

Regularized Exponentially Tilted Empirical Likelihood for Bayesian Inference

Eunseop Kim,^a Steven N. MacEachern,^b and Mario Peruggia^c

Abstract

Bayesian inference with empirical likelihood faces a challenge as the posterior domain is a proper subset of the original parameter space due to the convex hull constraint. We propose a regularized exponentially tilted empirical likelihood to address this issue. Our method removes the convex hull constraint using a novel regularization technique, incorporating a continuous exponential family distribution to satisfy a Kullback–Leibler divergence criterion. The regularization arises as a limiting procedure where pseudo-data are added to the formulation of exponentially tilted empirical likelihood in a structured fashion. We show that this regularized exponentially tilted empirical likelihood retains certain desirable asymptotic properties of (exponentially tilted) empirical likelihood and has improved finite sample performance. Simulation and data analysis demonstrate that the proposed method provides a suitable pseudo-likelihood for Bayesian inference. The implementation of our method is available as the R package **retel**. Supplementary materials for this article are available online.

Keywords: Bernstein–von Mises theorem; Convex hull; Entropy balancing; Kullback–Leibler divergence; Pseudo-data

1 Introduction

Statistical models defined through estimating equations and moment conditions allow semiparametric inferences on quantities of interest without distributional assumptions.

Empirical likelihood (EL) (Owen 1988, Qin & Lawless 1994), a popular approach in the

^aDepartment of Statistics, The Ohio State University, Columbus, OH. Corresponding author.

^bDepartment of Statistics, The Ohio State University, Columbus, OH.

^cDepartment of Statistics, The Ohio State University, Columbus, OH.

frequentist setting, enables nonparametric but still likelihood style inference. It shares many desirable properties with parametric likelihood, exhibiting Wilks’ phenomenon under mild conditions and allowing for Bartlett correction (DiCiccio et al. 1991). Moreover, confidence regions from EL have data-driven shapes and orientations.

EL is a member of the class of generalized empirical likelihoods (GEL) (Smith 1997, Newey & Smith 2004), which includes the exponential tilting of Efron (1981). Newey & Smith (2004) showed a duality between GEL and the class of minimum discrepancy methods (Cressie & Read 1984, Corcoran 1998). In this context, EL is formulated by finding a distribution supported on the sample that minimizes the Kullback–Leibler (KL) divergence to the empirical distribution, subject to moment constraints. Exponentially tilted empirical likelihood (ETEL) (Efron 1981, Jing & Wood 1996, Schennach 2005) is obtained by combining exponential tilting and EL, which minimizes the reverse KL divergence.

Bayesian analysis of EL poses a challenge, as posterior inference via Bayes’ Theorem requires a complete specification of the sampling distribution or the likelihood function. Lazar (2003) proposed using EL as a replacement for the likelihood function in Bayesian inference. Through simulation, she showed that the EL-posterior distributions can exhibit strong similarities to traditional posterior distributions. Schennach (2005) strengthened the case for these methods by showing that ETEL arises as the limit of nonparametric Bayesian procedures with a particular type of prior favoring entropy-maximizing distributions. Chib et al. (2018) established a Bernstein–von Mises theorem for the Bayesian ETEL-posterior distribution, ensuring that the frequentist coverage of credible sets is asymptotically correct. Similar asymptotic results for EL were established by Sueishi (2022).

However, both EL and ETEL have an inherent limitation in that they are only defined on a proper subset of the original parameter space due to the convex hull constraint or empty

set problem (Grendár & Judge 2009). By convention, the likelihoods are set to zero for parameter values that violate the convex hull constraint. For Bayesian inference, the zeroes in the likelihood imply a restricted posterior domain. This is conceptually unsatisfactory as, with a larger sample size, the convex hull may expand and the likelihood become positive. Additionally, as the restricted domain is often non-convex, (Chaudhuri et al. 2017), a more sophisticated posterior sampling scheme may be needed to fit the model.

To address these issues, various adjustments to EL have been suggested (Bartolucci 2007, Chen et al. 2008, Tsao & Wu 2013). Most relevant to our work, Chen et al. (2008) proposed the adjusted empirical likelihood (AEL), which adds a pseudo-observation in a way that satisfies the convex hull constraint for any given parameter value. This approach has been further developed by Emerson & Owen (2009) and Liu & Chen (2010), and it has been adapted for ETEL by Zhu et al. (2009) as the adjusted exponentially tilted empirical likelihood (AETEL).

In this paper, we propose a method to address the convex hull constraint for Bayesian ETEL. While previous proposals have primarily focused on EL and frequentist inference, our proposal builds upon the AEL framework, introducing notable distinctions. First, we extend the method to accommodate fractional observations, following the approach of (Hainmueller 2012). Second, we allow for the incorporation of multiple pseudo observations. Third, we pass to the limit, ensuring that the convex hull constraint is satisfied for all parameter values simultaneously. This resulting formulation naturally induces a form of regularization that removes the constraint. Our method’s main contributions encompass: (i) addressing the convex hull constraint for ETEL while retaining desirable asymptotic properties; (ii) enhancing stability and robustness of small-sample performance compared to existing methods; (iii) providing flexibility in Bayesian modelling and allowing one to

incorporate a novel form of prior information.

This paper is organized as follows. In [Section 2](#), we introduce the notation used in the paper and provide a brief overview of ETEL. Then, we propose a weighted version of ETEL that incorporates fractional pseudo-data with the maximum entropy reweighting scheme. In [Section 3](#), we propose inducing regularization on the formulation of ETEL, exploring two equivalent approaches: (i) a limiting procedure with fractional pseudo-data and (ii) direct incorporation of a continuous exponential family distribution in the minimization of the KL divergence. We derive asymptotic properties of the proposed methods. In [Section 4](#), we evaluate the performance of the methods through simulation studies. In [Section 5](#), we present an application to the estimation of median income for four-person families. Finally, we conclude with a discussion of directions for future research in [Section 6](#). The proofs of the theoretical results are provided in the supplementary materials. The implementation is available as the R package **retel**, along with the simulation and data analysis code, in a GitHub repository at <https://github.com/markean/retel>.

2 Weighted Exponentially Tilted Empirical Likelihood with Fractional Pseudo-Data

We begin by introducing ETEL, along with the setup and some notation. Let $\mathcal{D}_n = \{\mathbf{X}_i\}_{i=1}^n$ denote independent d_x -dimensional observations from a complete probability space $(\mathcal{X}, \mathcal{F}, P)$ satisfying the moment condition: $E_P[\mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta})] = \mathbf{0}$, where $\mathbf{g} : \mathbb{R}^{d_x} \times \Theta \mapsto \mathbb{R}^p$ is an estimating function with the true parameter value $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$. Consider a discrete probability distribution P_0 that is absolutely continuous with respect to the empirical distribution P_n . The KL divergence from P_n to P_0 is $D_{KL}(P_0 \parallel P_n) = \sum_{i=1}^n p_i \log(np_i)$,

where p_i are probabilities attached to the observations by P_0 . By minimizing the KL divergence subject to the constraints in the moment condition, we obtain a unique set of p_i and the associated distribution. For a given $\boldsymbol{\theta}$, the maximization problem

$$\max_{p_1, \dots, p_n} \left\{ \sum_{i=1}^n (-p_i \log(n p_i)) \mid \sum_{i=1}^n p_i \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}) = \mathbf{0}, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1 \right\}$$

yields a unique solution $(p_1(\boldsymbol{\theta}), \dots, p_n(\boldsymbol{\theta}))$, and ETEL is defined as $L_{ET}(\boldsymbol{\theta}) = \prod_{i=1}^n p_i(\boldsymbol{\theta})$.

By applying the method of Lagrange multipliers, we obtain

$$p_i(\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\lambda}_{ET}^\top \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}))}{\sum_{j=1}^n \exp(\boldsymbol{\lambda}_{ET}^\top \mathbf{g}(\mathbf{X}_j, \boldsymbol{\theta}))},$$

where $\boldsymbol{\lambda}_{ET} \equiv \boldsymbol{\lambda}_{ET}(\boldsymbol{\theta})$ solves the equation $n^{-1} \sum_{i=1}^n \exp(\boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta})) \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}) = \mathbf{0}$. The dual problem provides the solution: $\boldsymbol{\lambda}_{ET} = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} \sum_{i=1}^n \exp(\boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}))$. By construction, an M -estimator $\hat{\boldsymbol{\theta}}$ that solves $n^{-1} \sum_{i=1}^n \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}) = \mathbf{0}$ maximizes ETEL (Yiu et al. 2020).

In the Bayesian framework, ETEL can be used with a prior $\pi(\boldsymbol{\theta})$ to define the ETEL-posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D}_n) \propto \pi(\boldsymbol{\theta}) L_{ET}(\boldsymbol{\theta})$. Schennach (2005) showed that when all observations are distinct, $L_{ET}(\boldsymbol{\theta})$ can be obtained as the limit of a nonparametric Bayesian procedure. Her procedure involves assigning a mixture of uniform densities as a nonparametric prior on P that satisfies the moment condition and then marginalizing over the nuisance parameters. The convex hull constraint serves as the implicit constraint in the primal optimization problem, indicating that the interior of the convex hull of $\{\mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta})\}_{i=1}^n$, denoted by $\operatorname{Conv}_n(\boldsymbol{\theta})$, must contain $\mathbf{0}$. Consequently, the (posterior) domain of ETEL is restricted to $\Theta_n = \{\boldsymbol{\theta} \in \Theta : \mathbf{0} \in \operatorname{Conv}_n(\boldsymbol{\theta})\}$ so that even a 100% credible set may fail to contain $\boldsymbol{\theta}_0$. In general, Θ_n is nonconvex and is challenging to identify. Simulation methods to fit the models, such as Markov chain Monte Carlo or Hamiltonian Monte Carlo, require

long runs and may or may not be effective (Chaudhuri et al. 2017, Yu & Bondell 2023), leading to potential undercoverage issues and unreliable inference.

To address the convex hull constraint for EL, the AEL approach introduces a pseudo-observation that depends on $\boldsymbol{\theta}$. Here and throughout, we use $\mathbf{g}_i(\boldsymbol{\theta}) = \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta})$, $i = 1, \dots, n$, for notational convenience. The pseudo-observation has

$$\mathbf{g}_{n+1}(\boldsymbol{\theta}) = -\frac{a_n}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}), \quad (1)$$

where $a_n > 0$. Properties of the sequence a_n are used to establish asymptotic results. The addition of $\mathbf{g}_{n+1}(\boldsymbol{\theta})$ ensures that the convex hull constraint is satisfied for each $\boldsymbol{\theta} \in \Theta$.

Emerson & Owen (2009) and Liu & Chen (2010) proposed adding two pseudo-observations to improve the coverage accuracy of confidence regions obtained from AEL. Yu & Bondell (2023) established a Bernstein–von Mises theorem for Bayesian AEL. While the AEL approach is directly applicable to ETEL for fixing the convex hull constraint for a particular $\boldsymbol{\theta}$, it may introduce irregularities throughout Θ in the resulting posterior distribution when applied to Bayesian analysis, since it involves a preliminary entropy maximization step in constructing the likelihood function. Incorporating one or two pseudo-observations, specific to each $\boldsymbol{\theta}$, and treating them on par with actual observations may contribute to these irregularities.

As an initial step towards addressing the convex hull constraint for ETEL and establishing a connection with the regularization method discussed in Section 3, we propose a weighted exponentially tilted empirical likelihood (WETEL) approach with fractional pseudo-data. Our approach extends the AEL method by incorporating multiple pseudo-observations, in combination with the entropy balancing scheme of Hainmueller (2012). Entropy balancing is a data preprocessing technique used to achieve covariate balance in observational studies

with a binary treatment and in survey sampling. The preprocessing step involves applying a maximum-entropy reweighting scheme to ensure that the reweighted data satisfy a set of moment conditions. In the context of our framework, the pseudo-data can be seen as providing additional information for the analysis.

We introduce a fixed number, $m \in \mathbb{N}$, of pseudo-data denoted as $\mathbf{g}_{n+j}(\boldsymbol{\theta}) \in \mathbb{R}^p$ for $j = 1, \dots, m$. The use of the estimating function \mathbf{g} for the pseudo-data is for notational consistency. Apart from their dependence on $\boldsymbol{\theta}$, they need not necessarily be related to the observed data or estimating function. At this stage, we do not discuss any specific strategy for creating the pseudo-data. Instead, for our current purposes, we simply assume that the augmented data, comprising both the observed data and pseudo-data, satisfy the convex hull constraint.

Let w_i be the base weight for the i th observation in the augmented data, such that $\sum_{i=1}^N w_i = 1$, with $N = n + m$. We consider the following maximum-entropy reweighting scheme:

$$\max_{p_1, \dots, p_N} \left\{ \sum_{i=1}^N \left(-p_i \log \left(\frac{p_i}{w_i} \right) \right) \left| \sum_{i=1}^N p_i \mathbf{g}_i(\boldsymbol{\theta}) = \mathbf{0}, \quad p_i \geq 0, \quad \sum_{i=1}^N p_i = 1 \right. \right\}.$$

This scheme is equivalent to minimizing $D_{KL}(P_0 \parallel P_w)$ subject to the constraints above, where P_w is the weighted empirical distribution. Both P_0 and P_w are now supported on the augmented data. The objective function is modified to account for the weights and pseudo-data, and the moment condition is matched by the augmented data. The method of Lagrange multipliers yields

$$p_i(\boldsymbol{\theta}) = \frac{w_i \exp(\boldsymbol{\lambda}_{WET}^\top \mathbf{g}_i(\boldsymbol{\theta}))}{\sum_{j=1}^N w_j \exp(\boldsymbol{\lambda}_{WET}^\top \mathbf{g}_j(\boldsymbol{\theta}))},$$

where $\boldsymbol{\lambda}_{WET} = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} \sum_{i=1}^N w_i \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta}))$.

Next, building upon the weighted EL approach proposed by [Glenn & Zhao \(2007\)](#), we formulate the likelihood function as $L_{WET}(\boldsymbol{\theta}) = \prod_{i=1}^N p_i(\boldsymbol{\theta})^{Nw_i}$. Based on the inequality $\prod_{i=1}^N p_i(\boldsymbol{\theta})^{Nw_i} \leq \prod_{i=1}^N w_i^{Nw_i}$ for any solution $p_i(\boldsymbol{\theta})$, the likelihood ratio function of WETEL can be defined as $R_{WET}(\boldsymbol{\theta}) = \prod_{i=1}^N (p_i(\boldsymbol{\theta})/w_i)^{Nw_i}$. Consequently, the maximum WETEL estimator $\hat{\boldsymbol{\theta}}_w$ is obtained by solving the equation $\sum_{i=1}^N w_i \mathbf{g}_i(\boldsymbol{\theta}) = \mathbf{0}$. When using uniform weights with $w_i = 1/N$, the resulting WETEL reduces to ETEL with the pseudo-data included. However, in finite sample settings, the size of m relative to n , the pseudo-data specification, and the choice of weights can lead to substantial differences between WETEL and ETEL. To prevent this, we treat all pseudo-data as a single observation and assign fractional weights to them. Specifically, we set the weights as follows:

$$w_i = \begin{cases} \frac{1}{n+1} & (i = 1, \dots, n), \\ \frac{1}{m(n+1)} & (i = n+1, \dots, n+m). \end{cases} \quad (2)$$

This weight specification balances the contribution from the pseudo-data with the modified multiplier: $\boldsymbol{\lambda}_{WET} = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} \{\sum_{i=1}^n \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta})) + m^{-1} \sum_{i=n+1}^{n+m} \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta}))\}$.

Since WETEL is a generalization of ETEL with finite pseudo-data, it preserves the major asymptotic properties of ETEL as $n \rightarrow \infty$. Let $\mathbf{G} = E_P[\partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta}_0)]$, $\mathbf{V} = E_P[\mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)^\top]$, and $\boldsymbol{\Omega} = (\mathbf{G}^\top \mathbf{V}^{-1} \mathbf{G})^{-1}$, where $\partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta}_0)$ denotes the Jacobian matrix of $\mathbf{g}_i(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$. Moreover, the Euclidean norm for vectors is denoted by $|\cdot|$, and the Frobenius norm for matrices is denoted by $\|\cdot\|$. We also use $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to represent a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and χ_p^2 to represent a chi-square distribution with p degrees of freedom. We present the following technical conditions required to establish theoretical results.

Condition 1. The parameter space Θ is compact, with $\boldsymbol{\theta}_0$ an interior point of Θ and the unique solution to $E_P[\mathbf{g}_i(\boldsymbol{\theta})] = \mathbf{0}$.

Condition 2. With probability 1, $\mathbf{g}_i(\boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \Theta$, continuously differentiable in a neighborhood \mathcal{N} of $\boldsymbol{\theta}_0$, and $E_P[\sup_{\boldsymbol{\theta} \in \mathcal{N}} \|\partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta})\|] < \infty$.

Condition 3. $\text{rank}(\mathbf{G}) = \text{rank}(\mathbf{V}) = p$.

Condition 4. For some $\alpha > 3$, $E_P[\sup_{\boldsymbol{\theta} \in \Theta} |\mathbf{g}_i(\boldsymbol{\theta})|^\alpha] < \infty$.

These conditions are standard regularity conditions used to study the asymptotic behavior of GEL; see, for example, [Newey & Smith \(2004\)](#). We establish that the discrepancies between ETEL and WETEL, in terms of estimators and Lagrange multipliers, become asymptotically negligible.

Proposition 1. *Under [Conditions 1–4](#), $\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}} = o_P(n^{-1/2})$ and $\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) - \boldsymbol{\lambda}_{ET}(\boldsymbol{\theta}_0) = O_P(n^{-1})$.*

Consequently, WETEL shares with GEL first-order asymptotic properties.

Theorem 1. *Under [Conditions 1–4](#), $n^{1/2}(\widehat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_0)$ converges in distribution to $N(\mathbf{0}, \boldsymbol{\Omega})$ as $n \rightarrow \infty$, and $-2 \log R_{WET}(\boldsymbol{\theta}_0)$ converges in distribution to χ_p^2 as $n \rightarrow \infty$.*

3 Regularized Exponentially Tilted Empirical Likelihood

In [Section 2](#), we did not explicitly discuss the specification of pseudo-data for WETEL. When m is fixed, the convex hull constraint issue may arise in WETEL, where $\mathbf{0} \notin \text{Conv}_N(\boldsymbol{\theta})$ for certain values of $\boldsymbol{\theta}$, unless the pseudo-data are carefully specified. Even if we adopt a

strategy like the one in Equation (1), the limitation remains because the specification of pseudo-data, regardless of careful selection or the magnitude of m , still needs to depend on the observed data and parameter values. In this sense, the pseudo-data approach can be viewed as an ad-hoc solution that pragmatically addresses the issue but does not fully resolve the underlying challenge associated with a finite m .

In this section, we consider a procedure where m tends to infinity, enabling the pseudo-data to represent a continuous distribution in the limit. Since ETEL induces an exponential family of distributions supported on the data, a natural choice for the pseudo-data is a continuous exponential family distribution. To accomplish this, we introduce an auxiliary random variable $\tilde{\mathbf{g}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is assumed to be of full rank. The pseudo-data $\{\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \dots\}$ may be selected as appropriate quantiles of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, aiming to approximate the distribution as m increases. For the purposes of our discussion, we assume that the pseudo-data are independent samples from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, while treating the sample size n and the parameter $\boldsymbol{\theta}$ as fixed.

Using the fractional weights in Equation (2), we introduce a sequence of stochastic minimization problems for WETEL: $\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} c_m(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \in \mathbb{R}^p} \{d_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) + p_m(\boldsymbol{\lambda})\}$ for $m = 1, 2, \dots$, where $d_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{i=1}^n \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta}))$ and $p_m(\boldsymbol{\lambda}) = m^{-1} \sum_{j=1}^m \exp(\boldsymbol{\lambda}^\top \tilde{\mathbf{g}}_j)$. It follows from the independent sampling that $p_m(\boldsymbol{\lambda}) \rightarrow p(\boldsymbol{\lambda})$ with probability 1 as $m \rightarrow \infty$, where $p(\boldsymbol{\lambda}) = \exp(\boldsymbol{\lambda}^\top \boldsymbol{\mu} + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda} / 2)$ is the moment-generating function of $\tilde{\mathbf{g}}$. This suggests directly considering the following minimization problem:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} c(\boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \in \mathbb{R}^p} \{d_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) + p(\boldsymbol{\lambda})\}, \quad (3)$$

with the minimization performed after taking the limit. Then, the sequence of minimization problems can be viewed as a discretization of the population version of the minimization

problem. Such a setting can be commonly found in applications of stochastic programming (Wets 1974, Dupačová 1992), equipped with epi-convergence (Dupačová & Wets 1988, King & Wets 1991, Rockafellar & Wets 2009). We refer to this method as regularized exponentially tilted empirical likelihood (RETEL) and introduce the corresponding multiplier $\boldsymbol{\lambda}_{RET} = \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} \{d_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) + p(\boldsymbol{\lambda})\}$.

From the convexity and lower semicontinuity of $\exp(\boldsymbol{\lambda}^\top \widetilde{\boldsymbol{g}})$, it is shown that $p_m(\boldsymbol{\lambda})$ epi-converges to $p(\boldsymbol{\lambda})$ as $m \rightarrow \infty$ with probability 1 (see, for example, Artstein & Wets 1995, Theorem 2.3). This establishes the consistency of the minimizers with the following intermediate result:

Proposition 2. *Under Condition 2, with probability 1, the minimization problem in Equation (3) has a unique global minimizer $\boldsymbol{\lambda}_{RET}$ for each $\boldsymbol{\theta} \in \Theta$. Additionally, for any $\epsilon_m \downarrow 0$, we have $\lim_{m \rightarrow \infty} \{\epsilon_m\text{-argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} c_m(\boldsymbol{\lambda})\} = \{\boldsymbol{\lambda}_{RET}\}$, where $\{\epsilon_m\text{-argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} c_m(\boldsymbol{\lambda})\} = \{\boldsymbol{\lambda} \mid c_m(\boldsymbol{\lambda}) \leq \inf_{\boldsymbol{\lambda} \in \mathbb{R}^p} c_m(\boldsymbol{\lambda}) + \epsilon_m\}$.*

With probability 1, $\boldsymbol{\lambda}_{RET}$ is a limit point of the sequence of approximate solutions to the minimization problems. For any finite m , $\operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}^p} c_m(\boldsymbol{\lambda})$ may not exist with positive probability. However, the existence and uniqueness of $\boldsymbol{\lambda}_{RET}$ are guaranteed by the strict convexity of $p(\boldsymbol{\lambda})$, which acts as a penalty that regularizes $\boldsymbol{\lambda}$ and prevents $|\boldsymbol{\lambda}|$ from diverging, regardless of whether $\mathbf{0} \in \operatorname{Conv}_n(\boldsymbol{\theta})$. Figure 1 shows an example where λ_{WET} converges to λ_{RET} as a sequence of pseudo-data approximates a normal distribution.

The choice of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in $p(\boldsymbol{\lambda})$ depends on the requirements of a specific application, and each choice uniquely determines the shape and curvature of $p(\boldsymbol{\lambda})$. One simple option is to set $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, where \boldsymbol{I}_p denotes the $p \times p$ identity matrix. More generally, we can consider

$$p_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \tau_n \exp \left(\boldsymbol{\lambda}^\top \boldsymbol{\mu}_{n,\boldsymbol{\theta}} + \frac{1}{2} \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} \boldsymbol{\lambda} \right), \quad (4)$$

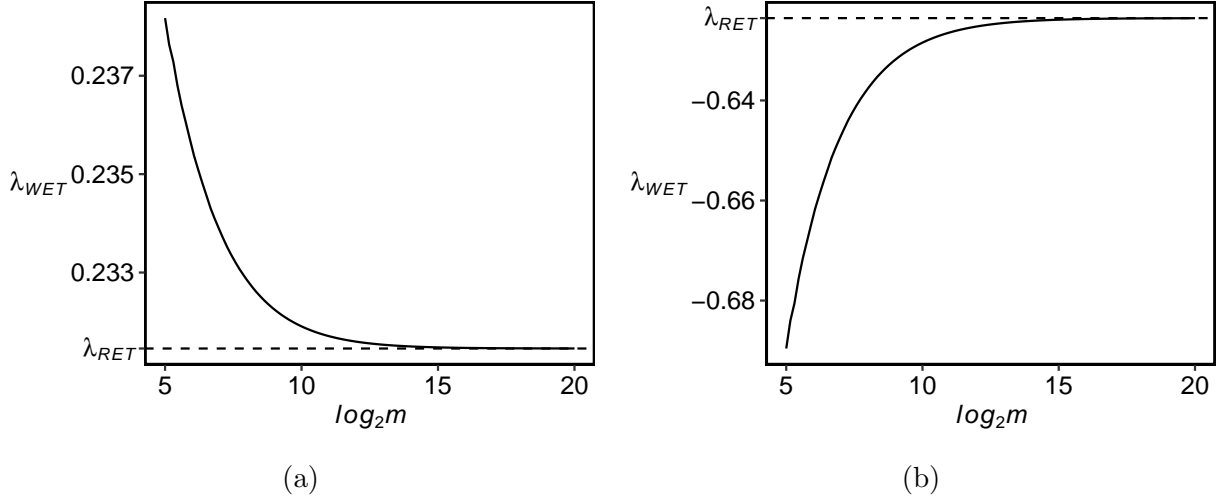


Figure 1. Plots of $\lambda_{WET}(\theta)$ versus $\log_2 m$ for the mean parameter θ . With two observations -2 and 2 fixed, the convex hull constraint is satisfied at $\theta = 1$ in (a) and violated at $\theta = 3$ in (b). For each m , the pseudo-data are generated as the $k/(m+1)$ quantile of the standard normal distribution for $k = 1, \dots, m$. When the convex hull constraint is satisfied, λ_{WET} converges faster to the respective λ_{RET} (horizontal dashed lines).

and the corresponding minimization problem in Equation (3) becomes:

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} c_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda} \in \mathbb{R}^p} \{d_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) + p_n(\boldsymbol{\theta}, \boldsymbol{\lambda})\}, \quad (5)$$

with the solution still denoted by $\boldsymbol{\lambda}_{RET}$. Here, $\tau_n > 0$ is a tuning parameter that controls the strength of $p(\boldsymbol{\lambda})$ as a penalty. The parameters $\boldsymbol{\mu}_{n,\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}}$, which may vary with n and $\boldsymbol{\theta}$, can be drawn from prior information, allowing for more flexibility in the regularization.

Note that the description of the regularization suppresses an implicit connection to $\boldsymbol{\theta}$. For example, when considering the mean parameter $\boldsymbol{\theta}$, setting $\boldsymbol{\mu}_{n,\boldsymbol{\theta}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} = \mathbf{I}_p$ corresponds to assuming a latent normal distribution $N(\boldsymbol{\theta}, \mathbf{I}_p)$ at each $\boldsymbol{\theta}$. On the other hand, changing to $\boldsymbol{\mu}_{n,\boldsymbol{\theta}} = \bar{\mathbf{X}} - \boldsymbol{\theta}$ introduces $N(\bar{\mathbf{X}}, \mathbf{I}_p)$ centered at the sufficient statistic $\bar{\mathbf{X}}$, making the regularization invariant with respect to $\boldsymbol{\theta}$. In this case, the two choices will lead to considerably different $\boldsymbol{\lambda}_{RET}$ for $\boldsymbol{\theta}$ lying outside the convex hull of the observed data.

From an operational perspective, any function $p(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}_{>0}$ that increases superlin-

early with $|\boldsymbol{\lambda}|$ can be considered to ensure a finite $\boldsymbol{\lambda}_{RET}$. This penalty method can also be extended to other GEL methods that share the Cressie–Read family of discrepancies. However, we focus on ETEL due to its connection to the exponential family it generates (Yiu et al. 2020) and to the auxiliary continuous exponential family distribution that is naturally introduced.

In the following, we present an alternative approach to formulating the minimization problem in Equation (5). This approach does not involve the concept of a sequence of procedures with pseudo-data but instead directly considers a mixture of a normal and a multinomial distribution supported on the data. For a given $\boldsymbol{\lambda}$, we apply exponential tilting to the $N(\boldsymbol{\mu}_{n,\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}})$ distribution of $\tilde{\mathbf{g}}$, resulting in the $\boldsymbol{\lambda}$ -tilted distribution $N(\boldsymbol{\mu}_{n,\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}}\boldsymbol{\lambda}, \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}})$. We denote the corresponding random variable as $\tilde{\mathbf{g}}_{\boldsymbol{\lambda}}$. To formulate the problem, we consider two probability distributions:

$$\tilde{P}_n = \frac{n}{n + \tau_n} P_n + \frac{\tau_n}{n + \tau_n} N(\boldsymbol{\mu}_{n,\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}}), \quad \tilde{P}_{\boldsymbol{\lambda}} = (1 - p_c) P_0 + p_c N(\boldsymbol{\mu}_{n,\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}}\boldsymbol{\lambda}, \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}}),$$

where each distribution is defined as a convex mixture of a discrete and a continuous distribution. The constant p_c in $\tilde{P}_{\boldsymbol{\lambda}}$ represents the probability assigned to the tilted distribution. The following result parallels the idea that $D_{KL}(P_0 \parallel P_n)$ is minimized by ETEL.

Proposition 3. *For any $\boldsymbol{\theta} \in \Theta$, the minimization problem in Equation (5) is the dual problem of minimizing $D_{KL}(\tilde{P}_{\boldsymbol{\lambda}} \parallel \tilde{P}_n)$ with respect to p_i , $i = 1, \dots, n$, and p_c , subject to $\sum_{i=1}^n p_i \mathbf{g}_i(\boldsymbol{\theta}) + p_c \mathbb{E}_{\tilde{P}_{\boldsymbol{\lambda}}}[\tilde{\mathbf{g}}_{\boldsymbol{\lambda}}] = \mathbf{0}$, $p_c \geq 0$, $p_i \geq 0$, and $\sum_{i=1}^n p_i + p_c = 1$.*

As a consequence, the optimal values of $p_i(\boldsymbol{\theta})$ and $p_c(\boldsymbol{\theta})$ can be expressed as follows:

$$p_i(\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\lambda}_{RET}^\top \mathbf{g}_i(\boldsymbol{\theta}))}{c_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})} \quad (i = 1, \dots, n), \quad p_c(\boldsymbol{\theta}) = \frac{p_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})}{c_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})},$$

where $\boldsymbol{\lambda}_{RET}$ is the solution to the equation:

$$\sum_{i=1}^n \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta})) \mathbf{g}_i(\boldsymbol{\theta}) + p_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) (\boldsymbol{\mu}_{n,\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} \boldsymbol{\lambda}) = \mathbf{0}. \quad (6)$$

The formulations presented above suggest the possibility of using other exponential family distributions without modification. However, in this context, we proceed with normal distributions since the focus is to expand the domain to the entire parameter space for any estimating function. Furthermore, the normal distribution has the unique property of being the maximum entropy distribution among all distributions with a given mean and covariance (Cover & Thomas 2006, Theorem 8.6.5).

Once we have determined $\boldsymbol{\lambda}_{RET}$, we define the likelihood and likelihood ratio functions as follows:

$$L_{RET}(\boldsymbol{\theta}) = p_c(\boldsymbol{\theta}) \prod_{i=1}^n p_i(\boldsymbol{\theta}), \quad R_{RET}(\boldsymbol{\theta}) = \left(\frac{n + \tau_n}{\tau_n} p_c(\boldsymbol{\theta}) \right) \prod_{i=1}^n (n + \tau_n) p_i(\boldsymbol{\theta}). \quad (7)$$

RETEL differs from penalty approaches for EL (Tang & Leng 2010, Leng & Tang 2012, Chang et al. 2018), where a penalty term is added to the empirical log-likelihood ratio to induce sparsity in the solution $\hat{\boldsymbol{\theta}}$. Instead, RETEL aims to regularize the behavior of the multiplier $\boldsymbol{\lambda}$ before computing the likelihood. With $\boldsymbol{\lambda}$ having a concrete interpretation as a tilting parameter in minimizing the KL divergence, RETEL is also distinct from the penalized EL approach of Bartolucci (2007). It is worth noting that RETEL shares some

connection with hybrid approaches that combine EL with a parametric likelihood (Qin 1994, Hjort et al. 2018). However, instead of directly multiplying ETEL by a parametric likelihood function, RETEL takes a more indirect approach by employing $p_c(\boldsymbol{\theta})$, which captures the effect from the assumed normal distribution.

To make RETEL more closely reflect the observed data, we can drop $p_c(\boldsymbol{\theta})$ from Equation (7) and define another version of RETEL with the following functions:

$$\tilde{L}_{RET}(\boldsymbol{\theta}) = \prod_{i=1}^n p_i(\boldsymbol{\theta}), \quad \tilde{R}_{RET}(\boldsymbol{\theta}) = \prod_{i=1}^n (n + \tau_n) p_i(\boldsymbol{\theta}). \quad (8)$$

Dropping $p_c(\boldsymbol{\theta})$ does not mean reverting to the original ETEL since $p_c(\boldsymbol{\theta})$ affects the other $p_i(\boldsymbol{\theta})$ such that $\sum_{i=1}^n p_i(\boldsymbol{\theta}) + p_c(\boldsymbol{\theta}) = 1$. The impact of $p_c(\boldsymbol{\theta})$ and the underlying normal distribution remains embedded in the procedure and cannot be entirely removed, although τ_n can control the degree of this effect. A larger value of τ_n assigns more probability to $p_c(\boldsymbol{\theta})$ relative to the other $p_i(\boldsymbol{\theta})$, resulting in a greater reliance on the $\boldsymbol{\lambda}$ -tilted distribution for inference. We distinguish between the two versions by using $RETEL_f$ and $RETEL_r$ to refer to the approaches using Equations (7) and (8), respectively.

To ensure that the same M -estimator $\hat{\boldsymbol{\theta}}$ of ETEL also maximizes RETEL, it is desirable to formulate RETEL in a way that preserves this property. This can be achieved by setting $\boldsymbol{\mu}_{n,\boldsymbol{\theta}} = \mathbf{0}$ or $\boldsymbol{\mu}_{n,\boldsymbol{\theta}} = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta})$ in Equation (6), which leads to $\boldsymbol{\lambda}_{RET}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and $R_{RET}(\hat{\boldsymbol{\theta}}) = \tilde{R}_{RET}(\hat{\boldsymbol{\theta}}) = 1$. This property of RETEL, where the M -estimator is naturally preserved, distinguishes it from WETEL and other methods that add finite pseudo-data. Figure 2 illustrates, with a single observation, the difference between $\log R_{RET}(\boldsymbol{\theta})$ and $\log \tilde{R}_{RET}(\boldsymbol{\theta})$ as τ_n increases.

Now, we establish that RETEL retains certain desirable asymptotic properties of EL and ETEL. We consider RETEL obtained from $\boldsymbol{\lambda}_{RET}$ in Equation (5). The following condition

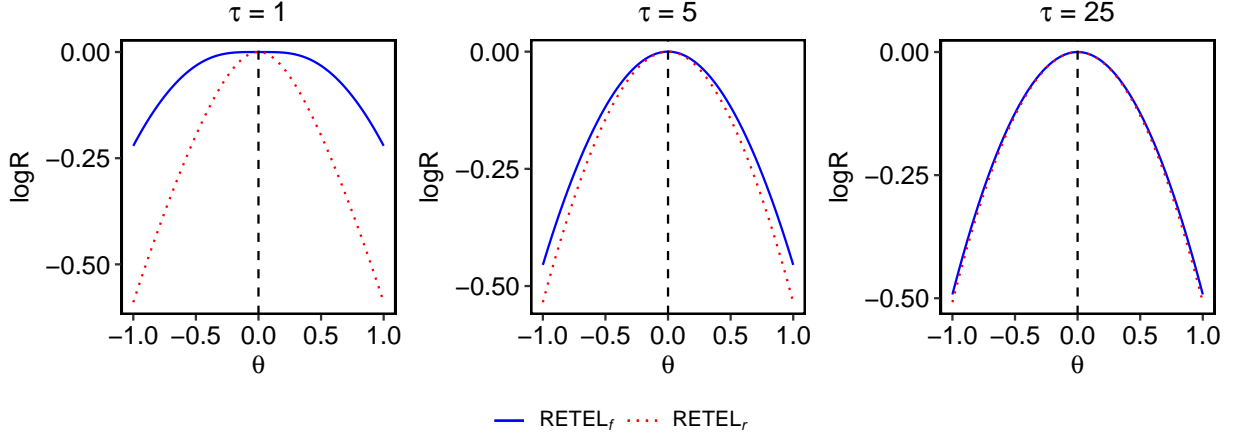


Figure 2. Plots of $\log R_{RET}(\boldsymbol{\theta})$ (solid blue lines) and $\log \tilde{R}_{RET}(\boldsymbol{\theta})$ (dashed red lines) for the mean parameter with varying $\tau_n \in \{1, 5, 25\}$. Both versions of RETEL achieve their maximum at the single data point 0 (vertical dashed line). Here, $\mu_{n,\theta}$ and $\Sigma_{n,\theta}$ are set to $-\theta$ and 1, respectively. The difference between the two versions diminishes as τ_n increases.

controls $p_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ in Equation (4). The condition ensures the asymptotic stability of the regularization when it depends on n and $\boldsymbol{\theta}$:

Condition 5. $\tau_n = O(\log n)$; $\boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} = \boldsymbol{\mu} + o_P(1)$ for some $\boldsymbol{\mu} \in \mathbb{R}^p$; $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0}$ is positive definite for any n with probability 1; and $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0} = \boldsymbol{\Sigma} + o_P(1)$ for some $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$.

Theorem 2. Under Conditions 1–5, $\log(R_{RET}(\boldsymbol{\theta}_0)/\tilde{R}_{RET}(\boldsymbol{\theta}_0)) = O_P(n^{-1/2})$. Additionally, both $-2 \log R_{RET}(\boldsymbol{\theta}_0)$ and $-2 \log \tilde{R}_{RET}(\boldsymbol{\theta}_0)$ converge in distribution to χ_p^2 .

As a consequence, the logarithms of the regularized methods are identical up to $O_P(n^{-1/2})$, and both methods exhibit Wilks' theorem. For Bayesian inference, we can obtain the Bernstein–von Mises result for both versions of RETEL.

Condition 6. The prior measure admits a density with respect to the Lebesgue measure. The density $\pi(\cdot)$ is continuous in Θ and is positive in a neighborhood of $\boldsymbol{\theta}_0$.

Condition 7. For any $\delta > 0$, there exists $\epsilon > 0$ such that

$$P \left(\sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \delta} \frac{1}{n} (\log L_{RET}(\boldsymbol{\theta}) - \log L_{RET}(\boldsymbol{\theta}_0)) \leq -\epsilon \right) \rightarrow 1.$$

Condition 6 and Condition 7 are regularity conditions to establish the Bernstein–von Mises theorem for EL and ETEL (Chib et al. 2018, Yu & Bondell 2023).

Theorem 3. *Under Conditions 1–7,*

$$\sup_{\mathcal{B}} |\pi(n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \mathcal{B} \mid \mathcal{D}_n) - N(\mathbf{0}, \boldsymbol{\Omega})(\mathcal{B})| \rightarrow 0$$

in probability, where $\pi(n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \mid \mathcal{D}_n)$ is the posterior distribution of $n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ obtained from RETEL, and $\mathcal{B} \in \Theta$ denotes any Borel set.

This result implies that, when the moment constraints are correctly specified, the total variation distance between the posterior distribution of $n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and $N(\mathbf{0}, \boldsymbol{\Omega})$ tends to zero in probability.

4 Simulation

4.1 Posterior Coverage

Monahan & Boos (1992) proposed examining the validity of a pseudo-likelihood $L(\theta)$ based on the coverage probabilities of posterior intervals. For a parameter $\theta \in \mathbb{R}$, let $\pi(\theta \mid x)$ be the posterior density obtained using $L(\theta)$ with an absolutely continuous prior density $\pi(\theta)$ and observed data x . For this pseudo-likelihood to be valid by coverage, posterior intervals should provide correct coverage probabilities. In particular, when (X, θ) is generated from the Bayesian model, the random variable $H = \int_{-\infty}^{\theta} \pi(t \mid X) dt$ should follow a uniform distribution $U(0, 1)$. This approach has been adopted for EL by Lazar (2003) and Cheng & Zhao (2019).

To investigate the validity of RETEL for Bayesian inference, we begin by simulating

Table 1. p -values from the Kolmogorov–Smirnov test for uniformity.

n	s	$\tau_n = 1$		$\tau_n = \log n$		ETEL	AETEL
		RETEL _f	RETEL _r	RETEL _f	RETEL _r		
5	1	< 0.001	0	< 0.001	< 0.001	0	0
	5	< 0.001	< 0.001	< 0.001	< 0.001	0	0
20	1	0.018	< 0.001	0.071	< 0.001	< 0.001	0.007
	5	0.038	< 0.001	0.098	< 0.001	< 0.001	0.064
50	1	0.612	0.248	0.655	0.303	0.221	0.425
	5	0.742	0.320	0.815	0.423	0.303	0.466
100	1	0.529	0.430	0.527	0.418	0.364	0.714
	5	0.513	0.367	0.544	0.369	0.323	0.781

a value of θ from a logistic distribution denoted as $\text{Logistic}(l, s)$, where l is the location parameter and s is the scale parameter. Next, we generate n observations from $N(\theta, 1)$ and compute H for the two versions of RETEL. Throughout the analysis, we employ $p_n(\theta, \lambda)$ in Equation (4) with $\mu_{n,\theta} = \bar{X} - \theta$ and $\Sigma_{n,\theta} = 1$ for the univariate mean parameter θ . For comparison purposes, we also compute H using ETEL and AETEL. Keeping l fixed at 0, we repeat this procedure 10,000 times for each combination of $n \in \{5, 20, 50, 100\}$, $s \in \{1, 5\}$, and $\tau_n \in \{1, \log n\}$. We approximate the posterior distributions on a grid of θ values. Using the computed H values, we conduct the Kolmogorov–Smirnov test to evaluate the uniformity of the distributions. The resulting p -values are reported in Table 1.

With a smaller sample size of $n = 20$, RETEL_f tends to show a closer conformity to $U(0, 1)$ compared to ETEL and AETEL. The impact of a larger prior variance ($s = 5$) and a larger $p_c(\theta)$ ($\tau_n = \log n$) becomes more apparent when $n = 20$. As the sample size increases, the differences between the posterior distributions of the methods become negligible. When $n = 100$, none of the methods exhibit a significant departure of H from $U(0, 1)$. In all settings, RETEL_f produces larger p -values than RETEL_r. Figure 3 displays

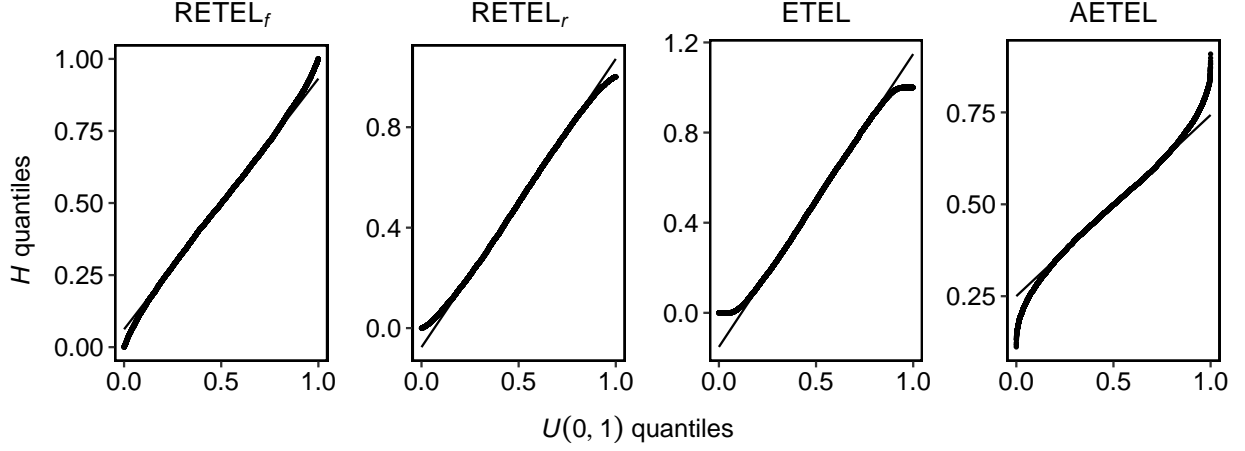


Figure 3. Quantile-quantile plots for the distribution of H versus $U(0, 1)$ when $n = 5$, $s = 5$, and $\tau_n = 1$. The light-tailed distribution from ETEL is due to the convex hull constraint. AETEL produces a heavier-tailed distribution than RETEL_f .

the quantile-quantile plots for the distribution of H versus $U(0, 1)$ when $n = 5$, $s = 5$, and $\tau_n = 1$. The plots highlight the differences in the tails of the distributions that are not apparent from the p -values alone. We emphasize that the Kolmogorov–Smirnov tests are based on a sample of 10,000 replicates and so are able to pick up quite small departures from the uniform distribution. All of the methods provide an excellent approximation to the null distribution when n is 50 or more. Additional plots for the full results are provided in Section 8 in the supplementary materials.

Next, we investigate the frequentist properties of the posterior intervals obtained from RETEL. We consider a true mean parameter value $\theta_0 = 0$ and generate n observations from $N(0, 1)$. Using the logistic prior distribution described earlier, we compute 95% posterior credible interval for θ using each of the four methods. This procedure is repeated 10,000 times for different combinations of $n \in \{5, 20, 50, 100\}$, $s \in \{0.5, 1, 5\}$, and $l \in \{0, 2\}$, while fixing τ_n at $\log n$. We then calculate the coverage rate and average length of the central credible intervals.

The results for $l = 0$ are presented in Table 2, where the prior mean matches the true parameter value. It can be seen that for all methods, as the sample size increases, the

intervals become shorter and the coverage rates approach the target of 95%. As s decreases, indicating stronger prior information on θ at 0, higher coverage rates and shorter intervals are obtained. The differences between the methods are most pronounced when $n = 5$. The intervals obtained from ETEL exhibit significantly lower coverage rates compared to the other methods. AETEL produces the widest intervals with coverage rates higher than the nominal level. The wider intervals and departure from the nominal coverage rate are related to the boundedness problem of AEL, which arises due to the addition of one pseudo-observation (Emerson & Owen 2009). On the other hand, RETEL yields higher coverage rates than the nominal level, but with much shorter intervals compared to AETEL. Within RETEL, RETEL_f produces wider intervals with higher coverage rates than RETEL_r , consistent with the findings from the plots in Figure 3.

Table 3 shows the results when $l = 2$, indicating a prior mean that is far from the true parameter value. In this case, the credible intervals tend to be wider with lower coverage rates. ETEL is relatively unaffected due to the convex hull constraint. However, the effect of different l values is noticeable for the other methods. Particularly when $n = 5$ and $s = 0.5$, the strong prior shifts the intervals toward 2. AETEL is the most affected, as its coverage rate is considerably lower than that of RETEL, even with wider intervals. To sum up, RETEL exhibits robust performance across various prior means and variances, demonstrating close-to-nominal posterior coverage rates with small sample sizes.

4.2 Expected Kullback–Leibler Divergence

The restricted posterior domain significantly affects Bayesian inference with EL and ETEL, especially when the sample size is small. In an example with only two observations, X_1 and X_2 , where the interest is in the mean parameter θ , the posterior domain shrinks to a

Table 2. Coverage rates (CR) and average lengths (Length) of 95% credible intervals when $l = 0$.

< $l = 0$ >		RETEL _f		RETEL _r		ETEL		AETEL	
n	s	CR	Length	CR	Length	CR	Length	CR	Length
5	0.5	99.0	1.690	94.4	1.385	79.2	1.128	100	2.572
	1	98.0	1.913	92.4	1.505	77.8	1.200	100	5.338
	5	97.3	2.031	91.3	1.561	77.2	1.230	100	9.367
20	0.5	94.9	0.826	94.0	0.803	93.1	0.790	96.3	0.886
	1	94.3	0.853	93.4	0.828	92.4	0.815	96.0	0.932
	5	94.2	0.863	93.1	0.837	92.2	0.824	96.1	0.965
50	0.5	95.0	0.540	94.8	0.534	94.5	0.530	95.8	0.558
	1	94.8	0.547	94.5	0.542	94.2	0.537	95.4	0.566
	5	94.7	0.550	94.4	0.544	94.1	0.540	95.3	0.569
100	0.5	94.9	0.386	94.8	0.385	94.6	0.380	95.2	0.394
	1	94.8	0.389	94.7	0.387	94.4	0.383	95.1	0.397
	5	94.8	0.390	94.6	0.388	94.3	0.384	95.1	0.398

Notes: CR is shown in percentage. The largest standard error of the lengths is 0.005 when $n = 5$ and $s = 5$.

Table 3. Coverage rates (CR) and average lengths (Length) of 95% credible intervals when $l = 2$.

< $l = 2$ >		RETEL _f		RETEL _r		ETEL		AETEL	
n	s	CR	Length	CR	Length	CR	Length	CR	Length
5	0.5	88.6	1.920	85.3	1.530	73.2	1.157	80.0	3.887
	1	96.5	1.969	90.8	1.534	76.9	1.207	100	6.174
	5	97.3	2.031	91.3	1.561	77.2	1.230	100	10.858
20	0.5	92.2	0.857	91.4	0.832	90.7	0.819	94.2	1.529
	1	93.7	0.858	92.9	0.833	92.1	0.820	96.6	0.995
	5	94.1	0.863	93.1	0.837	92.2	0.824	96.2	0.975
50	0.5	93.8	0.549	93.4	0.543	93.3	0.539	94.4	0.569
	1	94.6	0.549	94.3	0.543	94.1	0.538	95.2	0.568
	5	94.7	0.550	94.5	0.544	94.1	0.540	95.3	0.569
100	0.5	94.3	0.390	94.2	0.388	94.0	0.384	94.7	0.398
	1	94.6	0.390	94.4	0.388	94.2	0.384	95.0	0.398
	5	94.7	0.390	94.6	0.388	94.3	0.384	95.1	0.398

Notes: CR is shown in percentage. The largest standard error of the lengths is 0.006 when $n = 20$ and $s = 0.5$.

singleton as $|X_1 - X_2|$ decreases toward zero. This example illustrates a problematic aspect of EL and ETEL, where we have more definitive information on the parameter with fewer data.

More generally, consider a parametric model $\mathcal{M} = \{p(\mathbf{x} | \boldsymbol{\theta}) | \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ and a prior density $\pi(\boldsymbol{\theta})$. The expected information obtained from observing \mathbf{x} from \mathcal{M} can be measured using the expected KL divergence:

$$I(\pi | \mathcal{M}) = \int_{\mathcal{X}} D_{KL}(\pi(\cdot | \mathbf{x}) \| \pi(\cdot)) m(\mathbf{x}) d\mathbf{x},$$

where $\pi(\boldsymbol{\theta} | \mathbf{x}) = \pi(\boldsymbol{\theta})p(\mathbf{x} | \boldsymbol{\theta})/m(\mathbf{x})$ and $m(\mathbf{x}) = \int_{\Theta} \pi(\boldsymbol{\theta})p(\mathbf{x} | \boldsymbol{\theta})d\boldsymbol{\theta}$. Let $I(\pi | \mathcal{M}_n)$ denote the expected information obtained from the set of observations $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. It is expected that $I(\pi | \mathcal{M}_n)$ increases monotonically with n (Mantovan & Todini 2006). The following result, based on Berger et al. (2009, Theorem 3), illustrates this monotonicity property.

Proposition 4. *Let $\mathcal{M} = \{p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\theta}) | \mathbf{x}_1 \in \mathcal{X}, \mathbf{x}_2 \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\}$ be a model with a sufficient statistic $\mathbf{t} = \mathbf{t}(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{U}$. Suppose $\pi(\boldsymbol{\theta})$ is a strictly positive and continuous prior on Θ , where $\pi(\boldsymbol{\theta} | \mathbf{x}_1, \mathbf{x}_2) = \pi(\boldsymbol{\theta})p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\theta})/m(\mathbf{x}_1, \mathbf{x}_2)$ and $m(\mathbf{x}_1, \mathbf{x}_2) = \int_{\Theta} \pi(\boldsymbol{\theta})p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\theta})d\boldsymbol{\theta} < \infty$. Under Condition 1, if $\int_{\mathcal{U}} p(\mathbf{t} | \boldsymbol{\theta}) \log(p(\mathbf{t} | \boldsymbol{\theta})/p(\mathbf{t} | \boldsymbol{\theta}'))d\mathbf{t} < \infty$ for any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\theta}' \in \Theta$, then $I(\pi | \mathcal{M}_1) \leq I(\pi | \mathcal{M}_2) < \infty$.*

Based on the above proposition, the approximate validity of a pseudo-likelihood for Bayesian inference can be evaluated by examining whether it preserves the monotonicity property.

To examine the performance of RETEL compared to EL and ETEL, we consider two independent experiments where we obtain independent observations, denoted as X_{ij} for $i = 1, 2$ and $j = 1, \dots, n$, from the following hierarchical model:

$$X_{ij} | \theta_i \sim N(\theta_i, \sigma^2),$$

$$\theta_i \mid \mu \sim \text{Cauchy}(\mu, \gamma),$$

$$\mu \sim N(0, \tau^2).$$

We assume fixed values of $\sigma = 1$, $\gamma = 1$, and $\tau = 10$. We use a variety of empirical likelihoods in place of the normal density for X_{ij} . Our main focus is on the marginal posterior distribution of μ , with the density denoted by $\pi(\mu \mid \mathcal{D}_n)$. Given the values of θ_1 and θ_2 with $\Delta = |\theta_1 - \theta_2| > 2$, the Cauchy distribution for θ_1 and θ_2 yields two maximum likelihood estimates of μ given by $(\theta_1 + \theta_2)/2 \pm \sqrt{\Delta^2 - 1}$ (Dharmadhikari & Joag-Dev 1985). Consequently, when combined with the large standard deviation of the prior distribution for μ , the restricted posterior domain of θ_1 and θ_2 from EL and ETEL leads to a bimodal marginal posterior distribution for μ . This bimodality can potentially result in inflated values of $I(\pi \mid \mathcal{M}_n)$ for EL and ETEL, particularly when n is small.

The marginal likelihood, $m(\mathbf{x})$, for the four methods cannot be computed analytically. Instead, we can observe that $I(\pi \mid \mathcal{M})$ can be expressed as:

$$I(\pi \mid \mathcal{M}) = \int_{\Theta} \pi(\boldsymbol{\theta}) \left[\int_{\mathcal{X}} D_{KL}(\pi(\cdot \mid \mathbf{x}) \parallel \pi(\cdot)) p(\mathbf{x} \mid \boldsymbol{\theta}) d\mathbf{x} \right] d\boldsymbol{\theta},$$

where $D_{KL}(\pi(\cdot \mid \mathbf{x}) \parallel \pi(\cdot))$ is computed with respect to μ . Since our focus is on comparing $I(\pi \mid \mathcal{M}_n)$ for the methods, we fix θ_1 and θ_2 at -3 and 3 , respectively. For each method and $n \in \{2, 4, 6, 8, 10\}$, we estimate the inner integrand of $I(\pi \mid \mathcal{M}_n)$ through simulation using the following steps:

Step 1. Generate X_{1j} from $N(-3, 1)$ and X_{2j} from $N(3, 1)$ for $j = 1, \dots, n$.

Step 2. Generate 10,000 posterior samples of θ_1 , θ_2 , and μ with a random-walk Metropolis–Hastings algorithm.

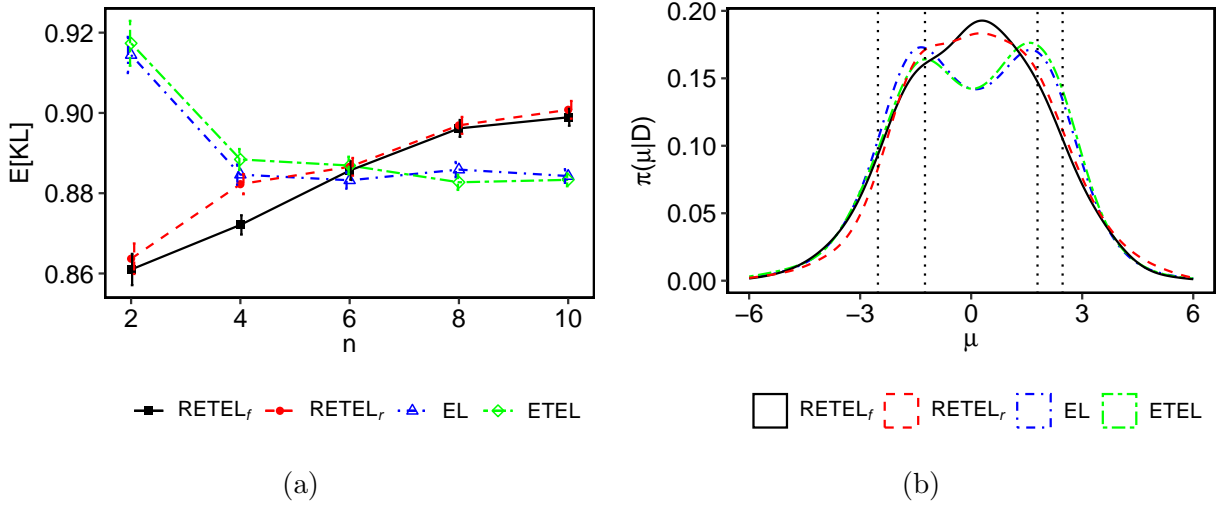


Figure 4. Plots of (a) the expected KL divergence (inner integrand) and (b) the marginal posterior density $\pi(\mu \mid \mathcal{D}_n)$ with $n = 2$. The error bars in (a) represent plus or minus one standard error. The vertical dotted lines in (b) indicate the four realized data points.

Step 3. Estimate $\pi(\mu \mid \mathcal{D}_n)$ from the posterior samples and compute $D_{KL}(\pi(\cdot \mid \mathcal{D}_n) \parallel \pi(\cdot))$ by numerical integration with adaptive quadrature.

Step 4. Repeat Steps 1–3 10,000 times and take the average of the estimates from Step 3.

Step 2 is implemented with two chains of length 5,000, ensuring that the potential scale reduction factor (Gelman & Rubin 1992) remains below 1.1 on average for each method. For the regularized methods, $\tau_n = 1$ is used when $n = 2$, and $\tau_n = \log n$ is used otherwise. We implement EL using the R package **melt** (Kim 2023).

The results are summarized in Figure 4. In Figure 4a, it can be seen that $I(\pi \mid \mathcal{M}_n)$ is the smallest when $n = 2$ for RETEL_f (0.861) and RETEL_r (0.864), and it increases monotonically as the sample size grows. RETEL_r tends to produce slightly larger $I(\pi \mid \mathcal{M}_n)$ compared to RETEL_f. On the other hand, EL and ETEL attain the largest $I(\pi \mid \mathcal{M}_n)$ when $n = 2$, with values of 0.914 and 0.917, respectively. The values of $I(\pi \mid \mathcal{M}_n)$ decrease as the sample size and the range of the data increase. EL and ETEL do not exhibit an upward trend in $I(\pi \mid \mathcal{M}_n)$ and, even as n moves toward 10, do not show a notable improvement.

This discrepancy is caused by the strong bimodality of $\pi(\mu \mid \mathcal{D}_n)$, as illustrated in [Figure 4b](#).

5 Application

We present an application of RETEL to the estimation of median 1989 income for four-person families by State in the USA. In the field of small area estimation ([Ghosh & Rao 1994](#), [Rao 2003](#)), the state-level direct estimates provided by the Census Bureau based on the Current Population Survey data may not be sufficiently accurate for some states due to limited sample sizes. To address this issue, Bayesian methods have been proposed, which incorporate additional information or related auxiliary variables specific to these small areas ([Fay & Herriot 1979](#), [Datta et al. 1996](#), [Ghosh et al. 1996](#)). In particular, EL has been applied to small area estimation in hierarchical Bayesian models ([Chaudhuri & Ghosh 2011](#), [Chaudhuri et al. 2017](#), [Jahan et al. 2022](#)).

Let Y_i , $i = 1, \dots, 51$, represent the direct estimate of the 1989 median income for four-person families in the i th state, including the District of Columbia. We also consider the direct estimate of the 1979 median income, denoted by X_{1i} , as an auxiliary variable. Additionally, following [Chung et al. \(2019\)](#), we incorporate the adjusted census median income denoted by X_{2i} , where $X_{2i} = (\text{PCI}_{i,1989}/\text{PCI}_{i,1979})X_{1i}$. Here, $\text{PCI}_{i,1979}$ and $\text{PCI}_{i,1989}$ refer to per capita income from the Bureau of Economic Analysis in 1979 and 1989, respectively. All variables are standardized to ensure numerical stability and facilitate illustration.

Similar to the generalized linear model approach of [Chaudhuri & Ghosh \(2011\)](#), we assume that the Y_i are conditionally independent given θ_i . Specifically, we assume:

$$\text{E}[Y_i \mid \theta_i] = \theta_i, \quad \text{Var}[Y_i \mid \theta_i] = V_i,$$

$$\begin{aligned}\theta_i \mid \boldsymbol{\beta}, \sigma^2 &\stackrel{\text{ind}}{\sim} N(\mathbf{X}_i^\top \boldsymbol{\beta}, \sigma^2), \\ \boldsymbol{\beta} \mid \sigma^2 &\sim N(\boldsymbol{\beta}_0, g\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}), \\ \sigma^2 &\sim \pi(\sigma^2 \mid \mathcal{D}_n) \propto \frac{1}{\sigma^2}.\end{aligned}$$

Here, $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $\mathbf{X}_i = (X_{1i}, X_{2i})$, and \mathbf{X} is the matrix with the i th row given by \mathbf{X}_i^\top . The sampling variance V_i is set to 1. We adopt the g -prior of Zellner (1988) for $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_0 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ and $g = 0.1$, where $\mathbf{Y} = (Y_1, \dots, Y_{51})$. For the likelihood function, we use RETEL_f, RETEL_r, EL, and ETEL with the bivariate estimating function $(Y_i - \theta_i, (Y_i - \theta_i)^2/V_i - 1)$.

For each method, we use a random-walk Metropolis–Hastings algorithm to draw posterior samples of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and σ^2 from four chains, each of length 250,000. The regularized versions employ $\mu_{n,\boldsymbol{\theta}} = n^{-1} \sum_{i=1}^n \mathbf{g}(Y_i, \boldsymbol{\theta})$, $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} = (n-1)^{-1} \sum_{i=1}^n \mathbf{g}(Y_i, \boldsymbol{\theta}) \mathbf{g}(Y_i, \boldsymbol{\theta})^\top$, and $\tau_n = \log n$, where $n = 51$. The maximum potential scale reduction factor of all the methods is 1.0119 for $\boldsymbol{\theta}$, 1.0006 for $\boldsymbol{\beta}$, and 1.0137 for σ^2 . We compute the 95% posterior credible interval for each θ_i and use the posterior median $\hat{\theta}_i$ as an estimate for Y_i . The performance of the methods is evaluated using the following metrics: average absolute deviation (AAD) $n^{-1} \sum_{i=1}^n |\hat{\theta}_i - Y_i|$, average absolute relative deviation (AARD) $n^{-1} \sum_{i=1}^n |(\hat{\theta}_i - Y_i)/Y_i|$, average squared deviation (ASD) $n^{-1} \sum_{i=1}^n (\hat{\theta}_i - Y_i)^2$, and average squared relative deviation (ASRD) $n^{-1} \sum_{i=1}^n ((\hat{\theta}_i - Y_i)/Y_i)^2$.

Table 4 provides the summary. The results show that RETEL demonstrates improvement over EL and ETEL, exhibiting smaller deviations in all metrics and providing more accurate estimates. Although RETEL has slightly longer intervals compared to EL and ETEL, RETEL_r performs the best among the methods in terms of accuracy. On the other hand, EL and ETEL exhibit nearly equivalent performances, aligning with the findings in

Table 4. Comparison of the accuracy of estimates of θ from the four methods. The average length of the 51 credible intervals is added for each method (Length).

Method	AAD	AARD	ASD	ASRD	Length
RETEL _f	0.276	0.922	0.112	4.234	3.757
RETEL _r	0.272	0.900	0.109	3.831	3.781
EL	0.280	0.947	0.112	4.587	3.755
ETEL	0.279	0.942	0.115	4.696	3.729

Chaudhuri & Ghosh (2011).

6 Discussion

Bayesian methods are fundamentally based on probability, with inference proceeding from the prior distribution to the posterior distribution via conditioning on the observed data. Bayesian versions of EL and ETEL place a prior distribution on a finite number of features of a nonparametric (and hence infinite dimensional) distribution and regard the remainder of the distribution itself as a nuisance parameter. The lack of a full probability model prevents one from integrating over the nuisance parameter. EL and ETEL replace the integration with a maximization, and this replacement produces artifacts that clash with known properties that all Bayesian methods must have.

The most striking departure from Bayesian behavior is the zeroing out of regions of the parameter space as one moves from prior distribution to posterior distribution, with the expectation that, as more data are collected, the zeroed out regions will reappear and be assigned positive probability. These regions concern the main parameters of interest—those that are represented by the estimating equations that give rise to EL and ETEL. The regions and behavior follow from the convex hull constraint. Another feature of Bayesian methods (and most other statistical methods) is that the data are the data. An observation, \mathbf{X}_i , may

come from a distribution that depends on an unknown parameter θ , but \mathbf{X}_i is not allowed to differ for different values of θ . Methods previously proposed to handle the convex hull constraint, such as AEL and AETEL, rely on pseudo-data that change with the parameter.

This paper has investigated a suite of methods to deal with the convex hull constraint without the need to invoke parameter-dependent pseudo-data. The first step was the development of WETEL as an extension of AEL and AETEL. WETEL accommodates fractional observations and reduces the dependence of pseudo-data on the parameter, allowing for a massive expansion of the convex hull while aligning the pseudo-data more closely with the observed data. As a subsequent step, WETEL leads to the regularization technique of RETEL by passing to a limit where pseudo-data are added in a particular way. We also provided a distinct derivation of RETEL as the solution to a KL divergence optimization problem involving a mixture of the empirical distribution and a continuous exponential family distribution.

The likelihood ratios from RETEL compare the constrained regularized likelihood to the unconstrained regularized likelihood. This is implicit in [Equations \(7\) and \(8\)](#). In essence, RETEL replaces the empirical distribution with a regularized empirical distribution before considering tilts that match constraints. This stabilizes the results, particularly for smaller sample sizes. It also appears to produce a posterior distribution that is less pathological and more amenable to traditional sampling techniques for model fitting. We showed that RETEL retains the desirable properties of EL and ETEL such as Wilks' and Bernstein–von Mises' theorems. The simulation and data analysis demonstrated that RETEL exhibits improved finite sample performance compared to EL and ETEL for Bayesian inference. Overall, our findings highlight the effectiveness of RETEL as a pseudo-likelihood for Bayesian inference in overcoming the convex hull constraint of EL and ETEL.

There are a number of reasons to replace integration in a Bayesian model with maximization. In addition to handling the nuisance parameter, maximization can be much quicker than integration. We suspect that an appropriate regularization in RETEL will bring the maximized version of the problem closer to a genuine Bayesian solution. This is a direction for future research. Another promising direction involves investigating whether RETEL retains the robust higher-order asymptotic properties of ETEL. [Jing & Wood \(1996\)](#) showed that ETEL is not Bartlett correctable. [Schennach \(2007\)](#) showed that the ETEL has robust higher-order asymptotic properties under model misspecification compared to the EL estimator. [Chib et al. \(2018\)](#) established Bernstein–von Mises results for ETEL under model misspecification. Further research is needed to determine the extent to which these properties hold for RETEL.

Supplementary Materials

The supplementary materials contain technical proofs and plots from simulations.

Funding

This work was supported by the National Science Foundation under Grants No. SES-1921523 and DMS-2015552.

References

Artstein, Z. & Wets, R. J.-B. (1995), ‘Consistency of minimizers and the SLLN for stochastic programs’, *Journal of Convex Analysis* **2**, 1–17.

- Bartolucci, F. (2007), ‘A penalized version of the empirical likelihood ratio for the population mean’, *Statistics & Probability Letters* **77**, 104–110.
- Berger, J. O., Bernardo, J. M. & Sun, D. (2009), ‘The formal definition of reference priors’, *The Annals of Statistics* **37**, 905–938.
- Chang, J., Tang, C. Y. & Wu, T. T. (2018), ‘A new scope of penalized empirical likelihood with high-dimensional estimating equations’, *The Annals of Statistics* **46**, 3185–3216.
- Chaudhuri, S. & Ghosh, M. (2011), ‘Empirical likelihood for small area estimation’, *Biometrika* **98**, 473–480.
- Chaudhuri, S., Mondal, D. & Yin, T. (2017), ‘Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation’, *Journal of the Royal Statistical Society, Series B* **79**, 293–320.
- Chen, J., Variyath, A. M. & Abraham, B. (2008), ‘Adjusted empirical likelihood and its properties’, *Journal of Computational and Graphical Statistics* **17**, 426–443.
- Cheng, Y. & Zhao, Y. (2019), ‘Bayesian jackknife empirical likelihood’, *Biometrika* **106**, 981–988.
- Chib, S., Shin, M. & Simoni, A. (2018), ‘Bayesian estimation and comparison of moment condition models’, *Journal of the American Statistical Association* **113**, 1656–1668.
- Chung, H. C., Datta, G. S. & Maples, J. (2019), *Estimation of Median Incomes of the American States: Bayesian Estimation of Means of Subpopulations*, Springer-Verlag, pp. 505–518.
- Corcoran, S. A. (1998), ‘Bartlett adjustment of empirical discrepancy statistics’, *Biometrika* **85**, 967–972.

- Cover, T. M. & Thomas, J. A. (2006), *Elements of Information Theory*, Wiley-Interscience.
- Cressie, N. & Read, T. R. (1984), ‘Multinomial goodness-of-fit tests’, *Journal of the Royal Statistical Society, Series B* **46**, 440–464.
- Datta, G., Ghosh, M., Nangia, N. & Natarajan, K. (1996), Estimation of median income of four-person families: a Bayesian approach, *in* ‘Bayesian analysis in statistics and econometrics: essays in honor of Arnold Zellner’, pp. 129–140.
- Dharmadhikari, S. & Joag-Dev, K. (1985), ‘Examples of nonunique maximum likelihood estimators’, *The American Statistician* **39**, 199–200.
- DiCiccio, T. J., Hall, P. & Romano, J. (1991), ‘Empirical likelihood is Bartlett-correctable’, *The Annals of Statistics* **19**, 1053–1061.
- Dupačová, J. (1992), ‘Epi-consistency in restricted regression models: The case of a general convex fitting function’, *Computational Statistics & Data Analysis* **14**, 417–425.
- Dupačová, J. & Wets, R. J.-B. (1988), ‘Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems’, *The Annals of Statistics* **16**, 1517–1549.
- Efron, B. (1981), ‘Nonparametric standard errors and confidence intervals’, *Canadian Journal of Statistics* **9**, 139–158.
- Emerson, S. C. & Owen, A. B. (2009), ‘Calibration of the empirical likelihood method for a vector mean’, *Electronic Journal of Statistics* **3**, 1161–1192.
- Fay, R. E. & Herriot, R. A. (1979), ‘Estimates of income for small places: An application of James–Stein procedures to census data’, *Journal of the American Statistical Association* **74**, 269–277.

- Gelman, A. & Rubin, D. B. (1992), ‘Inference from iterative simulation using multiple sequences’, *Statistical Science* **7**, 457–472.
- Ghosh, M., Nangia, N. & Kim, D. H. (1996), ‘Estimation of median income of four-person families: A Bayesian time series approach’, *Journal of the American Statistical Association* **91**, 1423–1431.
- Ghosh, M. & Rao, J. N. K. (1994), ‘Small area estimation: An appraisal’, *Statistical Science* **9**, 55–76.
- Glenn, N. & Zhao, Y. (2007), ‘Weighted empirical likelihood estimates and their robustness properties’, *Computational Statistics & Data Analysis* **51**, 5130–5141.
- Grendár, M. & Judge, G. (2009), ‘Empty set problem of maximum empirical likelihood methods’, *Electronic Journal of Statistics* pp. 1542–1555.
- Hainmueller, J. (2012), ‘Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies’, *Political Analysis* **20**, 25–46.
- Hjort, N. L., McKeague, I. W. & van Keilegom, I. (2018), ‘Hybrid combinations of parametric and empirical likelihoods’, *Statistica Sinica* **28**, 2389–2407.
- Jahan, F., Kennedy, D. W., Duncan, E. W. & Mengersen, K. L. (2022), ‘Evaluation of spatial Bayesian empirical likelihood models in analysis of small area data’, *PLOS ONE* **17**, 1–27.
- Jing, B.-Y. & Wood, A. T. A. (1996), ‘Exponential empirical likelihood is not Bartlett correctable’, *The Annals of Statistics* **24**, 365–369.
- Kim, E. (2023), ***melt**: Multiple Empirical Likelihood Tests*. R package version 1.10.0.

- King, A. J. & Wets, R. J.-B. (1991), ‘Epi-consistency of convex stochastic programs’, *Stochastics and Stochastic Reports* **34**, 83–92.
- Lazar, N. A. (2003), ‘Bayesian empirical likelihood’, *Biometrika* **90**, 319–326.
- Leng, C. & Tang, C. Y. (2012), ‘Penalized empirical likelihood and growing dimensional general estimating equations’, *Biometrika* **99**, 703–716.
- Liu, Y. & Chen, J. (2010), ‘Adjusted empirical likelihood with high-order precision’, *The Annals of Statistics* **38**, 1341–1362.
- Mantovan, P. & Todini, E. (2006), ‘Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology’, *Journal of Hydrology* **330**, 368–381.
- Monahan, J. F. & Boos, D. D. (1992), ‘Proper likelihoods for Bayesian analysis’, *Biometrika* **79**, 271–278.
- Newey, W. K. & Smith, R. J. (2004), ‘Higher order properties of GMM and generalized empirical likelihood estimators’, *Econometrica* **72**, 219–255.
- Owen, A. B. (1988), ‘Empirical likelihood ratio confidence intervals for a single functional’, *Biometrika* **75**, 237–249.
- Qin, J. (1994), ‘Semi-empirical likelihood ratio confidence intervals for the difference of two sample means’, *Annals of the Institute of Statistical Mathematics* **46**, 117–126.
- Qin, J. & Lawless, J. (1994), ‘Empirical likelihood and general estimating equations’, *The Annals of Statistics* **22**, 300–325.
- Rao, J. N. K. (2003), *Small Area Estimation*, Wiley-Interscience.
- Rockafellar, R. T. & Wets, R. J.-B. (2009), *Variational analysis*, Springer-Verlag.

- Schennach, S. M. (2005), ‘Bayesian exponentially tilted empirical likelihood’, *Biometrika* **92**, 31–46.
- Schennach, S. M. (2007), ‘Point estimation with exponentially tilted empirical likelihood’, *The Annals of Statistics* **35**, 634–672.
- Smith, R. J. (1997), ‘Alternative semi-parametric likelihood approaches to generalised method of moments estimation’, *The Economic Journal* **107**, 503–519.
- Sueishi, N. (2022), ‘Large sample justifications for the Bayesian empirical likelihood’, *Econometric Theory* **0**, 1–31.
- Tang, C. Y. & Leng, C. (2010), ‘Penalized high-dimensional empirical likelihood’, *Biometrika* **97**, 905–920.
- Tsao, M. & Wu, F. (2013), ‘Empirical likelihood on the full parameter space’, *The Annals of Statistics* **41**, 2176–2196.
- Wets, R. J.-B. (1974), ‘Stochastic programs with fixed recourse: The equivalent deterministic program’, *SIAM Review* **16**, 309–339.
- Yiu, A., Goudie, R. J. B. & Tom, B. D. M. (2020), ‘Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood’, *Biometrika* **107**, 857–873.
- Yu, W. & Bondell, H. D. (2023), ‘Variational Bayes for fast and accurate empirical likelihood inference’, *Journal of the American Statistical Association* **0**, 1–13.
- Zellner, A. (1988), ‘Bayesian analysis in econometrics’, *Journal of Econometrics* **37**, 27–50.
- Zhu, H., Zhou, H., Chen, J., Li, Y., Lieberman, J. & Styner, M. (2009), ‘Adjusted exponentially tilted likelihood with applications to brain morphology’, *Biometrics* **65**(3), 919–927.

Supplementary Material for “Regularized Exponentially Tilted Empirical Likelihood for Bayesian Inference”

Eunseop Kim,^a Steven N. MacEachern,^b and Mario Peruggia^c

We employ the same notation as in the main paper. We define the following quantities:

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbb{E}_P[\mathbf{g}_i(\boldsymbol{\theta})], \quad \mathbf{h}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}), \quad \mathbf{V}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}) \mathbf{g}_i(\boldsymbol{\theta})^\top.$$

Additionally, we use the notation “ \rightarrow_p ” and “ \rightarrow_d ” to denote convergence in probability and convergence in distribution, respectively. We restate [Conditions 1–7](#) below:

Condition 1. The parameter space Θ is compact, with $\boldsymbol{\theta}_0$ as an interior point of Θ and the unique solution to $\mathbb{E}_P[\mathbf{g}_i(\boldsymbol{\theta})] = \mathbf{0}$.

Condition 2. With probability 1, $\mathbf{g}_i(\boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \Theta$, continuously differentiable in a neighborhood \mathcal{N} of $\boldsymbol{\theta}_0$, and $\mathbb{E}_P[\sup_{\boldsymbol{\theta} \in \mathcal{N}} \|\partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta})\|] < \infty$.

Condition 3. $\text{rank}(\mathbf{G}) = \text{rank}(\mathbf{V}) = p$.

Condition 4. For some $\alpha > 3$, $\mathbb{E}_P[\sup_{\boldsymbol{\theta} \in \Theta} |\mathbf{g}_i(\boldsymbol{\theta})|^\alpha] < \infty$.

Condition 5. $\tau_n = O(\log n)$; $\boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} = \boldsymbol{\mu} + o_P(1)$ for some $\boldsymbol{\mu} \in \mathbb{R}^p$; $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0}$ is positive definite for any n with probability 1; and $\boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0} = \boldsymbol{\Sigma} + o_P(1)$ for some $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$.

^aDepartment of Statistics, The Ohio State University, Columbus, OH. Corresponding author.

^bDepartment of Statistics, The Ohio State University, Columbus, OH.

^cDepartment of Statistics, The Ohio State University, Columbus, OH.

Condition 6. The prior measure admits a continuous density $\pi(\boldsymbol{\theta})$ with respect to the Lebesgue measure, and $\pi(\boldsymbol{\theta})$ is positive in a neighborhood of $\boldsymbol{\theta}_0$.

Condition 7. For any $\delta > 0$, there exists $\epsilon > 0$ such that

$$P \left(\sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \delta} \frac{1}{n} (\log L_{RET}(\boldsymbol{\theta}) - \log L_{RET}(\boldsymbol{\theta}_0)) \leq -\epsilon \right) \rightarrow 1.$$

1 Proof of Proposition 1

We introduce the following pieces of notation:

$$\begin{aligned} \tilde{\mathbf{h}}_N(\boldsymbol{\theta}) &= \sum_{i=1}^N w_i \mathbf{g}_i(\boldsymbol{\theta}) = \frac{n}{n+1} \mathbf{h}_n(\boldsymbol{\theta}) + \frac{1}{m(n+1)} \sum_{i=n+1}^N \mathbf{g}_i(\boldsymbol{\theta}), \\ \tilde{\mathbf{V}}_N(\boldsymbol{\theta}) &= \sum_{i=1}^N w_i \mathbf{g}_i(\boldsymbol{\theta}) \mathbf{g}_i(\boldsymbol{\theta})^\top = \frac{n}{n+1} \mathbf{V}_n(\boldsymbol{\theta}) + \frac{1}{m(n+1)} \sum_{i=n+1}^N \mathbf{g}_i(\boldsymbol{\theta}) \mathbf{g}_i(\boldsymbol{\theta})^\top. \end{aligned}$$

Since the value of m is fixed, the terms $m^{-1} \sum_{i=n+1}^N \mathbf{g}_i(\boldsymbol{\theta})$ and $m^{-1} \sum_{i=n+1}^N \mathbf{g}_i(\boldsymbol{\theta}) \mathbf{g}_i(\boldsymbol{\theta})^\top$ are finite for each $\boldsymbol{\theta}$. By the weak law of large numbers and [Condition 1](#), we have

$$\mathbf{h}_n(\boldsymbol{\theta}_0) \rightarrow_p \mathbf{0}, \quad \tilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) \rightarrow_p \mathbf{0}, \quad \mathbf{V}_n(\boldsymbol{\theta}_0) \rightarrow_p \mathbf{V}, \quad \tilde{\mathbf{V}}_N(\boldsymbol{\theta}_0) \rightarrow_p \mathbf{V}.$$

Applying the uniform law of large numbers and [Condition 2](#), we obtain

$$\sup_{\boldsymbol{\theta} \in \mathcal{N}} \|\partial_{\boldsymbol{\theta}} \mathbf{h}_n(\boldsymbol{\theta}) - \mathbb{E}_P[\partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta})]\| \rightarrow_p 0, \quad \sup_{\boldsymbol{\theta} \in \mathcal{N}} \left\| \partial_{\boldsymbol{\theta}} \tilde{\mathbf{h}}_N(\boldsymbol{\theta}) - \mathbb{E}_P[\partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta})] \right\| \rightarrow_p 0.$$

By [Condition 3](#) and [Jacod & Sørensen \(2018, Theorem 2.5\)](#), there exist consistent estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_w$ such that $\hat{\boldsymbol{\theta}} \rightarrow_p \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_w \rightarrow_p \boldsymbol{\theta}_0$, and $\mathbf{h}_n(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ and $\tilde{\mathbf{h}}_N(\hat{\boldsymbol{\theta}}_w) = \mathbf{0}$ with probability approaching 1. Using [Condition 3](#) and the central limit theorem, we can show

that both $n^{1/2}\mathbf{h}_n(\boldsymbol{\theta}_0)$ and $n^{1/2}\mathbf{h}_N(\boldsymbol{\theta}_0)$ are stochastically bounded. As a result, $n^{1/2}|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|$ and $n^{1/2}|\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0|$ are also stochastically bounded (Jacod & Sørensen 2018, Theorem 2.9), which establishes $\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}} = o_P(n^{-1/2})$.

Next, the first-order condition for $\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)$ yields

$$\sum_{i=1}^N w_i \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)) \mathbf{g}_i(\boldsymbol{\theta}_0) = \mathbf{0}.$$

Using the weight specification for w_i in the main paper, it follows from Newey & Smith (2004, Lemma A2), together with Conditions 1–4, that $\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) = O_P(n^{-1/2})$ with probability approaching 1. Expanding the condition around $\boldsymbol{\lambda} = \mathbf{0}$, we obtain

$$\mathbf{0} = \widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) + \widetilde{\mathbf{V}}_N(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) + \mathbf{R}_N, \quad (1)$$

where

$$\mathbf{R}_N = \frac{1}{2} \sum_{i=1}^N w_i \exp(\boldsymbol{\xi}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)^\top \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)$$

for some $\boldsymbol{\xi}$ between $\mathbf{0}$ and $\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)$. Let

$$\mathbf{R}_n = \frac{1}{2(n+1)} \sum_{i=1}^n \exp(\boldsymbol{\xi}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)^\top \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)$$

and

$$\mathbf{R}_m = \frac{1}{2m(n+1)} \sum_{i=n+1}^N \exp(\boldsymbol{\xi}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)^\top \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0).$$

Then, combining $\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) = O_P(n^{-1/2})$, Condition 4, and Lemma A1 in Newey & Smith

(2004), we obtain $\mathbf{R}_N = \mathbf{R}_n + \mathbf{R}_m = O_P(n^{-1}) + O_P(n^{-2})$ and

$$\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) = -\tilde{\mathbf{V}}_N(\boldsymbol{\theta}_0)^{-1} \tilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) + O_P(n^{-1}). \quad (2)$$

By following the same steps for $\boldsymbol{\lambda}_{ET}$, we get $\boldsymbol{\lambda}_{ET}(\boldsymbol{\theta}_0) = -\mathbf{V}_n(\boldsymbol{\theta}_0)^{-1} \mathbf{h}_n(\boldsymbol{\theta}_0) + O_P(n^{-1})$.

Therefore, we have $\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) - \boldsymbol{\lambda}_{ET}(\boldsymbol{\theta}_0) = O_P(n^{-1})$. \square

2 Proof of Theorem 1

The saddle point problem of $\hat{\boldsymbol{\theta}}_w$ and $\boldsymbol{\lambda}_{WET}(\hat{\boldsymbol{\theta}}_w)$ leads to the following first-order conditions:

$$\begin{aligned} \sum_{i=1}^N w_i \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta})) \partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta})^\top \boldsymbol{\lambda} &= \mathbf{0}, \\ \sum_{i=1}^N w_i \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta})) \mathbf{g}_i(\boldsymbol{\theta}) &= \mathbf{0}. \end{aligned}$$

With Conditions 1–4, we can directly apply the results from Newey & Smith (2004, Theorem 3.2) and Zhu et al. (2009, Theorem 1). By using Equation (1) and expanding the conditions around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\lambda} = \mathbf{0}$, we obtain

$$\mathbf{W}_N \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}_0 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\tilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) \end{pmatrix} + o_P(n^{-1/2}),$$

where

$$\mathbf{W}_N = \begin{pmatrix} \mathbf{0} & \sum_{i=1}^N w_i \partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta}_0)^\top \\ \sum_{i=1}^N w_i \partial_{\boldsymbol{\theta}} \mathbf{g}_i(\boldsymbol{\theta}_0) & \tilde{\mathbf{V}}_N(\boldsymbol{\theta}_0) \end{pmatrix} \rightarrow_p \mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{G}^\top \\ \mathbf{G} & \mathbf{V} \end{pmatrix}.$$

Consequently, we have

$$\mathbf{W}^{-1} = \begin{pmatrix} -\mathbf{\Omega} & \mathbf{H} \\ \mathbf{H}^\top & \mathbf{P} \end{pmatrix},$$

where $\mathbf{H} = \mathbf{\Omega}\mathbf{G}^\top\mathbf{V}^{-1}$ and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{G}\mathbf{\Omega}\mathbf{G}^\top\mathbf{V}^{-1}$. Hence, we obtain

$$\widehat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_0 = -\mathbf{H}\widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) + o_P(n^{-1/2}),$$

and the first result follows by noting that $n^{1/2}\widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0)$ converges in distribution to $N(\mathbf{0}, \mathbf{V})$ and $\mathbf{H}\mathbf{V}\mathbf{H}^\top = \mathbf{\Omega}$. For the second result, observe that

$$\begin{aligned} & -2 \log \mathbf{R}_{WET}(\boldsymbol{\theta}_0) \\ &= -2N \sum_{i=1}^N w_i \left\{ \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) - \log \left(\sum_{i=1}^N w_i \exp \left(\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) \right) \right) \right\} \quad (3) \\ &= -2N \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) + 2N \log \left(\sum_{i=1}^N w_i \exp \left(\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) \right) \right) \end{aligned}$$

and

$$\begin{aligned} & \log \left(\sum_{i=1}^N w_i \exp \left(\boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) \right) \right) \\ &= \log \left(1 + \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) + \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{V}}_N(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) / 2 + o_P(n^{-1}) \right) \quad (4) \\ &= \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0) + \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{V}}_N(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) / 2 + o_P(n^{-1}). \end{aligned}$$

Substituting the expressions in Equation (4) and Equation (2) into Equation (3), we obtain

$$\begin{aligned} -2 \log \mathbf{R}_{WET}(\boldsymbol{\theta}_0) &= N \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{V}}_N(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{WET}(\boldsymbol{\theta}_0) + o_P(1) \\ &= N \widetilde{\mathbf{h}}_N(\boldsymbol{\theta}_0)^\top \widetilde{\mathbf{V}}_N(\boldsymbol{\theta}_0)^{-1} \widetilde{\mathbf{g}}_n^*(\boldsymbol{\theta}_0)^\top + o_P(1), \end{aligned}$$

and the result follows. □

3 Proof of Proposition 2

Fix any $\boldsymbol{\theta} \in \Theta$. From [Condition 2](#), $d_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is finite and continuous in $\boldsymbol{\lambda}$. Then the epi-convergence of $p_m(\cdot)$ to $p(\cdot)$ implies that $c_m(\cdot)$ epi-converges to $c(\cdot)$ as $m \rightarrow \infty$ with probability 1. Consider a lower level set $\mathcal{C} = \{\boldsymbol{\lambda} \in \mathbb{R}^p \mid c(\boldsymbol{\lambda}) \leq c(\mathbf{0}) = n + 1\}$. It can be seen that \mathcal{C} is closed and bounded since $c(\cdot)$ is lower semicontinuous and level-bounded. Thus, $c(\cdot)$ attains its minimum at a point $\boldsymbol{\lambda}_{RET} \in \mathcal{C}$, which is the unique global minimizer by the strict convexity of $c(\cdot)$. With $\min_{\boldsymbol{\lambda}} c(\boldsymbol{\lambda}) = c(\boldsymbol{\lambda}_{RET}) < \infty$, it follows from the basic properties of epi-convergence that $\limsup_{m \rightarrow \infty} \{\epsilon_m\text{-argmin}_{\boldsymbol{\lambda}} c_m(\boldsymbol{\lambda})\} \subset \text{argmin}_{\boldsymbol{\lambda}} c(\boldsymbol{\lambda})$ for any $\epsilon_m \downarrow 0$ as $m \rightarrow \infty$ ([Rockafellar & Wets 2009](#), Theorem 7.31). With probability 1, $\liminf_{m \rightarrow \infty} \{\epsilon_m\text{-argmin}_{\boldsymbol{\lambda}} c_m(\boldsymbol{\lambda})\}$ is nonempty, so the uniqueness of the solution completes the proof. \square

4 Proof of Proposition 3

We fix $\boldsymbol{\theta}$ and consider maximizing

$$\begin{aligned} -D_{KL}(\tilde{P}_{\boldsymbol{\lambda}} \parallel \tilde{P}_n) &= - \int_{\mathcal{X}} \log \left(\frac{\tilde{P}_{\boldsymbol{\lambda}}(d\omega)}{\tilde{P}_n(d\omega)} \right) \tilde{P}_{\boldsymbol{\lambda}}(d\omega) \\ &= - \sum_{i=1}^n p_i \log((n + \tau_n) p_i) - p_c \log \left(\frac{n + \tau_n}{\tau_n} p_c \right) - \frac{p_c}{2} \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}} \boldsymbol{\lambda}, \end{aligned}$$

subject to the moment constraint

$$\sum_{i=1}^n p_i \mathbf{g}_i(\boldsymbol{\theta}) + p_c \mathbb{E}_{\tilde{P}_{\boldsymbol{\lambda}}} [\tilde{\mathbf{g}}_{\boldsymbol{\lambda}}] = \sum_{i=1}^n p_i \mathbf{g}_i(\boldsymbol{\theta}) + p_c (\boldsymbol{\mu}_{n, \boldsymbol{\theta}} + \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}} \boldsymbol{\lambda}) = \mathbf{0}.$$

The Lagrangian associated with the constrained maximization problem is

$$L = - \sum_{i=1}^n p_i \log p_i - p_c \log p_c + p_c \log m - \frac{p_c}{2} \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} \boldsymbol{\lambda} \\ + \boldsymbol{\kappa}^\top \left(\sum_{i=1}^n p_i \mathbf{g}_i(\boldsymbol{\theta}) + p_c (\boldsymbol{\mu}_{n,\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} \boldsymbol{\lambda}) \right) + \nu \left(\sum_{i=1}^n p_i + p_c - 1 \right),$$

where $\boldsymbol{\kappa} \in \mathbb{R}^p$ and $\nu \in \mathbb{R}$ are Lagrange multipliers. Differentiating the Lagrangian expression with respect to each p_i and p_c , and equating the derivatives to zero, we have $\boldsymbol{\kappa} = \boldsymbol{\lambda}$ and

$$\nu = \sum_{i=1}^n p_i \log p_i + p_c \log p_c + p_c \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}} \boldsymbol{\lambda} / 2 - p_c \log \tau_n + 1.$$

After some algebra, it can be shown that

$$p_i(\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\lambda}_{RET}^\top \mathbf{g}_i(\boldsymbol{\theta}))}{c_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})} \quad (1, \dots, n), \quad p_c(\boldsymbol{\theta}) = \frac{p_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})}{c_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})},$$

where $c_n(\boldsymbol{\theta}, \boldsymbol{\lambda}_{RET})$ is the normalizing constant. This leads to solving the dual problem, and the result follows. \square

5 Proof of Theorem 2

We begin by establishing that $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) = O_P(n^{-1/2})$. Observe that

$$\begin{aligned} \frac{1}{n} c_n(\boldsymbol{\theta}_0, \mathbf{0}) &= \frac{1}{n} d_n(\boldsymbol{\theta}_0, \mathbf{0}) + \frac{1}{n} p_n(\boldsymbol{\theta}_0, \mathbf{0}) \\ &= 1 + \frac{\tau_n}{n} \\ &\geq \frac{1}{n} c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)), \end{aligned}$$

where the last inequality follows from the definition of $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)$. We perform a Taylor expansion of $c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))$ around $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) = \mathbf{0}$, yielding

$$\begin{aligned} \frac{1}{n} c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)) &= 1 + \frac{\tau_n}{n} + \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \left(\mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \right) \\ &\quad + \frac{1}{2n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \left(\sum_{i=1}^n \exp\left(\tilde{\boldsymbol{\lambda}}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)\right) \mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)^\top \right) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\quad + \frac{\tau_n}{2n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0}) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0), \end{aligned}$$

where $\tilde{\boldsymbol{\lambda}}$ lies between $\mathbf{0}$ and $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)$. Using the above expansion, we find

$$\begin{aligned} 0 &\geq \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \left(\mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \right) \\ &\quad - \frac{1}{2n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \left(\sum_{i=1}^n \left(-\exp\left(\tilde{\boldsymbol{\lambda}}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)\right) \right) \mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)^\top \right) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\quad + \frac{\tau_n}{2n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0}) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\geq \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \left(\mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \right) \\ &\quad - \frac{1}{2} \max_{1 \leq i \leq n} \left\{ -\exp\left(\tilde{\boldsymbol{\lambda}}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)\right) \right\} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\quad + \frac{\tau_n}{2n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0}) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\geq -|\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)| \left| \mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \right| \\ &\quad - \frac{1}{2} \max_{1 \leq i \leq n} \left\{ -\exp\left(\tilde{\boldsymbol{\lambda}}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)\right) \right\} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\quad + \frac{\tau_n}{4n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0}) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \\ &\geq -|\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)| \left| \mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \right| \\ &\quad + \frac{1}{4} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + \frac{\tau_n}{4n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0}) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0). \end{aligned}$$

The last inequality holds since $\max_{1 \leq i \leq n} \{-\exp(\tilde{\boldsymbol{\lambda}}^\top \mathbf{g}_i(\boldsymbol{\theta}_0))\} < -1/2$ with probability approaching 1 (Newey & Smith 2004, Lemma A1). Let $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) = |\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)| \boldsymbol{\xi}$ with $|\boldsymbol{\xi}| = 1$.

Rearranging the terms in the last inequality gives

$$\frac{1}{4} |\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)| \boldsymbol{\xi}^\top \left(\mathbf{V}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} (\boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} \boldsymbol{\mu}_{n,\boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0}) \right) \boldsymbol{\xi} \leq \left| \mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} \right|.$$

Since $\tau_n n^{-1} (\boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} \boldsymbol{\mu}_{n,\boldsymbol{\theta}_0}^\top + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0}) = O_P(n^{-1})$ by [Condition 5](#), [Condition 3](#) implies that

$$C |\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)| \leq \left| \mathbf{h}_n(\boldsymbol{\theta}_0) + \frac{\tau_n}{n} \boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} \right|$$

for some constant $C > 0$ with probability approaching 1. Thus, we have

$$\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) = O_P(n^{-1/2}). \quad (5)$$

Next, we rewrite the first-order condition for $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)$ as follows:

$$\frac{1}{n} \sum_{i=1}^n \exp(\boldsymbol{\lambda}^\top \mathbf{g}_i(\boldsymbol{\theta}_0)) \mathbf{g}_i(\boldsymbol{\theta}_0) + \frac{1}{n} p_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) (\boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0} \boldsymbol{\lambda}) = \mathbf{0}. \quad (6)$$

By considering [Condition 5](#) and [Equation \(5\)](#), we find that

$$\frac{1}{n} p_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)) (\boldsymbol{\mu}_{n,\boldsymbol{\theta}_0} + \boldsymbol{\Sigma}_{n,\boldsymbol{\theta}_0} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)) = O_P(n^{-1/2}).$$

Expanding the left-hand side of [Equation \(6\)](#) for $\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)$ around $\boldsymbol{\lambda} = \mathbf{0}$, we obtain

$$\mathbf{0} = \mathbf{h}_n(\boldsymbol{\theta}_0) + \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + R_1 + O_P(n^{-1/2}),$$

where

$$|R_1| \leq \frac{C}{n} \sum_{i=1}^n |\mathbf{g}_i(\boldsymbol{\theta}_0)|^3 |\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)|^2$$

for some constant $C > 0$ with probability approaching 1. From [Condition 4](#), it follows that

$R_1 = O_P(n^{-1})$ and

$$\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) = -\mathbf{V}_n(\boldsymbol{\theta}_0)^{-1} \mathbf{h}_n(\boldsymbol{\theta}_0) + O_P(n^{-1/2}). \quad (7)$$

Next, we evaluate the expressions:

$$\begin{aligned} \frac{d_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} &= \frac{1}{n + \tau_n} \sum_{i=1}^n \exp\left(\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0)\right) \\ &= \frac{n}{n + \tau_n} + \frac{n}{n + \tau_n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{h}_n(\boldsymbol{\theta}_0) \\ &\quad + \frac{n}{2(n + \tau_n)} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + R_2, \end{aligned}$$

where $R_2 = O_P(n^{-3/2})$. Similarly,

$$\begin{aligned} \frac{p_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} &= \frac{\tau_n}{n + \tau_n} \exp\left(\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} + \frac{1}{2} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)\right) \\ &= \frac{\tau_n}{n + \tau_n} \left(1 + \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} + \frac{1}{2} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + R_3\right) \end{aligned}$$

with $R_3 = O_P(n^{-1})$. From [Equation \(5\)](#), we get

$$\frac{p_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} = \frac{\tau_n}{n + \tau_n} + O_P(n^{-3/2}).$$

Putting the above expressions together, we have

$$\begin{aligned} \frac{c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} &= 1 + \frac{n}{n + \tau_n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{h}_n(\boldsymbol{\theta}_0) \\ &\quad + \frac{n}{2(n + \tau_n)} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + O_P(n^{-3/2}), \end{aligned}$$

and

$$\begin{aligned}
& \log \left(\frac{c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} \right) \\
&= \frac{n}{n + \tau_n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{h}_n(\boldsymbol{\theta}_0) \\
&\quad + \frac{n}{2(n + \tau_n)} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + O_P(n^{-3/2}) + O_P(n^{-2}) \\
&= \frac{n}{n + \tau_n} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{h}_n(\boldsymbol{\theta}_0) \\
&\quad + \frac{n}{2(n + \tau_n)} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + O_P(n^{-3/2}) \\
&= O_P(n^{-1/2}).
\end{aligned} \tag{8}$$

From Equation (8), it follows that

$$\begin{aligned}
& \log \left(\frac{R_{RET}(\boldsymbol{\theta}_0)}{\tilde{R}_{RET}(\boldsymbol{\theta}_0)} \right) \\
&= \log \left(\frac{n + \tau_n}{\tau_n} p_c(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)) \right) \\
&= \log \left(\frac{n + \tau_n}{c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))} \exp \left(\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} + \frac{1}{2} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) \right) \right) \\
&= \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\mu}_{n, \boldsymbol{\theta}_0} + \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \boldsymbol{\Sigma}_{n, \boldsymbol{\theta}_0} \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) - \log \left(\frac{c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} \right) \\
&= O_P(n^{-1/2}),
\end{aligned}$$

establishing the first result. For the second result, it suffices to show that $-2 \log \tilde{R}_{RET}(\boldsymbol{\theta}_0)$

converges in distribution to χ_p^2 . We have

$$\begin{aligned}
-2 \log \tilde{R}_{RET}(\boldsymbol{\theta}_0) &= -2 \sum_{i=1}^n \log((n + \tau_n) p_i(\boldsymbol{\theta}_0)) \\
&= -2 \sum_{i=1}^n \left(\boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{g}_i(\boldsymbol{\theta}_0) - \log \left(\frac{c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} \right) \right) \\
&= -2n \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{h}_n(\boldsymbol{\theta}_0) + 2n \log \left(\frac{c_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0))}{n + \tau_n} \right).
\end{aligned}$$

Applying Equation (8) and rearranging the terms with Equation (7), we obtain

$$\begin{aligned}
-2 \log \tilde{R}_{RET}(\boldsymbol{\theta}_0) &= -2 \left(\frac{n\tau_n}{n + \tau_n} \right) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{h}_n(\boldsymbol{\theta}_0) \\
&\quad + \left(\frac{n}{n + \tau_n} \right) n \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + o_P(1) \\
&= \left(\frac{n}{n + \tau_n} \right) n \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0) \boldsymbol{\lambda}_{RET}(\boldsymbol{\theta}_0) + o_P(1) \\
&= n \mathbf{h}_n(\boldsymbol{\theta}_0)^\top \mathbf{V}_n(\boldsymbol{\theta}_0)^{-1} \mathbf{h}_n(\boldsymbol{\theta}_0) + o_P(1) \\
&\rightarrow_d \chi_p^2.
\end{aligned}$$

This establishes the second result. \square

6 Proof of Theorem 3

The proof is based on the proofs in Chib et al. (2018, Theorem 2.1), Yiu et al. (2020, Theorem 2), and Yu & Bondell (2023, Lemma 2), with details omitted for brevity. By introducing the local parameter $\mathbf{s} = n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and applying a change of variables, we can express the posterior density as follows:

$$\begin{aligned}
\pi(n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \mid \mathcal{D}_n) &= \frac{\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s})}{\int \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) d\mathbf{s}} \\
&= \frac{\pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) \exp(\log L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) - \log L_{RET}(\boldsymbol{\theta}_0))}{\int \pi(\boldsymbol{\theta}_0 + n^{-1/2}\tilde{\mathbf{s}}) \exp(\log L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2}\tilde{\mathbf{s}}) - \log L_{RET}(\boldsymbol{\theta}_0)) d\tilde{\mathbf{s}}}.
\end{aligned}$$

We define $C_n = \int \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) \exp(\log L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) - \log L_{RET}(\boldsymbol{\theta}_0)) d\mathbf{s}$ and $f(\mathbf{s}) = (2\pi)^{-p/2} |\boldsymbol{\Omega}|^{-1/2} \exp(-\mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} / 2)$. Using Scheffé's lemma, our goal is to show that

$$\int \left| C_n^{-1} \pi(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s}) \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - f(\mathbf{s}) \right| d\mathbf{s} \rightarrow_p 0.$$

We observe that

$$\int \left| C_n^{-1} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - f(\mathbf{s}) \right| d\mathbf{s} \leq C_n^{-1} (I_1 + I_2),$$

where

$$I_1 = \int \left| \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \right| d\mathbf{s}$$

and

$$I_2 = \int \left| \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) - C_n f(\mathbf{s}) \right| d\mathbf{s}.$$

Then, it suffices to show that $I_1 \rightarrow_p 0$, which implies $C_n \rightarrow_p \pi(\boldsymbol{\theta}_0)(2\pi)^{p/2} |\boldsymbol{\Omega}|^{1/2}$ and $I_2 \rightarrow_p 0$.

Let $\delta > 0$ and $c > 0$. We partition the integration domain into three subsets: $A_1 = \{\mathbf{s} : |\mathbf{s}| > \delta n^{1/2}\}$, $A_2 = \{\mathbf{s} : c \log n^{1/2} < |\mathbf{s}| \leq \delta n^{1/2}\}$, and $A_3 = \{\mathbf{s} : |\mathbf{s}| \leq c \log n^{1/2}\}$. We begin with A_1 , where we have

$$\begin{aligned} & \int_{A_1} \left| \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \right| d\mathbf{s} \\ & \leq \int_{A_1} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \exp(\log L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) - \log L_{RET}(\boldsymbol{\theta}_0)) d\mathbf{s} \\ & \quad + \int_{A_1} \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) d\mathbf{s} \\ & \leq \int_{A_1} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \exp\left(n \sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq n^{-1/2} |\mathbf{s}|} \frac{1}{n} (\log L_{RET}(\boldsymbol{\theta}) - \log L_{RET}(\boldsymbol{\theta}_0))\right) d\mathbf{s} \\ & \quad + \int_{A_1} \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) d\mathbf{s} \\ & \leq \int_{A_1} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \exp\left(n \sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \delta} \frac{1}{n} (\log L_{RET}(\boldsymbol{\theta}) - \log L_{RET}(\boldsymbol{\theta}_0))\right) d\mathbf{s} \\ & \quad + \int_{A_1} \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) d\mathbf{s}. \end{aligned}$$

On the right-hand side of the last inequality above, the second integral goes to zero due to the properties of normal distributions. The first integral converges to zero in probability by [Condition 7](#).

We now focus on A_2 and express the integral as

$$\int_{A_2} \left| \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - \pi(\boldsymbol{\theta}_0) \exp \left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} \right) \right| d\mathbf{s} \leq T_1 + T_2,$$

where

$$T_1 = \int_{A_2} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \exp(\log L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) - \log L_{RET}(\boldsymbol{\theta}_0)) d\mathbf{s}$$

and

$$T_2 = \int_{A_2} \pi(\boldsymbol{\theta}_0) \exp \left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} \right) d\mathbf{s}.$$

Denoting $\sigma_{\min} > 0$ as the smallest eigenvalue of $\boldsymbol{\Omega}^{-1}$, for sufficiently large n and some constant $C > 0$, it follows that

$$\begin{aligned} T_2 &\leq \pi(\boldsymbol{\theta}_0) \int_{A_2} \exp(-\sigma_{\min} |\mathbf{s}|^2 / 2) d\mathbf{s} \\ &\leq \pi(\boldsymbol{\theta}_0) \exp \left(-\sigma_{\min} (c \log n^{1/2})^2 / 2 \right) \text{vol}(A_2) \\ &\leq \pi(\boldsymbol{\theta}_0) \exp(-\sigma_{\min} c^2 \log n / 4) \text{vol}(A_2) \\ &\leq C \pi(\boldsymbol{\theta}_0) n^{p/2 - \sigma_{\min} c^2 / 4}. \end{aligned}$$

As a result, $T_2 \rightarrow 0$ for sufficiently large c . Regarding T_1 , employing a Taylor expansion argument ([Chib et al. 2018](#), Lemma C.2) for $\log L_{RET}(\boldsymbol{\theta})$, combined with [Condition 5](#), leads to

$$\log L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) - \log L_{RET}(\boldsymbol{\theta}_0) = -\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} + R_n(\mathbf{s}),$$

where it can be shown that $R_n(\mathbf{s}) = O_P((|\mathbf{s}| + |\mathbf{s}|^2)n^{-1/2})$. Thus, there exists a constant $C > 0$ such that $|R_n(\mathbf{s})| \leq C(|\mathbf{s}| + |\mathbf{s}|^2)n^{-1/2}$ with arbitrarily high probability for large n .

For any $\delta_n \downarrow 0$:

$$\begin{aligned} \sup_{|\mathbf{s}| \leq \delta_n n^{1/2}} \frac{|R_n(\mathbf{s})|}{1 + |\mathbf{s}|^2} &\leq \sup_{|\mathbf{s}| \leq \delta_n n^{1/2}} \frac{C(|\mathbf{s}| + |\mathbf{s}|^2)n^{-1/2}}{1 + |\mathbf{s}|^2} \\ &\leq \sup_{|\mathbf{s}| \leq \delta_n n^{1/2}} \frac{2C|\mathbf{s}|}{n^{1/2}} \\ &\leq 2C\delta_n. \end{aligned}$$

For any $\epsilon > 0$ and $\eta > 0$, the results in [Andrews \(1994\)](#) imply that there exists $\delta > 0$ such that

$$\limsup_{n \rightarrow \infty} P \left(\sup_{|\mathbf{s}| \leq \delta n^{1/2}} \frac{|R_n(\mathbf{s})|}{1 + |\mathbf{s}|^2} > \epsilon \right) < \eta.$$

Moreover, this stochastic equicontinuity condition implies, as shown in [Chernozhukov & Hong \(2003\)](#), that

$$\limsup_{n \rightarrow \infty} P \left(\sup_{c \log n^{1/2} < |\mathbf{s}| \leq \delta n^{1/2}} \frac{|R_n(\mathbf{s})|}{|\mathbf{s}|^2} > \epsilon \right) < \eta$$

and

$$\limsup_{n \rightarrow \infty} P \left(\sup_{|\mathbf{s}| \leq c \log n^{1/2}} |R_n(\mathbf{s})| > \epsilon \right) = 0 \tag{9}$$

for some $c > 0$. Therefore, $|R_n(\mathbf{s})| \leq \sigma_{\min} |\mathbf{s}|^2/4$ for all $\mathbf{s} \in A_2$ with arbitrarily high probability for large n , and

$$\begin{aligned} T_1 &= \int_{A_2} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \exp \left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} + R_n(\mathbf{s}) \right) d\mathbf{s} \\ &\leq \sup_{\mathbf{s} \in A_2} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \int_{A_2} \exp(-\sigma_{\min} |\mathbf{s}|^2/2 + |R_n(\mathbf{s})|) d\mathbf{s} \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}) \int_{A_2} \exp(-\sigma_{\min} |\mathbf{s}|^2/4) d\mathbf{s}. \end{aligned}$$

Similar to T_2 , it follows from [Conditions 1](#) and [6](#) that $T_1 \rightarrow_p 0$.

Finally, we express the integral over A_3 as

$$\int_{A_3} \left| \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - \pi(\boldsymbol{\theta}_0) \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \right| d\mathbf{s} \leq T_3 + T_4,$$

where

$$T_3 = \int_{A_3} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left| \left(\frac{L_{RET}(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s})}{L_{RET}(\boldsymbol{\theta}_0)} \right) - \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \right| d\mathbf{s}$$

and

$$T_4 = \int_{A_3} |\pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) - \pi(\boldsymbol{\theta}_0)| \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) d\mathbf{s}.$$

We have $|\pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) - \pi(\boldsymbol{\theta}_0)| \exp(-\mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}/2) \rightarrow 0$ for any $\mathbf{s} \in A_1$, and

$$\sup_{\mathbf{s} \in A_3} |\pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) - \pi(\boldsymbol{\theta}_0)| \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \leq 2 \sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}),$$

which implies that $T_4 \rightarrow 0$. Moving on,

$$\begin{aligned} T_3 &= \int_{A_3} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \left| \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} + R_n(\mathbf{s})\right) - \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \right| d\mathbf{s} \\ &\leq \sup_{\mathbf{s} \in A_3} \pi(\boldsymbol{\theta}_0 + n^{-1/2} \mathbf{s}) \int_{A_3} \left| \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s} + R_n(\mathbf{s})\right) - \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) \right| d\mathbf{s} \\ &\leq \sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}) \int_{A_3} \exp\left(-\frac{1}{2} \mathbf{s}^\top \boldsymbol{\Omega}^{-1} \mathbf{s}\right) |\exp(R_n(\mathbf{s})) - 1| d\mathbf{s}. \end{aligned}$$

From [Equation \(9\)](#), we deduce that $\sup_{\mathbf{s} \in A_1} R_n(\mathbf{s}) \rightarrow_p 0$ and, consequently, that $T_3 \rightarrow_p 0$.

This completes the proof. \square

7 Proof of Proposition 4

It follows from the compactness of Θ under [Condition 1](#) and Lemma 1 of [Berger et al. \(2009\)](#)

that $I(\pi \mid \mathcal{M}_2) < \infty$. We write

$$\begin{aligned}
 I(\pi \mid \mathcal{M}_2) &= \int_{\mathcal{X}} \int_{\mathcal{X}} D_{KL}(\pi(\cdot \mid \mathbf{x}_1, \mathbf{x}_2) \parallel \pi(\cdot)) m(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2 \\
 &= \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\Theta} \pi(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{x}_2) \log \left(\frac{\pi(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{x}_2)}{\pi(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} m(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2 \\
 &= \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\Theta} \pi(\boldsymbol{\theta}) p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\theta})}{m(\mathbf{x}_1, \mathbf{x}_2)} \right) d\boldsymbol{\theta} d\mathbf{x}_1 d\mathbf{x}_2 \\
 &= \int_{\Theta} \pi(\boldsymbol{\theta}) \int_{\mathcal{X}} \int_{\mathcal{X}} p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\theta})}{m(\mathbf{x}_1, \mathbf{x}_2)} \right) d\mathbf{x}_1 d\mathbf{x}_2 d\boldsymbol{\theta}
 \end{aligned}$$

and

$$I(\pi \mid \mathcal{M}_1) = \int_{\Theta} \pi(\boldsymbol{\theta}) \int_{\mathcal{X}} p(\mathbf{x}_1 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_1 \mid \boldsymbol{\theta})}{m(\mathbf{x}_1)} \right) d\mathbf{x}_1 d\boldsymbol{\theta}.$$

Let

$$A_1 = \int_{\mathcal{X}} p(\mathbf{x}_1 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_1 \mid \boldsymbol{\theta})}{m(\mathbf{x}_1)} \right) d\mathbf{x}_1.$$

and

$$A_2 = \int_{\mathcal{X}} \int_{\mathcal{X}} p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\theta})}{m(\mathbf{x}_1, \mathbf{x}_2)} \right) d\mathbf{x}_1 d\mathbf{x}_2$$

It suffices to show that $A_1 \leq A_2$. To this end,

$$\begin{aligned}
 A_2 &= \int_{\mathcal{X}} \int_{\mathcal{X}} p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta}) p(\mathbf{x}_1 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta}) p(\mathbf{x}_1 \mid \boldsymbol{\theta})}{m(\mathbf{x}_2 \mid \mathbf{x}_1) m(\mathbf{x}_1)} \right) d\mathbf{x}_2 d\mathbf{x}_1 \\
 &= \int_{\mathcal{X}} \int_{\mathcal{X}} p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta}) p(\mathbf{x}_1 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta})}{m(\mathbf{x}_2 \mid \mathbf{x}_1)} \right) d\mathbf{x}_2 d\mathbf{x}_1 \\
 &\quad + \int_{\mathcal{X}} \int_{\mathcal{X}} p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta}) p(\mathbf{x}_1 \mid \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_1 \mid \boldsymbol{\theta})}{m(\mathbf{x}_1)} \right) d\mathbf{x}_2 d\mathbf{x}_1 \\
 &= \int_{\mathcal{X}} p(\mathbf{x}_1 \mid \boldsymbol{\theta}) \int_{\mathcal{X}} p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta})}{m(\mathbf{x}_2 \mid \mathbf{x}_1)} \right) d\mathbf{x}_2 d\mathbf{x}_1 + A_1.
 \end{aligned}$$

Thus, by Jensen's inequality, we have

$$\int_{\mathcal{X}} p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta}) \log \left(\frac{p(\mathbf{x}_2 \mid \mathbf{x}_1, \boldsymbol{\theta})}{m(\mathbf{x}_2 \mid \mathbf{x}_1)} \right) d\mathbf{x}_2 \geq 0,$$

which implies $A_1 \leq A_2$. □

8 Quantile-Quantile Plots

Figures 1–4 show quantile-quantile plots comparing the distribution of H to $U(0, 1)$ from the simulations in Section 4.1 of the main paper.

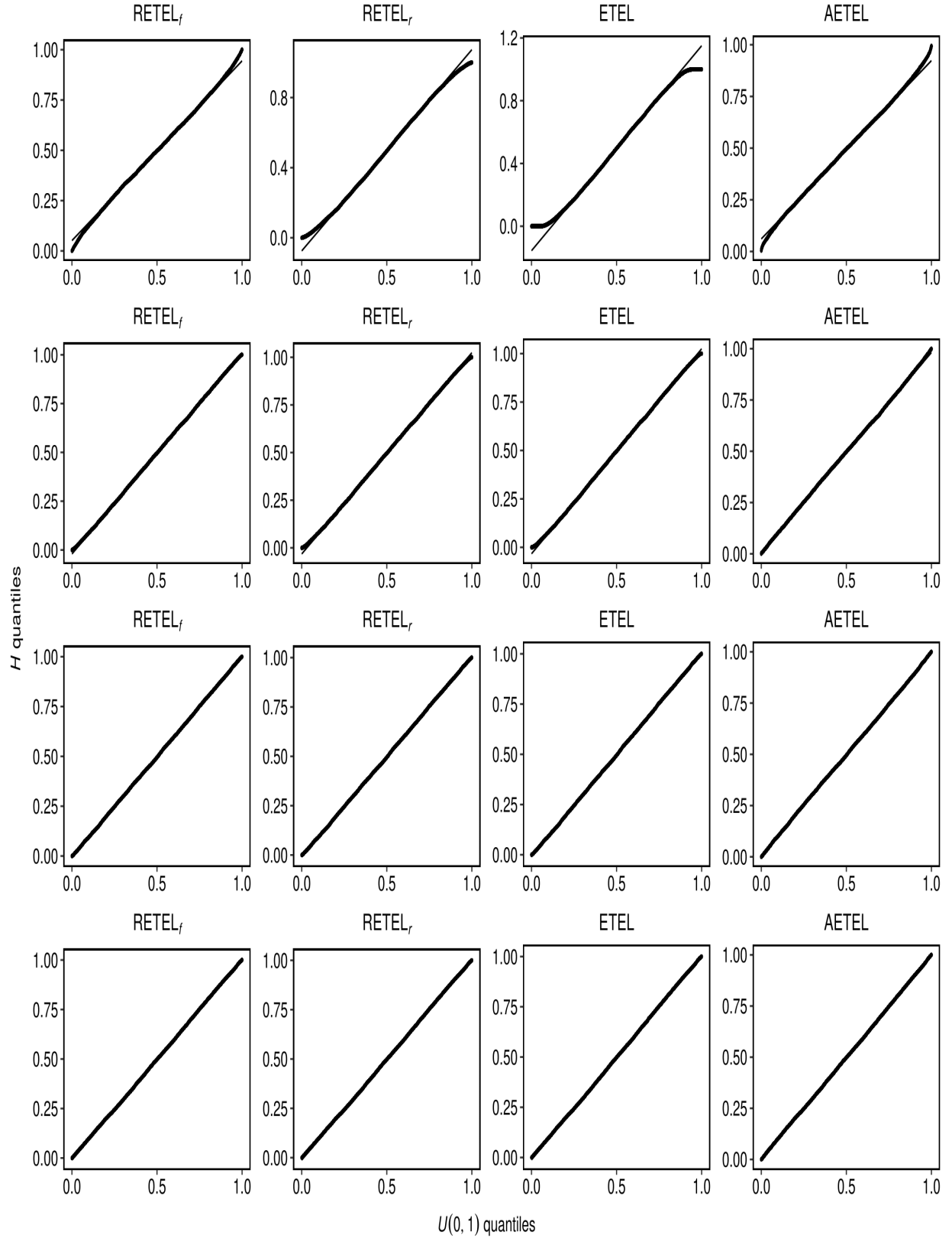


Figure 1. Quantile-quantile plots for the distribution of H versus $U(0, 1)$ under $s = 1$ and $\tau_n = 1$ for different sample sizes ($n = 5$ in the first row, $n = 20$ in the second row, $n = 50$ in the third row, and $n = 100$ in the fourth row).

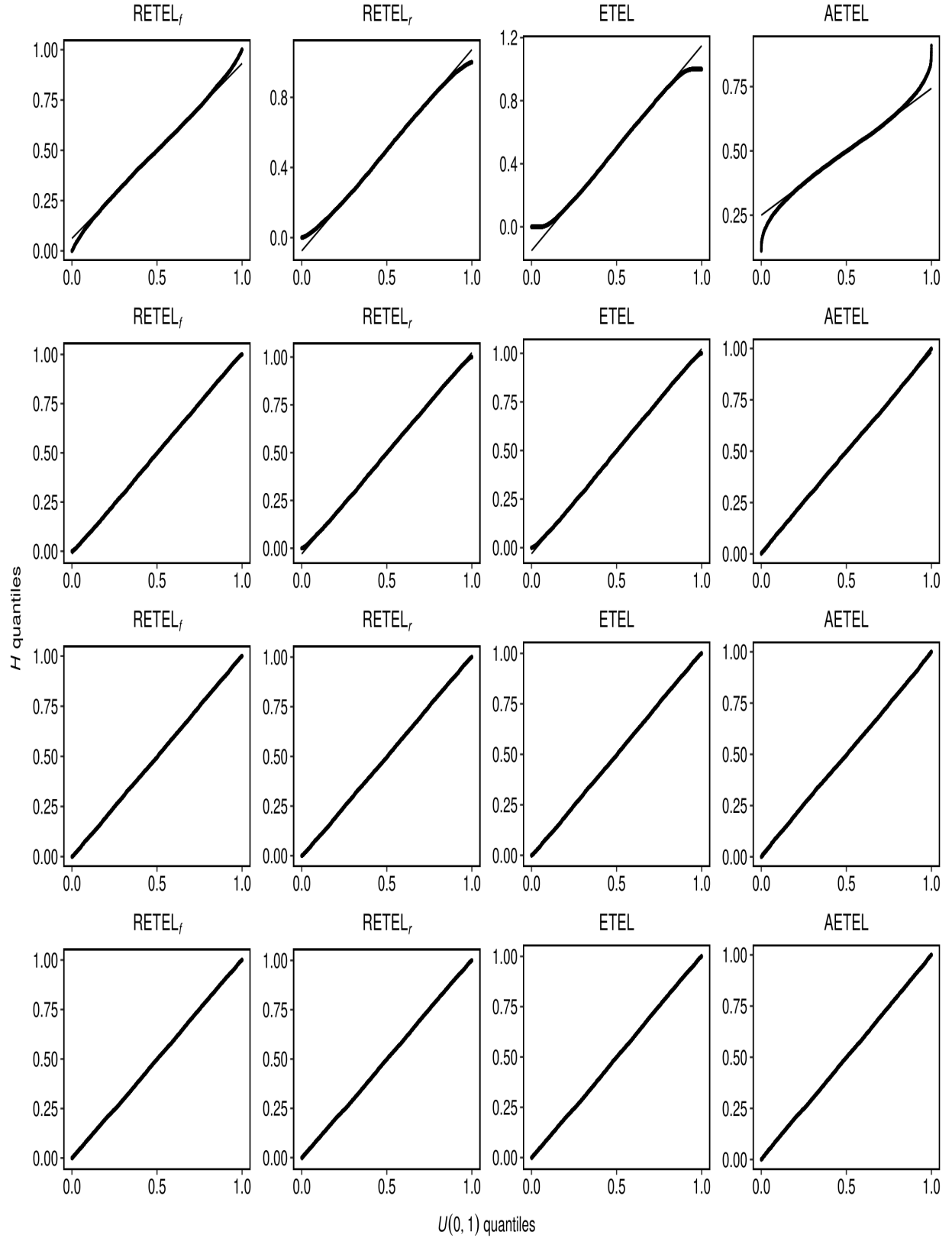


Figure 2. Quantile-quantile plots for the distribution of H versus $U(0, 1)$ under $s = 5$ and $\tau_n = 1$ for different sample sizes ($n = 5$ in the first row, $n = 20$ in the second row, $n = 50$ in the third row, and $n = 100$ in the fourth row).

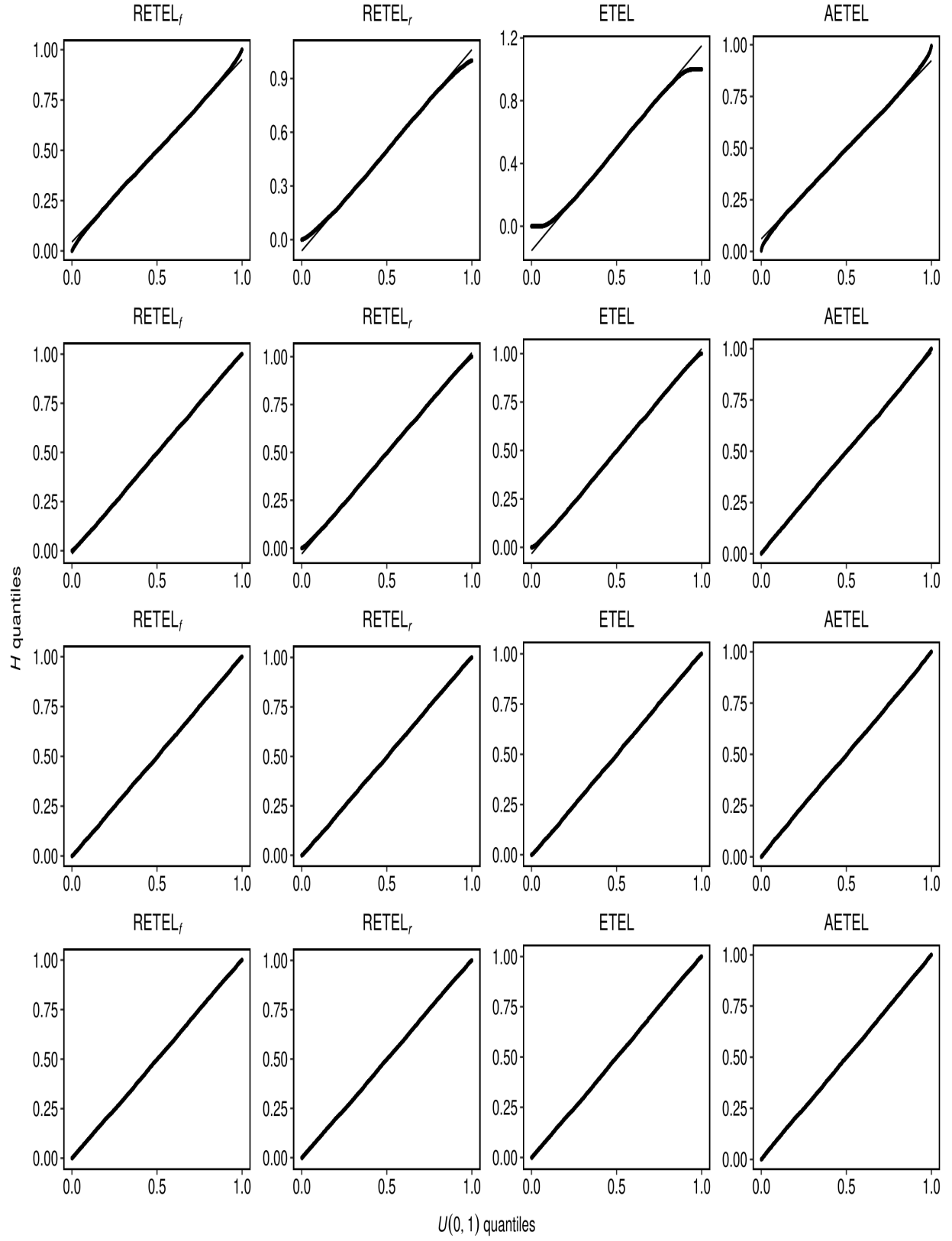


Figure 3. Quantile-quantile plots for the distribution of H versus $U(0, 1)$ under $s = 1$ and $\tau_n = \log n$ for different sample sizes ($n = 5$ in the first row, $n = 20$ in the second row, $n = 50$ in the third row, and $n = 100$ in the fourth row).

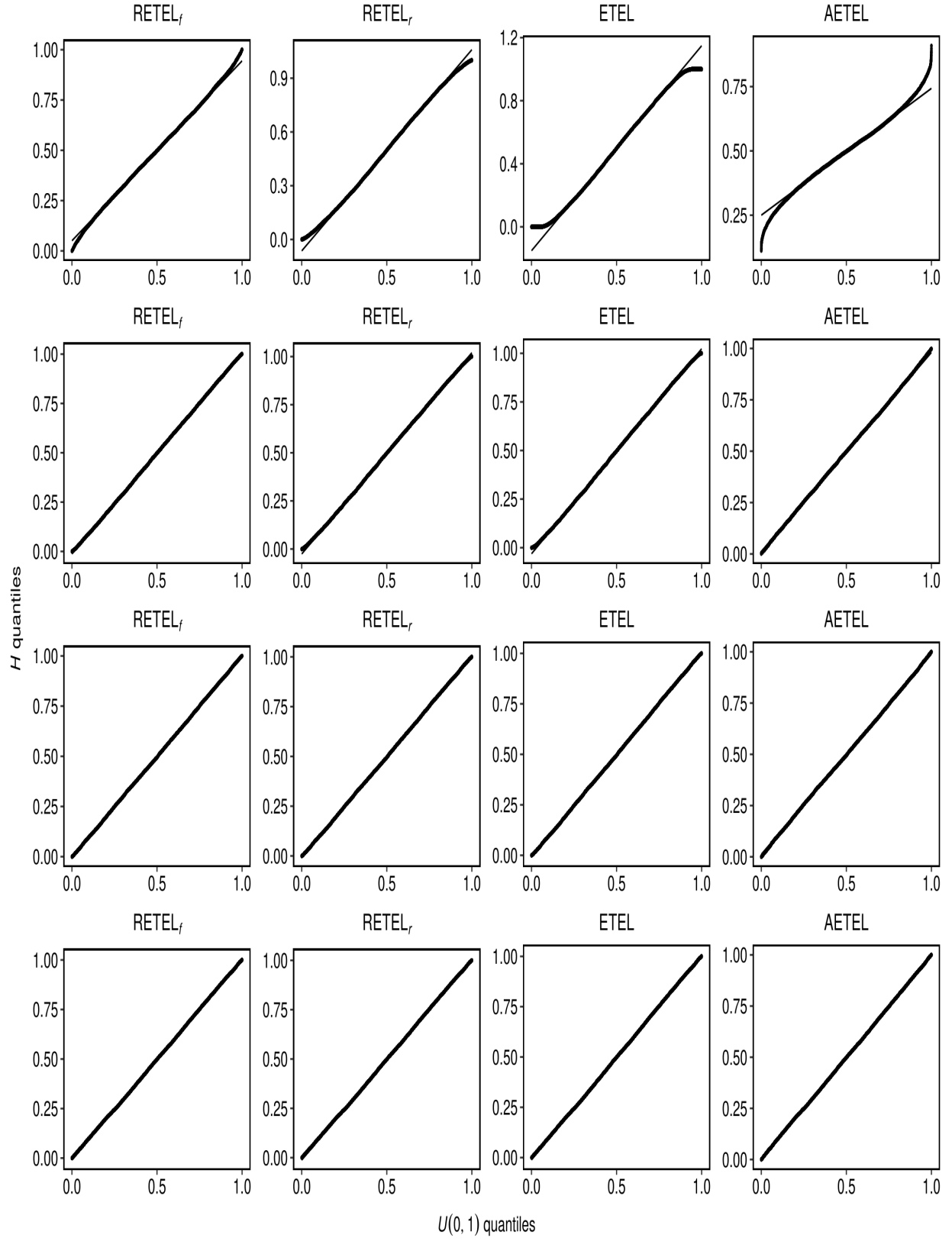


Figure 4. Quantile-quantile plots for the distribution of H versus $U(0, 1)$ under $s = 5$ and $\tau_n = \log n$ for different sample sizes ($n = 5$ in the first row, $n = 20$ in the second row, $n = 50$ in the third row, and $n = 100$ in the fourth row).

References

- Andrews, D. W. K. (1994), Empirical process methods in econometrics, *in* ‘Handbook of Econometrics’, Vol. 4, Elsevier, pp. 2247–2294.
- Berger, J. O., Bernardo, J. M. & Sun, D. (2009), ‘The formal definition of reference priors’, *The Annals of Statistics* **37**, 905–938.
- Chernozhukov, V. & Hong, H. (2003), ‘An MCMC approach to classical estimation’, *Journal of Econometrics* **115**(2), 293–346.
- Chib, S., Shin, M. & Simoni, A. (2018), ‘Bayesian estimation and comparison of moment condition models’, *Journal of the American Statistical Association* **113**, 1656–1668.
- Jacod, J. & Sørensen, M. (2018), ‘A review of asymptotic theory of estimating functions’, *Statistical Inference for Stochastic Processes* **21**, 415–434.
- Newey, W. K. & Smith, R. J. (2004), ‘Higher order properties of GMM and generalized empirical likelihood estimators’, *Econometrica* **72**, 219–255.
- Rockafellar, R. T. & Wets, R. J.-B. (2009), *Variational analysis*, Springer-Verlag.
- Yiu, A., Goudie, R. J. B. & Tom, B. D. M. (2020), ‘Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood’, *Biometrika* **107**, 857–873.
- Yu, W. & Bondell, H. D. (2023), ‘Variational Bayes for fast and accurate empirical likelihood inference’, *Journal of the American Statistical Association* **0**, 1–13.
- Zhu, H., Zhou, H., Chen, J., Li, Y., Lieberman, J. & Styner, M. (2009), ‘Adjusted exponentially tilted likelihood with applications to brain morphology’, *Biometrics* **65**(3), 919–927.