Bayesian Analysis (2026) **TBA**, Number TBA, pp. 1

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 [33]. 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, [13]) and empirical likelihood (EL, [24]), to KL-based exponentially tilted empirical likelihood (ETEL, [31]), to minimum

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

Wasserstein estimators [3]. 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, [14]), to doubly robust estimators that combine propensity score and outcome models, to more recent Wasserstein- and energy-distance-based distributional balancing [17, 30]. 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 [1], through synthetic likelihood [35, 27] 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.

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

	Moments or summaries	KL-related	Wasserstein or IPM
Robust inference (frequentist)	M-estimating equations; EL [24, 31]	Quasi-likelihood [34]	Distributionally robust optimization (DRO) [5]
Robust inference (Bayesian)	Gibbs posterior with general loss [15, 4]; Generalized posterior with estimating equations [8]	Coarsened posterior [20]; Safe Bayes [11]	Constrained transport posterior (this paper); MMD-based Gibbs posterior [7]
Semiparametric estimation	GMM [13]; Indirect inference [10]	Information criteria [16]	Minimum Wasserstein estimation [3]
Distribution balancing for causal inference	IPW [29]; Covariate balancing propensity score [14]	Entropy balancing [12]	IPM-based balancing [17]; Characteristic function distance [30]
Simulation-based inference	ABC with summary statistics [1]; Simulated method of moments [18]; Synthetic likelihood [35, 27];	Classifier-based likelihood ratio [9, 33]; Neural likelihood/ratio estimation (NLE/NRE) [25, 19]	Wasserstein ABC [2]; Sliced-Wasserstein inference; MMD-based ABC [26]

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. [32], 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 [28, 6] 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 $SW_p(P_n, P) = O_p(n^{-1/2})$ [23, 22], 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. [21].

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

[1] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate Bayesian computation in population genetics." *Genetics*, 162(4): 2025–2035. 2

[2] 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. 2

[3] — (2019). "On parameter estimation with the Wasserstein distance." *Information and Inference: A Journal of the IMA*, 8(4): 657–676. 2

[4] 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(5): 1103–1130. 2

[5] Blanchet, J., Murthy, K., and Zhang, F. (2022). "Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes." *Mathematics of Operations Research*, 47(2): 1500–1529. 2

[6] Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). "Sliced and Radon Wasserstein barycenters of measures." *Journal of Mathematical Imaging and Vision*, 51(1): 22–45. 3

[7] 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. 2

[8] Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation." *Journal of econometrics*, 115(2): 293–346. 2

[9] Cranmer, K., Pavez, J., and Louppe, G. (2015). "Approximating likelihood ratios with calibrated discriminative classifiers." *arXiv preprint arXiv:1506.02169*. 2

[10] Gourieroux, C., Monfort, A., and Renault, E. (1993). "Indirect inference." *Journal of applied econometrics*, 8(S1): S85–S118. 2

[11] Grünwald, P. and van Ommen, T. (2017). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it." *Bayesian Analysis*, 12(4): 1069–1103. 2

[12] Hainmueller, J. (2012). "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political analysis*, 20(1): 25–46. 2

[13] Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the econometric society*, 1029–1054. 1, 2

[14] Imai, K. and Ratkovic, M. (2014). "Covariate balancing propensity score." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 243–263. 2

[15] 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): 2207 – 2231. 2

[16] Kitamura, Y. and Stutzer, M. (1997). "An information-theoretic alternative to generalized method of moments estimation." *Econometrica: Journal of the Econometric Society*, 861–874. 2

[17] 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. 2

[18] McFadden, D. (1989). "A method of simulated moments for estimation of discrete response models without numerical integration." *Econometrica: Journal of the Econometric Society*, 995–1026. 2

[19] Miller, B. K., Weniger, C., and Forré, P. (2022). "Contrastive neural ratio estimation." In *Advances in Neural Information Processing Systems*, volume 35, 3262–3278. 2

[20] Miller, J. W. and Dunson, D. B. (2019). "Robust Bayesian inference via coarsening." *Journal of the American Statistical Association*. 2

[21] 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(2): 1006–1027. 4

[22] 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. 3

[23] 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. 3

[24] Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC. 1, 2

[25] 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. 2

[26] Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). "K2-ABC: Approximate Bayesian computation with kernel embeddings." In *Artificial intelligence and statistics*, 398–407. PMLR. 2

[27] Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). "Bayesian synthetic likelihood." *Journal of Computational and Graphical Statistics*, 27(1): 1–11. 2

[28] 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. 3

[29] Robins, J. M. and Rotnitzky, A. (1995). "Semiparametric efficiency in multivariate regression models with missing data." *Journal of the American Statistical Association*, 90(429): 122–129. 2

[30] Santra, D., Chen, G., and Park, C. (2026). "Distributional balancing for causal inference: A unified framework via characteristic function distance." *arXiv preprint arXiv:2601.15449*. 2

[31] Schennach, S. M. (2007). "Point estimation with exponentially tilted empirical likelihood." *The Annals of Statistics*, 35(2): 634–672. 1, 2

[32] Tomaselli, L., Ventura, V., and Wasserman, L. (2025). "Robust simulation based inference." *arXiv preprint arXiv:2508.02404*. 3

[33] Wang, Y., Kaji, T., and Rockova, V. (2022). "Approximate Bayesian computation via classification." *Journal of Machine Learning Research*, 23(350): 1–49. 1, 2

[34] Wedderburn, R. W. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method." *Biometrika*, 61(3): 439–447. 2

[35] Wood, S. N. (2010). "Statistical inference for noisy nonlinear ecological dynamic systems." *Nature*, 466(7310): 1102–1104. 2