Bayesian Analysis (2026) **21**, Number 2, pp. 937–1028

Robust Probabilistic Inference via a Constrained Transport Metric (with Discussion)

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, $\pi_{\mathrm{MCM}}(\psi \mid x_1, \ldots, x_n)$ 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

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

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

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

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

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

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.

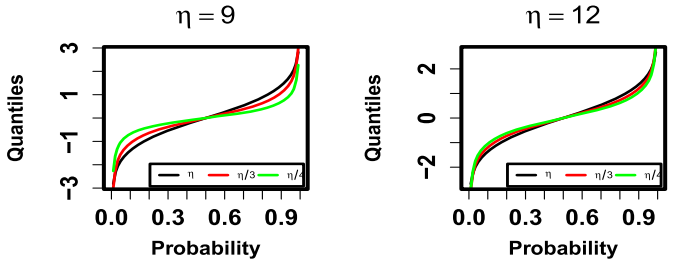

Figure 1: *The plots show that $MW_{2,\mathrm{EMM},\alpha}$ fails to capture the differences in the tail behavior of the probability distributions. Let $p_0 \equiv \sum_{k=1}^{K_0} s_{0k} t_\eta(m_{0k}, \Sigma_{0k})$, $p_0' \equiv \sum_{k=1}^{K_0} s_{0k} t_{\eta'}(m_{0k}, \Sigma'_{0k})$, $p_1 \equiv \sum_{k=1}^{K_1} s_{1k} \delta_{m_{1k}}$ 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

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

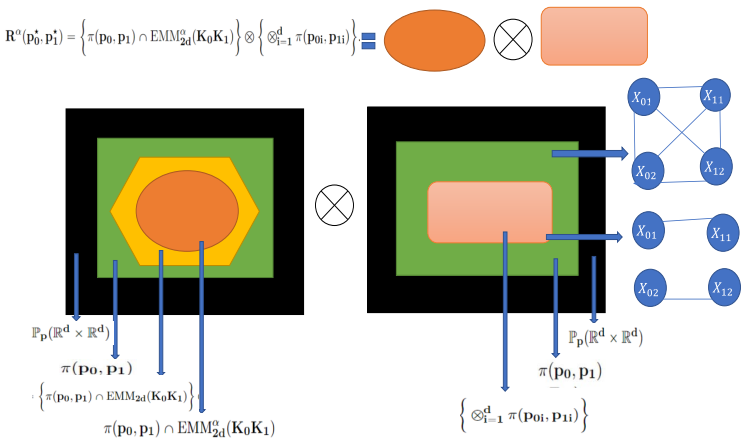

Figure 2: The augmentation and restriction scheme to construct $R^{\alpha}(p_0^{\star}, p_1^{\star})$, left: construction of $\left\{\mathcal{C}_{\mathrm{EMM},\alpha} = \mathcal{C}(p_0, p_1) \cap \mathrm{EMM}^{\alpha}_{2d,K_0,K_1}\right\}$, right: construction of $\left\{ \otimes_{i=1}^{d} \mathcal{C}(p_{0i}, p_{1i})\right\}$ 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 $\mathrm{KL}(\nu_{w,x} \,||\, \nu_{(1/n,\ldots,1/n)^\mathrm{T},x}) = \log n - (-\sum_{i=1}^n w_i \log w_i)$.

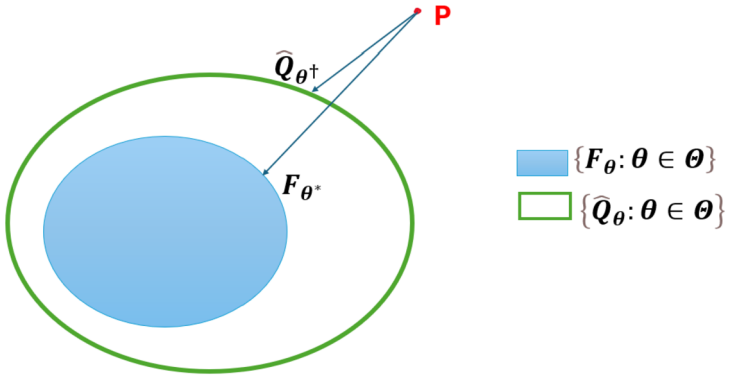

Figure 3: Here P denotes the unknown data generative model of interest. A proposed parametric class of models $\{F_\theta : \theta \in \Theta \subseteq \mathbf{R}^d\}$ aims to adequately describe P. The $\{\hat{Q}_\theta, \theta \in \Theta\}$ should be interpreted as an enlargement of the class $\{F_\theta, \theta \in \Theta\}$, 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-

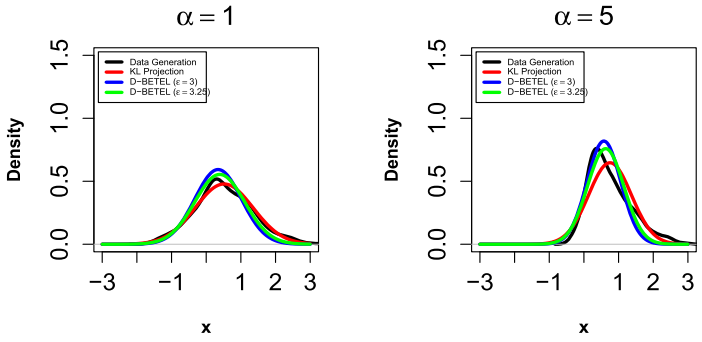

Figure 4: **Skew-normal distribution.** We plot the density estimates $f_{\theta^\star}$ corresponding to the MLE in (4.1) and $f_{\theta^\dagger}$ corresponding to the D-ETEL estimator in (4.3), where the data is generated from a univariate skew normal distribution (Azzalini and Dalla Valle, 1996) and 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 $\mu_h = (\mu_{h,1}, \ldots, \mu_{h,d})^\intercal \in \mathbf{R}^d$ 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

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

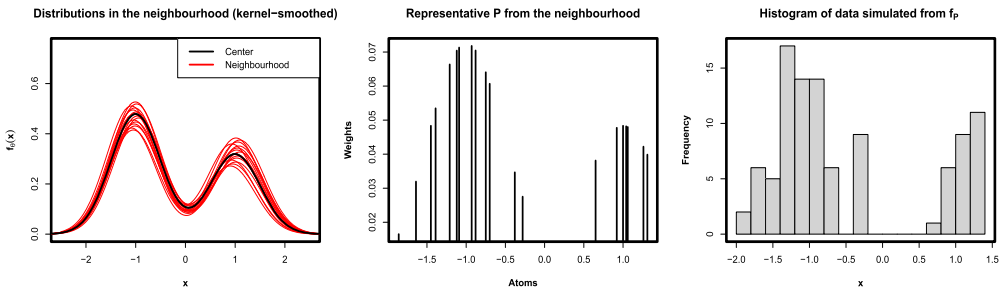

Figure 5: ***Left panel.*** *the pdf* $f_\theta(x)$ *corresponding to the centering distribution* $F_\theta \equiv 0.6 \times N(-1, 0.5^2) + 0.4 \times N(1, 0.5^2)$ *in black and representative discrete distributions in a D-neighborhood (with D chosen as* W_{AR} *introduced in Section 3) of* F_θ *in 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* $W^2_{\mathrm{AR}}(P, F_\theta) = 2.5$*,* ***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,

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

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.

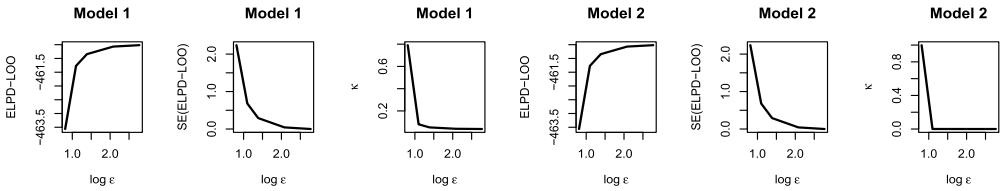

Figure 6: ***Hyper-parameter tuning for model based clustering with sample size*** $\mathbf{n = 100}$***, skewness parameter*** $\boldsymbol{\alpha} = (\mathbf{3.5}, \mathbf{3.5})^{\mathrm{T}}$**.** ELPD_ε gradually plateaus out and $\text{SE}(\text{ELPD}_\varepsilon) \downarrow 0$ as $\log \varepsilon \uparrow$ 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.

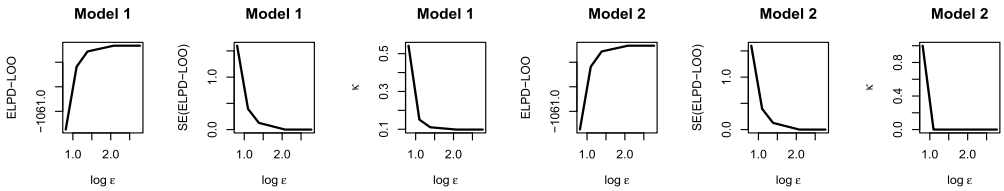

Figure 7: Hyper-parameter tuning for model based clustering with sample size $\mathbf{n = 200}$, skewness parameter $\boldsymbol{\alpha} = (\mathbf{2.5}, \mathbf{2.5})^{\mathrm{T}}$. ELPD_ε gradually plateaus out and $\text{SE}(\text{ELPD}_\varepsilon) \downarrow 0$ as $\log \varepsilon \uparrow$ for 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.

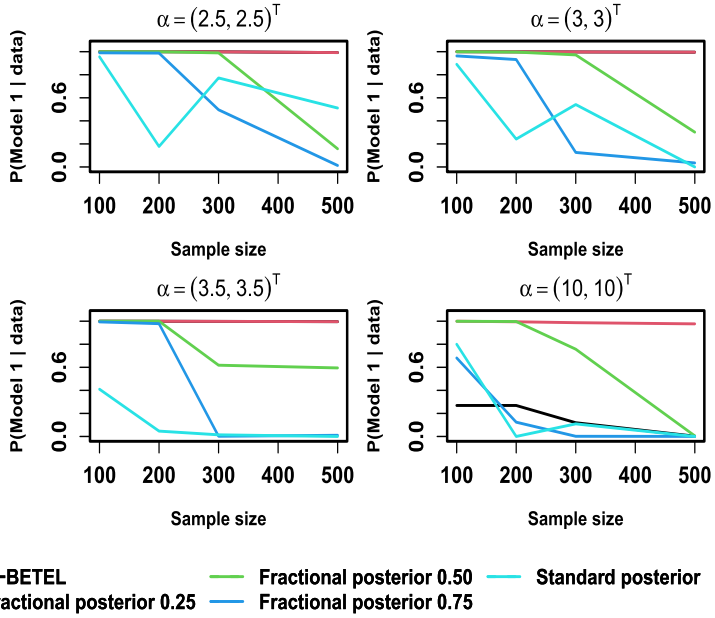

Figure 8: **Model based clustering.** We are comparing the Bayes factor for selecting the simpler model via D-BETEL, the standard posterior, and the fractional posterior (Miller and Dunson, 2019) with different temperatures, across varying values of skewness parameter α of the generating skew normal 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 938, 941, 953

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 937

Azzalini, A. and Dalla Valle, A. (1996). "The multivariate skew-normal distribution." *Biometrika*, 83(4): 715–726. URL https://doi.org/10.1093/biomet/83.4.715 950, 951, 957

Becker, Candès, and Grant (2011). "Templates for convex cone problems with applications to sparse signal recovery." *Mathematical Programming Computation*. 943

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. 962

Bion-Nadal, J. and Talay, D. (2019). "On a Wasserstein-type distance between solutions to stochastic differential equations." *The Annals of Applied Probability*. 945

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 943

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 950, 957

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 958

Cambanis, S., Huang, S. T., and Simons, G. (1981). "On the theory of elliptically contoured distributions." *Journal of Multivariate Analysis*, 11: 368–385. 944

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. 962

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. 962

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. 946

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 937

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 938

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 957

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 938, 941, 961

Chib, S., Shin, M., and Simoni, A. (2021). "Bayesian Estimation and Comparison of Conditional Moment Models."
URL https://arxiv.org/abs/2110.13531 938

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 943

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 939, 942, 945, 947, 948

De Blasi, P., Lijoi, A., and Prünster, I. (2013). "An asymptotic analysis of a class of discrete nonparametric priors." *Statistica Sinica*, 23: 1299–1321. 953

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 938

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 942, 945

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. 962

Du, W. and Wu, X. (2021). "Robust Fairness-aware Learning Under Sample Selection Bias."
URL https://arxiv.org/abs/2105.11570 937

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 937

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 943

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 938, 941, 942, 953, 956

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 937

Gerber, S. and Maggioni, M. (2017). "Multiscale Strategies for Computing Optimal Transport." 942

Gnedin, A. (2009). "Species sampling problems for Gibbs partitions with finitely many types." *Electronic Journal of Probability*, 14: no. 49, 1481–1499. 953

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. 953

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. 943

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 938

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. 962

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. 937

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 959

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 938

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 944

Hooker, G. and Vidyashankar, A. (2011). "Bayesian Model Robustness via Disparities."
URL https://arxiv.org/abs/1112.4213 938

Huber, P. J. (2011). "Robust statistics." In *International encyclopedia of statistical science*, 1248–1251. Springer. 937

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/87.2.371 943

Ishwaran, H. and Zarepour, M. (2002a). "Dirichlet prior sieves in finite normal mixtures." *Statistica Sinica*, 941–963. 954

Ishwaran, H. and Zarepour, M. (2002b). "Exact and approximate sum representations for the Dirichlet process." *Canadian Journal of Statistics*, 30(2): 269–283. 954

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 962

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 938

Johnson, S. G. (2022). "The NLopt nonlinear-optimization package." *The Comprehensive R Archive Network*. 943

Kitagawa, J., Mérigot, Q., and Thibert, B. (2017). "Convergence of a Newton algorithm for semi-discrete optimal transport." 942

Kleijn, B. J. and van der Vaart, A. W. (2012). "The Bernstein-von-Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381. 937

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 938, 949, 961

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 938, 941

Lazar, N. A. (2003). "Bayesian Empirical Likelihood." *Biometrika*, 90(2): 319–326. URL http://www.jstor.org/stable/30042042 938

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 939, 944

Lijoi, A., Prünster, I., and Rigon, T. (2020). "The Pitman–Yor multinomial process for mixture modelling." *Biometrika*, 107(4): 891–906. 963

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. 963

Liu, X., Kong, W., and Oh, S. (2021). "Differential privacy and robust statistics in high dimensions." URL https://arxiv.org/abs/2111.06578 937

McAuliffe, Blei, and Jordan (2006). "Nonparametric empirical Bayes for the Dirichlet process mixture model." *Statistics and Computing*, 1(2): 5–14. 943

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 937, 938, 950, 957, 958, 960, 971

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 942, 953, 954, 962

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 938

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/ 942

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. Wiley-Interscience. 944

Müller, P. and Quintana, F. A. (2004). "Nonparametric Bayesian data analysis." *Statistical science*, 19(1): 95–110. 938

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*, volume 1. Springer. 938

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 944

Nguyen, X. (2013). "Convergence of Latent Mixing Measures in Finite and Infinite Mixture Models." *The Annals of Statistics*, 41(1): 370–400. 962

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. 962

Ning, C. and Ma, X. (2023). "Data-Driven Bayesian Nonparametric Wasserstein Distributionally Robust Optimization."
URL https://arxiv.org/abs/2311.02953 962

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC. 938

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 939, 942

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. 953

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/ 943

Resnick, S. (2013). "A Probability Path." *Birkhäuser Boston*. 955

Schennach, S. M. (2005). "Bayesian exponentially tilted empirical likelihood." *Biometrika*, 92(1): 31–46.
URL https://doi.org/10.1093/biomet/92.1.31 938, 940, 953, 955, 961

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 949

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 937

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 942, 944

Teh, Y. W. (2010). "Dirichlet Process." *Encyclopedia of machine learning*, 1063: 280–287. 938, 941, 942, 956

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 956

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 938, 941

Villani, C. (2003). "Topics in Optimal Transportation." *American Mathematical Society*.
URL https://www.math.ucla.edu/~wgangbo/Cedric-Villani.pdf 939, 942, 949

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 937

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. 937

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50(1): 1–25.
URL http://www.jstor.org/stable/1912526 949, 961

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 937

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 956

Invited Discussion

Antonio R. Linero*

I congratulate Chakraborty et al. (2025) on this stimulating paper, which tries to implement the wisdom that "all models are wrong, but some are useful" (Box and Draper, 1987). How to do this is a central question in modern Bayesian practice: how should we perform Bayesian inference on an interpretable parametric model when we do not believe that the model is literally true? For most scientific problems, we cannot hope to fully reconstruct the data-generating mechanism in microscopic detail, nor do we need to. Rather, we usually want an interpretable, approximately correct model that is useful for guiding further scientific discoveries or policy decisions.

The immediate approach to performing Bayesian inference under model misspecification is to simply proceed as though the model were actually correct. It is well known that if one does this, then (i) the posterior of θ will tend to concentrate around the Kullback-Leibler projection of P onto $\{F_\theta : \theta \in \Theta\}$, but (ii) the resulting credible sets need not have good frequentist properties (White, 1982; Kleijn and van der Vaart, 2006, 2012) and (iii) it is not clear that the Kullback-Leibler projection is the "correct" projection to use. D-BETEL is especially interesting in this regard because, rather than robustifying inference around the Kullback-Leibler projection, it changes the estimand itself (see Section 4). The authors present evidence that this provides D-BETEL with robustness to model contamination and outliers.

D-BETEL has many points of contact with the rich literature on robust Bayesian methods. There are obvious connections to empirical likelihood methods (Owen, 2001), such as the Bayesian empirical likelihood of Lazar (2003) and the Bayesian exponentially tilted empirical likelihood (BETEL) of Schennach (2005); as an aside, I am glad that this underappreciated topic is being picked up again by both the authors and others (Chaudhuri et al., 2017; Chib et al., 2018, 2022; Tang and Yang, 2022; Kim et al., 2023; Yu and Bondell, 2024). I see BETEL and D-BETEL as accomplishing slightly different aims: BETEL addresses how to draw valid inference for a pre-specified projection parameter, whereas D-BETEL aims to soften the projection itself in a P-adaptive fashion. Another connection is with Bayesian nonparametrics, where a flexible prior is centered around a parametric guess or base measure, thereby permitting richer data-generating mechanisms while retaining some connection to the original model (Ferguson, 1973; Müller and Quintana, 2004; Dunson, 2010). The authors connect with Bayesian nonparametrics by identifying their approach with the limiting posterior of a Bayesian nonparametric prior; a similar argument for BETEL is given by Schennach (2005). A third connection is to generalized or pseudo-Bayesian updating, including Gibbs posteriors, power posteriors, coarsened posteriors, and related devices that seek robustness by relaxing the direct identification of posterior updating with the likelihood (Chernozhukov and Hong, 2003; Jiang and Tanner, 2008; Bissiri et al., 2016; Holmes and Walker, 2017; Grünwald

*Department of Statistics and Data Sciences, University of Texas at Austin, antonio.linero@austin.utexas.edu

and van Ommen, 2017; Miller and Dunson, 2019). Each of these approaches addresses parts of the problem, but leaves open others.

The organizing theme of my discussion will center on the following question: *what do I really want from approaches like D-BETEL?* What parameter should we target, how should we measure adequacy of the parametric model, and what kind of uncertainty quantification can we reasonably support? Chakraborty et al. (2025) raise interesting questions about all of these issues, and so provide a good opportunity to think carefully about them.

1 Criteria for Summary Models and D-BETEL

Consider a parametric model $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ that we wish to anchor inference to. Below I provide a highly opinionated wish list of things I would want an inference method to satisfy, given that I have chosen to anchor my inferences to $\mathcal{F}$:

A1 Good Target: The posterior should concentrate around some $\theta^\star$ associated with a "good" approximation $F_{\theta^\star}$ for the true data-generating mechanism P. If possible, this "good" $F_{\theta^\star}$ should have a clear interpretation.

A2 Adequacy Measure: An interpretable measure of adequacy that tells me the extent to which $F_{\theta^\star}$ fails to capture P. Such a measure is useful for determining whether $F_{\theta^\star}$ is a useful summary of P for whatever my purposes are. The adequacy measure might be an overall goodness-of-fit measure, but might also be targeted toward specific inferences.

A3 Calibrated Uncertainty: Ideally, we should have a Bernstein-von Mises theorem of the form

$$d\left(\pi\left\{n^{1/2}(\theta-\widehat{\theta})\in\cdot \mid x_1,\ldots,x_n\right\},\text{Normal}(0,V)\right)\xrightarrow{P}0$$

as $n\to\infty$, where $d(\cdot,\cdot)$ is a divergence measure on probability distributions over Θ, $\widehat{\theta}$ is a semiparametric-efficient estimator of the population-level estimand $\theta^\star$, and V is an appropriate semiparametric efficiency bound. This provides frequentist calibration in the sense that posterior credible intervals are asymptotically valid confidence intervals.

A4 Expandability: When our adequacy measure suggests that $F_{\theta^\star}$ is a poor summary of P, we would like to be able to make richer inferences. That is, we should be able to gracefully expand $\{F_\theta : \theta \in \Theta\}$ into a larger family $\{F_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$ that is more adequate.

A5 Computational Tractability: The method should be easy to implement for a large class of useful choices of $\mathcal{F}$.

A6 Robustness/Insensitivity: Inferences should be robust in the sense of not being sensitive to data contamination, and should be minimally sensitive to small deviations of P from $\mathcal{F}$.

Taking the achievement of these properties as a starting point, we can ask how well D-BETEL meets these criteria. What follows are loose speculations and questions to the authors.

The D-BETEL Target Chakraborty et al. (2025) argue that D-BETEL implicitly targets the parameter $\theta^\dagger$ defined by the following two-step procedure. First, they construct a P-adaptive parametric family $\mathcal{Q}_P = \{\widehat{Q}_\theta : \theta \in \Theta\}$ defined by $\widehat{Q}_\theta = \arg\min_Q[\mathrm{KL}(Q\|P) + \lambda^\star D(F_\theta, Q)]$. Then D-BETEL targets the $\theta^\dagger$ defined by the Kullback-Leibler projection of P onto $\mathcal{Q}_P$. Figure 3 of Chakraborty et al. (2025) suggests that the Kullback-Leibler projection onto $\mathcal{Q}_P$ will be closer to P than the Kullback-Leibler projection onto $\mathcal{F}$, and so there is an argument that targeting $\widehat{Q}_{\theta^\dagger}$ is an improvement over targeting the Kullback-Leibler projection $F_{\theta^\star}$.

Balanced against this, there may be reasons why one would prefer to target $\theta^\star$ over $\theta^\dagger$. One is related to the authors' discussion of generalized linear models. Suppose that we specify a generalized linear model with canonical link in an exponential dispersion family: $f(y \mid x, \beta) = \exp\left\{\frac{yx^\top\beta - b(x^\top\beta)}{\phi} + c(y;\phi)\right\}$. A desirable property of this model is that, even under misspecification, we can consistently recover β provided only that the *mean model* $E(Y_i \mid X_i = x) = b'(x^\top\beta)$ holds; for example, if we model count data using a Poisson loglinear model with $E(Y_i \mid X_i = x) = e^{x^\top\beta}$, then as long as this mean model is correctly specified the posterior distribution should concentrate around the true β_0, even if the outcome is not Poisson (of course, inferences may not be frequentist-calibrated). Additionally, even when the model is misspecified, the $\beta^\star$ corresponding to the Kullback-Leibler projection will have certain calibration properties: for example, when X_i has an intercept term, it is typically the case that $E(e^{X_i^\top\beta^\star}) = E(Y_i)$, so the fitted model is consistent with the marginal mean of Y_i (more generally, the errors are uncorrelated with the X_i's). These properties are desirable, but seem unlikely to be preserved by D-BETEL. Additionally, it is not clear what the appropriate interpretation of $\theta^\dagger$ is, as $F_{\theta^\dagger}$ being interpretable does not make $\widehat{Q}_{\theta^\dagger}$ interpretable. *Can the authors provide guidance for when one would prefer to target $\theta^\dagger$ over $\theta^\star$, or generally provide an argument for targeting $\theta^\dagger$ as a parameter of scientific interest?*

Uncertainty Quantification Chakraborty et al. (2025) do not speculate about whether the D-BETEL posterior satisfies a Bernstein-von Mises theorem (BVMT). On the one hand, it is natural to conjecture that D-BETEL would permit a BVMT given its proximity to BETEL, for which Chib et al. (2018) establishes a BVMT. On the other hand, the authors suggest in Section 4 that, as $\epsilon \downarrow 0$, we recover inference in the parametric model, and we know that inference based on a misspecified parametric model will not be calibrated in general. *Do the authors believe that a BVMT holds for D-BETEL and, if not, is there a variant of D-BETEL for which a BVMT* does *hold? Relatedly, what does optimal frequentist inference for $\theta^\dagger$ look like?*

Adequacy Measure D-BETEL comes with an adequacy measure $D(F_\theta, \nu) \leq \epsilon$, which for ANDREW measures how far the measure $\nu = \sum_i w_i \delta_{X_i}$ deviates in W_{AR} from F_θ.

The authors propose selecting ϵ using leave-one-out cross-validation, which can then also be used as an overall estimate of adequacy. If computationally feasible, it may also be useful to consider the "solution path" of the θ posterior distribution as ϵ increases, as this would allow us to understand how the posterior of θ changes as we tolerate larger and larger deviations from the parametric model. Such a tool would be useful because it is probably difficult to explain the meaning of the statement $D(F_\theta, \nu) \leq \epsilon$ to collaborators in a scientifically meaningful way. *Do the authors see W_{AR} or related distances in general as useful summaries of adequacy, and is it feasible or useful to look at the solution path of the posterior in ϵ instead?*

Computation A large obstacle to the routine use of D-BETEL is computation. Chakraborty et al. (2025) focus on the special case of elliptical mixture models (EMMs) and their restricted Wasserstein metric W^2_{AR}; this limits D-BETEL to the setting where $\mathcal{F}$ is an EMM. Despite this limitation, the authors are able to apply their methodology to a Poisson loglinear model by heuristically treating the score evaluations $\eta_i(\beta)$ as coming from an EMM, and so it seems that ANDREW is more broadly applicable than it would initially seem to be; on the other hand, it is unclear how generally one can apply this trick. I would be very excited to see the development of computational methods for D-BETEL in arbitrary models. *Are there computational directions for future work that would resolve these issues?*

Robustness One of the main motivations of Chakraborty et al. (2025) is to draw inferences that are robust to data contamination and outliers, and it is likely that D-BETEL performs particularly well on this criterion relative to competing methods. In their Supplementary Material, Chakraborty et al. (2025) argue that D-BETEL is robust to model contamination and offer a mixture of theoretical and empirical support for this claim. This is also the basis for their comparison with BETEL in Section 6.2. *Is there a path toward a complete explanation of the robustness properties of D-BETEL?*

Expandability One possible area for future work is to expand D-BETEL to allow for the adaptive choice of the underlying parametric model $\mathcal{F}$. If the true data-generating mechanism is sufficiently far away from $\mathcal{F}$, then rather than robustifying inference to this fact it may be more insightful to replace $\mathcal{F}$ with a more complex model. The goal here would not be to capture the true data-generating mechanism — I agree with the authors that this goal is "often futile" — but rather to pivot to a different parametric model when $\mathcal{F}$ is inadequate. Alternatively, we may have multiple *qualitatively different* models $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_M$, and hope to both (i) select the best summarizing model and (ii) remain robust to misspecification of that model. *Is there a feasible way to anchor D-BETEL to multiple parametric models at once, or use it to decide between parametric models?*

2 Comparison to Other Approaches

For the sake of putting my assessment of D-BETEL on A1–A6 in context, Table 1 lists the performance of some competing approaches along A1–A6 as well. The values in

Method	Target	BVMT	Measure	Computation	Expandability	Robust
$\mathcal{M}$-completion	✓	≈	✓	✓	✓	✗
Bayesian Bootstrap	✓	✓	≈	✓	?	≈
BETEL	✓	✓	?	✗	?	≈
D-BETEL	—	—	—	—	—	—

Table 1: This table represents the author's personal opinion about the current state of affairs at the time of writing this paper. Possible values are usually satisfactory (✓), sometimes satisfactory (≈), in need of improvement (✗), or uncertain (?).

the table represent my personal opinion, with the entries for D-BETEL left blank for comment by the authors.

$\mathcal{M}$-Completion Bernardo and Smith (1994) divide views of Bayesian model selection into the $\mathcal{M}$-closed, $\mathcal{M}$-open, and $\mathcal{M}$-*completed* perspectives. In the $\mathcal{M}$-completed perspective, one constructs a reference model $\mathcal{G}$ that is large enough to contain both the family $\mathcal{F}$ and the true data-generating mechanism P. Given such a reference model, we can place a prior on $\mathcal{G}$ and explicitly target the best summarization of P by a member of $\mathcal{F}$; typically this would be the Kullback-Leibler projection $\theta^\star = \arg\min_\theta \mathrm{KL}(P \| F_\theta)$, which can also serve as a measure of summarization adequacy, but conceptually one could also target the parameter $\theta^\dagger$ defined by D-BETEL. Workflows for basing inference on posterior projections within the $\mathcal{M}$-completed framework are given, for example, by McLatchie et al. (2025); Woody et al. (2021). Bernstein-von Mises-type results are also available for the $\mathcal{M}$-completion approach, even under nonparametric models (Xie and Xu, 2021; Linero, 2025), although the conditions required are quite strong and require estimating nuisance functions at relatively fast rates (by contrast, BETEL and the Bayesian bootstrap have BVM results without requiring estimation of nuisance functions at all). Whether $\mathcal{M}$-completion is a viable strategy computationally depends on the feasibility of sampling from a posterior over $\mathcal{G}$ and computing the projection. It is also an expansive strategy in the sense that we can consider possibly many summarizing models $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_M$ and actively explore which one provides the best model summary relative to the reference model. I would not describe the $\mathcal{M}$-completion approach as being particularly robust, however; all of the usual questions of misspecification of the likelihood and prior are moved from $\mathcal{F}$ to $\mathcal{G}$, and specifying reasonable priors on very large model spaces is challenging. As argued by McLatchie et al. (2025), the feasibility of this strategy is mainly determined by our ability to construct a strong reference model and a suitable prior over it.

Bayesian Bootstrap An extreme take on the $\mathcal{M}$-completion approach is to place no restrictions whatsoever on the reference $\mathcal{G}$, with a Bayesian bootstrap used for inference (Rubin, 1981). One can then define summarizing models as minimizing a specified loss function: $\theta^\star = \arg\min_\theta \int \ell(\theta, x)\, F(dx)$ where $F = \sum_i w_i\, \delta_{X_i}$ and $(w_1, \ldots, w_n) \sim \mathrm{Dirichlet}(1, \ldots, 1)$ in the posterior. Lyddon et al. (2019) refer to this strategy as the loss-likelihood bootstrap, and when $\ell(\theta, x) = -\log f_\theta(x)$ we expect it to target the same Kullback-Leibler projection that is targeted by BETEL. The Bayesian bootstrap scores

well on most of our desired criteria. It has a good target, satisfies a Bernstein-von Mises result under quite general conditions, and exact samples can often be drawn from the posterior. While not intrinsically robust, this can be partially controlled by specifying a robust loss function $\ell(\theta, x)$ rather than the log-likelihood — perhaps it could even be chosen to target $\theta^\dagger$. The fact that the posterior is based on an improper prior, however, makes me suspicious of taking at face value the realized quantity $\min_\theta \int \ell(\theta, x)\, F(dx)$ as a measure of adequacy, or of trying to use it to decide whether to expand the model.

Score-Based BETEL The usual BETEL can be anchored to a parametric model F_θ by taking the associated moment conditions to be the score equations $E\left\{\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right\} = 0$. This targets the $\theta^\star$ corresponding to the Kullback-Leibler projection of P onto $\mathcal{F}$. BETEL provides a reasonable target parameter $\theta^\star$ and, as shown by Chib et al. (2018), BETEL satisfies a Bernstein-von Mises theorem under suitable conditions. In generalized linear models, the score equations can continue to identify a meaningful parameter even under substantial distributional misspecification, so BETEL inherits some of the robustness of estimating-equation methods. BETEL can also be expected to work well under more general conditions than $\mathcal{M}$-completion, and so is more robust, although Chakraborty et al. (2025) show that it is less robust than D-BETEL to data contamination. Its weaker points, in my view, are the lack of an adequacy measure, limited expandability, and limited computational tractability. There is no intrinsic scalar measure of adequacy analogous to D-BETEL's ϵ, although motivated by Chakraborty et al. (2025) one might consider using Wasserstein distance as a possible diagnostic. The method also does not itself say when the parametric anchor should be enlarged, although again Chib et al. (2018) show that one can test for the inclusion or exclusion of additional moment conditions. Finally, evaluating the BETEL likelihood requires solving an optimization problem at each proposed value of θ, which makes sampling from the posterior challenging. Even so, my guess is that BETEL will generally be easier to implement than D-BETEL.

3 Concluding Remarks

D-BETEL is compelling because it tries to combine several things that are rarely available at once: an interpretable parametric anchor, an intrinsic notion of adequacy, robustness to data contamination, and a justification as an approximation to genuine nonparametric Bayesian inference. At the same time, the paper leaves several important questions open. When should one prefer the target $\theta^\dagger$ to the usual projection parameter $\theta^\star$? What kind of semiparametrically valid uncertainty quantification, if any, can be established for the resulting posterior? How broadly can the method be extended beyond the current EMM/ANDREW setting while retaining its practical appeal? I therefore see Chakraborty et al. (2025) as a persuasive proposal for a new design principle that I hope will be built upon in future work.

References

Bernardo, J. M. and Smith, A. F. (1994). *Bayesian Theory* volume 586. Wiley Online Library. doi:10.1002/9780470316870. 974

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 78, 1103–1130. doi:10.1111/rssb.12158. 970

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons. MR0861118. 970

Chakraborty, A., Bhattacharya, A., and Pati, D. (2025). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis* 1, 1–33. Advance online publication. doi:10.1214/25-BA1535. 970, 971, 972, 973, 975

Chaudhuri, S., Mondal, D., and Yin, T. (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 79, 293–320. doi:10.1111/rssb.12164. 970

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics* 115, 293–346. doi:10.1016/S0304-4076(03)00100-3. 970

Chib, S., Shin, M., and Simoni, A. (2018). Bayesian estimation and comparison of moment condition models. *Journal of the American Statistical Association* 113, 1656–1668. doi:10.1080/01621459.2017.1358172. 970, 972, 975

Chib, S., Shin, M., and Simoni, A. (2022). Bayesian estimation and comparison of conditional moment models. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 84, 740–764. doi:10.1111/rssb.12484. 970

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.), *Bayesian Nonparametrics*, 223–273. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802478.008. 970

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1, 209–230. doi:10.1214/aos/1176342360. 970

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis* 12, 1069–1103. doi:10.1214/17-BA1085. 970

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* 104, 497–503. doi:10.1093/biomet/asx010. 970

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics* 36, 2207–2231. doi:10.1214/07-AOS547. 970

Kim, E., MacEachern, S. N., and Peruggia, M. (2023). Regularized exponentially tilted empirical likelihood for Bayesian inference. arXiv:2312.17015. 970

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* 34, 837–877. doi:10.1214/009053606000000029. 970

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics* 6, 354–381. doi:10.1214/12-EJS675. 970

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika* 90, 319–326. doi:10.1093/biomet/90.2.319. 970

Linero, A. R. (2025). Defensive model expansion for robust Bayesian inference. arXiv:2510.09598. doi:10.48550/arXiv.2510.09598. 974

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* 106, 465–478. doi:10.1093/biomet/asz006. 974

McLatchie, Y., Rögnvaldsson, S., Weber, F., and Vehtari, A. (2025). Advances in projection predictive inference. *Statistical Science* 40, 128–147. doi:10.1214/24-sts949. 974

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association* 114, 1113–1125. doi:10.1080/01621459.2018.1469995. 970

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* 19, 95–110. doi:10.1214/088342304000000017. 970

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC. 970

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics* 9, 130–134. MR0600538. 974

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika* 92, 31–46. doi:10.1093/biomet/92.1.31. 970

Tang, R. and Yang, Y. (2022). Bayesian inference for risk minimization via exponentially tilted empirical likelihood. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 84, 1257–1286. doi:10.1111/rssb.12510. 970

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25. doi:10.2307/1912526. 970

Woody, S., Carvalho, C. M., and Murray, J. S. (2021). Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics* 30, 144–161. doi:10.1080/10618600.2020.1796684. 974

Xie, F. and Xu, Y. (2021). Bayesian projected calibration of computer models. *Journal of the American Statistical Association* 116, 1965–1982. doi:10.1080/01621459.2020.1753519. 974

Yu, W. and Bondell, H. D. (2024). Variational Bayes for fast and accurate empirical likelihood inference. *Journal of the American Statistical Association* 119, 1089–1101. doi:10.1080/01621459.2023.2169701. 970

Invited Discussion

Yuexi Wang*

I congratulate the authors on the thought-provoking paper. Their constrained transport construction offers a principled way to robustify Bayesian inference while retaining an interpretable centering model, and it also naturally connects to a broader literature in statistics and machine learning.

In this discussion, we first review the proposed method in that broader literature, emphasizing its potential relevance beyond the scenarios considered in the paper. We then discuss two possible extensions: an interpretation of D-BETEL as implicit model expansion and a possible alternative inside the ANDREW metric through sliced Wasserstein comparisons.

1 Distribution Matching as a Unifying Theme

A common theme in many areas of statistics and econometrics is the need to compare or align distributions, especially in settings with intractable likelihoods, missing data, or model misspecification. Here we organize the related literature along two axes: the discrepancy used to measure distributional similarity, and the inferential problem in which the discrepancy is used.

Along the first axis, the earlier developments begin by comparing selected properties of distributions, such as moments or other summary statistics chosen by convention or expert knowledge. While this reduces the dimensionality of the matching problem, performance depends on whether the summaries retain sufficient information relevant to the target parameter and information loss is often unavoidable. A second generation of methods resorts to the Kullback-Leibler (KL) divergence or likelihood ratio, since these quantities connect naturally to full likelihood inference. When exact computation of KL is infeasible, modern machine-learning methods provide flexible nonparametric alternatives through the classification trick, which approximates the likelihood ratio by a binary classifier (Wang et al., 2022). However, they can become unsuitable when the model and empirical distribution have different supports, such as when comparing a discrete distribution to a continuous one. This has motivated the adoption of Wasserstein distances and related integral probability metrics (IPMs) such as the energy distance and maximum mean discrepancy (MMD), which are less sensitive to support mismatch and often more stable for empirical comparisons. These developments are at the heart of the current work.

Along the second axis, analogous progressions have unfolded across several research areas, summarized in Table 1. In semiparametric inference, the development has proceeded from generalized method of moments (GMM, (Hansen, 1982)) and empirical likelihood (EL, (Owen, 2001)), to KL-based exponentially tilted empirical likelihood

*Department of Statistics, University of Illinois Urbana-Champaign, yxwang99@illinois.edu

	Moments or summaries	KL-related	Wasserstein or IPM
Robust inference (frequentist)	M-estimating equations; EL (Owen, 2001; Schennach, 2007)	Quasi-likelihood (Wedderburn, 1974)	Distributionally robust optimization (DRO) (Blanchet et al., 2022)
Robust inference (Bayesian)	Gibbs posterior with general loss (Jiang and Tanner, 2008; Bissiri et al., 2016); Generalized posterior with estimating equations (Chernozhukov and Hong, 2003)	Coarsened posterior (Miller and Dunson, 2019); Safe Bayes (Grünwald and van Ommen, 2017)	Constrained transport posterior (this paper); MMD-based Gibbs posterior (Chérief-Abdellatif and Alquier, 2020)
Semiparametric estimation	GMM (Hansen, 1982); Indirect inference (Gourieroux et al., 1993)	Information criteria (Kitamura and Stutzer, 1997)	Minimum Wasserstein estimation (Bernton et al., 2019b)
Distribution balancing for causal inference	IPW (Robins and Rotnitzky, 1995); Covariate balancing propensity score (Imai and Ratkovic, 2014)	Entropy balancing (Hainmueller, 2012)	IPM-based balancing (Kong et al., 2023); Characteristic function distance (Santra et al., 2026)
Simulation-based inference	ABC with summary statistics (Beaumont et al., 2002); Simulated method of moments (McFadden, 1989); Synthetic likelihood (Wood, 2010; Price et al., 2018);	Classifier-based likelihood ratio (Cranmer et al., 2015; Wang et al., 2022); Neural likelihood/ratio estimation (NLE/NRE) (Papamakarios et al., 2019; Miller et al., 2022)	Wasserstein ABC (Bernton et al., 2019a); Sliced-Wasserstein inference; MMD-based ABC (Park et al., 2016)

Table 1: A non-exhaustive list of distribution matching ideas across research areas.

(ETEL, (Schennach, 2007)), to minimum Wasserstein estimators (Bernton et al., 2019b). In causal inference, especially in observational studies, balancing the pre-treatment covariate distribution for different treatment groups is often essential for recovering causal estimates such as the average treatment effect. Methods have evolved from inverse probability weighting (IPW, (Imai and Ratkovic, 2014)), to doubly robust estimators that combine propensity score and outcome models, to more recent Wasserstein- and energy-distance-based distributional balancing (Kong et al., 2023; Santra et al., 2026). Similarly, simulation-based inference, whose core challenge is measuring distributional similarity between simulated and observed data, has progressed from approximate Bayesian computation (ABC) with summary statistics (Beaumont et al., 2002), through synthetic likelihood (Wood, 2010; Price et al., 2018) and classifier-based KL discrepancies, to Wasserstein and energy distances applied directly to empirical samples.

The paper fits naturally into this progression under the robust inference framework. Moreover, it proposes a novel computational approach for conducting inference conditional on a Wasserstein neighborhood of the model, which is often either intractable

or computationally prohibitive. With its computational advantage, the proposed metric can be applied to a wide range of problems where distribution matching is needed, such as the ones mentioned above.

2 D-BETEL as Implicit Model Expansion

While parametric models are rarely exact, in many applications, the parametric family F_θ is regarded as a working (centering) submodel embedded in a larger semiparametric model for the true data-generating distribution. This work provides a novel way to leverage the constrained transport metric to implicitly expand the model family, allowing for a more flexible approximation of the true distribution while retaining enough structure for tractable inference.

The authors point out that the population-level target of the distributionally constrained model, which defines the likelihood of D-BETEL, can be represented through an enlargement of F_θ as

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

This objective balances the data-fidelity term $\mathrm{KL}(Q\|P)$ against the model-proximity term $D(F_\theta, Q)$. A related, but more explicit, model expansion is considered in robust simulation-based inference by Tomaselli et al. (2025), with the baseline likelihood multiplied by an exponential tilt

$$f_{\theta,\beta}(x) \propto f_\theta(x) \exp\{\beta^T b(x)\},$$

where $b(x) = (b_1(x), \ldots, b_k(x))$ is a vector of fixed functions and $f_{\theta,\beta} = f_\theta$ when $\beta = 0$.

This comparison suggests two possible directions for future research. First, one could study whether the constrained-transport construction can be applied to simulation-based inference problems, providing a data-driven way to expand the model rather than specifying the form of expansion in advance. Second, one could explore directly approximating $\hat{Q}_\theta$ by a flexible tilted family, possibly with nonparametric basis functions, and use this approximation to implement D-BETEL beyond the elliptical mixture models considered in the paper.

3 Extension to Sliced Wasserstein Distance

Another natural question is whether it is possible to extend the distance metric in D-BETEL to sliced Wasserstein distance (Rabin et al., 2011; Bonneel et al., 2015) and its variants, which have been widely used in distribution-matching problems because of their computational efficiency and fast convergence rates.

Sliced Wasserstein Distance The sliced Wasserstein distance (SWD) of order p is defined by averaging one-dimensional Wasserstein distances over all projection directions,

$$\mathrm{SW}_p^p(P, Q) = \int_{S^{d-1}} W_p^p\big(\mathcal{P}_\xi P,\, \mathcal{P}_\xi Q\big)\, d\sigma(\xi), \tag{1}$$

where σ is the uniform measure on S^{d-1}. Since the projection $\mathcal{P}_\xi$ is a linear map, the projected distribution $\mathcal{P}_\xi P$ is a one-dimensional distribution, and the Wasserstein distance $W_p(\mathcal{P}_\xi P, \mathcal{P}_\xi Q)$ can be computed efficiently by sorting the projected samples.

Adopting SWD in Defining the Misspecification Neighborhood Since SWD also satisfies the key ingredients for the metric D in the D-BETEL framework, one can consider replacing $D = W_{AR}$ with $D = SW_p$ (with appropriate modifications) in the D-BETEL framework.

A statistical advantage of SW_p over W_p in moderate-to-high dimensions is its faster convergence rate: the plug-in estimator satisfies $SW_p(P_n, P) = O_p(n^{-1/2})$ (Nadjahi et al., 2019, 2020), in contrast to the curse-of-dimensionality rate $W_p(P_n, P) = O_p(n^{-1/d})$ for the full Wasserstein distance. This also suggests that the radius ε of the misspecification neighborhood can be calibrated at the parametric rate $O(n^{-1/2})$ for SW_p, while W_p suggests a slower rate $O(n^{-1/d})$ to achieve the same statistical guarantees. A similar phenomenon is studied in the context of distributionally robust optimization by Montiel Olea et al. (2026).

However, a direct application of SWD in the D-BETEL framework would not feature the same tractability as the modified W2 metric $\mathrm{MW}^2_{2,\mathrm{EMM},\alpha}$, since the convenient explicit parameter-space decomposition under the EMM family will not hold for SWD. Such an extension would require careful modification of SWD to achieve computational tractability while retaining the statistical advantages, such as introducing an entropy regularization term. Additionally, using SWD would also complicate the analytical transparency in parameter space.

Adopting SWD for Joint Tail Matching One limitation of the modified W_2 distance is that its expression relies solely on the first- and second-order moments of F_θ, which may not be sufficient to capture the tail behavior. The authors propose ANDREW as a modification to $\mathrm{MW}^2_{2,\mathrm{EMM},\alpha}$ to match the quantiles of the marginal distributions, in order to capture the tail behavior.

This is another place one might consider replacing the matching on the marginals in ANDREW with matching on the sliced marginals, which would be more flexible in terms of capturing the joint tail behavior. For example, let $t_\nu(0, \Sigma)$ denote a bivariate Student-t distribution with $\nu > 2$ degrees of freedom, location zero, and scale matrix Σ. Consider

$$p_0 = t_\nu(0, \Sigma_0), \qquad p_0' = t_\nu(0, \Sigma_0'), \qquad \Sigma_0 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \Sigma_0' = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

These two distributions have the same coordinate-wise marginals and the same trace of the scale matrix, but their joint tail dependence differs along diagonal directions. In the D-BETEL setting where the second argument is the weighted empirical measure $\nu_{w,x}$, the closed-form ANDREW distance depends on the component means, covariance trace, and coordinate-wise marginal quantiles; hence it might not be able to distinguish p_0 from

p_0' in this example. Sliced marginal matching, however, would distinguish them through projections such as $(1,1)/\sqrt{2}$ and $(1,-1)/\sqrt{2}$. Therefore, replacing or augmenting the marginal matching in ANDREW with sliced marginal matching would be more flexible in terms of capturing joint tail structure while preserving the paper's idea of using tractable one-dimensional quantile comparisons.

In summary, we view the constrained transport construction as an important contribution to robust inference. By connecting model-based inference with distributional matching, the paper opens several promising directions for future work, including more flexible implicit model expansions and variants based on sliced Wasserstein comparisons. We look forward to seeing how these ideas develop in broader inferential settings.

References

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035. doi:10.1093/genetics/162.4.2025. 979

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019a). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 81, 235–269. doi:10.1111/rssb.12312. 979

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019b). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA* 8, 657–676. doi:10.1093/imaiai/iaz003. 979

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 78, 1103–1130. doi:10.1111/rssb.12158. 979

Blanchet, J., Murthy, K., and Zhang, F. (2022). Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research* 47, 1500–1529. doi:10.1287/moor.2021.1178. 979

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51, 22–45. doi:10.1007/s10851-014-0506-3. 980

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, 1–21. PMLR. MR4147569. 979

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics* 115, 293–346. doi:10.1016/S0304-4076(03)00100-3. 979

Cranmer, K., Pavez, J., and Louppe, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. arXiv:1506.02169. doi:10.48550/arXiv.1506.02169. 979

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* 8, S85–S118. doi:10.1002/jae.3950080507. 979

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis* 12, 1069–1103. doi:10.1214/17-BA1085. 979

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20, 25–46. doi:10.1093/pan/mpr025. 979

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 1029–1054. doi:10.2307/1912775. 978, 979

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 76, 243–263. doi:10.1111/rssb.12027. 979

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics* 36, 2207–2231. doi:10.1214/07-AOS547. 979

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 861–874. doi:10.2307/2171942. 979

Kong, I., Park, Y., Jung, J., Lee, K., and Kim, Y. (2023). Covariate balancing using the integral probability metric for causal inference. In *International Conference on Machine Learning*, 17430–17461. PMLR. 979

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 995–1026. doi:10.2307/1913621. 979

Miller, B. K., Weniger, C., and Forré, P. (2022). Contrastive neural ratio estimation. In *Advances in Neural Information Processing Systems*, volume 35, 3262–3278. 979

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association* 114, 1113–1125. doi:10.1080/01621459.2018.1469995. 979

Montiel Olea, J. L., Rush, C., Velez, A., and Wiesel, J. (2026). The distributionally robust prediction error of the LASSO and related estimators. *The Annals of Statistics* 54, 1006–1027. doi:10.1214/25-AOS2599. 981

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. (2020). Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, 20802–20812. 981

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. (2019). Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 32. 981

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC. doi:10.1201/9781420036152. 978, 979

Papamakarios, G., Sterratt, D., and Murray, I. (2019). Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 837–848. PMLR. 979

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, 398–407. PMLR. 979

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics* 27, 1–11. doi:10.1080/10618600.2017.1302882. 979

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446. Springer. doi:10.1007/978-3-642-24785-9_37. 980

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 90, 122–129. doi:10.1080/01621459.1995.10476494. 979

Santra, D., Chen, G., and Park, C. (2026). Distributional balancing for causal inference: A unified framework via characteristic function distance. arXiv:2601.15449. doi:10.48550/arXiv.2601.15449. 979

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics* 35, 634–672. doi:10.1214/009053606000001208. 979

Tomaselli, L., Ventura, V., and Wasserman, L. (2025). Robust simulation based inference. arXiv:2508.02404. doi:10.48550/arXiv.2508.02404. 980

Wang, Y., Kaji, T., and Rockova, V. (2022). Approximate Bayesian computation via classification. *Journal of Machine Learning Research* 23, 1–49. MR4577789. 978, 979

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika* 61, 439–447. doi:10.1093/biomet/61.3.439. 979

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* 466, 1102–1104. doi:10.1038/nature09319. 979

Invited Discussion

Jonathan H. Huggins*

1 Introduction

We thank Chakraborty et al. (2025) for a stimulating contribution to one of the most important foundational questions in contemporary Bayesian inference: how should we represent the belief that a parametric model is close to, but only *approximately*, correct? Answering this question forces a confrontation with the fundamental goals of the statistical analysis — specifically, how to trade off statistical efficiency (for example, in a predictive sense) against robustness and interpretability (Li et al., 2026; Miller and Dunson, 2019; Shmueli, 2010). For D-BETEL, the trade-off is mediated by a global tolerance parameter ϵ that governs how close the empirical likelihood must be to the parametric family $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ under the data-informed augmented-restricted Wasserstein metric. Yet how to choose ϵ is a problem-specific question.

Similar issues arise in related work such as the coarsened posterior approach of Miller and Dunson (2019), where the trade-off is made by a different but conceptually similar mechanism. Rather than conditioning on the observed data exactly, one conditions on the event that the data distribution lies within a KL distance R of the model — informally, $\mathrm{KL}(\widehat{F}_n \,\|\, F_\theta) < R$ — with R assigned an exponential prior of rate α. The inverse rate captures the tolerance to misspecification, so large α encodes less tolerance to misspecification. The coarsened posterior can be approximated using the power likelihood $F_\theta^{\zeta_n}$ with $\zeta_n = \alpha/(\alpha + n)$ in place of F_θ. The optimistically weighted likelihood (OWL; Dewaskar et al., 2025) method fits into the same general framework, resulting in a different form of robustified likelihood with a single tunable robustness parameter.

Because all these methods expose a single scalar tolerance to the user, the natural follow-on is how to select it in a task-appropriate manner. In robust regression and prediction, the right tolerance is the one that controls expected predictive loss on held-out data, and selector based on the expected log probability density (ELPD) or cross-validation error is appropriate. For unsupervised representation learning — mixture modeling, factor analysis, and so on — the goal is instead to robustly recover an *interpretable* model whose component structure does not overfit to small deviations from the parametric form. The right tolerance in this regime is the one that suppresses spurious latent structure (Li et al., 2026; Miller and Dunson, 2019; Moran and Aragam, 2025). I focus on the unsupervised setting throughout this discussion, as it is where the choice of ϵ matters most while also being the hardest to calibrate. I compare D-BETEL to the coarsened posterior throughout, and leave a systematic comparison to other methods such as OWL as an important direction for future work.

*Department of Mathematics & Statistics and Faculty of Computing & Data Sciences, Boston University, huggins@bu.edu

2 D-BETEL Versus the Calibrated Coarsened Posterior

In the model-selection experiment of §6.1, Chakraborty et al. (2025) compare against a power posterior with fixed power; this does not, however, match the theory or methods of Miller and Dunson (2019) for the coarsened posterior, where the power must be tuned in a sample-size dependent manner. I therefore reproduce the experiment, adding the calibrated coarsened posterior as a comparator with the power parameter ζ selected per replicate by Miller and Dunson's elbow rule. Throughout, I use the published D-BETEL code from Chakraborty et al. (2025) with two modifications:

1. I draw Σ from a state-dependent Wishart proposal in both the M_1 and M_2 chains and evaluate the Metropolis–Hastings acceptance probability under the same kernel. The released code uses a fixed-scale Wishart proposal in the M_1 chain but evaluates the acceptance probability under a state-dependent Wishart. The difference does not appear to have much affect on the practical behavior of the chains.

2. I report D-BETEL's $P(M_K \mid X)$ using the Laplace approximation. The released code cites Chib and Jeliazkov (2001) but evaluates $\text{logsumexp}_t \ell_{(t)}$, the log-sum of D-BETEL posterior sample log-likelihoods, which is not a consistent estimator of $\log m(X)$ (Friel and Wyse, 2012; Llorente et al., 2023). I also implemented the Chib and Jeliazkov (2001) estimator but found it to be unstable.

Figure 1 adds the calibrated coarsened posterior (purple) to my reproduction of the panels of Chakraborty et al. (2025, Figure 8). D-BETEL concentrates near $P(M_1 \mid X) = 1$ across the entire skew range, and the calibrated coarsened posterior behaves similarly, with $P(M_1 \mid X) \approx 0.95$ across all (γ, n) cells. The gap to a single bivariate Gaussian in Chakraborty et al.'s Figure 8 — where the standard posterior and fixed-η fractional posteriors collapse onto M_2 as n grows — is therefore largely a calibration gap rather than a structural one.

The data panels in Figure 2 provide some insight: at $n = 500$, the $\gamma = 2.5$ data look like a mildly skewed unimodal cloud, and a two-component normal mixture fit assigns roughly equal weight to two largely-overlapping components rather than recovering two distinct sub-populations. Both D-BETEL and the calibrated coarsened posterior are robust to this type of model deviation.

3 The Challenge of Heterogeneous Misspecification

In Li et al. (2026), we identify a structural failure mode of the coarsened posterior: when misspecification is unevenly distributed across mixture components, a globally defined scalar tolerance can overfit to dominant misspecified components. This issue can be observed in the case of data from a univariate skew-normal mixture

$$x_i \sim \tfrac{1}{2}\,\text{SN}(-3, 1, \gamma_1) + \tfrac{1}{2}\,\text{SN}(3, 1, \gamma_2),$$

fit with a K-component Gaussian mixture model and compared in two regimes: a `same` regime ($\gamma_1 = \gamma_2 = -10$) in which both clusters are equally misspecified, and

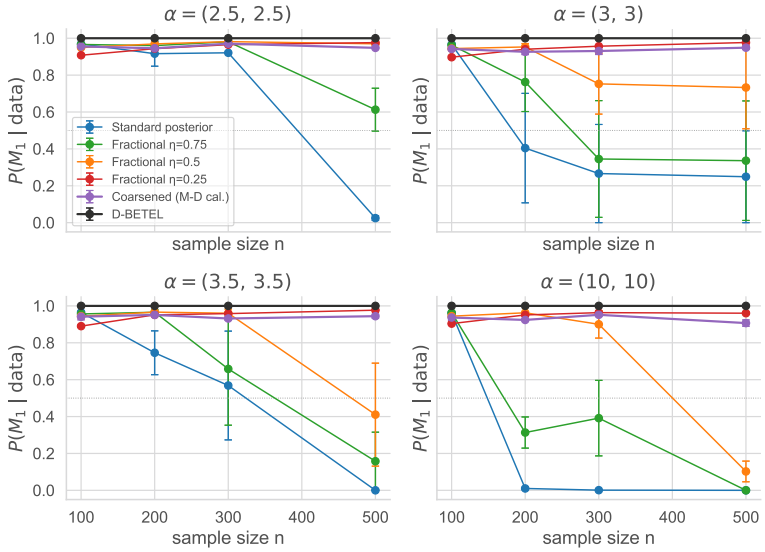

Figure 1: $P(M_1 \mid X)$ as a function of n on the bivariate skew-normal data of Chakraborty et al. (2025, §6.1), for skew $\gamma \in \{2.5, 3, 3.5, 10\}$. Error bars are ± 1 standard error across three replicates. The standard posterior (blue) and fixed-$\eta = 0.5, 0.75$ fractional posteriors (green, orange) collapse toward M_2 as n grows. D-BETEL (black) puts $P(M_1 \mid X) \approx 1$ across the full skew range (Laplace approximation at the chain's MAP, used uniformly across K). The calibrated coarsened posterior of Miller and Dunson (2019) (purple) also concentrates on M_1 at $P(M_1 \mid X) \approx 0.91$–$0.97$ across all (γ, n) cells.

a `different` regime ($\gamma_1 = -1$, $\gamma_2 = -10$) in which one cluster is mildly skewed and the other strongly. I fit D-BETEL and the calibrated coarsened posterior on this benchmark for $K \in \{1, 2, 3\}$ at $n \in \{200, 500, 1000\}$ with three replicates each, with α selected once per replicate using the Miller and Dunson elbow rule applied to the phase plot of posterior-mean log-likelihood against $E[k_{2\%}]$.

Figure 3 reports the posterior model probabilities $P(M_K \mid X)$ under a uniform prior on $K \in \{1, 2, 3\}$. The calibrated coarsened posterior spreads its mass between $K = 2$ and $K = 3$ in both regimes; the $K = 2$ share is roughly stable around 0.42–0.45 in the `same` regime and around 0.31–0.44 in the `different` regime. In the `different` regime $P(M_3 \mid X)$ averages about 0.61 (range 0.56–0.69), consistent with the early heterogeneous-misspecification overfitting predicted by Li et al. (2026) but not yet showing the runaway collapse when $n \approx 10^4$. The standard posterior, by contrast, has $P(M_3 \mid X) \approx 1$ in both regimes by $n = 500$, so the global tolerance α is preventing the full collapse even where it cannot prevent the local drift toward $K = 3$.

D-BETEL goes the other way: $P(M_1 \mid X) \approx 1$ in every (n, regime) cell of the grid, with the marginal likelihood gap between $K = 1$ and $K = 2$ on the order of 14 nats. This is the same behavior D-BETEL showed seen in Figure 1: the slack in

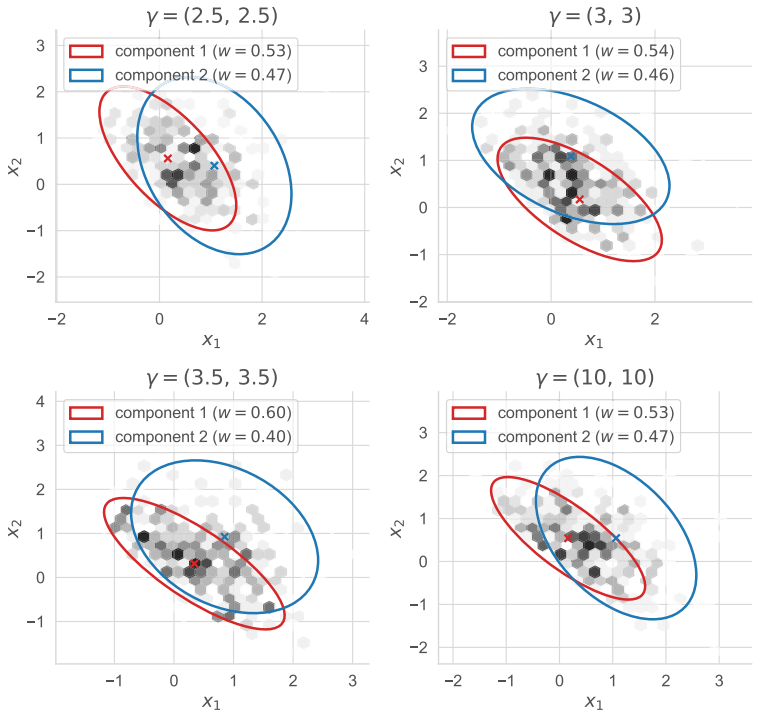

Figure 2: One replicate of the bivariate skew-normal data at $n = 500$ for each skew setting in Figure 1 (greyscale hexbins), with the two-component-mixture standard posterior fit overlaid as 2σ ellipses. In every panel the two components carry roughly equal mass and their 2σ ellipses largely overlap — so even when the standard posterior strongly prefers M_2, the two components do not define visibly distinct "sub-populations".

the optimal transport absorbs the inter-cluster gap rather than requiring a $K = 2$ mixture to explain it. Hence D-BETEL avoids over-fragmentation under heterogeneous misspecification, instead preferring $K = 1$ in *both* regimes — including the `same` regime, where the calibrated coarsened posterior correctly puts substantial mass on $K = 2$.

Using a predictive-accuracy criterion results in a slightly more complicated picture. Applied to D-BETEL, the Pareto-smoothed importance-sampled leave-one-out (LOO) estimator (ELPD-LOO, Vehtari et al., 2017) tends to prefer smaller K as sample size increases, An interesting follow-on question is whether D-BETEL continues to concentrate at $K = 1$ in the $n \approx 10^4$ regime, where Li et al. (2026) document the coarsened posterior's heterogeneous-misspecification overfitting most clearly.

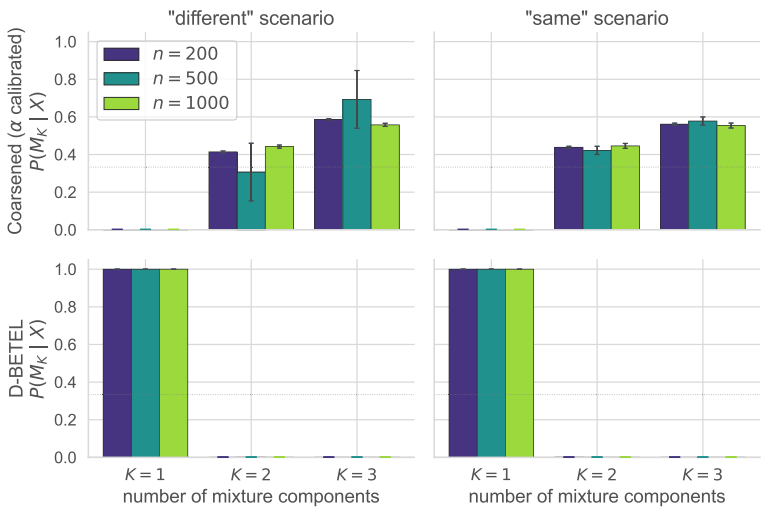

Figure 3: Posterior model probabilities $P(M_K \mid X)$ on the `same` (left column) and `different` (right column) regimes of the skew-mixture stress test of Li et al. (2026), for the calibrated coarsened posterior (top row) and D-BETEL (bottom row). Probabilities are computed under a uniform prior over $K \in \{1, 2, 3\}$. Bars are mean across three replicates with ± 1 standard-error bars; bar colors encode sample size n. The dotted line marks the uniform value $1/3$. Tick marks at 0 indicate K values whose posterior mass is below 0.01. D-BETEL $\log Z$ is computed via the Laplace approximation at the chain's MAP across all K with a $\log K!$ correction (§2). The coarsened posterior spreads mass between $K = 2$ and $K = 3$, shifting toward $K = 3$ as n grows in the `different` regime, then back partway at $n = 1000$. D-BETEL collapses onto $K = 1$ at every n in both regimes.

4 A Real Data Application

The real-data benchmark used by Miller and Dunson (2019) (and Li et al. (2026)) is the FlowCAP-I GvHD flow-cytometry dataset of Brinkman et al. (2007): 12 four-dimensional samples of $n \approx 13{,}000$–$33{,}000$ mononuclear cells, each manually partitioned into 3–5 clinically meaningful clusters. The model misspecification is quite severe for these datasets: cells within a cluster form non-Gaussian shapes, the manual boundaries are non-elliptical, and there are many outliers. Following Miller and Dunson (2019), calibration is performed on datasets 1–6 and evaluation on 7–12 using the F-measure between inferred hard cluster assignments and manual clustering. I fit D-BETEL at $K \in \{3, 4, 5\}$ (spanning the manual cluster counts) on each dataset subsampled to 3000 cells for computational tractability, with ϵ selected by ELPD-LOO on the same fixed grid as in Section 3.

The baseline numbers in Table 1 reproduce the qualitative picture from Miller and Dunson (2019, Fig. 4): the standard posterior overfits (mean test F-measure 0.56, 13–14 active components on average), and the calibrated coarsened posterior recovers most

Dataset	Manual K	Standard	Coarsened (shared)	Coarsened (oracle)	D-BETEL (ELPD-LOO)	D-BETEL (oracle K)
7	5	0.52	0.90	0.94	0.48	0.49
8	4	0.54	0.93	0.93	0.40	0.66
9	5	0.68	0.95	0.96	0.51	0.57
10	3	0.57	1.00	1.00	0.51	0.59
11	3	0.54	0.68	1.00	0.68	0.82
12	3	0.53	0.75	1.00	0.85	0.85
Mean		**0.56**	**0.87**	**0.97**	**0.57**	**0.66**

Table 1: Test-set F-measure on FlowCAP-I GvHD datasets 7–12. Baseline columns (Standard, Coarsened) use $K = 20$ with the Miller and Dunson (2019) latent-assignment Gibbs sampler; α for "shared" is selected by maximum mean training F-measure on $\{20, 50, \ldots, 600\}$ over datasets 1–6, and "oracle" uses the per-dataset best α on the same grid. D-BETEL is fit at $K \in \{3, 4, 5\}$ on each dataset subsampled to 3000 cells with ϵ selected by ELPD-LOO; "ELPD-LOO" picks K per dataset by ELPD-LOO and "oracle K" picks the per-dataset best K. "Manual K" is the number of non-zero clusters in the manual clustering.

of the manual labeled structure under either a shared α (0.87) or per-dataset oracle α (0.97).[1] D-BETEL reaches mean test F-measure 0.57 at the ELPD-LOO-selected K (always $K = 5$) and 0.66 under oracle $K \in \{3, 4, 5\}$ — substantially below the calibrated coarsened posterior.

The relationship between this benchmark and the synthetic experiments is more complex, and illustrates the importance of the model selection approach. The oracle K (the value that maximizes test F-measure) picks $K = 3$ on 4/6 datasets, including ds07 and ds09 where the manual partition has $K = 5$; this is consistent with the underfitting seen in the synthetic experiments. The ELPD-LOO, by contrast, picks $K = 5$ on every dataset (orange stars in Figure 4), which is consistent with optimizing for predictive accuracy and hence overfitting. However, it is possible that, as in the skew mixture example, the predictive measure will further favor small K. In addition to further experimentation, these results underline the need for investigations into the statistical properties of the robust estimators, with the ultimate goal of providing principled guidance.

5 Conclusion

Taken together, the experiments above suggest that D-BETEL is robust against overfitting, even under heterogeneous misspecification, but users must carefully consider the choice of ϵ and model-selection rule to avoid underfitting. This is an example of the classical robustness versus sensitivity trade-off. A related case is that non-parametric

[1] Miller and Dunson (2019) report $\alpha = 200$; my training selector picks $\alpha = 250$. The two are close on training (0.94 vs 0.92 mean F-measure) but $\alpha = 200$ generalizes slightly better on test (0.91 vs 0.87), so my selector mildly overfits training (Figure 5).

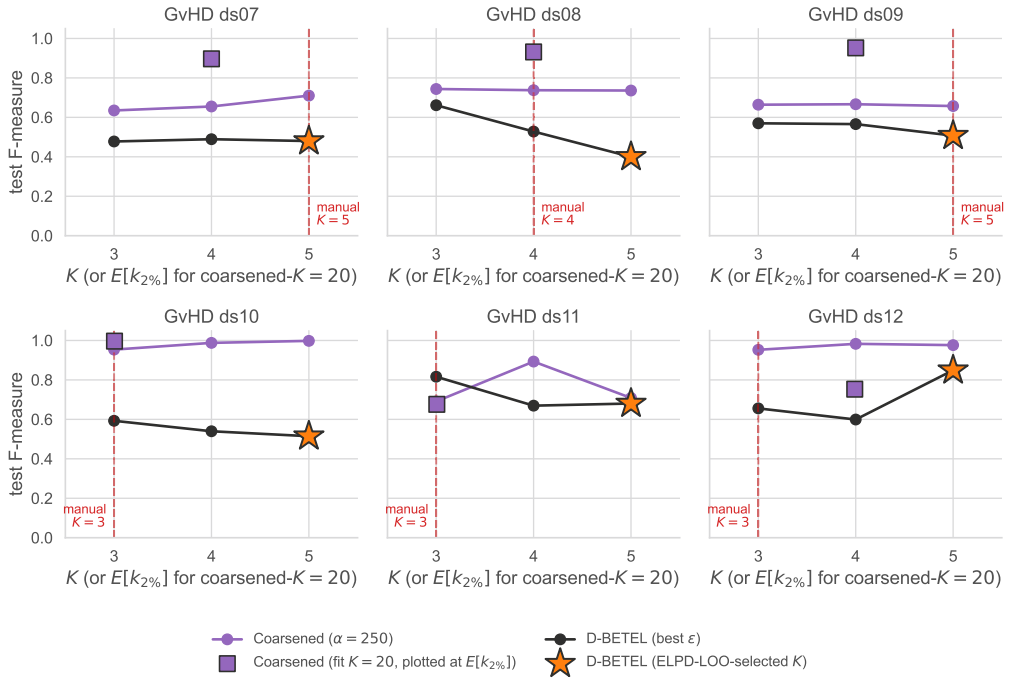

Figure 4: Per-dataset test F-measure on GvHD datasets 7–12 for D-BETEL (black) and the calibrated coarsened posterior ($\alpha = 250$, purple). The coarsened posterior is fit at $K \in \{3, 4, 5\}$ (purple line) and at $K = 20$ (purple square, plotted at the posterior-mean effective component count $E[k_{2\%}]$ using the sparsity-inducing Dirichlet prior). D-BETEL is fit at $K \in \{3, 4, 5\}$; the orange star marks the K selected by ELPD-LOO on the D-BETEL chains. Red dashed line marks the manual cluster count.

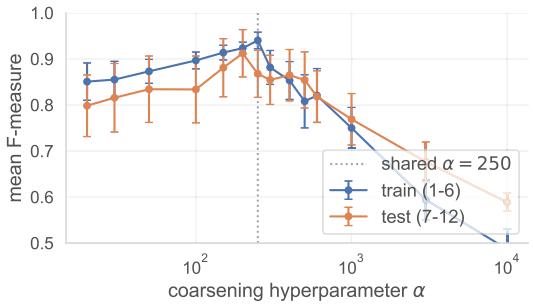

Figure 5: Calibration curve for α on the FlowCAP-I GvHD benchmark of Section 4: mean training and test F-measure as α varies over the M-D fine grid. The selected shared α is the dotted gray vertical line.

component selection methods (for example, the elbow rule (Thorndike, 1953), silhouette (Rousseeuw, 1987), gap statistic (Tibshirani et al., 2001), and parallel analysis (Dobriban, 2020; Horn, 1965)) tend to under-select model complexity under misspecification, while likelihood-based information criteria (Akaike, 1974; Schwarz, 1978; Spiegelhalter et al., 2002)) tend to over-select. How D-BETEL balances these trade-offs depends the selection rule for ϵ and the construction of the underlying metric. The proposed augmented-restricted Wasserstein distance opens up a rich design space of transport-based robustness metrics, and tailoring them to the forms of misspecification that arise in practice is an exciting research direction.

Another important direction is scaling D-BETEL up to the regimes where robustness tends to matter most, typically with parameter dimensions in the thousands and observations in the tens-to-hundreds of thousands. The structure of the formulation already points to two complementary approaches. First, an entropic-OT inner solver via Sinkhorn iterations (Cuturi, 2013) is a natural way to accelerate the LP solving step. Second, replacing the expensive inner evaluation with a likelihood (or log-posterior) surrogate has worked well for other expensive-likelihood problems (Conrad et al., 2016; Järvenpää et al., 2021; Kandasamy et al., 2017; Kennedy and O'Hagan, 2001; Wang and Li, 2018; Wilkinson, 2014). I look forward to seeing the continued work in this are by the authors and others!

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716–723. doi:10.1109/tac.1974.1100705. 992

Brinkman, R. R., Gasparetto, M., Lee, S.-J. J., Ribickas, A. J., Perkins, J., Janssen, W., et al. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation* 13, 691–700. doi:10.1016/j.bbmt.2007.02.002. 989

Chakraborty, A., Bhattacharya, A., and Pati, D. (2025). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis*. Advance publication. 985, 986, 987

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* 96, 270–281. doi:10.1198/016214501750332848. 986

Conrad, P. R., Marzouk, Y. M., Pillai, N. S., and Smith, A. (2016). Accelerating asymptotically exact MCMC for computationally intensive models via local approximations. *Journal of the American Statistical Association* 111, 1591–1607. doi:10.1080/01621459.2015.1096787. 992

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, Vol. 26, 2292–2300. 992

Dewaskar, M., Tosh, C., Knoblauch, J., and Dunson, D. B. (2025). Robustifying likelihoods by optimistically re-weighting data. *Journal of the American Statistical Association* 120, 1787–1798. doi:10.1080/01621459.2025.2468012. 985

Dobriban, E. (2020). Permutation methods for factor analysis and PCA. *The Annals of Statistics* 48. doi:10.1214/19-AOS1907. 992

Friel, N. and Wyse, J. (2012). Estimating the evidence – a review. *Statistica Neerlandica* 66, 288–308. doi:10.1111/j.1467-9574.2011.00515.x. 986

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi:10.1007/BF02289447. 992

Järvenpää, M., Gutmann, M. U., Vehtari, A., and Marttinen, P. (2021). Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. *Bayesian Analysis* 16, 147–178. doi:10.1214/20-BA1200. 992

Kandasamy, K., Schneider, J., and Póczos, B. (2017). Query efficient posterior estimation in scientific experiments via Bayesian active learning. *Artificial Intelligence* 243, 45–56. doi:10.1016/j.artint.2016.11.002. 992

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 63, 425–464 doi:10.1111/1467-9868.00294. 992

Li, J., Nguyen, N., Lai, M., Paschalidis, I. C., and Huggins, J. H. (2026). Robust model selection for discovery of latent mechanistic processes. arXiv:2602.22062. doi:10.48550/arXiv.2602.22062. 985, 986, 987, 988, 989

Llorente, F., Martino, L., Delgado, D., and López-Santiago, J. (2023). Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review* 65, 3–58. doi:10.1137/20M1310849. 986

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association* 114, 1113–1125. doi:10.1080/01621459.2018.1469995. 985, 986, 987, 989, 990

Moran, G. and Aragam, B. (2025). Towards interpretable deep generative models via causal representation learning. *Journal of the American Statistical Association* 121, 259–275. doi:10.1080/01621459.2026.2620154. 985

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65. doi:10.1016/0377-0427(87)90125-7. 992

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464. doi:10.1214/aos/1176344136. 992

Shmueli, G. (2010). To explain or to predict? *Statistical Science* 25, 289–310. doi:10.1214/10-STS330. 985

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Methodological* 64, 583–639. doi:10.1111/1467-9868.00353. 992

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika* 18, 267–276. doi:10.1007/bf02289061. 992

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B, Methodological* 63, 411–423. doi:10.1111/1467-9868.00293. 992

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27, 1413–1432. doi:10.1007/s11222-016-9696-4. 988

Wang, H. and Li, J. (2018). Adaptive Gaussian process approximation for Bayesian inference with expensive likelihood functions. *Neural Computation* 30, 3072–3094. doi:10.1162/neco_a_01127. 992

Wilkinson, R. D. (2014). Accelerating ABC methods using Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, Vol. 33, 1015–1023. PMLR. 992

Invited Discussion

Marco Avella Medina*, Juraj Marusic†, and Elvezio Ronchetti‡

1 Introduction

The authors are to be congratulated for their paper, which lies in the broad area of robust Bayesian statistics. We like the idea of defining an empirical likelihood (EL) optimization problem by constraining the discrete probability distribution of the weights to lie in a neighborhood of the postulated model F_θ. This constraint replaces the constraint defined by moment conditions, which is typically used in this setup. While in some fields (e.g. economics) models are often specified only through moment conditions, setting a constraint defined through a neighborhood of the assumed model is a very natural idea, which has been exploited in classical robust statistics. We will discuss some connections to ideas from the frequentist robust statistics literature and point out some challenges that might lead to interesting future research directions. Before delving into the robust statistics arguments we would like to highlight some additional benefits of the Bayesian ETEL (BETEL) construction.

2 Connections to saddlepoint approximations

We want to point to a connection that might shed some light on the choice of the weights (ETEL versus EL) and justify our preference for the authors' choice of exponential weights. In ETEL and in this paper the weights are obtained minimizing (under constraints) the (backward) Kullback-Leibler divergence between $\hat{F}_0 = (w_1, \ldots, w_n)$ and the empirical distribution $\hat{F} = (1/n, \ldots, 1/n)$, i.e.

$$d_{KL}(\hat{F}_0, \hat{F}) = \sum_{i=1}^{n} w_i \log[w_i/(1/n)].$$

This leads to the exponential weights. Notice that in EL the weights are obtained by minimizing (under constraint) the (forward) Kullback-Leibler divergence between $\hat{F}_0 = (w_1, \ldots, w_n)$ and the empirical distribution $\hat{F} = (1/n, \ldots, 1/n)$, i.e.

$$d_{KL}(\hat{F}, \hat{F}_0) = \sum_{i=1}^{n} \frac{1}{n} \log[(1/n)/w_i].$$

If the objective function (the Kullback-Leibler divergence) is used to define a test statistic, both lead to test statistics which are asymptotically χ^2-distributed under the null

*Department of Statistics, Columbia University, New York, NY 10027, USA, marco.avella@columbia.edu

†Department of Statistics, Columbia University, New York, NY 10027, USA, juraj.marusic@columbia.edu

‡Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Switzerland, Elvezio.Ronchetti@unige.ch

hypothesis. However, the error of the asymptotic approximation in the latter is absolute as in the classical parametric counterpart, the log-likelihood ratio test, whereas the error in the former is relative, providing an important improvement in the tails for the computation of critical values. This is due to the fact that the test statistic obtained minimizing the backward Kullback-Leibler divergence is the nonparametric version of the so-called saddlepoint test (see Robinson et al. (2003)), which inherits this property from (empirical) saddlepoint approximations. More details are provided in Monti and Ronchetti (1993); Ronchetti and Welsh (1994); Ronchetti (2020). In light of these connections, the BETEL construction seems to also be the more satisfactory choice from a frequentist standpoint.

3 Choice of the distance

The crucial point for robustness is the choice of the distance defining the neighborhood of the model. In classical robust statistics the Prokhorov metric has been often advocated (see Hampel et al. (1986, p. 96)), because it captures the effects of small changes on many observations (rounding effects) and large changes of a few observations (outliers) and it formalizes the fact that the model is only an idealized approximation of the underlying chance mechanism. However, from a technical point of view this distance is not easy to work with, compared to e.g. Huber's gross-error model. In the setup considered by the authors the situation is complicated by the fact that a distance between a discrete and a continuous distribution is needed. We understand that the Wasserstein distance could be a possible feasible choice, but we are not convinced that it captures the deviations observed on real data containing outliers. We note that there have been some proposals to make the Wasserstein distance more robust to outliers by modifying the optimal transport problem used to define it (Mukherjee et al., 2021; Nietert et al., 2026; La Vecchia and Liu, 2025). It seems to us that the robust entropic Wasserstein distance approach of La Vecchia and Liu (2025) has the potential to further robustify the distributionally contrainted BETEL (D-BETEL) methodology while keeping its computational feasibility.

4 Robustness analysis

It would be useful to perform a standard robustness analysis on the functional defined by (4.2) and (4.3) in Chakraborty et al. (2026). For instance, what are its influence function and its breakdown point? If the corresponding natural estimator $\hat{\theta}^{\dagger}$ of the functional $\theta^{\dagger}$ is asymptotically normal, its influence function could, in principle, be read off from the linear term leading to the asymptotic normality.

In principle one could for example try to adapt the robustness analysis discussed in Marusic et al. (2025) in the context of generalized posteriors. However, one is immediately faced with a series of obstacles that do not seem to be easy to address at this stage. Indeed, their characterizations of the posterior influence function and posterior breakdown point for generalized posteriors critically leverage the fact that for the class of M-posteriors discussed in that work the data enter the posterior distribution through

a smooth function of a linear functional of the empirical distribution function F_n. It is not clear to us how to extend this type of analysis to the D-BETEL posterior defined by (2.4)–(2.5) in Chakraborty et al. (2026). We note that even in the perhaps simpler frequentist world, there are not many formal robustness assessments of EL and ETEL estimators. One could expect such parameter estimators to be robust in the sense of the influence function if their corresponding moment equations are bounded since asymptotic linear representations are linear in these equations (Newey and Smith, 2004). To the best of our knowledge, the breakdown properties of EL and ETEL estimators have not been formally established in the literature.

5 Connections to Bayesian data-reweighting

A complementary line of work on robust Bayesian inference under model contamination is the Bayesian data-reweighting framework of Wang et al. (2017). Their construction augments the likelihood by raising each observation's contribution to a latent power, $p(y, \beta, w) \propto \pi(\beta)\, \pi_w(w) \prod_{i=1}^{n} \ell(y_i \mid \beta)^{w_i}$, with a Beta or Gamma prior π_w softly favouring $w_i \approx 1$; inference proceeds by sampling (β, w) jointly, and contaminated observations are identified by the contraction of $\mathbb{E}[w_i \mid y]$ toward zero. Both methods reweight observations to mitigate model mismatch, but the contrasts are sharp. In Wang et al. (2017) the weights are latent random variables regularized toward one by an explicit prior, whereas in D-BETEL they are deterministic functions of θ obtained from an entropy-maximization problem on the simplex with a Wasserstein constraint between F_θ and the weighted empirical distribution. The two views of an outlier differ accordingly: Wang et al. (2017) flag observation i pointwise, because $\ell(y_i \mid \beta)$ is small for plausible β, whereas the D-BETEL construction flags it distributionally, downweighting i not because it fits poorly in isolation but because doing so is the cheapest way (in entropy) to pull the empirical distribution into an ε-ball around F_θ. Marusic et al. (2025) showed that the Bayesian data reweighting approach can be interpreted from a generalized posterior (M-posterior) perspective which shows that the resulting posterior distribution of θ given the observed data is robust in the sense of the influence function and breakdown point.

6 Numerical experiments

6.1 Non-convexity of the D-BETEL loss landscape

The log empirical likelihood $L_{\mathrm{DCM}}(\theta)$ defined in equation (2.4) of Chakraborty et al. (2026) is, in general, neither concave nor continuous in θ. To illustrate this point we consider a simple one dimensional normal model where $\theta = (\mu, \sigma^2)$. Figure 1 illustrates this with a single contaminating observation: the feasibility set of the inner convex program partitions the parameter space into disjoint connected components, with a primary "clean-cluster" basin around the bulk of the data and a secondary "outlier" basin near the contaminating point. Separating the two is a band of θ-values for which no reweighting of the empirical measure can bring it within Wasserstein distance ε

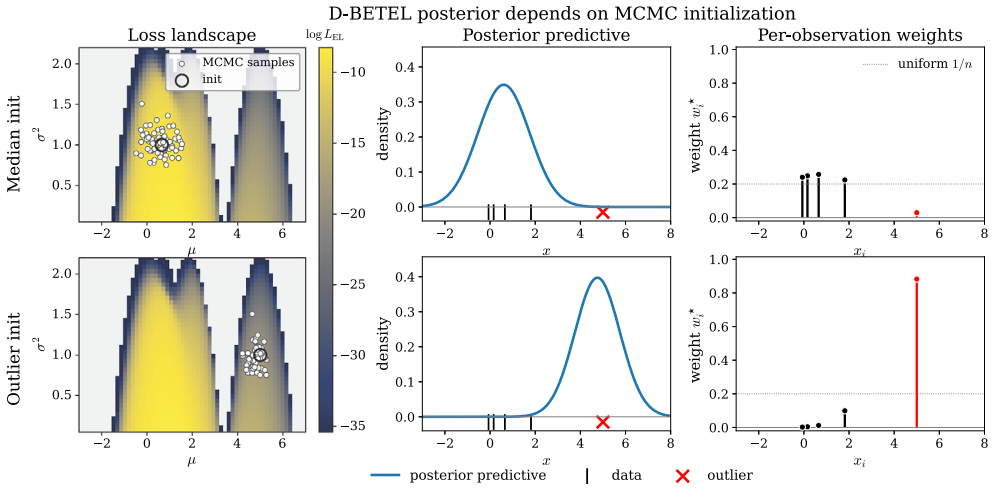

Figure 1: D-BETEL on a corrupted one-dimensional dataset ($n = 5$, $\varepsilon = 2.25$): the MCMC chain is structurally trapped in whichever feasible basin contains its initialization. Top row: initializing at the median places the chain in the clean-cluster basin and downweights the outlier ($w^\star_{\text{outlier}} \approx 0$). Bottom row: initializing near the outlier places it in a second basin that instead downweights the clean observations. Columns, from left to right: the log empirical likelihood $\log L_{\mathrm{DCM}}(\mu, \sigma^2)$ with MCMC samples (white) and initialization (black ring); the posterior predictive density overlaid on the observed data points; and the per-observation weights $w_i^\star$.

of $\mathcal{N}(\mu, \sigma^2)$. On this band the inner program is infeasible, $L_{\mathrm{DCM}}(\theta) = -\infty$, and the quasi-posterior assigns zero mass.

Implications for MCMC initialization A Metropolis–Hastings chain with local proposals cannot cross the infeasibility band: any proposal landing in it evaluates to $-\infty$ and is rejected. The chain is therefore trapped in whichever basin contains its starting point. As Figure 1 shows, initializing at $(\mathrm{med}(X_i), 1)$ yields a posterior concentrated on the clean cluster, whereas initializing $(5, 1)$ yields one that downweights the clean observations instead; both are valid samples from different local modes of the posterior, differing only in which observations they treat as outliers. Initializing at the MLE—itself sensitive to contamination—is particularly hazardous, as the EL constraint may be infeasible from the outset. We therefore recommend initializing the MCMC chain at an outlier robust estimator.

Connection to redescending M-estimators The nonconvex landscape described above seems to mirror the well-known multi-mode behaviour of redescending M-estimators in classical robust statistics, such as Tukey's biweight loss, whose objective is non-convex and admits multiple local minima. The MLE of the Cauchy location model is a special case of a redescending M-estimator for which the number of non-global minima can

be described very precisely (Reeds, 1985). In a Bayesian context the recent analysis of Agrawal et al. (2026) extends some of these results by showing that posteriors in heavy-tailed Bayesian models develop a local mode (a *micromode*) around each sufficiently isolated observation, with domains of attraction that grow polynomially in n. It would be interesting to study whether the nonconvexity of D-BETEL can be related to these recent developments.

6.2 Poisson regression

We revisit the Poisson regression example considered by the authors in Section 6.2 to compare D-BETEL with a well-understood robust generalized Bayesian approach. We show that in this case a simple alternative can be obtained combining the robust quasilikelihood loss of Cantoni and Ronchetti (2001) with the M-posterior construction. One can expect good robustness properties of this approach given the results of Marusic et al. (2025). More precisely, we consider a generalized Bayesian posterior (M-posterior) of the form

$$\pi(\beta \mid y) \;\propto\; \pi(\beta)\, \exp\Big(-\omega \sum_{i=1}^{n} \ell_i(\beta) \Big), \tag{6.1}$$

where, in place of the Poisson log-likelihood, we use the robust loss of Cantoni and Ronchetti (2001): compute each observation's Pearson residual $r_i(\beta) = (y_i - \mu_i(\beta))/\sqrt{\mu_i(\beta)}$ with $\mu_i(\beta) = \exp(x_i^\top \beta)$ and set the loss $\ell_i(\beta)$ to be such that

$$\nabla \ell_i(\beta) = \psi_c(r_i(\beta)) \sqrt{\mu_i(\beta)}\, x_i - a(\beta),$$

where $a(\beta)$ is a Fisher consistency constant; cf. equation (5) in Cantoni and Ronchetti (2001). We consider the Huber loss for ρ_c, which caps the influence of any single observation at $\pm c$.

		Standard posterior			Huber M-posterior		
p	θ	$\lVert\theta - \hat\theta\rVert_1$	HPD	(cov)	$\lVert\theta - \hat\theta\rVert_1$	HPD	(cov)
0.10	β_0	0.38	0.12	(0.00)	0.00	0.20	(0.90)
	β_1	0.05	0.02	(0.00)	0.00	0.03	(0.98)
0.12	β_0	0.39	0.11	(0.00)	0.01	0.22	(0.88)
	β_1	0.06	0.02	(0.00)	0.00	0.04	(0.94)
0.15	β_0	0.45	0.11	(0.00)	0.01	0.26	(0.86)
	β_1	0.06	0.02	(0.02)	0.00	0.04	(0.92)

Table 1: Huber M-posterior with Cantoni–Ronchetti correction on the D-BETEL Poisson example. HPD widths scaled by $\sqrt{n}$.

Table 1 reports the same metrics as in the main manuscript for the Huber M-posterior on the corrupted dataset. At the textbook tuning constant $c = 1.345$, the posterior mean is accurate across all contamination levels, but the credible intervals are not well calibrated in the frequentist sense. The variance-matching adjustment of Holmes and Walker (2017) widens the intervals and partially restores coverage.

References

Agrawal, S., Grazzi, S., and Roberts, G. O. (2026). On micromodes in Bayesian posterior distributions and their implications for MCMC. arXiv:2602.06931. 998

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association* 96, 1022–1030. doi:10.1198/016214501753209004. 998

Chakraborty, A., Bhattacharya, A., and Pati, D. (2026). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis* 21. doi:10.1214/25-BA153. 995, 996

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley. MR0829458. 995

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* 104, 497–503. doi:10.1093/biomet/asx010. 998

La Vecchia, D. and Liu, H. (2025). E-ROBOT: A dimension-free method for robust statistics and machine learning via Schrödinger bridge. arXiv:2509.11532. 995

Marusic, J., Avella Medina, M., and Rush, C. (2025). A theoretical framework for M-posteriors: Frequentist guarantees and robustness properties. arXiv:2510.01358. 995, 996, 998

Monti, A. C. and Ronchetti, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators. *Biometrika* 80, 329–338. doi:10.1093/biomet/80.2.329. 995

Mukherjee, D., Guha, A., Solomon, J. M., Sun, Y., and Yurochkin, M. (2021). Outlier-robust optimal transport. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 7850–7860. 995

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72, 219–255. doi:10.1111/j.1468-0262.2004.00482.x. 996

Nietert, S., Cummings, R., and Goldfeld, Z. (2026). Robust estimation under the Wasserstein distance. *Journal of Machine Learning Research*. To appear. 995

Reeds, J. A. (1985). Asymptotic number of roots of Cauchy location likelihood equations. *The Annals of Statistics* 13, 775–784. doi:10.1214/aos/1176349554. 998

Robinson, J., Ronchetti, E., and Young, G. A. (2003). Saddlepoint approximations and tests based on multivariate M-estimators. *The Annals of Statistics* 31, 1154–1169. doi:10.1214/aos/1059655909. 995

Ronchetti, E. (2020). Accurate and robust inference. *Econometrics and Statistics* 14, 74–88. doi:10.1016/j.ecosta.2019.12.003. 995

Ronchetti, E. and Welsh, A. H. (1994). Empirical saddlepoint approximations for multivariate M-estimators. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 56, 313–326. doi:10.1111/j.2517-6161.1994.tb01980.x. 995

Wang, Y., Kucukelbir, A., and Blei, D. M. (2017). Robust probabilistic modeling with Bayesian data reweighting. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, 3646–3655. 996

Contributed Discussion

Khai Nguyen*

We congratulate Chakraborty et al. (2025) (hereafter AAD) on the impressive idea of using an exponentially tilted empirical likelihood to define a robust extension of a parametric family via a distributional constraint with a Wasserstein variant, which is both elegant and technically compelling. Given parameters of interest θ and a weighted empirical distribution of the observed data $\nu_{w,x} = \sum_{i=1}^n w_i \delta_{x_i}$, AAD propose the following likelihood function:

$$\mathrm{L}_{\mathrm{DCM}}(\theta) := \left\{ \prod_{i=1}^n w_i : \arg\max_{w} \prod_{i=1}^n w_i^{-w_i}, w_i > 0, \sum_{i=1}^n w_i = 1, \mathcal{D}(F_\theta, \nu_{w,x}) \leq \epsilon \right\}, \quad (1)$$

where $F_\theta : \theta \in \Theta \subseteq \mathbb{R}^d$ with $d \geq 1$ is a parametric family of distributions, and $\epsilon > 0$ is a concentration parameter which controls fidelity to the centering model. Completing the model with a prior $\pi(\theta)$, the corresponding posterior distribution is: $\pi_{\mathrm{DCM}}(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta)\mathrm{L}_{\mathrm{DCM}}(\theta)$. By evaluating $\mathrm{L}_{\mathrm{DCM}}(\theta)$ using a Lagrange multiplier, AAD can proceed with posterior inference as usual. For the choice of the family of F_θ, AAD suggests employing the Elliptical Mixture Model (EMM). AAD then propose a new augmented and restricted Wasserstein metric between EMMs. Given p_0 and p_1 are two EMMs with $p_j = \sum_{k=1}^{K_j} s_{jk} ED_h(m_{jk}, \Sigma_{jk})$, the key idea is to augment p_j to $p_j^\star = p_j \otimes \tilde{p}_j$ (with $\tilde{p}_j = p_{j0} \otimes \ldots \otimes p_{jd}$ being independent products of marginals of p_j) and to restrict the coupling to be the product of an entropic regularized EMM (in 2d dimensions) and independent couplings of marginals of p_0 and p_j. The final form of the distance (Theorem 1 in Chakraborty et al. (2025)) can be written as

$$\mathcal{D}(p_0, p_1) = MW_2^2(p_0, p_1; \alpha) + \sum_{k=1}^d \int_0^1 (F_{0k}^{-1}(t) - F_{1k}^{-1}(t))^2 \mathrm{d}t, \quad (2)$$

where $MW_2^2(p_0, p_1; \alpha)$ is the entropic regularized mixture Wasserstein (Delon and Desolneux, 2020; Cuturi, 2013), and $F_{0k}^{-1}(t)$ and $F_{1k}^{-1}(t)$ are the quantile functions of the k-th marginals of p_0 and p_1, respectively.

Starting from (2), we would like to propose some extensions of the distance that are computationally efficient and interpretable. First, we recognize that the second term in (2) as a degenerate version of sliced Wasserstein (SW) distance (Rabin et al., 2011; Nguyen, 2025) between p_0 and p_1 (scaled by d). In particular, SW distance between p_0 and p_1 can be written as:

$$SW_2^2(p_0, p_1) = \mathbb{E}_{v \sim \sigma(\mathbb{S}^{d-1})} \left[W_2^2(P_v \sharp p_0, P_v \sharp p_1) \right], \quad (3)$$

where $P_v(x) = v^\top x$, $P_v \sharp p_j$ is the push-forward of p_j through P_v, $W_2^2(P_v \sharp p_0, P_v \sharp p_1) = \int_0^1 (F_{v^\top X_0}^{-1}(t) - F_{v^\top X_1}^{-1}(t))^2 \mathrm{d}t$, $\sigma(\mathbb{S}^{d-1})$ is a distribution on the unit hypersphere of dimension $d-1$, and $F_{v^\top X_0}^{-1}(t)$ and $F_{v^\top X_1}^{-1}(t)$ are the quantile functions of $v^\top X_0$ and $v^\top X_1$,

*Department of Statistics and Data Sciences, University of Texas at Austin, khainb@utexas.edu

with $X_0 \sim p_0$ and $X_1 \sim p_1$, respectively. In (2), σ is chosen to be the empirical distribution on basis vectors $v_1, \ldots, v_d$ with $v_k = (0, \ldots, 0, \underset{k\text{-th}}{1}, 0, \ldots, 0)$ for $k = 1, \ldots, d$ rather than the uniform distribution on $\mathbb{S}^{d-1}$ (Rabin et al., 2011) and other choices (Nguyen et al., 2021; Nguyen and Ho, 2023). This is implicit in the construction of the augmentation $\tilde{p}_j = p_{j0} \otimes \ldots \otimes p_{jd}$. The conventional SW can be obtained if we replace the augmentation by $\tilde{p}_j = \mathcal{R}_\sigma[p_j]$, where $\mathcal{R}_\sigma$ is the Weighted Radon Transform (Nguyen, 2025; Natterer, 2001) of p_j. In particular, $\mathcal{R}_\sigma$ maps a distribution on $\mathbb{R}^d$ to a distribution on $\mathbb{R} \times \mathbb{S}^{d-1}$. For convenience, we give the definition of the transform of the density function implied by $\mathcal{R}_\sigma$ as follows: $\mathcal{R}_\sigma[f_p] = \int_{\mathbb{R}^d} f_p(x) f_\sigma(\theta) \delta(t - v^\top x) \mathrm{d}x$, where f_p and f_σ are density functions of p and σ with respect to Lebesgue measure, and $\mathcal{R}_\sigma[f_p]$ is the density of $\mathcal{R}_\sigma[p_j]$. This choice allows dependency between augmented marginals. Moreover, with the entropic regularization, $\mathcal{D}$ in (2) is not a fully valid metric due to violation of the identity of indiscernibles property, i.e., $\mathcal{D}(p_0, p_1) = 0 \nRightarrow p_0 = p_1$. By converting the second term to SW distances, we can reclaim the identity property while the computational complexity remains.

Secondly, we want to discuss a potential replacement of the first term in (2). With entropic regularization, the time complexity of $MW_2^2(p_0, p_1; \alpha)$ is $\mathcal{O}(K_1 K_2)$ (Altschuler et al., 2017), which is still expensive with large numbers of components K_1 and K_2. As introduced in Nguyen and Mueller (2026), there are options that are based on comparing mixing measures of EMMs. Let $\bar{p}_j = \sum_{k=1}^{K_j} s_{jk} \delta_{(m_{jk}, \Sigma_{jk})}$ be the mixing measure of the EMM p_j. If the EMM family is identifiable, i.e., $\bar{p}_0 = \bar{p}_1 \iff p_0 = p_1$ (e.g., Gaussian, Student-t, and other EMMs depending on h), we can define a distance between $\bar{p}_0$ and $\bar{p}_1$ as a proxy for a distance between p_0 and p_1. For example, we can use $SW_2^2(V\sharp\bar{p}_0, V\sharp\bar{p}_1)$ with V being a transformation to turn the covariance matrices into vectors (Nguyen and Mueller, 2026). This allows reducing the time complexity to $\mathcal{O}[(K_1 + K_2)\log(K_1 + K_2)]$. Considering that the mixing measures are measures on the product of manifolds, i.e., Euclidean space and the manifold of symmetric (semi-)positive definite matrices, a better option is Mixture Sliced Wasserstein (Nguyen and Mueller, 2026):

$$\text{Mix-SW}_2^2(p_0, p_1) = \mathbb{E}_{(w,v,A)\sim\sigma(\mathbb{S}\times\mathbb{S}^{d-1}\times\mathcal{S}_d(\mathbb{R}))}\left[W_2^2(P_{w,v,A}\sharp\bar{p}_0, P_{w,v,A}\sharp\bar{p}_1)\right], \tag{4}$$

where $\mathcal{S}_d(\mathbb{R})$ is the set of symmetric matrices of $d \times d$ dimensions, σ is a chosen distribution e.g., uniform, and $P_{w,v,A}(\mu, \Sigma) = w_1\langle\mu, v\rangle + w_2 \,\text{Trace}(A \log \Sigma)$ is the generalized geodesic projection. Mix-SW is a valid metric with the time complexity is still $\mathcal{O}[(K_1 + K_2)\log(K_1 + K_2)]$, however, the distance is more geometrically meaningful. We can further exploit a property of EMMs. In particular, we know that a linear projection of an EMM is also an EMM. Let $X \sim ED_h(m, \Sigma)$, which admits the characteristic function $t \to \exp(it^\top m)h(t^\top \Sigma t)$, we have $v^\top X \sim ED_h(v^\top m, v^\top \Sigma v)$ with the characteristic function $t \to \exp(itv^\top m)h(t^2 v^\top \Sigma v)$. Therefore, we have $P_v\sharp p_j = \sum_{k=1}^{K_j} s_{jk} ED_h(v^\top m_{jk}, v^\top \Sigma_{jk} v)$ with $P_v(x) = v^\top x$, which admits the mixing measure $P'_v\sharp\bar{p}_j = \sum_{k=1}^{K_j} s_{jk}\delta_{(v^\top m_{jk}, v^\top \Sigma_{jk} v)}$ with $P'_v(m, \Sigma) = (v^\top m, v^\top \Sigma v)$. Combining with the generalized geodesic projection for one-dimensional mixing measures, we can obtain Sliced Mixture Wasserstein (Nguyen and Mueller, 2026):

$$\text{SMix-W}_2^2(p_0, p_1) = \mathbb{E}_{(w,v)\sim\sigma(\mathbb{S}\times\mathbb{S}^{d-1})}\left[W_2^2(P_{w,v}\sharp\bar{p}_0, P_{w,v}\sharp\bar{p}_1)\right], \tag{5}$$

where $P_{w,v}(m, \Sigma) = w_1 v^\top m + w_2 \log(\sqrt{v^\top \Sigma v})$ (square root for $v^\top \Sigma v$ is optional). Compared to Mix-SW, SMix-W reduces one projection parameter $A \in \mathcal{S}_d(\mathbb{R})$ while remaining the same time complexity of $\mathcal{O}[(K_1 + K_2)\log(K_1 + K_2)]$ and metricity.

References

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in Neural Information Processing Systems* 30. 1001

Chakraborty, A., Bhattacharya, A., and Pati, D. (2025). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis* 1, 1–33. doi:10.1214/25-BA1535. 1000

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems* 26. 1000

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences* 13, 936–970. doi:10.1137/19M1301047. 1000

Natterer, F. (2001). Inversion of the attenuated Radon transform. *Inverse Problems* 17, 113. doi:10.1088/0266-5611/17/1/309. 1001

Nguyen, K. (2025). An introduction to sliced optimal transport: Foundations, advances, extensions, and applications. *Foundations and Trends® in Computer Graphics and Vision* 17, 171–391. doi:10.1561/0600000119. 1000, 1001

Nguyen, K. and Ho, N. (2023). Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems* 36, 18046–18075. doi:10.52202/075280-0793. 1001

Nguyen, K., Ho, N., Pham, T., and Bui, H. (2021). Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=QYjO70ACDK 1001

Nguyen, K. and Mueller, P. (2026). Summarizing nonparametric Bayesian mixture posteriors–sliced optimal transport metrics for Gaussian mixtures. *Journal of Computational and Graphical Statistics*, 1–22 (just-accepted). doi:10.1080/10618600.2026.2634831. 1001

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446. Springer. 1000, 1001

Contributed Discussion

Marta Catalano* and Sirio Legramanti†

We congratulate the authors on this paper, in which they propose a distance-based Bayesian exponentially-tilted empirical likelihood (D-BETEL) approach. As a distance, they define a new variant of the Wasserstein metric, termed ANDREW, which is tailored to elliptical mixture models.

The proposed D-BETEL may be viewed as a likelihood-free Bayesian procedure, in the same broad spirit as Approximate Bayesian Computation (ABC; Marin et al., 2012), especially if we consider recent, discrepancy-based, ABC variants; see, e.g., Park et al. (2016), Jiang (2018), Bernton et al. (2019), Nguyen et al. (2020), Frazier (2020), Fujisawa et al. (2021), Legramanti et al. (2025). Notably, D-BETEL departs from traditional moment-based BETEL in a way that parallels how discrepancy-based ABC has moved away from summary statistics, often also related to moments. Both D-BETEL and ABC replace the likelihood with a term involving a discrepancy between an empirical distribution supported on the observed data and a parametric family F_θ. There are, however, important differences. In ABC, the discrepancy is typically computed between the empirical distributions (or some summary statistics) of the observed data and of synthetic data generated from F_θ. In contrast, in D-BETEL the discrepancy is evaluated between F_θ and a *weighted* discrete distribution supported on the observed data. More precisely, D-BETEL constructs a pseudo-likelihood by finding the maximum-entropy weighted empirical distribution supported on the observed sample that lies within an ε-neighborhood of F_θ under the chosen discrepancy. This entropy-based reweighting scheme provides a direct robustness mechanism, in which observations that are difficult to reconcile with F_θ are adaptively downweighted.

The connection between D-BETEL and ABC suggests two natural directions for further investigation. First, D-BETEL could borrow some ideas from the abovementioned recent advancements in ABC, which may be useful when simulation from F_θ is possible but the corresponding density, or the distance from F_θ to a discrete measure, is not available in a tractable form. In such settings, one could consider a simulation-based version of D-BETEL in which F_θ is replaced by a Monte Carlo empirical approximation. Naturally, this would require careful choice of the empirical-to-empirical discrepancy, since this discrepancy would be evaluated repeatedly inside the entropy-constrained optimization over weights. In this simulation-based D-BETEL variant, integral probability semimetrics (IPS), and in particular the maximum mean discrepancy (MMD; Muandet et al., 2017), appear to be natural candidates. Such discrepancies have proved effective within ABC (Legramanti et al., 2025), are directly computable between empirical measures, and may offer computational and robustness advantages in settings where a model-specific transport discrepancy such as ANDREW is not available. Moreover, MMD has been shown to yield estimators with robustness properties under outlier

*Department of AI, Data and Decision Sciences, Luiss University, Italy, mcatalano@luiss.it
†Department of Economics, University of Bergamo, Italy, sirio.legramanti@unibg.it

contamination (Alquier and Gerber, 2024). Combining such a robust discrepancy with the entropy-based reweighting mechanism of D-BETEL could provide an additional source of robustness, which would then arise both from the choice of discrepancy and the adaptive downweighting of outliers. It would be interesting to understand whether MMD-based or, more generally, IPS-based versions of D-BETEL can deliver posterior robustness properties comparable to those obtained with ANDREW.

Second, D-BETEL may provide useful ideas for ABC. In fact, computational scalability is a central challenge in ABC, which typically requires tens of thousands of discrepancy evaluations, and the ANDREW discrepancy is explicitly designed to balance multivariate geometry, tail sensitivity, and computational efficiency. Namely, in ANDREW, a tractable restricted transport component is enriched by a marginal quantile augmentation, which adds coordinate-wise tail information while preserving efficient computation. Since this added tail information is marginal, a natural extension would be to investigate whether joint tail dependence could also be incorporated, for example through copula- or projection-based terms. Moreover, it would be of interest to assess whether ANDREW (or related augmented and restricted transport discrepancies) can be successfully employed within ABC, and whether they yield posterior concentration properties similar to those proved by Bernton et al. (2019) and Legramanti et al. (2025) for the Wasserstein distance and generic integral probability semimetrics. This would clarify whether the computational advantages that make ANDREW effective within D-BETEL can also benefit ABC and, more broadly, likelihood-free inference.

References

Alquier, P. and Gerber, M. (2024). Universal robust regression via maximum mean discrepancy. *Biometrika* 111, 71–92. doi:10.1093/biomet/asad031. 1004

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, 235–269. doi:10.1111/rssb.12312. 1003, 1004

Frazier, D. T. (2020). Robust and efficient approximate Bayesian computation: A minimum distance approach. arXiv:2006.14126. MR4147553. 1003

Fujisawa, M., Teshima, T., Sato, I., and Sugiyama, M. (2021). γ-ABC: Outlier-robust approximate Bayesian computation based on a robust divergence estimator. In *International Conference on Artificial Intelligence and Statistics*, 1783–1791. PMLR. 1003

Jiang, B. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International Conference on Artificial Intelligence and Statistics*, 1711–1721. PMLR. doi:10.1201/b22400. 1003

Legramanti, S., Durante, D., and Alquier, P. (2025). Concentration of discrepancy-based approximate Bayesian computation via Rademacher complexity. *The Annals of Statistics* 53, 37–60. doi:10.1214/24-aos2453. 1003, 1004

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing* 22, 1167–1180. doi:10.1007/s11222-011-9288-2. 1003

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* 10, 1–141. doi:10.1561/2200000060. 1003

Nguyen, H. D., Arbel, J., Lü, H., and Forbes, F. (2020). Approximate Bayesian computation via the energy statistic. *IEEE Access* 8, 131683–131698. doi:10.1109/ACCESS.2020.3009878. 1003

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, 398–407. PMLR. 1003

Contributed Discussion

Francesco Gaffi* and Beatrice Franzolini†

We congratulate the authors for an original and stimulating contribution. The proposed D-BETEL framework offers an elegant and effective compromise between robustness and interpretability, where inference remains focused on a finite-dimensional parameter θ, indexing an interpretable parametric model F_θ, while still avoiding a full specification of the data-generating distribution. The procedure is particularly appealing in that inference results from a prior, specified only on the parameter of interest θ, and a modified likelihood contribution, obtained by finding the maximum-entropy reweighting of the empirical distribution of the observations that lies within a neighborhood of F_θ. Hence, values of θ are favored when this distance-based compatibility with F_θ can be achieved with little departure from the uniform empirical weights. In this sense, the modified likelihood rewards models F_θ that explain the bulk of the data while still allowing atypical observations to be downweighted. We view the nonparametric Bayesian interpretation in Section 5, which is shown to arise in an appropriate asymptotic regime, as a distinctive strength of this approach. Unlike many alternative pseudo-likelihood or generalized Bayes approaches, D-BETEL is shown to arise as a limiting marginal posterior under a suitable nonparametric hierarchical construction, thereby providing a genuine probabilistic motivation for its use.

A feature we find especially interesting is that the proposed approach allows the prior distribution to be defined directly on the finite-dimensional parameter of interest θ, while still encompassing data-generating processes more general than the parametric family F_θ. This aspect brings the proposal into close contact with recent developments in Bayesian nonparametrics, where prior information is incorporated by specifying the distribution of selected functionals of a random probability measure. In Gaffi et al. (2025), the emphasis is on prior elicitation within a given class: once this class has been chosen, the base measure is selected so that a linear functional, typically the mean, has a prescribed distribution, while preserving the structure that allows one to rely on standard inferential techniques for that class. In Kessler et al. (2015), a complementary strategy is adopted: a canonical nonparametric prior is modified so that the induced marginal distribution of a functional matches a desired one, resulting in a prior that generally lies outside the original class and requires adapted posterior sampling methods. In both cases, the underlying idea is that prior information is more naturally expressed in terms of functionals than in terms of the full infinite-dimensional object.

D-BETEL differs from these approaches in how this idea is implemented, but appears to be driven by a closely related principle. Rather than specifying or modifying a prior so as to match the distribution of a functional, it introduces a constraint through a discrepancy functional,

$$Q \mapsto D(F_\theta, Q),$$

*University of Bergamo, Italy, francesco.gaffi@unibg.it
†University of Milano-Bicocca, Italy, beatrice.franzolini@unimib.it

and selects admissible distributions accordingly. This functional is nonlinear and defined through an optimization over couplings, and therefore lies outside the class of functionals typically considered in the literature. The distinction is not only formal. Even in relatively simple settings, moving from linear to nonlinear functionals entails a substantial increase in complexity: for instance, quadratic functionals of random objects driven by completely random measures already require nontrivial tools based on Wiener–Itô decompositions and asymptotic analysis, as shown in Peccati and Prünster (2008), and explicit distributional control is generally difficult to obtain. In this sense, D-BETEL can be interpreted as extending the functional-based perspective toward functionals that capture global features of the distribution, such as overall shape and tail behavior.

At the same time, the analogy remains structurally strong. As in the approaches mentioned above, the infinite-dimensional problem is effectively controlled through a lower-dimensional quantity, and prior information is incorporated indirectly through the behavior of a functional. The nonparametric Bayesian representation provided in the paper further clarifies this point: the procedure can be embedded in a hierarchical model involving a random mixing distribution, although this equivalence holds in an asymptotic regime where nuisance parameters are marginalized out. In practice, the method avoids working with the infinite-dimensional object altogether, and the induced prior on it is only implicit, but the overall mechanism remains closely aligned with a functional specification viewpoint.

This perspective suggests a natural extension. The constraint $D(F_\theta, Q) \leq \varepsilon$ imposes a sharp boundary on the value of a nonlinear functional. One may instead consider specifying a prior distribution on $D(F_\theta, \tilde{P})$ for a random probability measure $\tilde{P}$, thereby replacing the hard constraint with a probabilistic one. Such a formulation would be closer in spirit to approaches based on fixing the distribution of a functional, while potentially avoiding boundary effects induced by the constraint. It might also provide a route to an exact nonparametric Bayesian representation, rather than one obtained through asymptotic equivalence, by explicitly incorporating prior beliefs on the extent of departure from the centering model.

From this point of view, the contribution of the paper is twofold. On the one hand, it provides a practically effective method for robust inference that bypasses the direct handling of infinite-dimensional parameters. On the other hand, it points toward a broader perspective in which nonlinear functionals, such as transport-based discrepancies, play a central role in linking parametric structure and nonparametric flexibility. While the literature on linear functionals offers useful guidance, extending these ideas to more complex functionals remains largely open. The present work shows that this direction is not only conceptually meaningful but also may admit concrete implementation, and therefore deserves further investigation.

References

Gaffi, F., Lijoi, A., and Prünster, I. (2025). Random probability measures with fixed mean distributions. *The Annals of Applied Probability* 35, 2239–2261. doi:10.1214/24-AAP2130. 1006

Kessler, D. C., Hoff, P. D., and Dunson, D. B. (2015). Marginally specified priors for nonparametric Bayesian estimation. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* 77, 35–58. doi:10.1111/rssb.12059. 1006

Peccati, G. and Prünster, I. (2008). Linear and quadratic functionals of random hazard rates: An asymptotic analysis. *The Annals of Applied Probability* 18, 1910–1943. doi:10.1214/07-AAP509. 1007

Contributed Discussion

Gourab Mukherjee*

Robust empirical Bayes prediction using D-BETEL

The authors are to be congratulated for a timely contribution to robust statistical inference. In Chakraborty et al. (2025), they propose D-BETEL, a distributionally robust Bayesian exponentially tilted empirical likelihood method that replaces exact parametric specification with a constrained transport neighborhood around a working model. The resulting procedure is both computationally tractable and conceptually appealing.

We consider empirical Bayes (EB) prediction in generalized linear mixed models (GLMMs) under acute data scarcity, where only a small subset of units contributes repeated observations and most units are observed at most once. Such regimes arise in modern prediction problems and are especially challenging because limited replication, severe imbalance, and possible covariate shift make the latent-effects distribution difficult to estimate reliably from the data Sarkar et al. (2026). In these settings, shrinkage is essential for stable prediction, since unit-specific estimates would otherwise be too noisy. Empirical Bayes is therefore attractive because it borrows strength across units while adapting to the limited data available for each one. We study whether the robustness of D-BETEL can be leveraged for EB prediction under data scarcity, imbalance, and covariate shift, and whether its performance is close to that of widely used EB methods based on nonparametric maximum likelihood estimation (NPMLE) Jiang and Zhang (2009) and g-modeling Narasimhan and Efron (2020).

In Section 3.2 of Chakraborty et al. (2025), the authors consider a Poisson GLM example. Here we study a related problem, but with random effects whose distribution is unknown and must be estimated. For $i = 1, \ldots, n$, consider $Y_{ik} = \text{Poisson}(x_{ik}^\top \beta + b_i), k = 1, \ldots, K_i$, and $\tilde{Y}_{i\tilde{k}} = \text{Poisson}(\tilde{x}_{i\tilde{k}}^\top \beta + b_i), \tilde{k} = 1, \ldots, \tilde{K}_i$. The random effects b_i are i.i.d. from g, β is a common fixed effect, and both are assumed to be invariant across training and test samples. The goal is to predict $\{\tilde{Y}_{i\tilde{k}}\}$ based on the observed $\{Y_{ik}\}$. NPMLE estimates g by marginal likelihood over a discrete support, while g-modeling fits g within a low-dimensional parametric family on a fixed grid Jiang and Zhang (2009); Narasimhan and Efron (2020). D-BETEL instead starts from a working family of priors and replaces exact fit by a transport-relaxed entropy criterion Chakraborty et al. (2025). On the common support grid used here, the difference is therefore between exact mixture fitting and transport-relaxed fitting. Table 1 summarizes the results from three simulation experiments comparing the performance of g-modeling, NPMLE, D-BETEL, and the unshrunken estimator. For Experiments 1 and 2, we report estimation error in the parameters. In Experiment 1, we consider Gaussian random effects and, across Cases A–E, examine different data-scarce imbalanced designs. The ratio of observations to parameters is 1, 2, 1.5, 1.2, and 1.4, respectively. The first two designs are balanced, whereas Cases C, D, and E exhibit low, moderate, and high imbalance, respectively. Experiments

*Department of Data Sciences and Operations, University of Southern California, gourab@usc.edu

2A–2C all have ratio equal to 1 and consider, respectively, a two-component Gaussian mixture, a heavy-tailed t distribution, and a contaminated prior with 10% large positive outliers as the random-effects distribution. Experiments 3A–3D consider the settings of Experiment 2 together with Gaussian random effects in Case A. They evaluate prediction under covariate shift, where future covariates follow a different distribution from the training covariates, and prediction accuracy is measured using log-scale mean error.

Experiment	ORACLE	g-modeling	NPMLE	D-BETEL	Unshrunken
1A	0.7782 (0.0881)	0.9525 (0.1783)	1.0195 (0.2636)	1.0966 (0.1244)	2.1301 (0.2001)
1B	0.7399 (0.0865)	0.8192 (0.1233)	0.8279 (0.1652)	1.0734 (0.1122)	1.6000 (0.1804)
1C	0.7127 (0.0803)	0.8248 (0.0990)	0.9487 (0.2070)	1.1201 (0.0854)	1.9337 (0.1614)
1D	0.7550 (0.0572)	0.9048 (0.1478)	1.0145 (0.3020)	1.1252 (0.1014)	2.0772 (0.1628)
1E	0.7346 (0.0808)	0.8856 (0.1561)	1.0080 (0.3826)	1.0803 (0.1089)	2.0961 (0.1600)
2A	1.6199 (0.2294)	1.6003 (0.2572)	1.6279 (0.2844)	1.8131 (0.2394)	2.3792 (0.3159)
2B	5.8146 (8.9803)	5.9295 (9.0564)	6.0208 (8.9841)	6.2165 (9.0033)	6.5493 (8.9969)
2C	1.6418 (0.3968)	1.4505 (0.2834)	1.5721 (0.2895)	1.9656 (0.2028)	2.3048 (0.2338)
3A	0.0086 (0.0016)	0.0625 (0.0051)	0.0622 (0.0049)	0.0622 (0.0050)	—
3B	0.0095 (0.0024)	0.0616 (0.0049)	0.0612 (0.0050)	0.0611 (0.0050)	—
3C	0.8331 (0.5800)	0.8809 (0.5783)	0.8806 (0.5778)	0.8806 (0.5776)	—
3D	0.2186 (0.1026)	0.2678 (0.1009)	0.2676 (0.1008)	0.2676 (0.1008)	—

Table 1: Mean (SD) loss across 25 replicates. Experiments 1–2 report mean absolute estimation error. Experiment 3 reports the log mean absolute prediction error under covariate shift.

All code for these results is available at https://github.com/gmukherjee/dbetel-eb. Across all regimes, D-BETEL yields performance that is close to that of state-of-the-art EB methods and is substantially better than the unshrunken MLE. In D-BETEL, the transport radius ϵ controls the degree of shrinkage; in our table we used $\epsilon = 0.35$ with a Wasserstein metric. Improved tuning may further enhance performance. In addition, using the full subject-level response vector in D-BETEL leads to a multivariate Wasserstein transport problem that is computationally expensive. The results reported here are instead based on a reduced model using summary statistics, which is computationally simpler but less informative; this suggests there is further room for improvement.

References

Chakraborty, A., Bhattacharya, A., and Pati, D. (2025). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis*, 1–33. Advance publication. doi:10.1214/25-BA1535. 1009

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics* 37, 1647–1684. doi:10.1214/08-AOS638. 1009

Narasimhan, B. and Efron, B. (2020). deconvolveR: A G-modeling program for deconvolution and empirical Bayes estimation. *Journal of Statistical Software* 94, 1–20. doi:10.18637/jss.v094.i11. 1009

Sarkar, A., Mukherjee, G., and Yano, K. (2026). Empirical Bayes predictive density estimation under covariate shift in large imbalanced linear mixed models. arXiv:2603.27984. 1009

Contributed Discussion

Lekha Patel*

I congratulate the authors on an excellent contribution at the intersection of empirical likelihood methods, optimal transport, and Bayesian nonparametrics. The formulation's most consequential contribution, in my view, is the explicit decoupling of the parameter of interest θ from the infinite-dimensional nuisance parameter $\xi^\star$ via a θ-dependent slice of the joint prior. This preserves the interpretation of θ within a parametric model whilst permitting the centering measure F_θ to be *approximately* correct. In this discussion piece, I focus on the implications of this perspective for Bayesian inverse problems (BIP), in which structural model misspecification is essentially endemic, and subsequently suggest extensions that would broaden the framework's reach.

Distributionally–constrained Bayesian Exponentially Tilted Empirical Likelihood (D-BETEL) as a robust likelihood for BIP Consider the canonical setup $y = \mathcal{G}(\theta) + \eta$, where $\mathcal{G} : \Theta \to \mathcal{Y}$ is a parameter-to-observable map and η is a noise process inducing a parametric family F_θ (Stuart, 2010). The realized forward map almost never coincides with $\mathcal{G}$ in practice, since aspects such as numerical discretization, unresolved subgrid physics, surrogate or reduced-order approximations, and experimental artifacts each introduce systematic structural (model) error. Standard responses, namely the Bayesian approximation error of Kaipio and Somersalo (2007) and the calibration framework of Kennedy and O'Hagan (2001), both presuppose a tractable parametric structure on the discrepancy that is typically Gaussian.

D-BETEL offers a qualitatively different alternative, in that by placing the discrepancy at the level of *distributions* and constraining the weighted empirical distribution $\nu_{w,x}$ to lie in an ANDREW-ball of F_θ, the data generating mechanism is permitted to depart from the centering model in manners that need not be tied to a particular parametric error structure. The asymptotic equivalence established in the paper's Theorem 2, which shows that the marginal posterior on θ arises as the limit of a nonparametric Bayesian model centered on F_θ, is therefore especially attractive in BIP, where one rarely trusts the model-form fully but does trust that θ is well-defined within it. I anticipate that this construction will sit naturally alongside, and in some regimes supplant, Kennedy and O'Hagan-style discrepancy modelling. Critically, unlike many pseudo likelihood-based robust Bayesian methods, D-BETEL retains a valid generative-model interpretation via Theorem 2, supporting the posterior predictive checks and forward uncertainty propagation that scientific applications typically demand.

Centering by a simulator: extending D-BETEL to likelihood-free settings In many BIP, F_θ is accessible only through a simulator (e.g., a Partial Differential Equation (PDE) solver or black-box surrogate), so that f_θ is unavailable in closed form. Exponentially Tilted Empirical Likelihoods (ETEL) are attractive here, since they only

*Scientific Machine Learning Department, Sandia National Laboratories, lpatel@sandia.gov

require a way to compare $\nu_{w,x}$ with F_θ rather than a closed-form density. This suggests a natural simulator-replaced variant in which F_θ is approximated by a Monte Carlo measure $\widehat{F}_{\theta,J} = J^{-1}\sum_{j=1}^{J}\delta_{\mathcal{G}(\theta;\xi_j)}$, although the precise relationship between this empirical-centering construction and the ANDREW formulation in Theorem 3 would require care. Here f_θ acts only as a sampler, sidestepping the high-dimensional density estimation burdens common in simulation-based inference (Cranmer et al., 2020). Because $\widehat{F}_{\theta,J}$ is itself discrete, the construction reduces to entropic-regularized OT between two empirical measures, bypassing the elliptical-mixture closed form of Theorem 3 in favor of a single Sinkhorn solve that is computationally favorable, but forfeiting the analytic tail correction encoded by the marginal quantile term. The tradeoff thus shifts from likelihood evaluation to repeated empirical transport, making the scaling of J, n, and the observation dimension central. Whether the equivalence in Theorem 2 lifts to this simulator-replaced regime would be of considerable interest, and I would especially welcome the authors' view on the conditions on $J = J(N)$ under which the marginal posterior under $\widehat{F}_{\theta,J}$ first converges to a finite-J simulator-centered D-BETEL posterior as $N \to \infty$, and subsequently to the exact F_θ-centered D- BETEL posterior as $J \to \infty$. I suspect this may require uniform stability of the feasible weight set $\{w : D[F_\theta, \nu_{w,x}] \leq \varepsilon\}$ under Monte Carlo perturbations of the centering measure, in the spirit of the empirical-Wasserstein concentration literature (Bernton et al., 2019).

Functional observations and ANDREW on infinite-dimensional spaces A practical ceiling on D-BETEL in BIP is that ANDREW is built for elliptical mixtures on $\mathbb{R}^d$, whereas many scientific observables are functional (e.g. spatial fields, spectra, time-series traces). Direct Wasserstein computation on function spaces is prohibitive, and naive vectorization through a fixed grid affects the geometry that motivated a transport metric to begin with. Two extensions strike me as natural: first a spectral-truncation analogue in which the marginal univariate Wasserstein terms in (2.14) act on Karhunen–Loève coefficients of the function-valued samples, and subsequently a coupling family parameterized by neural operators (Kovachki et al., 2023), which would inherit mesh invariance and integrate naturally with PDE-driven forward models. I wonder whether the authors view the *augment-then-restrict* strategy as general-purpose in this direction, or whether the identifiability condition underlying Lemma S9 places restrictions in the functional setting.

Tuning ε under limited replication The $\mathrm{ELPD}_\varepsilon$-based scheme (where ELPD refers to Expected Log point-wise Predictive Density) is elegant and has a clear leave-one-out interpretation when the data are exchangeable. In many BIP, however, the dataset of interest is effectively a single high-dimensional observation, such as one snapshot of a field, a single trace of a time-series, or a single inversion target, and consequently leave one-out no longer provides effective replication. The hyper-parameter ε may then need calibration by other means, for example with posterior predictive checks against simulator replicates, coverage diagnostics on synthetic twinned experiments, or cross validation across independent inversion datasets when available. The framework can plausibly accommodate such alternatives with minor modification, but the choice of calibration objective is non-trivial and I would thus value the authors' perspective on

principled defaults. More broadly, the interplay between an interpretable parametric centering and a flexible, transport-controlled departure from it is a promising modelling primitive for the many BIP that today rely on ad hoc discrepancy adjustments. I thank the authors for a stimulating contribution and look forward to their response.

References

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, Methodological* 81, 235–269. doi:10.1111/rssb.12312. 1012

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America* 117, 30055–30062. doi:10.1073/pnas.1912789117. 1012

Kaipio, J. and Somersalo, E. (2007). Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics* 198, 493–504. doi:10.1016/j.cam.2005.09.027. 1011

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B, Methodological* 63, 425–464. doi:10.1111/1467-9868.00294. 1011

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2023). Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research* 24, 1–97. MR4582511. 1012

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica* 19, 451–559. doi:10.1017/S0962492910000061. 1011

Contributed Discussion

Abhishek Roy*

1 Introduction

In this paper the authors make an innovative and thought-provoking contribution to robust Bayesian inference. The paper introduces a principled framework, Bayesian Exponentially tilted Empirical Likelihood with Distributional constraints (D-BETEL), that combines exponentially tilted empirical likelihood with a transport-based proximity constraint around a parametric centering family. The resulting methodology occupies an interesting middle ground between parametric Bayesian modeling preserving interpretability and fully nonparametric Bayesian procedures, while avoiding explicit infinite-dimensional nuisance parameterization.

We found several aspects of the paper particularly appealing. First, the proposed framework enlarges the effective model class through a distributional neighborhood around a parametric family. The weighted distribution Q provides data fidelity while allowing the interpretable reference model F_θ to remain relatively insensitive to outliers through the constraint $D(Q, F_\theta) \leq \varepsilon$. Second, the proposed ANDREW metric is technically elegant. The augmentation–restriction construction balances expressiveness and computational tractability in transport-based inference, while the marginal transport augmentation compensates for the loss of tail sensitivity induced by restricting couplings to elliptical mixture structures. The resulting metric inherits Sinkhorn-style scalability while remaining sensitive to tail behavior. Third, we appreciate the asymptotic nonparametric Bayes interpretation developed in Section 5. Unlike many generalized Bayes or pseudo-likelihood methods, D-BETEL is derived as a marginalization limit of a constrained hierarchical nonparametric model, providing a concrete probabilistic foundation.

While we view the paper as an important contribution, we would like to discuss two complementary extensions that may further broaden the scope of the framework.

2 Geometry-Aware D-BETEL

High-dimensional data in many scientific fields lie, at least approximately, on a low-dimensional Riemannian manifold, e.g., shape data in computational anatomy (Dryden and Mardia, 2016), directional data in environmental and geological science (Mardia and Jupp, 2009), and covariance matrices in brain imaging and diffusion tensor MRI (Pennec et al., 2006). In each case, inference method that ignores the manifold structure can produce statistically inefficient and scientifically uninterpretable results. Extending D-BETEL to manifold-valued data is therefore an important and natural direction.

*Department of Statistics, Texas A&M University, abhishekroy@tamu.edu

A manifold-aware formulation equips D-BETEL with robustness against off-manifold outliers. Suppose the centering family F_θ is supported intrinsically on a manifold $\mathcal{M}$, while the observed samples satisfy

$$x_i = m_i + \epsilon_i, \qquad m_i \in \mathcal{M},$$

where ϵ_i represents ambient perturbations transverse to the manifold. An intrinsic geodesic cost-based transport on $\mathcal{M}$ cannot explain the off-manifold perturbations (Pennec et al., 2006). Consequently, observations with large normal-direction perturbations become more expensive to match to a manifold-supported distribution, and the entropy-maximizing D-BETEL reweighting would naturally tend to assign them smaller weights. This also improves identifiability of F_θ. Without geometric constraints, the enlargement distribution Q_θ may absorb substantial off-manifold noise while remaining close to the empirical distribution, allowing even a poorly specified manifold-supported F_θ to appear adequate. For example, if the intrinsic data lie near a circle $\mathcal{M} \subset \mathbb{R}^2$ with radial noise contamination, Euclidean enlargements may spread mass into the ambient space to fit the noisy cloud. In contrast, intrinsic transport constraints prevent Q_θ from explaining arbitrary off-manifold perturbations, forcing F_θ to better capture the underlying geometry.

Three components of the framework require non-trivial adaptation. The *centering distribution* F_θ must be intrinsic to the manifold $\mathcal{M}$, and retain the low-dimensional interpretability. For example, von Mises–Fisher and Kent distributions on spheres (Mardia and Jupp, 2009), and log-Euclidean or affine-invariant Gaussians on Symmetric Positive Definite (SPD) manifolds (Pennec et al., 2006). A practical manifold-aware extension of the *transport metric* ANDREW would be needed. A potential alternative choice is Riemannian sliced Wasserstein distances Rustamov and Majumdar (2023); Quellmalz et al. (2023); Bonet et al. (2025). The *enlargement class* would also need to respect manifold geometry. A principled manifold-aware extension would likely require a combination of reweighting and projection or the class of enlargements need to be constrained to remain near the underlying manifold. In the above discussion, the underlying manifold is assumed to be known. However, a practical thorough work would require a joint estimation of the manifold as well.

3 Flow-Based Enlargement Class

A key contribution of the paper is the ANDREW metric, which combines transport-based geometry with computational tractability. To make distance evaluation feasible, the parametric centering family F_θ, and the restricted coupling class used to compute the transport discrepancy $D(F_\theta, \nu_{w,x})$ have been specialized to elliptical mixture models (EMM). The enlargement itself, however, remains the weighted empirical distribution

$$\nu_{w,x} = \sum_{i=1}^{n} w_i \delta_{x_i}, \qquad w \in \Delta_n.$$

This is elegant and computationally effective, but it restricts the coupling class that may fail to capture nonlinear geometric differences between F_θ and the reweighted

empirical distribution $\nu_{w,x}$ such as asymmetric tail couplings or higher-order dependence patterns requiring highly nonlinear mass rearrangements. To address this, we propose to leverage modern flow-based generative models (Papamakarios et al., 2019) based enlargement class. Recent flow matching methods (Lipman et al., 2023; Onken et al., 2021) are explicitly designed to approximate transport maps and probability flows between distributions, making them particularly well-suited to the transport-based setting of D-BETEL. We observe that the population-level target $\hat{Q}_\theta$ is defined as the optima over the space of probability measures (equation (4.2) of the paper). From this perspective, a flow-based construction provides a smooth surrogate for the population enlargement, capable of representing continuous nonlinear perturbations of F_θ beyond discrete reweighting. The tradeoff is that the enlargement class becomes much richer, making regularization essential to preserve identifiability of θ.

Specifically, define the flow-based enlargement $Q_\theta^\phi := (T_\phi)_\# F_\theta$, where $T_\phi : \mathbb{R}^d \to \mathbb{R}^d$ is the solution at time $t = 1$ of the neural ODE

$$\frac{d}{dt} X_t = v_\phi(X_t, t), \qquad X_0 \sim F_\theta,$$

and $v_\phi : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a neural velocity field. Because the flow class is highly expressive, Q_θ^ϕ can approximate the data distribution P for a wide range of θ, which would cause F_θ to lose its role as an identifiable scientific reference model. Anchoring the enlargement to θ requires a regularizer that controls the displacement from F_θ.

Motivated by the dynamic characterization of $W_2^2(F_\theta, Q_\theta^\phi)$ as the minimum kinetic energy over all admissible probability paths and velocity fields connecting F_θ and Q_θ^ϕ, we propose a continuous-time analog of D-BETEL's transport constraint $D(F_\theta, Q) \le \varepsilon$ where we learn ϕ for fixed θ by minimizing

$$\mathcal{L}_\phi(\theta) = \mathcal{SW}^2\Big(Q_\theta^\phi, P_n\Big) + \lambda \int_0^1 \mathbb{E}_{X_t \sim Q_{\theta,t}^{(\phi)}} \|v_\phi(X_t, t)\|^2 \, dt,$$

where $\mathcal{SW}^2$ denotes the squared sliced Wasserstein distance and P_n is the empirical distribution of the observed data. The first term keeps Q_θ^ϕ close to the empirical distribution. The second term, motivated by the Benamou–Brenier characterization of $W_2^2(F_\theta, Q_\theta^\phi)$, regularizes the displacement of Q_θ^ϕ from F_θ. Note that $\mathcal{SW}^2\Big(Q_\theta^\phi, P_n\Big)$ is well-defined and admits efficient gradient-based optimization with respect to ϕ. Profiling out ϕ for each θ defines the flow-based pseudo-likelihood

$$L_{\mathrm{CNF}}(\theta) = \exp\Big(-\min_\phi \mathcal{L}_\phi(\theta)\Big),$$

and combining with a prior $\pi(\theta)$ gives the robust posterior

$$\pi_{\mathrm{CNF}}(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta)\, L_{\mathrm{CNF}}(\theta).$$

This construction treats ϕ as a nuisance parameter that is profiled out for each θ evaluation, in the same spirit as D-BETEL profiles out the weights w through the constrained

entropy maximization. Unlike D-BETEL, however, the present pseudo-likelihood is not currently derived from a nonparametric Bayesian formulation analogous to Theorem 2 of the paper, and establishing such guarantees remains an important open problem. A further practical challenge is that minimizing $\mathcal{L}_\phi(\theta)$ for each Markov Chain Monte Carlo (MCMC) proposal can be expensive; a natural remedy is to amortize the optimization by conditioning the flow v_ϕ directly on θ and learning a shared transport model across parameter values.

Concluding Remarks

Overall, we believe this paper makes a significant contribution to robust Bayesian methodology by proposing a novel methodology that achieves robustness while preserving interpretability. Beyond the extensions we mention above, we believe this idea can impact several other directions such as causal inference under covariate shift, and interpretable machine learning.

References

Bonet, C., Drumetz, L., and Courty, N. (2025). Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds. *Journal of Machine Learning Research* 26, 1–76. MR4873583. 1015

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*. John Wiley & Sons. doi:10.1002/9781119072492. 1014

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=PqvMRDCJT9t. 1016

Mardia, K. V. and Jupp, P. E. (2009). *Directional Statistics*. John Wiley & Sons. MR1828667. 1014, 1015

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. (2021). OT-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 9223–9232. 1016

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. arXiv:1912.02762. MR4253750. 1016

Pennec, X., Fillard, P., and Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision* 66, 41–66. doi:10.1007/s11263-005-3222-z. 1014, 1015

Quellmalz, M., Beinert, R., and Steidl, G. (2023). Sliced optimal transport on the sphere. *Inverse Problems* 39, 105005. doi:10.1088/1361-6420/acf156. 1015

Rustamov, R. M. and Majumdar, S. (2023). Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *International Conference on Machine Learning*, 29388–29415. PMLR. 1015

Contributed Discussion

Hien Duy Nguyen[*,†], TrungTin Nguyen[‡], Julyan Arbel[§], and Florence Forbes[¶]

We congratulate the authors on their interesting contribution. We shall restrict ourselves to discussing the tolerance parameter ε in the D-BETEL construction in Chakraborty et al. (2025). Let Π denote the prior measure corresponding to π. The D-BETEL posterior is

$$\Pi_n^D(d\theta \mid x_{1:n}) \propto L_n^D(\theta)\,\Pi(d\theta), \qquad L_n^D(\theta) = \prod_{i=1}^{n} w_i^\star(\theta),$$

where

$$w^\star(\theta) \in \arg\max_{w \in \mathcal{S}_{n-1}} \left\{\mathcal{H}(w) : D(F_\theta, \nu_{w,x}) \le \varepsilon\right\}, \qquad \nu_{w,x} = \sum_{i=1}^{n} w_i \delta_{x_i}.$$

Assume that this likelihood is measurable in θ on the sets considered below. Although D-BETEL is not an ABC posterior in the algorithmic sense, it uses a tolerance to update through a discrepancy. Thus the fixed-tolerance, $n \to \infty$ analysis of ABC coarsened and pseudo-posterior measures in Nguyen et al. (2025) is relevant to the statistical role of ε. Let $X_1, X_2, \ldots$ be $\mathcal{X}$-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $X_i : \Omega \to \mathcal{X}$, and suppose they are i.i.d. with common law P_0. For each n, define $\nu_n(\omega) = n^{-1} \sum_{i=1}^{n} \delta_{X_i(\omega)}$, writing ν_n when ω is understood.

The following proposition records a simple fixed-tolerance implication of the D-BETEL entropy maximization. On a set K that lies strictly inside an ε-fit region and on which the empirical discrepancy converges uniformly, D-BETEL eventually supplies no data-dependent distinction among parameters. That is, all relative posterior mass within K is inherited from Π. Fixed-tolerance D-BETEL should therefore not generally be interpreted as point-consistent inside a strict ε-fit region. It is better viewed as a coarsened discrepancy posterior, in the sense of Miller and Dunson (2019), whose fixed tolerance induces ε-identification regions.

Proposition (Fixed-tolerance likelihood non-identifiability in D-BETEL)**.** *Let $K \subset \Theta$ be measurable with $\Pi(K) > 0$. Suppose that, for some $\eta > 0$,*

$$\sup_{\theta \in K} D(F_\theta, P_0) \le \varepsilon - \eta, \qquad \sup_{\theta \in K} |D(F_\theta, \nu_n) - D(F_\theta, P_0)| \longrightarrow 0 \quad a.s.$$

Then there exists an event $\Omega_K \in \mathcal{F}$ with $\mathbb{P}(\Omega_K) = 1$ such that, for every $\omega \in \Omega_K$, there is $N_K(\omega) < \infty$ for which, for all $n \ge N_K(\omega)$, $L_n^D(\theta) = n^{-n}$ for every $\theta \in K$, and $\Pi_n^D(B \mid K, X_{1:n}(\omega)) = \Pi(B)/\Pi(K)$ for every measurable $B \subset K$.

[*]La Trobe Univ., Victoria, Australia, h.nguyen5@latrobe.edu.au
[†]Kyushu Univ., Fukuoka, Japan
[‡]Queensland Univ. of Technology, Brisbane, Australia
[§]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France
[¶]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France

Proof. Let Ω_K be the probability-one event on which the uniform convergence assumption holds. For $\omega \in \Omega_K$, choose $N_K(\omega)$ so that, for all $n \geq N_K(\omega)$, the displayed supremum error is at most η. Then $\sup_{\theta \in K} D(F_\theta, \nu_n(\omega)) \leq \sup_{\theta \in K} D(F_\theta, P_0) + \eta \leq \varepsilon$. Hence the uniform weight vector $u_n = (1/n, \ldots, 1/n)$ is feasible for every $\theta \in K$, with $x_i = X_i(\omega)$, since $\nu_{u_n,x} = \nu_n(\omega)$. For $w \in \mathcal{S}_{n-1}$, $\mathcal{H}(w) = \log n - \sum_{i=1}^n w_i \log(w_i/(1/n))$, with $0 \log 0 = 0$. The sum is the Kullback–Leibler divergence from w to u_n, which is non-negative and vanishes only when $w = u_n$. Thus u_n is the unique entropy maximizer on the simplex and on any constrained subset containing it. Consequently $w^\star(\theta) = u_n$ and $L_n^D(\theta) = n^{-n}$ for every $\theta \in K$. For every measurable $B \subset K$,

$$\Pi_n^D(B \mid K, X_{1:n}(\omega)) = \frac{\int_B L_n^D(\theta)\, \Pi(d\theta)}{\int_K L_n^D(\theta)\, \Pi(d\theta)} = \frac{\Pi(B)}{\Pi(K)},$$

where the denominator is positive because $\Pi(K) > 0$. □

In the ABC setting, this distinction is formalized in Nguyen et al. (2025). There, $D_\infty(\theta_0, \theta)$ is the almost-sure large-sample limit of the discrepancy between data generated at θ_0 and synthetic data generated at θ. Theorem 1 of Nguyen et al. (2025) shows, under compactly supported weights and regularity assumptions, that for every prescribed neighborhood of $\Theta_0 = \{\theta : D_\infty(\theta_0, \theta) = 0\}$ one can choose a fixed tolerance so that fixed-tolerance ABC posterior measures converge to full mass on that neighborhood. Under misspecification, Remark 4 of Nguyen et al. (2025) then replaces Θ_0 by the corresponding minimal discrepancy set. The proposition above gives a D-BETEL analogue: within a region strictly satisfying the tolerance constraint, posterior updating preserves prior relative probabilities for all sufficiently large n on an event of probability one.

The authors' nonparametric Bayes interpretation is complementary rather than a substitute for such sample-size asymptotics. Their Theorem 2 gives an auxiliary limit with the sample size and tolerance held fixed, whereas the separate large-sample question asks what D-BETEL learns about θ as $n \to \infty$ and how this depends on whether the tolerance is held fixed or allowed to shrink. A large-sample theory for a shrinking-tolerance regime would clarify that the tolerance is not only a computational or robustness tuning parameter, but also a statistical coarsening parameter determining the granularity of the target. We do not have a general approach to the analysis of this more difficult limit regime but hope that the authors may consider the challenge in their subsequent work.

References

Chakraborty, A., Bhattacharya, A., and Pati, D. (2025). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis*, 1–33. Advance publication. doi:10.1214/25-BA1535. 1018

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association* 114, 1113–1125. doi:10.1080/01621459.2018.1469995. 1018

Nguyen, H. D., Nguyen, T. T., Arbel, J., and Forbes, F. (2025). Revisiting concentration results for approximate Bayesian computation. *Bayesian Analysis*, 1–23. Advance publication. doi:10.1214/25-BA1520. 1018, 1019

Rejoinder

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

1 Response to invited and contributed discussions

We sincerely thank all the discussants for their deeply thoughtful and constructive discussion of our work (Chakraborty et al., 2025). Their comments help clarify and enhance the scope and positioning of our work, while also highlighting several important future directions. A recurring theme across many of the discussions is the broader question of how Bayesian inference should be conducted when statistical models are viewed not as exact truths, but rather as scientifically interpretable approximations. This perspective is particularly relevant in many modern high-dimensional or simulation-based inference (SBI) settings where exact specification of the data-generating mechanism is often unrealistic and some degree of misspecification or contamination from the posited model is inevitable. Since its publication, we have continued to reflect on various open questions emanating from our work, and we were gratified to see many of these themes appearing in the discussions. In a recent preprint (Jacobs et al., 2026), we have made progress towards addressing some of these questions, and we will draw upon this work at places in our response.

Our response to the discussants' comments follow, which we organize around a number of unifying themes emerging from the discussions.

1.1 Target of estimation, model expansion, and generalized Bayesian perspectives

A number of discussants (Avella Medina et al., 2026; Huggins, 2026; Linero, 2026; Wang, 2026) touched upon a fundamental point concerning the precise inferential target of Distributionally-constrained Bayesian Exponentially-tilted Empirical Likelihood (D-BETEL) and how it differs from classical likelihood-based inference under misspecification. It is well-known that in likelihood-based inference, the *target of estimation* $\theta^\star$ is the Kullback–Leibler (KL) projection, assumed unique for the purpose of this discussion,

$$\theta^\star := \arg\min_{\theta\in\Theta} \mathrm{KL}(P||F_\theta) = \arg\max_{\theta\in\Theta} E_{X\sim P} \log f_\theta(X), \tag{1}$$

where P is the true data distribution, and $\{F_\theta : \theta \in \Theta\}$ is the statistical model, i.e., a parametric family of distributions indexed by parameter θ. Maximum likelihood estimation or usual Bayesian estimation asymptotically target θ^* under mild assumptions

*Global Statistical Sciences, Eli Lilly and Company (Formerly at Department of Statistics, Texas A&M University), abhisek_chakraborty@tamu.edu
†Department of Statistics, Texas A&M University, anirbanb@stat.tamu.edu
‡Department of Statistics, University of Wisconsin, Madison, dpati2@wisc.edu

(Kleijn and Van der Vaart, 2012). In the well-specified setting, i.e., $P = F_{\theta^\dagger}$ for some $\theta^\dagger \in \Theta$, the KL projection $\theta^\star = \theta^\dagger$. However, even if P is a 'small' contamination of $F_{\theta^\dagger}$, $\theta^\star$ can be substantially different from $\theta^\dagger$, depending upon the nature of the contamination. For example, in a classical Huber-contamination model (Huber, 1992), $P = (1-\varepsilon)F_{\theta^\dagger} + \varepsilon H$ for some $\varepsilon \in (0,1)$ and probability distribution H, even though P is close to $F_{\theta^\dagger}$ in total variation (TV) distance for ε small, $\theta^\star$ can be substantially different from $\theta^\dagger$ depending on H. This well-known phenomenon renders likelihood based estimation non-robust. An overarching goal of D-BETEL is to holistically provide robustness under general contamination regimes. As elucidated in Section 4 of Chakraborty et al. (2025), D-BETEL (with tuning parameter λ) targets

$$\theta^\dagger_\lambda = \arg\max_{\theta\in\Theta} E_{X\sim P}[\log \widehat{Q}_{\theta,\lambda}(X)], \tag{2}$$

where

$$\widehat{Q}_{\theta,\lambda} = \arg\min_{Q \ll P} \big[\mathrm{KL}(Q \,||\, P) + \lambda \mathrm{D}(F_\theta, Q)\big]. \tag{3}$$

Accordingly, the target induced by D-BETEL is intentionally different from the classical KL projection. Rather than identifying the best approximation under exact likelihood geometry, D-BETEL seeks a parameter whose induced distribution remains geometrically close to a suitably reweighted version of the observed data. This viewpoint naturally connects to the discussants' observations regarding implicit model expansion. As emphasized by Wang (2026) and Linero (2026) in their discussions, the constrained transport formulation can be viewed as implicitly enlarging the model family without explicitly specifying a semiparametric extension. The transport geometry permits the working model to flexibly adapt to localized deviations from the parametric family while still remaining centered around interpretable structure. We also believe this interpretation connects D-BETEL to generalized Bayesian updating, Bayesian nonparametrics, and robust empirical likelihood methodology. In particular, the adaptive reweighting induced by the constrained transport in (3) resembles a localized semiparametric enrichment of the parametric model. Gaffi and Franzolini (2026) draw interesting connections between D-BETEL and recent developments in Bayesian nonparametrics where priors on random probability measures are constructed to possess pre-specified marginals of linear functionals (Gaffi et al., 2025).

The question of uncertainty quantification via establishing Bernstein–von Mises (BvM) results for the D-BETEL posterior was also discussed Linero (2026). This is certainly an important and interesting direction we wish to address in future work. For fixed λ, we conjecture that the D-BETEL posterior will asymptotically assume a Gaussian shape, centered at $\theta^\dagger_\lambda$. A central question then would be to characterize the posterior covariance, and connect it with the appropriate asymptotic sandwich covariance structure under model misspecification. A second layer of complexity enters the picture when one considers a data-driven choice $\hat{\lambda}$ of the tuning parameter λ, for example the predictive criterion described in Section 5 of Chakraborty et al. (2025), or other variants as in our more recent work (Jacobs et al., 2026). Assuming that $\hat{\lambda}$ approaches some stable limit λ^*, which in itself would be non-trivial to establish, an interesting

question then would be to characterize the population target $\theta^{\dagger}_{\lambda^*}$, that is $\theta^{\dagger}_{\lambda}$ in (2) evaluated at $\lambda = \lambda^*$. Specifically, it would be interesting to formulate contamination mechanisms generalizing the Huber contamination model so that under 'small' contamination, $\theta^{\dagger}_{\lambda^*} = \theta^{\dagger}$. In our recent preprint (Jacobs et al., 2026), we formulated one such contamination mechanism which we call the Geometric+Huber (G+H) contamination model:

$$P = (1-\varepsilon)Q + \varepsilon H, \qquad W_2(Q, F_{\theta^\dagger}) \leq \rho.$$

This framework simultaneously captures classical Huber contamination, smooth geometric perturbations, and combinations of both. We conjecture that for small (ε, ρ) and possibly under suitable identifiability conditions on H, D-BETEL can recover $\theta^{\dagger}$ in the G+H model. However, establishing this would require substantial work. In Jacobs et al. (2026), we instead consider a minimum divergence estimation framework together with weighted likelihood/bootstrap uncertainty quantification. This places the resulting procedure much closer to a classical robust M-estimation framework (Rousseeuw et al., 2005) where asymptotic sandwich covariance behavior is expected under standard regularity conditions.

We also thank Huggins (2026) for bringing up the under-fragmentation tendency of D-BETEL. Importantly, we do not view this tendency toward a single cluster as merely an artifact of the method. Rather, it is a direct consequence of the modeling framework underlying our robustness formulation. As mentioned above, our uncertainty class is characterized through a (G+H) contamination/transport neighborhood, which is intentionally designed to absorb certain forms of distributional deviation from the nominal model. In particular, a mixture of skew-normal components can, after suitable transport and contamination adjustments, lie close to a single normal distribution. From this perspective, the apparent "under-fragmentation" reflects the fact that the observed heterogeneity is being interpreted as distributional misspecification rather than evidence for additional latent clusters. Consequently, it is not surprising that the proposed method favors a single cluster in such settings.

1.2 Connections to ABC and computational considerations

Several discussants (Catalano and Legramanti, 2026; Patel, 2026; Wang, 2026) emphasized the broader connection between D-BETEL and likelihood-free Bayesian inference such as approximate Bayesian computation (ABC), and the potential of cross-pollination of ideas between these literatures. ABC (Beaumont et al., 2002; Marin et al., 2012; Sisson et al., 2018) and its more modern moniker, simulation-based inference (SBI; Cranmer et al. (2020)), tackle statistical problems where the probability model f_θ is implicitly defined in terms of simulations, often available in the form of an expensive computer code, and the likelihood function is otherwise analytically intractable. In earlier versions of ABC, regions of high posterior probability were identified based on closeness of simulated data (often summarized in the form of certain carefully chosen *summary statistics*) under putative parameter values with the observed data. In their discussion, Catalano and Legramanti (2026) note an interesting parallel in the ways

that the ABC literature has gradually evolved from using (moment-based) summary statistics to distributional discrepancy measures, and the way in which D-BETEL departs from traditional BETEL based on moment constraints. Indeed, both D-BETEL and discrepancy-based ABC methods (Bernton et al., 2019; Legramanti et al., 2025) involve a distributional discrepancy measure between a discrete distribution supported on the observed data and F_θ. Since the Wasserstein distance or the modified transport metric ANDREW proposed in Chakraborty et al. (2025) can be approximated based on samples from F_θ, one can naturally envision an extension of D-BETEL to ABC/SBI settings. This perspective motivated our recent work (Jacobs et al., 2026) on likelihood-free simulation-based inference under a minimum divergence estimation framework (Basu et al., 2011), with the divergence measure motivated by the construction in (3). In Jacobs et al. (2026), only simulations from the model are required, inference is performed through robust minimum transport divergence estimation, and uncertainty quantification is obtained through a bootstrap procedure. We believe the resulting framework provides a robust alternative to existing ABC and SBI methodologies, particularly in settings involving geometric misspecification or contamination, and contributes to the emerging literature on robust ABC/SBI (Chérief-Abdellatif and Alquier, 2020; Dellaporta et al., 2022; Bharti et al., 2026).

On a somewhat related note, several discussants (Avella Medina et al., 2026; Catalano and Legramanti, 2026; Huggins, 2026; Linero, 2026) commented about computational tractability and scalability of D-BETEL. Since each likelihood evaluation of D-BETEL requires solving a transport problem, it is indeed important to characterize the computational complexity of likelihood evaluations as a function of the data dimension as well as the parameter dimension. In Chakraborty et al. (2025), we mostly used off-the-shelf computational tools for computing Wasserstein distances and ANDREW. Given we specifically require Wasserstein distances between a discrete and continuous distribution, there is a natural connection to the semi-discrete optimal transport literature. We have investigated the computational aspect more thoroughly in Jacobs et al. (2026), since in ABC/SBI settings, it is important to characterize the number of simulations required from the possibly expensive simulator. In Jacobs et al. (2026), we work directly with a robust semi-constrained Wasserstein divergence (although the computational architectures should be extendable to ANDREW) under the assumption that only simulations from the model are available. A key technical contribution is a new reformulation of the semi-discrete robust transport problem:

$$\ell_\lambda(P_n, F_\theta) = \sup_{g \in \mathbb{R}^n} \mathbb{E}_{X \sim F_\theta}[h(X, g)],$$

where the objective is concave in the dual variable. This reformulation permits stochastic subgradient ascent and leads to a scalable simulation-based inference pipeline requiring only the ability to simulate from the model. We establish explicit non-asymptotic convergence guarantees for the resulting stochastic optimization procedure. In particular, after s iterations, the approximation error decreases roughly at rate $O\big(\sqrt{\frac{n}{s}}\big)$, up to logarithmic factors, while computational complexity scales linearly as $O(ns)$. Consequently, choosing $s \asymp n\epsilon^{-2}$ yields approximation accuracy of order $\tilde{O}(\epsilon)$ with runtime $O\Big(\frac{n^2}{\epsilon^2}\Big)$. To the best of our knowledge, this is the first convergence guarantee for semi-discrete

robust semi-constrained optimal transport in the setting where one of the distributions is continuous and accessible only through simulation.

1.3 Choice of the transport metric and robustness geometry

Another major theme across the discussions concerns the choice of transport metric (Huggins, 2026; Wang, 2026; Avella Medina et al., 2026; Nguyen, 2026; Catalano and Legramanti, 2026) and the geometry used to define robustness. A major motivation for our use of transport geometry was the desire to simultaneously capture small geometric perturbations affecting many observations together with contamination through a smaller number of large outliers. Classical likelihood-based discrepancies often become unstable when the empirical and model distributions differ geometrically or possess support mismatch. Wasserstein-type discrepancies remain meaningful in these settings and retain direct geometric interpretability. At the same time, we fully agree with the discussants that classical Wasserstein distance itself is sensitive to sufficiently large outliers. This was one of the motivations behind the constrained transport formulation and entropy-based reweighting mechanism used in D-BETEL, which effectively allows contaminating mass to be downweighted before geometric comparison occurs.

We also appreciated the discussants' suggestions regarding sliced Wasserstein distances, entropy-regularized transport, robustified Wasserstein formulations, and more classical metrics such as the Prokhorov metric. We agree that sliced Wasserstein distances possess attractive statistical and computational properties, particularly in moderate and high dimensions, and may better capture certain forms of joint dependence and tail behavior. However, one of the primary reasons we adopted the current transport construction was computational tractability and analytical transparency. Directly replacing the current discrepancy with sliced Wasserstein distance would likely sacrifice some of the explicit decompositions and optimization properties that make the current framework computationally feasible. Nonetheless, we believe hybrid constructions combining robust transport geometry, sliced projections, entropy regularization, and contamination-aware reweighting represent promising future directions.

In particular, Nguyen (2026) interestingly identified that the marginal Wasserstein component of the original ANDREW distance is a degenerate sliced Wasserstein distance (Rabin et al., 2011) based only on coordinate-axis projections, and proposed replacing it with a genuine sliced Wasserstein term. This modification retains computational scalability while better capturing multivariate dependence and restoring the identity-of-indiscernibles property. Further, the discussant suggested replacing the entropy regularized mixture Wasserstein term by distances defined directly on the mixing measures of elliptical mixture models, such as Mixture Sliced Wasserstein (Mix-SW) or Sliced Mixture Wasserstein (SMix-W). These alternatives retain metricity while reducing the computational complexity from $O(K_1K_2)$ to $O((K_1 + K_2)\log(K_1 + K_2))$ by operating on the lower-dimensional space of mixture components. We appreciate these suggestions which are certainly interesting directions to pursue. In most practical applications of ANDREW, the number of mixture components is typically very small ($K_1 = 2$ or 3), and the extent of computational gains in practice will need empirical investigation.

1.4 Hyperparameter calibration and robustness tuning

Mechanics of tuning the hyperparameter λ was another prevalent point of discussion (Huggins, 2026; Gaffi and Franzolini, 2026; Linero, 2026; Nguyen et al., 2026). We agree with the discussants that calibrating the robustness parameter λ plays an important role. In Chakraborty et al. (2025), we proposed an out-of-sample predictive criterion for tuning λ. Further research is required to better understand the inner mechanisms of this predictive criterion. We agree that a predictive criterion alone can potentially overfit, especially in unsupervised settings involving misspecification. However, coming up with a universally applicable criterion may be rather difficult, and ensemble approaches (for example, the BMA-inspired approach proposed in Chakraborty et al. (2025)) also need careful investigation. In our recent work (§3.3 of Jacobs et al. (2026)), we developed a new data-driven λ-selection procedure directly tied to the underlying robustness objective under G+H contamination. The key observation is that λ controls the degree of reweighting of the empirical distribution. When λ is too small, the divergence behaves similarly to standard Wasserstein distance and inherits its lack of robustness to contamination; when λ is too large, the loss landscape degenerates and inference becomes uninformative. Our new approach studies the behavior of

$$W_2(F_{\widehat{\theta}_\lambda}, \widehat{Q}_\lambda(\widehat{\theta}_\lambda))$$

across a logarithmically spaced grid of λ values, where $\widehat{Q}_\lambda$ is the transport-induced reweighted empirical measure. As λ increases, contaminating observations become progressively down-weighted, and once the contaminating mass is largely removed, further increases yield diminishing returns. This naturally produces an elbow or U-shaped diagnostic curve, and the selected λ corresponds to the transition point. Importantly, this calibration procedure is directly tied to the geometric+Huber contamination interpretation of the method rather than purely predictive performance. We also note here that Gaffi and Franzolini (2026) raise an interesting point regarding a soft version of the hard constraint $D(F_\theta, Q) \leq \varepsilon$, potentially induced through a prior specification, so that posterior inference on the extent of departure from the centering model can be carried out.

Relatedly, Nguyen et al. (2026) establish an interesting proposition showing that, under sufficiently strong interior-feasibility conditions, the entropy-maximization problem that sits at the hearth of D-BETEL, asymptotically admits the uniform weights as its optimizer. However, our primary motivation concerns robust inference under persistent misspecification, where the transport constraint typically remains active and continues to induce informative non-uniform reweighting even asymptotically.

1.5 Extensions of D-BETEL and open questions

Finally, the discussions highlighted several additional promising directions for future research.

One recurring theme concerns the relationship between D-BETEL and Bayesian data reweighting approaches (Avella Medina et al., 2026; Wang, 2026). While both

frameworks achieve robustness through adaptive reweighting, Bayesian data reweighting treats weights as latent random variables regularized through a prior, whereas D-BETEL induces weights deterministically through entropy-constrained transport optimization. This leads to different notions of outliers: local likelihood incompatibility versus global geometric incompatibility.

Several discussants also raised important questions regarding posterior influence functions, breakdown point analysis, multi-modality and non-convexity, redescending behavior, and the relationship to generalized posterior robustness theory (Avella Medina et al., 2026; Linero, 2026). We agree these remain important open problems. In particular, we believe the multi-modal geometry observed in robust transport-based inference is closely related to the broader phenomenon that aggressively robust procedures often necessarily induce nonconvex optimization landscapes. At the same time, the newer convex dual stochastic optimization framework substantially improves computational tractability even when multi-modality remains present at the statistical level.

Finally, several discussants (Mukherjee, 2026; Roy, 2026; Patel, 2026; Catalano and Legramanti, 2026) highlighted exciting extensions involving more complex data domains, structured and dependent data, high-dimensional transport geometry, semi-parametric expansions, robust nonparametric likelihood formulations, and broader connections to generalized Bayesian inference. Roy (2026) highlights two particularly interesting extensions of D-BETEL. First, they argue that extending D-BETEL to manifold-valued data may provide robustness against off-manifold perturbations while preserving intrinsic geometric structure in applications such as shape analysis, directional data, and diffusion tensor imaging. Second, they propose enriching the enlargement class through modern flow-based generative models, allowing D-BETEL to capture non-linear geometric perturbations and higher-order transport structure beyond discrete empirical reweighting. Mukherjee (2026) considered empirical Bayesian inference in generalized linear mixed models under severe data scarcity and imbalance, where reliable estimation of latent random effects becomes particularly challenging. The discussant emphasizes that stable shrinkage and robustness to covariate shift are especially important in such regimes, motivating the use of robust D-BETEL-type formulations for empirical Bayes prediction. Patel (2026) proposes D-BETEL as a robust alternative for Bayesian inverse problems, where structural misspecification in the forward map often renders classical Gaussian discrepancy formulations inadequate. The discussant also highlights extensions of ANDREW to functional and infinite-dimensional settings, including spectral representations and neural-operator parameterizations capable of preserving transport geometry for complex scientific observables.

We are deeply grateful to all discussants for these insightful comments and believe they significantly strengthened both the technical and conceptual foundations of the work.

References

Avella Medina, M., Marusic, J., and Ronchetti, E. (2026). Discussion of the paper "Robust probabilistic inference via a constrained transport metric" by A. Chakraborty, A. Bhat-

tacharya, and D. Pati. *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1020, 1023, 1024, 1025, 1026

Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. CRC Press. doi:10.1201/b10956. 1023

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035. doi:10.1093/genetics/162.4.2025. 1022

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, 235–269. doi:10.1111/rssb.12312. 1023

Bharti, A., Dellaporta, C., Hikida, Y., and Briol, F.-X. (2026). Amortised and provably-robust simulation-based inference. arXiv:2602.11325. 1023

Catalano, M. and Legramanti, S. (2026). Contributed discussion to "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1022, 1023, 1024, 1026

Chakraborty, A., Bhattacharya, A., and Pati, D. (2025). Robust probabilistic inference via a constrained transport metric. *Bayesian Analysis*, 1–33. doi:10.1214/25-BA1535. 1020, 1021, 1023, 1025

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, 1–21. PMLR. MR4147569. 1023

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences* 117, 30055–30062. doi:10.1073/pnas.1912789117. 1022

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022). Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, 943–970. PMLR. 1023

Gaffi, F. and Franzolini, B. (2026). Contributed discussion on "Robust probabilistic inference via a constrained transport metric" by A. Chakraborty, A. Bhattacharya, and D. Pati. *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1021, 1025

Gaffi, F., Lijoi, A., and Prünster, I. (2025). Random probability measures with fixed mean distributions. *The Annals of Applied Probability* 35, 2239–2261. doi:10.1214/24-AAP2130. 1021

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution*, 492–518. Springer. doi:10.1007/978-1-4612-4380-9_35. 1021

Huggins, J. H. (2026). Invited discussion of "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1020, 1022, 1023, 1024, 1025

Jacobs, P., Patel, L., Bhattacharya, A., and Pati, D. (2026). Robust simulation based inference through robust optimal transport. arXiv:2605.18741. 1020, 1021, 1022, 1023, 1025

Kleijn, B. J. and Van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. doi:10.1214/12-EJS675. 1021

Legramanti, S., Durante, D., and Alquier, P. (2025). Concentration of discrepancy-based approximate Bayesian computation via Rademacher complexity. *The Annals of Statistics* 53, 37–60. doi:10.1214/24-aos2453. 1023

Linero, A. R. (2026). Invited discussion. *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1020, 1021, 1023, 1025, 1026

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing* 22, 1167–1180. doi:10.1007/s11222-011-9288-2. 1022

Mukherjee, G. (2026). Discussion of "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1026

Nguyen, H. D., Nguyen, T., Arbel, J., and Forbes, F. (2026). Discussion of "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1025

Nguyen, K. (2026). Discussion of "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1024

Patel, L. (2026). Invited discussion of "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1022, 1026

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In Kuijper, A., Bredies, K., Pock, T., and Bischof, H. (eds.), *Scale Space and Variational Methods in Computer Vision*, volume 6667 of *Lecture Notes in Computer Science*, 435–446. Springer. doi:10.1007/s10851-015-0579-7. 1024

Rousseeuw, P. J., Hampel, F. R., Ronchetti, E. M., and Stahel, W. A. (2005). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons Incorporated. doi:10.1002/9781118186435. 1022

Roy, A. (2026). Discussion on "Robust probabilistic inference via a constrained transport metric". *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1026

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press. doi:10.1201/9781315117195. 1022

Wang, Y. (2026). Invited discussion. *Bayesian Analysis*. TBA, Number TBA, pp. 1–TBA. 1020, 1021, 1022, 1024, 1025