct draws of the mixture weights from a Dirichlet distribution, commonly used in the finite dimensional version of the Dirichlet process (Ishwaran and Zarepour, 2002a,b), or in the mixture of finite mixtures setup (Miller and Harrison, 2018). The distributional specification $\pi_{\infty,N}$ for $\xi^\star$ in (5.2) induces a mixture of finite mixtures (MFM; Miller and Harrison (2018)) for $P^{(N)}$ given by $P^{(N)} = \sum_{k=1}^{\infty} p(k) \left[\sum_{h=1}^{k} (b_h/N)\delta_{\mu_h}\right]$. This completes the construction of the mixing measure $P^{(N)}$.

Next, we construct a joint prior on $(\xi^\star, \theta)$ by first specifying a prior distribution $\pi(\cdot)$ on θ, and then the conditional distribution of $\xi^* \mid \theta$ by restricting the distribution $\pi_{\infty,N}(\cdot)$ to the slice $A_{\varepsilon,N}(\theta) := \{\xi^\star : \mathrm{D}(P^{(N)}, F_\theta) < \varepsilon\}$ defined on the support of $\xi^\star$, where the metric D and the scalar $\varepsilon > 0$ are as in (2.2). Thus, $\pi_{\varepsilon,N}(\xi^\star \mid \theta) \propto \pi_{\infty,N}(\xi^\star)\ 1_{A_{\varepsilon,N}(\theta)}(\xi^\star)$, that is, given a specific value of θ, only draws from the unconditional prior $\pi_{\infty,N}$ are retained for which $P^{(N)}$ and F_θ are ε-close under the metric D. The joint prior on $(\xi^\star, \theta)$ can therefore be expressed as

$$\pi_{\varepsilon,N}(\xi^\star, \theta) \propto \pi(\theta)\pi_{\varepsilon,N}(\xi^\star \mid \theta). \tag{5.3}$$

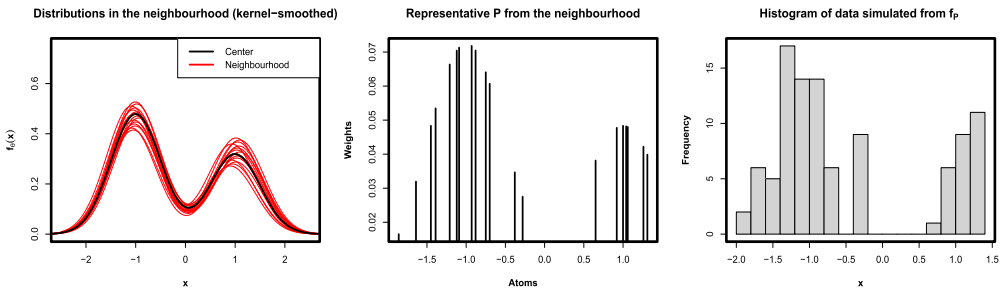

Figure 5: ***Left panel.*** *the pdf* $f_\theta(x)$ *corresponding to the centering distribution* $F_\theta \equiv 0.6 \times N(-1, 0.5^2) + 0.4 \times N(1, 0.5^2)$ *in black and representative discrete distributions in a D-neighborhood (with D chosen as* W_{AR} *introduced in Section 3) of* F_θ *in red after kernel smoothing,* ***Middle panel.*** *one particular* $P = \sum_{h=1}^{k} \pi_h \delta_{\mu_h}$ *with* $k = 20$ *in the* W_{AR}*-neighborhood of* F_θ *with* $W^2_{\text{AR}}(P, F_\theta) = 2.5$*,* ***Right panel.*** *histogram of a random sample of size* 100 *drawn from* $f_P(x) = \int f(x \mid \eta) P(d\eta)$ *with* $\tau = 10^2$ *in equation (5.1).*

This completes our hierarchical specification. Figure 5 presents a schematic of the hierarchical model in equations (5.1)–(5.3).

Combining the joint prior $\pi_{\varepsilon,N}(\xi^\star, \theta)$ with the generative model in (5.1), one obtains the joint posterior distribution $\pi_{\varepsilon,N}(\theta, \xi^\star \mid x_{1:n})$ of $(\theta, \xi^\star)$. A fully Bayesian analysis of the posterior of $\pi_{\varepsilon,N}(\theta, \xi^\star \mid x_{1:n})$ entails traversing the high-dimensional parameter space of $(\xi^\star, \theta)$ to simultaneously learn $(P^{(N)}, \theta)$. Instead, motivated by Schennach (2005), we marginalize $\pi_{\varepsilon,N}(\theta, \xi^\star \mid x_{1:n})$ with respect to nuisance parameters $\xi^\star$ to obtain the marginal posterior $\pi_{\varepsilon,N}(\theta \mid x_{1:n})$, that enables us to access targeted inference on the parameter of interest θ. We shall operate in an asymptotic regime motivated by Schennach (2005), where we let the hyper parameters $\tau \equiv \tau(N), p \equiv p(N)$ and the base-measure $H^{(N)}$ to evolve with N. Under this environment, we show that the marginal posterior $\pi_{\varepsilon,N}(\theta \mid x_{1:n})$ converges to (2.5) as $N \to \infty$.

Theorem 2. *Fix the concentration parameter* $\varepsilon > 0$ *and sample size* n*. Suppose* $H^{(N)}$*,* p *and* τ *satisfy the assumptions stated in Supplementary Section (3.1). Then the marginal posterior* $\pi_{\varepsilon,N}(\theta \mid x_{1:n})$ *defined after (5.1)–(5.3) converges point-wise in* θ *to the D-BETEL posterior in equation (2.5) as* $N \to \infty$*,*

$$|\pi_{\varepsilon,N}(\theta \mid x_{1:n}) - \pi(\theta \mid x_{1:n})| \to 0 \quad \textit{as} \quad N \to \infty.$$

An application of Scheffe's theorem (Resnick, 2013) yields,

$$\|\pi_{\varepsilon,N}(\cdot \mid x_{1:n}) - \pi(\cdot \mid x_{1:n})\|_{\text{TV}} := \frac{1}{2} \int_\theta |\pi_{\varepsilon,N}(\theta \mid x_{1:n}) - \pi(\theta \mid x_{1:n})| d\theta \to 0 \text{ as } N \to \infty.$$

Due to space limitations, a detailed description of the asymptotic framework and the proof of the theorem are deferred to the Supplementary Section 3. We conclude the section by presenting a scheme for the tuning of the ε parameter.

Hyper-parameter tuning

A key revelation from our presentation on the non-parametric Bayes interpretation of D-BETEL is that the hyper-parameter ε bears clear similarity to the concentration parameter in a Dirichlet process (Ferguson, 1973; Teh, 2010), as it determines how tightly $\nu_{w,x}$ sits around F_θ with respect to D. Since the D-BETEL formulation enjoys concrete probabilistic interpretation, we are able to provide a principled guideline for hyper-parameter ε. To that end we recall that, given $x_1, \ldots, x_n$, the Bayesian leave-one-out estimate of out-of-sample predictive fit (Vehtari et al., 2016; Yao et al., 2018) is $\text{ELPD}_\varepsilon = \sum_{i=1}^n \log \pi(x_i \mid x_{-i})$, where $\pi(x_i \mid x_{-i}) = \int \pi(x_i \mid \theta)\ \pi(\theta \mid x_{-i})\ d\theta$ is the leave-one-out predictive density given the data without the i-th data point, and the corresponding standard error is $\text{SE}[\text{ELPD}_\varepsilon] = \sqrt{n}\sqrt{\text{Var}[\log \pi(x_1|x_{-1}), \ldots, \log \pi(x_n|x_{-n})]}$. When ε is too large, the distance-based restriction does not kick in and estimated $\text{SE}[\text{ELPD}_\varepsilon]$ is close to 0. So, we consider a decreasing sequence of ε values, say $\varepsilon_1, \ldots, \varepsilon_h$, such that $\varepsilon_i > \varepsilon_j\ \forall\ 1 \le i < j \le h$. A general strategy to select the sequence is to first consider a grid over powers of 2 and then use a finer grid in the interval where ELPD_ε undergoes steep change. Suppose ε_{h_0} is the largest value of ε for which the distance-based restriction is active. Then, borrowing from the idea of pseudo Bayesian model averaging (Yao et al., 2018), our estimate of the model parameter θ is

$$\hat{\theta}_{\text{MA}} = \sum_{i=h_0}^{h} \kappa_i \hat{\theta}_i, \quad \text{with} \quad \kappa_i = \frac{\exp(-\text{ELPD}_{\varepsilon_i})}{\sum_{j=h_0}^{h} \exp(-\text{ELPD}_{\varepsilon_j})}, \tag{5.4}$$

where $\hat{\theta}_i$ and $\exp(-\text{ELPD}_{\varepsilon_i})$ are the parameter estimate and estimated ELPD at $\varepsilon = \varepsilon_i$, respectively. From the definition of ELPD, we can interpret it as a measure of the extent of unequal weighting of the observations. In the presence of contamination, our approach of selecting the hyper-parameter promotes unequal weighting of the observations to ensure – under weighting of outlying observations, and over-weighting observations around the "center". This inbuilt mechanism of ensuring immunity against outliers while maintaining a valid generative model interpretation is what sets our method apart from lot of the existing pseudo-likelihood based approaches. Finally, it is also important to point out that, although $\hat{\theta}_{\text{MA}}$ in equation (5.4) is calculated via an weighted average, in practice $\hat{\theta}_{\text{MA}}$ and the associated highest posterior density (HPD) set typically degenerate to those corresponding to a handful of values of ε. Thus this procedure inherits the generative model interpretation of D-BETEL.

We now have all the necessary ingredients for D-BETEL, and we illustrate the proposed methodology in a number of specific applications. All the examples in the following section use D-BETEL in (2.4) with our proposed transport metric ANDREW in (3.2).

6 Applications

6.1 Model based clustering

Motivated by the model-based clustering example in Miller and Dunson (2019); see also Cai et al. (2020a); we generate data from a bivariate skew-normal distribution (Azzalini

and Dalla Valle, 1996) with probability density function $f(x) = 2\phi(x)\Phi(\alpha^{\mathrm{T}}x),\ x \in \mathbf{R}^2$, with the two-dimensional skewness parameter $\alpha \neq (0,0)$ to imitate a situation where the underlying true distribution is bivariate normal in the presence of mild contamination. In this sub-section, firstly, we wish to demonstrate that D-BETEL is resistant to presence of mild perturbations in the data generating mechanism and can adequately describe the above set up with a bivariate normal centering, without resorting to more complex centering distributions. In particular, we want to demonstrate that, when the extent of skewness in the data generating mechanism is small, D-BETEL would still be able to model the data well with centering distribution $F_\theta \equiv \mathrm{N}_2(\mu, \Sigma)$. Of course, as the extent of skewness increases, D-BETEL would prefer $F_\theta \equiv \omega \mathrm{N}_2(\mu_1, \Sigma_1) + (1-\omega)\mathrm{N}_2(\mu_2, \Sigma_2)$ as the centering family, compared to $F_\theta \equiv \mathrm{N}_2(\mu, \Sigma)$. Secondly, we shall showcase all our tools in action on this simple example, and skip some of these details in later sections.

Throughout this example, for the purposes of model comparison via marginal likelihood, we follow the approach in Chib and Jeliazkov (2001) to approximate the log marginal density $\log m(x \mid \mathbf{M})$ of a model $\mathbf{M}$ via $\log m(x \mid \mathbf{M}) = \log f(x \mid \mathbf{M}, \theta^*) + \log \pi(\theta^* \mid \mathbf{M}) - \log \pi(\theta^* \mid x, \mathbf{M})$, where $\log f(x \mid \mathbf{M}, \theta^*)$ and $\log \pi(\theta^* \mid \mathbf{M})$ are, respectively, the log-likelihood and log prior of the model $\mathbf{M}$ at θ^*, preferably a high-density point.

We generate data from a bivariate skew-normal distribution with varying value of skewness parameter $\alpha \neq (0,0)$. We choose sample sizes $n \in \{100,\ 200,\ 300,\ 500\}$, and set $\alpha = (2.5, 2.5)^{\mathrm{T}}$, $(3.0, 3.0)^{\mathrm{T}}$, $(3.5, 3.5)^{\mathrm{T}}$ – giving us 12 simulation set-ups in total. First, we compare the following two fully parametric models: (i) $\mathbf{M}_1$, which models the data as independent draws from $F_\theta \equiv \mathrm{N}_2(\mu, \Sigma)$ with $\theta = (\mu, \Sigma)^{\mathrm{T}}$, and imposes a diffuse $\mathrm{N}_2(0, 10^3\mathrm{I}_2)$ prior on μ and $\mathrm{Wishart}_2(\nu_0, V_0)$ prior on Σ^{-1}, independently. (ii) $\mathbf{M}_2$, which used a mixture normal model $F_\theta \equiv \omega\, \mathrm{N}_2(\mu_1, \Sigma_1) + (1-\omega)\, \mathrm{N}_2(\mu_2, \Sigma_2)$ with $\theta = (\omega, \mu_1, \Sigma_1, \mu_2, \Sigma_2)^{\mathrm{T}}$, and imposes independent diffuse $\mathrm{N}_2(0, 10^3\mathrm{I}_2)$ priors on μ_1, μ_2, an $\mathrm{U}(0,1)$ prior on ω, and independent $\mathrm{Wishart}_2(\nu_0, V_0)$ priors on Σ_1^{-1} and Σ_2^{-1}. To explore the high-density neighborhoods of the posterior distributions, we use coordinate-wise Metropolis–Hastings updates. For smaller sample sizes, the simpler model $\mathbf{M}_1$ provides higher marginal likelihood compared to $\mathbf{M}_2$. However, as the sample size grows, the more complex model $\mathbf{M}_2$ predictably starts being preferred, refer to Figure 8 which plots the posterior model probability of $\mathbf{M}_1$ as a function of sample size. Next, we consider the D-BETEL counterparts of $\mathbf{M}_1$ and $\mathbf{M}_2$, which we refer to as $\mathbf{M}_1^\star$ and $\mathbf{M}_2^\star$, respectively, with $\mathbf{M}_1^\star$ using a single normal distribution $F_\theta \equiv \mathrm{N}_2(\mu, \Sigma)$ with $\theta = (\mu, \Sigma)^{\mathrm{T}}$ as the centering distribution, and $\mathbf{M}_2^\star$ centered around $F_\theta \equiv \omega \mathrm{N}_2(\mu_1, \Sigma_1) + (1-\omega)\mathrm{N}_2(\mu_2, \Sigma_2)$ with $\theta = (\omega, \mu_1, \Sigma_1, \mu_2, \Sigma_2)^{\mathrm{T}}$. We use same the prior specification and MH sampling scheme as before.

First, we showcase our data driven approach to tune the hyper-parameter ε for both $\mathbf{M}_1^\star$ and $\mathbf{M}_2^\star$. Figures 6 and 7 present plots for $\mathrm{ELPD}_\varepsilon$, $\mathrm{SE}(\mathrm{ELPD}_\varepsilon)$ and κ, defined in Section 5, as functions of $\log \varepsilon$ for two particular combination of (n, α) values. We considered a grid of ε values over powers of 2 and then use a finer grid in the interval where $\mathrm{ELPD}_\varepsilon$ undergoes steep change. For sufficiently large value of ε, the distance based constraint practically becomes inactive, and consequently $\mathrm{ELPD}_\varepsilon$ plateaus out and $\mathrm{SE}(\mathrm{ELPD}_\varepsilon) \downarrow 0$. Finally, we obtain the D-BETEL based parameter estimates $\hat{\theta}_{\mathrm{MA}}$ as delineated in Section 5. Although $\hat{\theta}_{\mathrm{MA}}$ in equation (5.4) is calculated via an weighted

average, in practice $\hat{\theta}_{\text{MA}}$ typically degenerates to estimates corresponding to a handful of values of ε, as apparent in the plot of κ as a function of $\log \varepsilon$ in Figures 6, 7. We observe similar pattern for the remaining combinations of (n, α) values, and refrain from presenting them here in order to avoid repetitiveness.

Figure 8 presents the posterior probability of selecting the simpler model with only one bivariate normal component under the standard posterior, a fractional posterior with varying temperature parameters, and D-BETEL. For the standard posterior, the posterior probability of selecting the simpler model $\mathbf{M}_1$ drop below 0.5 as sample size increases. On the contrary, D-BETEL is more resistant towards presence of mild skewness in the data generating mechanism, and still prefer the simpler model across the sample sizes we considered. The fractional posterior (Miller and Dunson, 2019) approach with the temperature parameter decreasing with sample size is expected to enjoy similar numerical results. In fact, the benefits of the fractional posterior (Miller and Dunson, 2019) is already apparent in our simulations with the choice of the temperature parameter equal to 0.25. However, unless the temperature parameter of the fractional posterior is chosen to be appropriately small, it cannot reliably estimate the number of components in finite mixture models, under mild model mis-specification (Cai et al., 2020b). Finally, a comparison of computational times for the D-BETEL, the standard posterior, and the fractional posterior-based approaches is presented in Table 1.

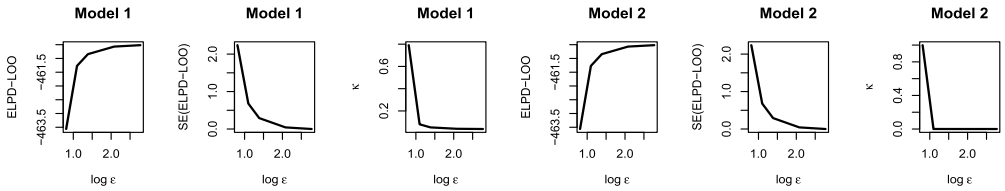

Figure 6: ***Hyper-parameter tuning for model based clustering with sample size*** $\mathbf{n = 100}$***, skewness parameter*** $\boldsymbol{\alpha} = (\mathbf{3.5}, \mathbf{3.5})^{\mathrm{T}}$**.** ELPD_ε gradually plateaus out and $\text{SE}(\text{ELPD}_\varepsilon) \downarrow 0$ as $\log \varepsilon \uparrow$ for both the models. Consequently, weights κ corresponding to a handful of ε values contribute meaningfully to the weighted sum in $\hat{\theta}_{\text{MA}}$ and rest are ≈ 0.

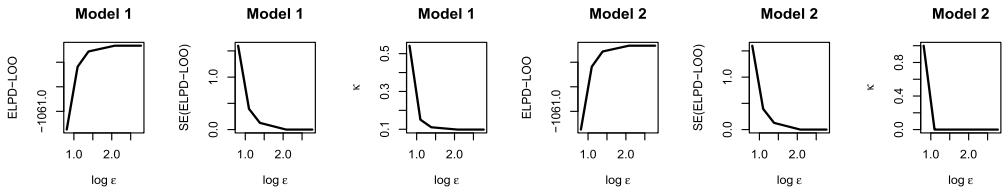

Figure 7: Hyper-parameter tuning for model based clustering with sample size $\mathbf{n = 200}$, skewness parameter $\boldsymbol{\alpha} = (\mathbf{2.5}, \mathbf{2.5})^{\mathrm{T}}$. ELPD_ε gradually plateaus out and $\text{SE}(\text{ELPD}_\varepsilon) \downarrow 0$ as $\log \varepsilon \uparrow$ for both the models. Consequently, weights κ corresponding to a handful of ε values contribute meaningfully to the weighted sum in $\hat{\theta}_{\text{MA}}$ and rest are ≈ 0.

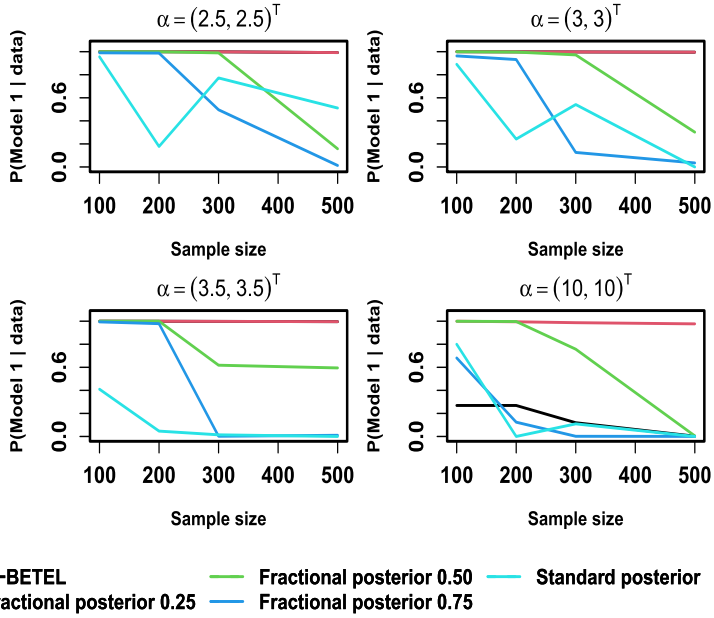

Figure 8: **Model based clustering.** We are comparing the Bayes factor for selecting the simpler model via D-BETEL, the standard posterior, and the fractional posterior (Miller and Dunson, 2019) with different temperatures, across varying values of skewness parameter α of the generating skew normal distribution and sample sizes. The top left panel is for $\alpha = (2.5, 2.5)^{\mathrm{T}}$, the top right panel is for $\alpha = (3, 3)^{\mathrm{T}}$, the bottom left panel is for $\alpha = (3.5, 3.5)^{\mathrm{T}}$, and the bottom right panel is for $\alpha = (10, 10)^{\mathrm{T}}$. In presence of mild contamination, unlike the standard posterior, D-BETEL and fractional posterior with low temperature still prefer the simpler model across the sample sizes. However, when the extent of misspecification increases, that is, the skewness parameter $\alpha = (10, 10)^{\mathrm{T}}$, then D-BETEL chooses the more complicated model to account for the perturbation, as expected.

	$n = 100$	$n = 200$	$n = 300$	$n = 500$
D-BETEL	79.8	87.6	93.1	106.1
Standard Posterior	26.9	27.4	27.6	27.9
Fractional Posterior (0.25)	32.6	32.6	32.1	32.5
Fractional Posterior (0.5)	33.7	31.8	36.7	33.7
Fractional Posterior (0.75)	31.5	32.2	31.7	30.4

Table 1: **(Time comparison for model based clustering).** We are comparing the average time (in seconds) required for comparing the two model via D-BETEL, the standard posterior, and the fractional posterior (Miller and Dunson, 2019) with different temperatures. We report run-times (in seconds) averaged over different values of the skewness parameter $\alpha \in \{(2.5, 2.5)^{\mathrm{T}}, (3, 3)^{\mathrm{T}}, (3.5, 3.5)^{\mathrm{T}}, (10, 10)^{\mathrm{T}}\}$ of the generating skew normal distribution, for varying sample sizes.

6.2 Generalized linear regression

In many scientific applications, we are constrained to operate under stringent modeling assumptions derived from the domain knowledge or common practice in the field. For example, in a count regression problem, collaborators may require us to assume that the conditional distribution of the counts follow a Poisson or a negative binomial distribution. In such cases, the parameters in the postulated model carry interpretability to the domain experts. Then, using a moment condition based model as an alternative to a fully parametric model is not desirable, even if the postulated parametric model is inadequate. But D-ETEL provides a viable alternative in such cases.

Suppose we observe data $\{(y_i, x_i) \in \mathbf{R} \times \mathbf{R}^d\}_{i=1}^n$ on a response variable y and covariates x for n individuals. In generalized linear regression set up, we model the response by an exponential family distribution:

$$f(y_i \mid \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right],$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are known functions such that $m_i = b'(\theta_i)$, $\sigma_i^2 = \phi b''(\theta_i)$ are, respectively, the mean and variance of the distribution, and there exists a one-to-one continuously differentiable link function $g(\cdot)$ such that $g^{-1}(x_i^{\mathrm{T}}\beta) = b'(\theta_i)$. The log-likelihood of the parameter of interest β is $l(\beta \mid x, y) = \sum_{i=1}^n l_i(\beta \mid x_i, y_i) = \sum_{i=1}^n \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \log c(y_i, \phi)\right]$, where θ_i is a function of m_i. The corresponding Fisher's score function $S = (S_0,\ S_1, .., \ S_d)^{\mathrm{T}}$: $S_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \left[\frac{(y_i - m_i)}{a(\phi)} \frac{1}{V_i} \frac{\partial m_i}{\partial \beta_j}\right] = 0$ with $V_i = \frac{\partial m_i}{\partial \theta_i} = b''(\theta_i)$. For simplicity of exposition, we express $S = \sum_{i=1}^n \eta_i$, where $\eta_{ij} = \frac{\partial l_i}{\partial \beta_j}, \eta_i = (\eta_{i0},\ \eta_{i1}\ , \ldots,\ \eta_{id})^{\mathrm{T}}$, $i = 1, \ldots, n,\ j = 0, \ldots, d$. The score statistic S is asymptotically normal with mean 0 (Haynes, 2013) and n_i captures the deviation from 0 for the i-th observation. With that intuition, to conduct robust Bayesian inference on β, we posit D-BETEL on $\{\eta_i\}_{i=1}^n$ with a finite mixture of $(d+1)$-variate normal densities, that is, $F_\theta \equiv \sum_{j=1}^K \pi_j\ \mathrm{N}(\mu_j, \Sigma_j)$ with $\theta = (\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)^{\mathrm{T}}$ such that $\sum_{j=1}^K \pi_j\mu_j = 0$ as our choice for centering parametric guess.

We generate data from a Poisson random effects model, $\log(m_i) = \beta_0 + \beta_1 x_i + h_i$, $y_i \sim \mathrm{Poisson}(m_i)$, $i = 1, \ldots, n$, where $\beta_0 = 5$, $\beta_1 = 1$, $x_i \sim \mathrm{N}(5, 1)$ and $h_i \sim (1-p)\ \mathbf{1}\{0\} + p\ \mathrm{N}(1, 0.1^2)$, in order to mimic a situation where a small proportion of outliers are present in the data-set. We place flat $\mathrm{N}(0, 100^2)$ priors on $\log\sigma_1^2(\beta)$, $\log\sigma_2^2(\beta)$, β_0 and β_1 and a $\mathrm{U}(-1, 1)$ prior on $\rho(\beta)$, independently. We devise a Metropolis–Hastings algorithm to update $g(\beta) = (\log\sigma_1^2(\beta),\ \log\sigma_2^2(\beta),\ \rho(\beta), \beta_0,\ \beta_1)^{\mathrm{T}}$ at $(t+1)$th iteration using the 1-step proposal scheme: $g^{(t+1)}(\beta) \sim \mathrm{N}_5\big(g^{(t)}(\beta),\ k\ \nabla g(\hat{\beta}_m)\ \mathrm{I}^{-1}(\hat{\beta}_m)\ \nabla^T g(\hat{\beta}_m)\big)$, where $\mathrm{I}(\hat{\beta}_m))$ is the Fisher's information matrix evaluated at the maximum likelihood estimator $\hat{\beta}_m$ of β, and k is a tuning parameter.

Under model misspecification, the maximum likelihood estimator of θ converges to the KL projection of true data generating mechanism within $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$, as the sample size $n \to \infty$, modulo appropriate regularity conditions (White, 1982).

Moreover, a usual Bayesian posterior also contracts around the same target parameter under similar regularity conditions (Kleijn and van der Vaart, 2006). So, it is instructive to compare the inference based on D-BETEL, with that based on the standard Bayesian approach. Additionally, the we include comparison with BETEL with moment conditions derived from score equations of the Poisson regression model to numerically demonstrate that such a procedure, as expected, is not capable of handling model misspecification. In the context of generalized linear regression framework, setting up meaningful moment conditions is somewhat unwieldy. Consequently, such moment conditions derived from the score equations are common in the literature.

In Table 2, we expand on the performance of D-BETEL for varying extent of perturbations in the data generating mechanism with sample size $n = 100$, relative to popular practical approaches. In particular, we compare D-BETEL against a standard posterior as well as Bayesian analysis with the estimating equations set to $\mathrm{E}[\partial \log l(\beta \mid X, Y)/\partial\beta] = 0$ to infer about the parameter β. In particular, the Bayesian inference based on the moment restrictions is conducted via Bayesian exponentially tilted empirical likelihood (Schennach, 2005; Chib et al., 2018). From Table 2, we report the L_1 error of posterior means, length of the HPD sets and associated coverage probabilities (within braces) for D-BETEL and competing approaches. It is evident that D-BETEL is more resistant towards presence of outliers when compared with the standard Bayesian and MCM based approaches, across all the sample sizes and proportion of contamination in the data sets that we considered. Also, D-BETEL provides slightly wider credible sets compared to the standard posterior based approach, while maintaining high coverage probability. Additional simulation results for $n = 250, 500$ are presented in the Supplementary Section 4.

		D-BETEL		Standard posterior		MCM	
p	θ	$\|\|\theta - \hat{\theta}\|\|_1$	HPD	$\|\|\theta - \hat{\theta}\|\|_1$	HPD	$\|\|\theta - \hat{\theta}\|\|_1$	HPD
0.10	β_0	0.02	0.18 (1.00)	0.35	0.12 (0.20)	0.41	0.22 (0.10)
	β_1	0.01	0.02 (1.00)	0.07	0.02 (0.20)	0.06	0.04 (0.22)
0.12	β_0	0.02	0.22 (1.00)	0.31	0.12 (0.35)	0.47	0.35 (0.14)
	β_1	0.01	0.04 (1.00)	0.08	0.02 (0.59)	0.06	0.06 (0.32)
0.15	β_0	0.06	0.22 (0.94)	0.47	0.11 (0.00)	0.54	0.23 (0.06)
	β_1	0.01	0.04 (0.94)	0.06	0.02 (0.00)	0.09	0.05 (0.18)

Table 2: **Generalized linear regression (Poisson regression).** Here the **sample size** n **is** 100. We compare standard posterior yielded from the fully parametric model, moment condition model (MCM) based on the maximum likelihood equations, and D-BETEL based parameter estimates over 50 replicated simulations with proportion of outlier $p = 0.10, 0.12, 0.15$. D-BETEL is more resistant towards presence of outliers all values of p considered, however it provides slightly wider 95% credible sets while maintaining the high coverage probability. Additional simulation results for $n = 250, 500$ is presented in sequel.

7 Discussion

Generative probabilistic models are immensely popular in applications as they provide a general recipe for statistical inference using the maximum likelihood or Bayesian framework. However, it is also well understood that the resulting inference can crucially depend on the modeling assumptions. In this article, we introduced a flexible Bayesian semi-parametric modeling framework D-BETEL, and demonstrated it's utility to conduct robust inference under perturbations of the data-generating mechanism. D-BETEL is endowed with a fully data-driven hyper-parameter tuning scheme, and enjoys a valid generative model interpretation, which is scarce in pseudo-likelihood based robust Bayesian methods. R scripts to reproduce the results presented in the article are available at zovialpapai/D-BETEL.

While semi-parametric in nature, a particularly attractive feature of D-BETEL is that the user only needs to specify a plausible family of probability models F_θ for the data along with a prior distribution for the parameter of interest θ, and does not need to explicitly model departures from the parametric guess as is typical with nonparametric Bayesian techniques, all nuisance parameters are implicitly marginalized out and a marginal posterior for θ is returned. It remains possible to retrieve a discretized estimate of the generating distribution to allow a more fine-grained analysis of how the data departs from the parametric guess. The proposed approach is also very general, while we have illustrated its usage for i.i.d. and independent non-i.i.d (i.n.i.d.) setups, extensions to broader classes of dependent data models should be straightforward. Studying theoretical properties of D-BETEL, especially second-order properties, is an interesting avenue for future work.

While developing the methodology, we proposed a general framework to devise expressible and computationally efficient optimal transport metrics. We believe this framework may have far-reaching utility beyond the scope of the current article, since optimal transport metric has become increasingly popular in the context of Bayesian analysis in recent years. On the theory side, optimal transport has been utilized in studying convergence properties of latent mixing measures (Nguyen, 2013), posterior concentration of the base probability measure of a Dirichlet measure (Nguyen, 2016), posterior contraction in finite mixture of regression models (Do et al., 2025), posterior contraction in Gaussian mixture models (Guha et al., 2023), to name a few. On the methodological side, the Wasserstein metric has been adopted in measuring dependence in Bayesian non-parametric models (Catalano et al., 2021, 2024), approximate Bayesian computation (Bernton et al., 2019), Bayesian non-parametric distributionally robust optimization (Ning and Ma, 2023), memory efficient and minimax distribution estimation (Jacobs et al., 2023), etc. While these applications exhibit substantial promise for future development, yet the usage of transport metrics remain underexplored within the Bayesian framework.

Finally, in this article, we demonstrated that D-BETEL is asymptotically equivalent to a hierarchical setup similar to the mixture of finite mixture models (Miller and Harrison, 2018). A potential future work may explore if one can device similar techniques to conduct targeted inference on parameters of interest when considering other

non-parametric Bayesian priors, e.g. Pitman-Yor multinomial prior (Lijoi et al., 2020), normalized infinitely divisible multinomial processes (Lijoi et al., 2023), etc.

Funding

Drs. Bhattacharya and Pati acknowledge NSF DMS-1916371 and NSF DMS-2210689 for partially funding the project.

Supplementary Material

Supplementary Material to "Robust probabilistic inference via a constrained transport metric" (DOI: 10.1214/25-BA1535SUPP; .pdf). Supplementary section 1 presents the proof of Theorem 1 on derivation of the modified optimal transport metric ANDREW, and related auxiliary results. Supplementary Section 2 presents a heuristic justification of the robustness of D-ETEL. Supplementary Section 3 provides proofs of theorems supporting a nonparametric Bayesian interpretation of the proposed D-BETEL methodology. In Supplementary Section 4, additional simulation results for the generalized linear regression setting with outliers, corresponding to sample sizes $n = 250$ and 500, are presented. In Supplementary Section 5, we compare D-BETEL with $\mathrm{D} = \mathrm{MW}_2$ and $\mathrm{D} = W_{\mathrm{AR}}$ on a generalized linear regression task.

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2(6): 1152–1174. URL https://doi.org/10.1214/aos/1176342871 2, 5, 16

Avella-Medina, M. (2021). "Privacy-Preserving Parametric Inference: A Case for Robust Statistics." *Journal of the American Statistical Association*, 116(534): 969–983. URL https://doi.org/10.1080/01621459.2019.1700130 1

Azzalini, A. and Dalla Valle, A. (1996). "The multivariate skew-normal distribution." *Biometrika*, 83(4): 715–726. URL https://doi.org/10.1093/biomet/83.4.715 14, 15, 19

Becker, Candès, and Grant (2011). "Templates for convex cone problems with applications to sparse signal recovery." *Mathematical Programming Computation*. 7

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). "Approximate Bayesian Computation with the Wasserstein Distance." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2): 235–269. Published: 17 February 2019. 25

Bion-Nadal, J. and Talay, D. (2019). "On a Wasserstein-type distance between solutions to stochastic differential equations." *The Annals of Applied Probability*. 9

Birgin, E. and Martínez, J. (2008). "Improving ultimate convergence of an augmented Lagrangian method." *Optimization Methods and Software*, 23(2): 177–195. URL https://doi.org/10.1080/10556780701577730 7

Cai, D., Campbell, T., and Broderick, T. (2020a). "Finite mixture models do not reliably learn the number of components."
URL https://arxiv.org/abs/2007.04470 14, 19

Cai, D., Campbell, T., and Broderick, T. (2020b). "Power posteriors do not reliably learn the number of components in a finite mixture."
URL https://openreview.net/pdf?id=BRb4tLp6A3o 21

Cambanis, S., Huang, S. T., and Simons, G. (1981). "On the theory of elliptically contoured distributions." *Journal of Multivariate Analysis*, 11: 368–385. 8

Catalano, M., Lavenant, H., Lijoi, A., and Prünster, I. (2024). "A Wasserstein Index of Dependence for Random Measures." *Journal of the American Statistical Association*, 119(547): 2396–2406. 25

Catalano, M., Lijoi, A., and Prünster, I. (2021). "Measuring Dependence in the Wasserstein Distance for Bayesian Nonparametric Models." *The Annals of Statistics*, 49(5): 2916–2947. 25

Chakraborty, A., Bhattacharya, B., Pati, D. (2025). Supplementary Material to "Robust probabilistic inference via a constrained transport metric." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1535SUPP. 10

Chen, Z., Yu, P., and Haskell, W. B. (2019). "Distributionally robust optimization for sequential decision-making." *Optimization*, 68(12): 2397–2426.
URL https://doi.org/10.1080/02331934.2019.1655738 1

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation." *Journal of Econometrics*, 115(2): 293–346.
URL https://ideas.repec.org/a/eee/econom/v115y2003i2p293-346.html 2

Chib, S. and Jeliazkov, I. (2001). "Marginal Likelihood From the Metropolis–Hastings Output." *Journal of the American Statistical Association*, 96(453): 270–281.
URL https://doi.org/10.1198/016214501750332848 20

Chib, S., Shin, M., and Simoni, A. (2018). "Bayesian Estimation and Comparison of Moment Condition Models." *Journal of the American Statistical Association*, 113(524): 1656–1668.
URL https://doi.org/10.1080/01621459.2017.1358172 2, 5, 24

Chib, S., Shin, M., and Simoni, A. (2021). "Bayesian Estimation and Comparison of Conditional Moment Models."
URL https://arxiv.org/abs/2110.13531 2

Conn, A. R., Gould, N. I. M., and Toint, P. (1991). "A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds." *SIAM Journal on Numerical Analysis*, 28(2): 545–572.
URL https://doi.org/10.1137/0728030 7

Cuturi, M. (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran

Associates, Inc.
URL https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf 3, 6, 9, 11, 12

De Blasi, P., Lijoi, A., and Prünster, I. (2013). "An asymptotic analysis of a class of discrete nonparametric priors." *Statistica Sinica*, 23: 1299–1321. 16

De Blasi, P. and Walker, S. G. (2013). "Bayesian Asymtotics Wth Misspecified Models." *Statistica Sinica*, 23(1): 169–187.
URL http://www.jstor.org/stable/24310519 2

Delon, J. and Desolneux, A. (2020). "A Wasserstein-type distance in the space of Gaussian Mixture Models." *SIAM Journal on Imaging Sciences*, 13(2): 936–970.
URL https://hal.archives-ouvertes.fr/hal-02178204 6, 9

Do, D., Do, L., and Nguyen, X. (2025). "Strong Identifiability and Parameter Learning in Regression with Heterogeneous Response." *Electronic Journal of Statistics*, 19(1): 131–203. 25

Du, W. and Wu, X. (2021). "Robust Fairness-aware Learning Under Sample Selection Bias."
URL https://arxiv.org/abs/2105.11570 1

Dwork, C. and Lei, J. (2009). "Differential privacy and robust statistics." *STOC '09: Proceedings of the forty-first annual ACM symposium on Theory of computing*, 371–380.
URL https://dl.acm.org/doi/10.1145/1536414.1536466 1

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430): 577–588.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550 7

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209–230.
URL https://doi.org/10.1214/aos/1176342360 2, 5, 6, 16, 19

Fiksel, e J., Datta, A., Amouzou, A., and Zeger, S. (2021). "Generalized Bayes Quantification Learning under Dataset Shift." *Journal of the American Statistical Association*, 0(0): 1–19.
URL https://doi.org/10.1080/01621459.2021.1909599 1

Gerber, S. and Maggioni, M. (2017). "Multiscale Strategies for Computing Optimal Transport." 6

Gnedin, A. (2009). "Species sampling problems for Gibbs partitions with finitely many types." *Electronic Journal of Probability*, 14: no. 49, 1481–1499. 16

Gnedin, A. and Pitman, J. (2005). "Exchangeable Gibbs partitions and Stirling triangles." *Zapiski Nauchnykh Seminarov POMI*, 325: 83–102. Translated in Journal of Mathematical Sciences, 138(3): 5674–5685, 2006. 16

Grant, M. and Boyd, S. (2008). "Graph implementations for nonsmooth convex programs." In Blondel, V., Boyd, S., and Kimura, H. (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, 95–110. Springer-Verlag Limited. http://stanford.edu/~boyd/graph_dcp.html. 7

Grünwald, P. and van Ommen, T. (2017). "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It." *Bayesian analysis*.
URL https://pure.uva.nl/ws/files/22184651/1510974325.pdf 2

Guha, A., Ho, N., and Nguyen, X. (2023). "On Excess Mass Behavior in Gaussian Mixture Models with Orlicz-Wasserstein Distances." In Kamsetty, S., Koyejo, O., Jegelka, S., Sabato, S., and von Luxburg, U. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 11847–11870. PMLR. 25

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. (2018). "Co-teaching: Robust training of deep neural networks with extremely noisy labels." In *NeurIPS*, 8535–8545. 1

Haynes, W. (2013). *Maximum Likelihood Estimation*, 1190–1191. New York, NY: Springer New York.
URL https://doi.org/10.1007/978-1-4419-9863-7_1235 23

Holmes, C. C. and Walker, S. G. (2017). "Assigning a value to a power likelihood in a general Bayesian model." *Biometrika*, 104(2): 497–503.
URL https://doi.org/10.1093/biomet/asx010 2

Holzmann, H., Munk, A., and Gneiting, T. (2006). "Identifiability of Finite Mixtures of Elliptical Distributions." *Scandinavian Journal of Statistics*, 33(4): 753–763.
URL http://www.jstor.org/stable/4616956 8

Hooker, G. and Vidyashankar, A. (2011). "Bayesian Model Robustness via Disparities."
URL https://arxiv.org/abs/1112.4213 2

Huber, P. J. (2011). "Robust statistics." In *International encyclopedia of statistical science*, 1248–1251. Springer. 1

Ishwaran, H. and Zarepour, M. (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." *Biometrika*, 87(2): 371–390.
URL https://doi.org/10.1093/biomet/87.2.371 7

Ishwaran, H. and Zarepour, M. (2002a). "Dirichlet prior sieves in finite normal mixtures." *Statistica Sinica*, 941–963. 17

Ishwaran, H. and Zarepour, M. (2002b). "Exact and approximate sum representations for the Dirichlet process." *Canadian Journal of Statistics*, 30(2): 269–283. 17

Jacobs, P. M., Patel, L., Bhattacharya, A., and Pati, D. (2023). "Memory Efficient And Minimax Distribution Estimation Under Wasserstein Distance Using Bayesian Histograms."
URL https://arxiv.org/abs/2307.10099 25

Jiang, W. and Tanner, M. A. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining." *The Annals of Statistics*, 36(5).
URL http://dx.doi.org/10.1214/07-AOS547 2

Johnson, S. G. (2022). "The NLopt nonlinear-optimization package." *The Comprehensive R Archive Network*. 7

Kitagawa, J., Mérigot, Q., and Thibert, B. (2017). "Convergence of a Newton algorithm for semi-discrete optimal transport." 6

Kleijn, B. J. and van der Vaart, A. W. (2012). "The Bernstein-von-Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381. 2

Kleijn, B. J. K. and van der Vaart, A. W. (2006). "Misspecification in infinite-dimensional Bayesian statistics." *The Annals of Statistics*, 34(2): 837–877.
URL https://doi.org/10.1214/009053606000000029 2, 13, 24

Lavine, M. (1994). "More Aspects of Polya Tree Distributions for Statistical Modelling." *The Annals of Statistics*, 22(3): 1161–1176.
URL https://doi.org/10.1214/aos/1176325623 2, 5

Lazar, N. A. (2003). "Bayesian Empirical Likelihood." *Biometrika*, 90(2): 319–326.
URL http://www.jstor.org/stable/30042042 2

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. (2019). "Tree-Sliced Variants of Wasserstein Distances." In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
URL https://proceedings.neurips.cc/paper/2019/file/2d36b5821f8affc6868b59dfc9af6c9f-Paper.pdf 3, 8

Lijoi, A., Prünster, I., and Rigon, T. (2020). "The Pitman–Yor multinomial process for mixture modelling." *Biometrika*, 107(4): 891–906. 26

Lijoi, A., Prünster, I., and Rigon, T. (2023). "Finite-dimensional discrete random structures and Bayesian clustering." *Journal of the American Statistical Association*, 119(546): 929–941. 26

Liu, X., Kong, W., and Oh, S. (2021). "Differential privacy and robust statistics in high dimensions."
URL https://arxiv.org/abs/2111.06578 1

McAuliffe, Blei, and Jordan (2006). "Nonparametric empirical Bayes for the Dirichlet process mixture model." *Statistics and Computing*, 1(2): 5–14. 7

Miller, J. W. and Dunson, D. B. (2019). "Robust Bayesian Inference via Coarsening." *Journal of the American Statistical Association*, 114(527): 1113–1125. PMID: 31942084.
URL https://doi.org/10.1080/01621459.2018.1469995 1, 2, 14, 19, 21, 22

Miller, J. W. and Harrison, M. T. (2018). "Mixture Models With a Prior on the Number of Components." *Journal of the American Statistical Association*, 113(521): 340–356. PMID: 29983475.
URL https://doi.org/10.1080/01621459.2016.1255636 6, 16, 17, 25

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (2017). "Robust and Scalable Bayes via a Median of Subset Posterior Measures." *Journal of Machine Learning Research*, 18(124): 1–40.
URL http://jmlr.org/papers/v18/16-655.html 2

Mirebeau, J.-M. (2015). "Discretization of the 3d monge-ampere operator, between wide stencils and power diagrams." *ESAIM: Mathematical Modelling and Numerical Analysis – Modélisation Mathématique et Analyse Numérique*, 49(5).
URL http://www.numdam.org/articles/10.1051/m2an/2015016/ 6

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. Wiley-Interscience. 8

Müller, P. and Quintana, F. A. (2004). "Nonparametric Bayesian data analysis." *Statistical science*, 19(1): 95–110. 2

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*, volume 1. Springer. 2

Muzellec, B. and Cuturi, M. (2018). "Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions." In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
URL https://proceedings.neurips.cc/paper/2018/file/b613e70fd9f59310cf0a8d33de3f2800-Paper.pdf 8

Nguyen, X. (2013). "Convergence of Latent Mixing Measures in Finite and Infinite Mixture Models." *The Annals of Statistics*, 41(1): 370–400. 25

Nguyen, X. (2016). "Borrowing Strength in Hierarchical Bayes: Posterior Concentration of the Dirichlet Base Measure." *Bernoulli*, 22(3): 1535–1571. Originally listed as Technical Report No. 532, Department of Statistics, University of Michigan, January 2013. 25

Ning, C. and Ma, X. (2023). "Data-Driven Bayesian Nonparametric Wasserstein Distributionally Robust Optimization."
URL https://arxiv.org/abs/2311.02953 25

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC. 2

Panaretos, V. M. and Zemel, Y. (2019). "Statistical Aspects of Wasserstein Distances." *Annual Review of Statistics and Its Application*, 6(1): 405–431.
URL http://dx.doi.org/10.1146/annurev-statistics-030718-104938 3, 6

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." In *Annals of Probability*, volume 25, 855–900. Institute of Mathematical Statistics. 16

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/ 7

Resnick, S. (2013). "A Probability Path." *Birkhäuser Boston*. 18

Schennach, S. M. (2005). "Bayesian exponentially tilted empirical likelihood." *Biometrika*, 92(1): 31–46.
URL https://doi.org/10.1093/biomet/92.1.31 2, 4, 16, 18, 24

Schennach, S. M. (2007). "Point estimation with exponentially tilted empirical likelihood." *The Annals of Statistics*, 35(2): 634–672.
URL https://doi.org/10.1214/009053606000001208 13

Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D., and Goldstein, T. (2020). "Adversarially robust transfer learning." In *International Conference on Learning Representations*.
URL https://openreview.net/forum?id=ryebG04YvB 1

Taskesen, B., Shafieezadeh-Abadeh, S., and Kuhn, D. (2022). "Semi-discrete optimal transport: hardness, regularization and numerical solution." *Mathematical Programming*, 1033–1106.
URL https://doi.org/10.1007/s10107-022-01856-x 6, 8

Teh, Y. W. (2010). "Dirichlet Process." *Encyclopedia of machine learning*, 1063: 280–287. 2, 5, 6, 19

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). "Bayesian Leave-One-Out Cross-Validation Approximations for Gaussian Latent Variable Models." *Journal of Machine Learning Research*, 17(103): 1–38.
URL http://jmlr.org/papers/v17/14-540.html 19

Verdinelli, I. and Wasserman, L. (1998). "Bayesian goodness-of-fit testing using infinite-dimensional exponential families." *The Annals of Statistics*, 26(4): 1215–1241.
URL https://doi.org/10.1214/aos/1024691240 2, 5

Villani, C. (2003). "Topics in Optimal Transportation." *American Mathematical Society*.
URL https://www.math.ucla.edu/~wgangbo/Cedric-Villani.pdf 3, 6, 13

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. (2020a). "Robust Optimization for Fairness with Noisy Protected Groups." In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, 5190–5203. Curran Associates, Inc.
URL https://proceedings.neurips.cc/paper/2020/file/37d097caf1299d9aa79c2c2b843d2d78-Paper.pdf 1

Wang, Z., Hu, G., and Hu, Q. (2020b). "Training Noise-Robust Deep Neural Networks via Meta-Learning." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4523–4532. 1

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50(1): 1–25.
URL http://www.jstor.org/stable/1912526 13, 23

Xu, H. and Mannor, S. (2010). "Distributionally Robust Markov Decision Processes." In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates,

Inc.
URL https://proceedings.neurips.cc/paper/2010/file/19f3cd308f1455b3fa09a282e0d496f4-Paper.pdf 1

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis*, 13(3): 917–1007.
URL https://doi.org/10.1214/17-BA1091 19