Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **87**, 1127–1149
https://doi.org/10.1093/jrsssb/qkaf009
Advance access publication 6 March 2025
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series B Statistical Methodology

Bayesian penalized empirical likelihood and Markov Chain Monte Carlo sampling

Jinyuan Chang[1,2], Cheng Yong Tang[3] and Yuanzheng Zhu[4]

[1]Big Data Laboratory on Financial Security and Behavior (MOE Philosophy and Social Sciences Laboratory), Southwestern University of Finance and Economics, Chengdu, Sichuan, China
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[3]Department of Statistics, Operations, and Data Science, Temple University, Philadelphia, PA, USA
[4]Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

Address for correspondence: Yuanzheng Zhu, Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan, China. Email: zhuyz@swufe.edu.cn

Abstract

In this study, we introduce a novel methodological framework called Bayesian penalized empirical likelihood (BPEL), designed to address the computational challenges inherent in empirical likelihood (EL) approaches. Our approach has two primary objectives: (i) to enhance the inherent flexibility of EL in accommodating diverse model conditions, and (ii) to facilitate the use of well-established Markov Chain Monte Carlo sampling schemes as a convenient alternative to the complex optimization typically required for statistical inference using EL. To achieve the first objective, we propose a penalized approach that regularizes the Lagrange multipliers, significantly reducing the dimensionality of the problem while accommodating a comprehensive set of model conditions. For the second objective, our study designs and thoroughly investigates two popular sampling schemes within the BPEL context. We demonstrate that the BPEL framework is highly flexible and efficient, enhancing the adaptability and practicality of EL methods. Our study highlights the practical advantages of using sampling techniques over traditional optimization methods for EL problems, showing rapid convergence to the global optima of posterior distributions and ensuring the effective resolution of complex statistical inference challenges.

Keywords: Bayesian methods, Bernstein–von Mises theorem, estimating equations, MCMC, penalized empirical likelihood

1 Introduction

Empirical likelihood (EL) (Owen, 2001) is a versatile and flexible tool for statistical inference, providing a framework that accommodates broadly defined model conditions. Unlike traditional likelihood approaches, EL does not require the explicit specification of probability distributions governing the data generation process (DGP). This inherent flexibility offers numerous practical advantages, such as the ability to incorporate a wide range of model specifications and prior knowledge, making it highly adaptable for integrating information from multiple data sources. Additionally, EL retains key benefits of its parametric likelihood counterpart, including efficiency (in the semiparametric sense) and the convenience of conducting hypothesis tests and estimating confidence sets through the Wilks-type likelihood ratio framework.

Recent developments in EL approaches have a focus on addressing the challenges posed by complex high-dimensional data. To handle the complexities arising from various model conditions, researchers have explored regularization techniques applied to the Lagrange multipliers associated with EL or the empirical versions of moment conditions, aiming to achieve enhanced model parsimony. In Shi (2016), a two-step procedure is introduced. The first step involves employing a

Received: October 29, 2023. **Revised:** January 17, 2025. **Accepted:** February 12, 2025

'relaxed' EL that incorporates specific inequality constraints in its formulation. The second step includes moment selection and bias correction. Chaussé (2017) addresses a continuum of moment conditions where the numerical optimization problem becomes ill-conditioned. To resolve this, a penalty on the continuous version of the Lagrange multiplier's counterpart is proposed and investigated. Chang et al. (2018) proposes a method to penalize the magnitudes of both the Lagrange multiplier and the model parameters, specifically to tackle high-dimensional model parameters under complex conditions. More recently, Chang et al. (2021) explores the projection of high-dimensional moment conditions onto lower-dimensional spaces to facilitate statistical inference for specific components of model parameters and to assess model specification validity. Besides addressing the challenge of handling many moment conditions, the development of EL approaches that incorporate penalties on model parameters to promote parsimonious structures can effectively manage high-dimensional problems, as discussed in Tang and Leng (2010), Leng and Tang (2012), Chang et al. (2015), and Chang et al. (2023).

The synergy of Bayesian methodologies with traditional likelihoods has consistently demonstrated its effectiveness. Leveraging advances in sampling techniques, Bayesian approaches have established their significance in tackling a wide array of challenges across various domains. This is particularly valuable when dealing with intricate statistical problems where maximizing or even computing the objective function becomes infeasible. The amalgamation of Bayesian principles with EL shows great promise in practical applications. This integration enhances the adaptability and robustness of the Bayesian framework, enabling the creation of statistical models that can accommodate a wide range of scenarios. Recent developments in the realm of Bayesian EL (BEL) methods are evident in a growing body of literature; see Lazar (2003), Rao and Wu (2010), Chaudhuri and Ghosh (2011), Yang and He (2012), Mengersen et al. (2013), Chib et al. (2018), Cheng and Zhao (2019), Zhao et al. (2020), Tang and Yang (2022), and Yu and Bondell (2024).

The class of EL approaches often encounters significant challenges due to substantial computational complexity, which frequently presents barriers in practice. These difficulties primarily arise from the nonconvex nature of the objective function and the potential nonconvexity of its support. As the complexity of the model increases with additional parameters and conditions, these computational obstacles become more severe. Thus, developing computationally efficient strategies is crucial to address these challenges. Indeed, as demonstrated in Chaussé (2017) and related works, solving the associated optimization problem of penalized EL (PEL) can be a dauntingly difficult task. In our study, we demonstrate that, when combined with the Bayesian framework, sampling schemes offer promising alternatives. Once successfully drawn, samples from the posterior distribution can be used to develop the estimator.

In recent research, sampling techniques, often perceived as computationally demanding alternatives to optimization methods, demonstrate remarkable efficiency in approximating target distributions, outperforming optimization alternatives in handling nonconvex problems; see Ma et al. (2019). While sampling techniques offer a promising approach within the framework of BEL, there exist numerous challenges associated with devising these computational schemes. On one hand, EL has the potential to leverage information from various model conditions, leading to more precise estimates of unknown model parameters. However, the inclusion of a large number of these conditions introduces additional complexities, both in theory and practical implementation. Indeed, the dimensionality of the problem remains a central obstacle in EL approaches, as elaborated in Hjort et al. (2009). Furthermore, the incorporation of an increasing number of moment conditions can substantially amplify the nonconvex nature of the associated optimization problems, making the development of an effective sampling scheme increasingly more challenging. As underscored in Chaudhuri et al. (2017), traditional Markov Chain Monte Carlo (MCMC) techniques encounter significant hurdles when applied to BEL due to the intricate and nonconvex characteristics of the parameter space in which new samples are generated.

Our research aims to establish an innovative methodological framework, guided by two primary objectives: (i) our approach maintains the inherent flexibility and adaptability of EL, allowing for the incorporation of broad model conditions; and (ii) our framework provides convenient access to well-established MCMC computing schemes, streamlining practical implementations. To address the first objective and mitigate challenges stemming from numerous model conditions, we propose a penalized approach. By penalizing the magnitudes of the Lagrange multipliers used

in evaluating EL at specific model parameter values, we create an effective mechanism similar to moment selection. This approach reduces the problem's dimensionality while still leveraging the potential efficiency gains from a comprehensive set of model conditions. For the second objective, our approach effectively overcomes the obstacles associated with devising sampling schemes for applying Bayesian approaches, thanks to the efficient dimensionality reduction achieved through PEL. In our study, we demonstrate the practicality of our framework using two well-established sampling methods: the popular Metropolis–Hastings (M-H) sampling and the influential adaptive multiple importance sampling (AMIS) technique for approximate Bayesian computations.

Our study makes several noteworthy contributions, in addition to the methodological advancement mentioned earlier. On a theoretical level, our analysis establishes the properties of the BPEL estimator, allowing for an exponentially increasing number of model conditions, thereby enabling unprecedented adaptability in practical applications. Furthermore, we develop theory that guarantees the convergence of the two showcased sampling schemes, thereby ensuring the validity of BPEL in statistical inference. Our study reinforces the observations made in a recent study by Ma et al. (2019) that sampling techniques offer compelling alternatives to optimization methods in addressing computationally demanding problems. Our theoretical results and numerical studies demonstrate that sampling schemes converge rapidly to stationary distributions centred around the true global optimizer. In contrast, optimization methods often require more time and can become trapped at local peaks, limiting their ability to locate the true optimum.

The rest of this article are structured as follows. Section 2 delves into the framework of BPEL and introduces two MCMC algorithms. Numerical studies and real data analysis for an international trade dataset are presented in Sections 3 and 4, respectively. Section 5 comprehensively develops the properties and theoretical guarantees of the proposed methods. Some discussions are provided in Section 6, while all technical proofs are available in the online supplementary material. The code for implementing our proposed methods is available at the GitHub repository: https://github.com/JinyuanChang-Lab/BayesianPenalizedEL.

1.1 Notation

For any positive integer q, write $[q] = \{1, \ldots, q\}$ and let $\mathbf{I}_q$ be the $q \times q$ identity matrix. Denote by $I(\cdot)$ the indicator function. Let $\mathrm{vech}(\cdot)$ be an operator that stacks the columns of the lower triangular part of its argument square matrix. For a q-dimensional vector $\mathbf{a} = (a_1, \ldots, a_q)^\top$, we use $|\mathbf{a}|_2 = (\sum_{i=1}^q a_i^2)^{1/2}$ and $\mathrm{supp}(\mathbf{a}) = \{i \in [q] : a_i \neq 0\}$ to denote its L_2-norm and support, respectively. Let $\mathcal{U}(a, b)$ be the uniform distribution among (a, b), and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Denote by $\mathcal{T}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate Student's distribution with k degrees of freedom, mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$. For two positive real-valued sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n \leq c_0$ for some positive constant c_0, $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously, and $a_n \ll b_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$.

2 Methodology

2.1 Penalized empirical likelihood

Let $\mathcal{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ represent a set of d-dimensional independent and identically distributed observations, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \boldsymbol{\Theta}$ be a p-dimensional parameter. Here, the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^p$ is a compact set. The information regarding the model parameter $\boldsymbol{\theta}$ is gathered through a set of unbiased moment conditions $\mathbb{E}\{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta}_0)\} = \mathbf{0}$, where $\mathbf{g}(\cdot\,;\cdot) = \{g_1(\cdot\,;\cdot), \ldots, g_r(\cdot\,;\cdot)\}^\top \in \mathbb{R}^r$ is referred to as the estimating function, and the true, yet unknown value $\boldsymbol{\theta}_0$ is situated within the interior of $\boldsymbol{\Theta}$.

In existing studies, it has been typically required that $r \geq p$ for the identification of $\boldsymbol{\theta}_0$. When p and r are fixed constants, the EL with the estimating function $\mathbf{g}(\cdot\,;\cdot)$ considered in Qin and Lawless (1994) can be formulated as

$$\mathrm{EL}(\boldsymbol{\theta}) = \exp\left[-n\log n - \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})} \sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}\right], \tag{1}$$

where $\hat{\Lambda}_n(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^r : \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta}) \in \mathcal{V} \text{ for any } i \in [n]\}$ with an open interval $\mathcal{V}$ containing zero. The standard EL estimator for $\boldsymbol{\theta}_0$ is defined as $\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{EL}(\boldsymbol{\theta})$, which is equivalent to solving the corresponding dual problem:

$$\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}. \tag{2}$$

The estimator $\tilde{\boldsymbol{\theta}}_n$ exhibits several desirable properties: (i) it is $\sqrt{n}$-consistent, (ii) it possesses asymptotic normality, and (iii) it attains the semiparametric efficiency bound of Godambe and Heyde (1987). However, in high-dimensional scenarios, the literature has highlighted the challenge of accommodating a diverging r. This issue is discussed in works such as Donald et al. (2003), Chen et al. (2009), Hjort et al. (2009), Leng and Tang (2012) and Chang et al. (2015). To elaborate, it is generally required that $r \ll n^{1/2}$ for the consistency and $r \ll n^{1/3}$ for the asymptotic normality of $\tilde{\boldsymbol{\theta}}_n$. These constraints on the diverging rate of r pose significant challenges when dealing with high-dimensional estimating equations.

To address scenarios where $r \gg n$ and p remains fixed, we investigate the PEL estimator for $\boldsymbol{\theta}_0$ as follows:

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\} - n \sum_{j=1}^{r} P_v(|\lambda_j|) \right], \tag{3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^\top$, and $P_v(\cdot)$ is a penalty function with the tuning parameter v. Given a penalty function $P_v(\cdot)$ with the tuning parameter v, we define $\rho(t; v) = v^{-1} P_v(t)$ for $t \in [0, \infty)$ and $v \in (0, \infty)$. For $P_v(\cdot)$ in (3), we consider the following class of penalty functions:

$$\begin{aligned} \mathscr{P} = \{P_v(\cdot) : &\rho(t; v) \text{ is increasing in } t \in [0, \infty) \text{ and has continuous derivative} \\ &\rho'(t; v) \quad \text{for any } t \in (0, \infty) \text{ with } \rho'(0^+; v) \in (0, \infty), \text{ where} \\ &\rho'(0^+; v) \text{ is independent of } v\}. \end{aligned} \tag{4}$$

The class $\mathscr{P}$ is broad and general, encompassing commonly used penalty functions. Theorem 1 in Section 5.1 demonstrates that the PEL estimator $\hat{\boldsymbol{\theta}}_n$ follows an asymptotically normal distribution and accommodates exponentially diverging r with respect to n.

To practically implement (3), we encounter a two-layer optimization problem for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\lambda} \in \mathbb{R}^r$. Let

$$f_n(\boldsymbol{\lambda}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\} - \sum_{j=1}^{r} P_v(|\lambda_j|). \tag{5}$$

Since $n^{-1} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}$ is concave in $\boldsymbol{\lambda}$, the inner optimization layer of (3), which seeks $\boldsymbol{\lambda}$ given $\boldsymbol{\theta}$ by maximizing $f_n(\boldsymbol{\lambda}; \boldsymbol{\theta})$, can be efficiently implemented even for large r when $P_v(\cdot)$ is chosen as a convex function, such as the L_1 penalty. The main challenge is the outer optimization layer of (3), which seeks the optimizer $\hat{\boldsymbol{\theta}}_n$. This is difficulty due to the nonconvex nature of the problem, making it NP-hard to find global minima (Jain & Kar, 2017). As a result, this complexity often leads to computational inefficiency and a higher likelihood of converging to local optima.

2.2 Bayesian penalized empirical likelihood

We are motivated to explore an alternative approach using sampling techniques to solve the nonconvex problem associated with PEL. Indeed, as an efficient alternative for addressing nonconvex optimization problems, Ma et al. (2019) has demonstrated that solving these issues with MCMC techniques can yield highly effective results. Their findings indicate that the computational complexity of sampling algorithms exhibits linear scalability with the model dimension, in contrast to the exponential scaling of optimization algorithms in nonconvex settings.

Applying sampling techniques to EL in conjunction with a Bayesian framework emerges as a compelling approach. For EL($\boldsymbol{\theta}$) defined as (1), let $\pi_0(\cdot)$ represent a prior distribution for $\boldsymbol{\theta}$. Then, the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{X}_n)$ is proportional to $\pi_0(\boldsymbol{\theta}) \times \text{EL}(\boldsymbol{\theta})$. In cases where r and p are fixed constants, $\pi(\boldsymbol{\theta} \mid \mathcal{X}_n)$ converges to a Gaussian distribution with mean being the standard EL estimator $\tilde{\boldsymbol{\theta}}_n$ defined as (2). Consequently, when samples are successfully drawn from the posterior distribution, their sample mean can serve as an estimator for $\boldsymbol{\theta}_0$.

As the model's complexity increases, BEL faces challenges. In this study, we explore a scenario with high-dimensional model conditions ($r \gg n$), while keeping p fixed. The flexibility by allowing large number r also brings significant challenges. For example, as demonstrated in Tsao (2004), as $n \to \infty$, $\mathbb{P}\{\text{EL}(\boldsymbol{\theta}) = 0\} \to 1$ for any $\boldsymbol{\theta}$ in a small neighbourhood of $\boldsymbol{\theta}_0$ if $r/n \geq 0.5$. Such degeneration renders EL($\boldsymbol{\theta}$) inapplicable in this scenario. To handle diverging r, we propose to replace EL($\boldsymbol{\theta}$) by

$$\text{PEL}_v(\boldsymbol{\theta}) = \exp\left(- n \log n - \max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\} - n \sum_{j=1}^{r} P_v(|\lambda_j|) \right] \right), \tag{6}$$

where $P_v(\cdot)$ is a penalty function with the tuning parameter v. Since adding the penalty term $P_v(\cdot)$ encourages sparse Lagrange multiplier $\boldsymbol{\lambda}$, the PEL effectively performs a selection of the model conditions at each given $\boldsymbol{\theta}$. We then consider the BPEL with a prior distribution $\pi_0(\cdot)$, which leads to a posterior distribution defined as

$$\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n) \propto \pi_0(\boldsymbol{\theta}) \times \text{PEL}_v(\boldsymbol{\theta}) \times I(\boldsymbol{\theta} \in \boldsymbol{\Theta}). \tag{7}$$

Our BPEL connects with and differs from the so-called Gibbs posterior in the literature of Bayesian methods (Bissiri et al., 2016; Tang & Yang, 2022). On one hand, they share a common foundation with the Gibbs posterior in that both are built upon generic loss functions. The key difference lies in the device each utilizes: EL employs an appropriate multinomial likelihood, $(p_1, \ldots, p_n)$ with $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$, subject to a broad class of model conditions. In contrast, the Gibbs posterior uses a 'pseudo-likelihood' proportional to the exponential loss. Furthermore, the inclusion of the penalty on the Lagrange multiplier helps achieve substantial dimension reduction of the problem, which is key in handling high-dimensional problems with many moment conditions. As shown in our numerical studies in Section 3 and Section A.3 (online supplementary material), MCMC schemes developed from the proposed BPEL demonstrate compelling performance in their finite sample accuracy in approximating the posterior distributions.

Our theory, as elaborated in Section 5.2, establishes the fundamental properties of BPEL. Theorem 2 in Section 5.2 demonstrates that the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$ defined as (7) exhibits a Gaussian limiting distribution centred around the PEL estimator $\hat{\boldsymbol{\theta}}_n$ as defined in (3). Additionally, we define the expected value as

$$\mathbb{E}_{\boldsymbol{\theta} \sim \pi^\dagger}(\boldsymbol{\theta}) = \int_{\mathbb{R}^p} \boldsymbol{\theta} \pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)\, \mathrm{d}\boldsymbol{\theta}. \tag{8}$$

Corollary 1 in Section 5.2 suggests that $\hat{\boldsymbol{\theta}}_n$ can be effectively approximated by $\mathbb{E}_{\boldsymbol{\theta} \sim \pi^\dagger}(\boldsymbol{\theta})$ with an approximation error that diminishes faster than $n^{-1/2}$. This validates the approach to obtain $\hat{\boldsymbol{\theta}}_n$: generating samples from the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$ and then using the associated sample mean to approximate $\hat{\boldsymbol{\theta}}_n$. In Section 2.3, we will introduce two algorithms designed for implementing BPEL.

The impact of prior specification on the properties of resulting estimators is a notable area of research. For instance, Vexler et al. (2014) explores this in the context of EL. In various scenarios, the choice of prior can enhance desirable properties of the estimator derived from the posterior distribution, such as sparsity, as discussed in Narisetty and He (2014), Castillo et al. (2015), and Ouyang and Bondell (2023). Given the two primary goals of our study—developing BPEL and investigating it with MCMC—we use a noninformative prior in our numerical demonstrations. As detailed in Section A.1 (online supplementary material), we examined the effects of different prior specifications. The overall finding is intuitive: when the prior is specified closer to the

true value, the resulting estimator performs better compared to using a noninformative prior. Conversely, if the prior is specified further from the true value, the performance of the estimator deteriorates and becomes less competitive.

2.3 MCMC algorithms

2.3.1 Algorithm 1

In recent decades, MCMC sampling methods have achieved significant success and have garnered influential applications across diverse fields. For an extensive overview of this body of work, we refer to the monograph by Brooks et al. (2011) and reference therein. The M-H algorithm family plays a central role in the practical implementation of MCMC techniques, serving as a cornerstone in the toolbox of statisticians and data scientists.

Our first algorithm explores the utilization of the M-H algorithm for BPEL. To accomplish this, we begin by specifying a proposal distribution with a density function denoted as $\phi(\cdot \mid \mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^p$. Subsequently, we employ the M-H algorithm to generate samples from the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$, as defined in (7). The specific steps for this process are detailed in Algorithm 1.

At each iteration k, Algorithm 1 begins with a state $\boldsymbol{\theta}^k \in \boldsymbol{\Theta}$. In the proposal step, it generates a new parameter $\boldsymbol{\vartheta}^{k+1}$ from the proposal distribution centred at $\boldsymbol{\theta}^k$, denoted by $\phi(\cdot \mid \boldsymbol{\theta}^k)$. Following this, in the accept-reject step, Algorithm 1 decides whether to accept $\boldsymbol{\vartheta}^{k+1}$ with a probability denoted as α^{k+1}. This crucial step ensures that the Markov chain, guided by Algorithm 1, remains within the valid parameter space $\boldsymbol{\Theta}$. Consequently, it expedites the convergence of the resulting chain towards its stationary distribution, which is the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$. There exist various approaches for selecting the proposal distribution with density $\phi(\cdot \mid \cdot)$, including methods like the symmetric Metropolis algorithm, random walk M-H, and the independence sampler, as detailed by Roberts and Rosenthal (2004).

2.3.2 Algorithm 2

Another widely used MCMC technique is importance sampling (Hesterberg, 1995; Ripley, 2006). This method involves generating samples from a proposal distribution and then applying importance weights to these samples to account for the disparities between the proposal distribution and the target distribution. In practical applications, recycling successive samples often proves to be an effective strategy (Marin et al., 2019), particularly when the computation of importance weights is computationally intensive. In this context, Cornuet et al. (2012) introduces the AMIS algorithm, which combines various importance sampling methods with adaptive techniques. The integration

Algorithm 1 M-H algorithm to generate samples from $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$

Input: the proposal distribution with density $\phi(\cdot \mid \cdot)$, the number of iteration K, an initial point $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$.

for $k = 0, 1, \ldots, K-1$ **do**

Proposal step:

generate $\boldsymbol{\vartheta}^{k+1}$ from the proposal distribution with density $\phi(\boldsymbol{\vartheta} \mid \boldsymbol{\theta}^k)$.

Accept-reject step:

compute

$$\alpha^{k+1} = \begin{cases} \min\left\{1, \frac{\pi^\dagger(\boldsymbol{\vartheta}^{k+1} \mid \mathcal{X}_n)\phi(\boldsymbol{\theta}^k \mid \boldsymbol{\vartheta}^{k+1})}{\pi^\dagger(\boldsymbol{\theta}^k \mid \mathcal{X}_n)\phi(\boldsymbol{\vartheta}^{k+1} \mid \boldsymbol{\theta}^k)}\right\}, & \text{if } \boldsymbol{\vartheta}^{k+1} \in \boldsymbol{\Theta} \text{ with } \pi^\dagger(\boldsymbol{\theta}^k \mid \mathcal{X}_n)\phi(\boldsymbol{\vartheta}^{k+1} \mid \boldsymbol{\theta}^k) \neq 0, \\ 1, & \text{if } \boldsymbol{\vartheta}^{k+1} \in \boldsymbol{\Theta} \text{ with } \pi^\dagger(\boldsymbol{\theta}^k \mid \mathcal{X}_n)\phi(\boldsymbol{\vartheta}^{k+1} \mid \boldsymbol{\theta}^k) = 0, \\ 0, & \text{if } \boldsymbol{\vartheta}^{k+1} \notin \boldsymbol{\Theta}. \end{cases}$$

generate $u \sim \mathcal{U}(0, 1)$.

if $u \leq \alpha^{k+1}$, **then** $\boldsymbol{\theta}^{k+1} \leftarrow \boldsymbol{\vartheta}^{k+1}$, **else** $\boldsymbol{\theta}^{k+1} \leftarrow \boldsymbol{\theta}^k$.

end for

Output: $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$.

Algorithm 2 An MAMIS algorithm to generate the weighted samples with respect to $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$

Input: the proposal distribution admits density $\varphi(\cdot\,; \boldsymbol{\zeta})$ with the parameter $\boldsymbol{\zeta} \in \mathbb{R}^s$, an initial parameter $\hat{\boldsymbol{\zeta}}_1$, an explicitly known function $\mathbf{h}: \mathbb{R}^p \mapsto \mathbb{R}^s$, the number of iteration K and the increasing sampling sizes $\{N_1, \ldots, N_K\}$.

for $k \in [K]$ **do**

  for $i \in [N_k]$ **do**

    Proposal step:

    generate $\boldsymbol{\theta}_i^k$ from the proposal distribution with density $\varphi(\boldsymbol{\theta}; \hat{\boldsymbol{\zeta}}_k)$.

    compute the importance weight $\omega_i^k = \pi^{\dagger}(\boldsymbol{\theta}_i^k \mid \mathcal{X}_n)/\varphi(\boldsymbol{\theta}_i^k; \hat{\boldsymbol{\zeta}}_k)$.

  end for

  update the parameter of the proposal distribution: $\hat{\boldsymbol{\zeta}}_{k+1} = N_k^{-1} \sum_{i=1}^{N_k} \omega_i^k \mathbf{h}(\boldsymbol{\theta}_i^k)$.

end for

for $k \in [K]$ **do**

  for $i \in [N_k]$ **do**

    Recycling process:

    update the importance weight $\omega_i^k = \pi^{\dagger}(\boldsymbol{\theta}_i^k \mid \mathcal{X}_n)/\{S_K^{-1} \sum_{l=1}^{K} N_l \varphi(\boldsymbol{\theta}_i^k; \hat{\boldsymbol{\zeta}}_l)\}$ with $S_K = N_1 + \cdots + N_K$ if $\boldsymbol{\theta}_i^k \in \boldsymbol{\Theta}$.

  end for

end for

Output: the weighted samples $(\boldsymbol{\theta}_1^1, \omega_1^1), \ldots, (\boldsymbol{\theta}_{N_1}^1, \omega_{N_1}^1), \ldots, (\boldsymbol{\theta}_1^K, \omega_1^K), \ldots, (\boldsymbol{\theta}_{N_K}^K, \omega_{N_K}^K)$.

of the AMIS approach with EL, as shown in Mengersen et al. (2013), is particularly compelling. To ensure the consistency of AMIS, Marin et al. (2019) introduces a modified variant called modified AMIS (MAMIS) with a simpler recycling strategy compared to AMIS.

We present and investigate an MAMIS algorithm, as outlined in Algorithm 2, specifically designed for computing BPEL. This algorithm operates in a scenario where a density function $\varphi(\cdot\,; \boldsymbol{\zeta})$ is defined, with $\boldsymbol{\zeta}$ representing a parameter in $\mathbb{R}^s$, and where an explicit function $\mathbf{h}: \mathbb{R}^p \mapsto \mathbb{R}^s$ is known. This configuration allows us to generate weighted samples that effectively capture the characteristics of the posterior distribution $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$, as defined in (7).

Algorithm 2 generates a sequence of samples while progressively adjusting the parameter $\boldsymbol{\zeta} \in \mathbb{R}^s$ involved in the proposal distribution. At each iteration k of Algorithm 2, the new value for the parameter $\boldsymbol{\zeta}$ of the proposal distribution is determined based on the most recent N_k samples drawn. This represents the primary distinction between the MAMIS algorithm by Marin et al. (2019) and the AMIS algorithm by Cornuet et al. (2012). Specifically, MAMIS updates the proposal distribution parameter using only the last N_k samples at iteration k, while AMIS updates this parameter by considering all past $\sum_{j=1}^{k} N_j$ samples. The end product output of Algorithm 2 is generated by updating the importance weights for all samples produced during the recycling process.

2.4 Sampling vs. optimizations

We advocate the utilization of sampling techniques as a practical and efficient alternative to optimization methods for addressing computationally challenging PEL problems. Specifically for obtaining the estimator $\hat{\boldsymbol{\theta}}_n$ as defined in (3), we can rely on samples $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K$ generated from the M-H algorithm (see Algorithm 1), estimating $\mathbb{E}_{\boldsymbol{\theta} \sim \pi^{\dagger}}(\boldsymbol{\theta})$, as defined in (8), by computing the sample mean, i.e. $K^{-1} \sum_{k=1}^{K} \boldsymbol{\theta}^k$. When employing the MAMIS algorithm (see Algorithm 2) and completing K iterations, the estimator for $\mathbb{E}_{\boldsymbol{\theta} \sim \pi^{\dagger}}(\boldsymbol{\theta})$ is determined as a weighted average:

$$\widehat{\mathbb{E}}_{\pi^{\dagger},K}(\boldsymbol{\theta}) = \frac{1}{S_K} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \omega_i^k \boldsymbol{\theta}_i^k, \tag{9}$$

where $S_K = N_1 + \cdots + N_K$.

Our theory in Section 5.2 supports the use of sampling algorithms as efficient alternatives. For the M-H algorithm, Theorem 3 in Section 5.2 demonstrates that, conditional on $\mathcal{X}_n$, the average $K^{-1}\sum_{k=1}^{K}\boldsymbol{\theta}^k$ converges almost surely to $\mathbb{E}_{\boldsymbol{\theta}\sim\pi^\dagger}(\boldsymbol{\theta})$ as $K\to\infty$. For the MAMIS algorithm, Theorem 4 in Section 5.2 establishes that, conditional on $\mathcal{X}_n$, $\widehat{\mathbb{E}}_{\pi^\dagger,K}(\boldsymbol{\theta})$ in (9) converges almost surely to $\mathbb{E}_{\boldsymbol{\theta}\sim\pi^\dagger}(\boldsymbol{\theta})$ as $K\to\infty$. These results, combined with Corollary 1 in Section 5.2, validate the properties of BPEL estimators obtained through these established sampling techniques. Another consideration in Algorithms 1 and 2 is the choice of the initial point, denoted, respectively, as $\boldsymbol{\theta}^0$ and $\hat{\boldsymbol{\zeta}}_1$. Our theoretical analyses only require $\boldsymbol{\theta}^0\in\boldsymbol{\Theta}$ satisfying $\pi^\dagger(\boldsymbol{\theta}^0\mid\mathcal{X}_n)>0$ and do not impose any restriction on $\hat{\boldsymbol{\zeta}}_1$; see Theorems 3 and 4 in Section 5.2 for details. Our empirical simulation studies in Section 3 consistently demonstrate the proposed algorithms' robust performance, irrespective of the initial value chosen. Notice that the performance of the optimization methods for the nonconvex optimization problems usually depends crucially on the choice of the initial point. The combination of theoretical analysis and empirical evidence underscores that, in comparison to competing optimization methods, these sampling-based approaches offer significant advantages in terms of convergence speed, stability across replications, and resilience to variations in initial values. This reaffirms the benefits of incorporating BPEL into the methodology.

The M-H and MAMIS algorithms each have their strengths. M-H is easy to implement, but high rejection rates can reduce its efficiency, especially with a poorly tuned proposal distribution. MAMIS, while requiring more effort—particularly in computing importance weights—offers improved sampling efficiency and is less sensitive to the proposal distribution, making it ideal for complex posterior distributions. Choosing between these algorithms depends on the specific problem and the balance between implementation ease and sampling efficiency.

3 Numerical studies

3.1 Data generation process

We conduct simulation studies to empirically assess the performance of our proposed methods. For the DGP, we adopt the structural equation $y_i=\hbar(\mathbf{u}_i^\top\boldsymbol{\theta}_0)+e_i^{(0)}$, $i\in[n]$, where $\hbar:\mathbb{R}\mapsto\mathbb{R}$ is a continuous function, $e_i^{(0)}$ is the error, and $\mathbf{u}_i=(u_{i,1},u_{i,2})^\top$ represents two endogenous variables. The set of all instrumental variables (IVs) is denoted as $\mathbf{z}_i=(z_{i,1},\ldots,z_{i,r})^\top$ for $i\in[n]$. The true reduced-form equations for the endogenous variables are specified as $u_{i,1}=0.5z_{i,1}+0.5z_{i,2}+e_i^{(1)}$ and $u_{i,2}=0.5z_{i,3}+0.5z_{i,4}+e_i^{(2)}$, where $(e_i^{(1)},e_i^{(2)})$ represents the random errors. Essentially, each of the two endogenous variables is influenced by only two IVs. All IVs are selected orthogonal to the error term $e_i^{(0)}$. Hence, we have $\mathbb{E}\{y_i-\hbar(\mathbf{u}_i^\top\boldsymbol{\theta}_0)\mid\mathbf{z}_i\}=\mathbf{0}$, which implies that $\boldsymbol{\theta}_0$ can be identified by the r unbiased moment conditions $\mathbb{E}\{\mathbf{g}(\mathbf{x}_i;\boldsymbol{\theta}_0)\}=\mathbf{0}$, where $\mathbf{g}(\mathbf{x}_i;\boldsymbol{\theta})=\{y_i-\hbar(\mathbf{u}_i^\top\boldsymbol{\theta})\}\mathbf{z}_i$ with $\mathbf{x}_i=(y_i,\mathbf{u}_i^\top,\mathbf{z}_i^\top)^\top$. In the DGP, we generate $\mathbf{z}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I}_r)$, and

$$\begin{pmatrix}e_i^{(0)}\\e_i^{(1)}\\e_i^{(2)}\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}0.43&0.3&0.3\\0.3&0.34&0.09\\0.3&0.09&0.34\end{pmatrix}\right).$$

We set $\boldsymbol{\theta}_0=(0.5,0.5)^\top$ and consider two selections for the link function $\hbar(\cdot)$: (i) the linear case with $\hbar(\nu)=\nu$, and (ii) the nonlinear case with $\hbar(\nu)=\sin\nu$.

3.2 Sampling efficiency and stability

We begin by demonstrating the improvement in sampling efficiency achieved through the use of PEL. In this context, we generate data following the DGP with linear link function $\hbar(\cdot)$ by setting $n=120$ and varying r in the range [50, 1,000]. We aim to sample from two posterior distributions $\pi_0(\boldsymbol{\theta})\times\mathrm{EL}(\boldsymbol{\theta})$ and $\pi_0(\boldsymbol{\theta})\times\mathrm{PEL}_v(\boldsymbol{\theta})$, where $\mathrm{EL}(\boldsymbol{\theta})$ and $\mathrm{PEL}_v(\boldsymbol{\theta})$ are, respectively, given in (1) and (6). Evaluating $\mathrm{PEL}_v(\boldsymbol{\theta})$ involves an optimization problem that solves for $\boldsymbol{\lambda}$ by maximizing the objective function $f_n(\boldsymbol{\lambda};\boldsymbol{\theta})$ defined as (5) at given $\boldsymbol{\theta}$. To ensure the attainment of a sparse Lagrange multiplier

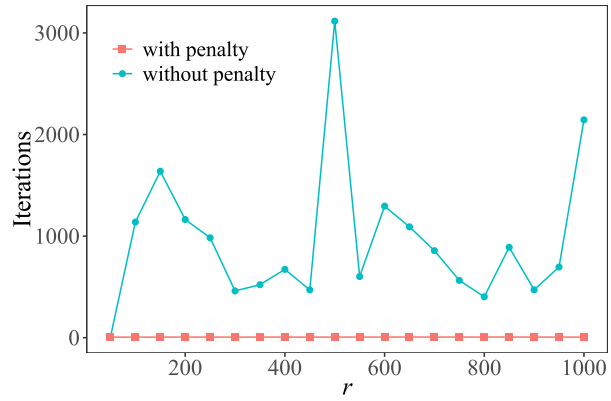

Figure 1. The average number of iterations over 500 runs required to obtain five valid samples.

and maintain the convexity of the objective function, we select $P_v(\cdot)$ as the L_1 penalty function. In practice, since the prior information about the true parameter $\boldsymbol{\theta}_0$ is typically unavailable, we select $\pi_0(\cdot)$ as the improper uniform prior. We implement Algorithm 1 to sample from both posterior distributions using identical settings, employing a proposal distribution $\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}_p)$ with $\sigma^2 = 10^{-4}$ and initializing from $\boldsymbol{\theta}^0 = (0.3, 0.3)^\top$. In the case of PEL, we set the tuning parameter $v = 0.03$ involved in $\mathrm{PEL}_v(\boldsymbol{\theta})$.

To compare efficiency, we measure the number of iterations required to obtain the same number of accepted samples. Figure 1 illustrates the average number of iterations needed over 500 runs to accept five samples for different values of r, thereby providing a comparison between using EL and PEL within a Bayesian framework. The sampling efficiency of Algorithm 1 when using $\mathrm{PEL}_v(\boldsymbol{\theta})$ is notably superior to that achieved with $\mathrm{EL}(\boldsymbol{\theta})$. The selection of σ^2 within the proposal distribution closely influences the acceptance rate in each step of the M-H algorithm. With our small choice of σ^2 in the simulation, the M-H algorithm should efficiently generate valid samples. It is worth highlighting that the acceptance rate remains consistently high and stable when using PEL across all r settings. In contrast, when employing EL without any penalty, it may require thousands more iterations to achieve the same number of accepted samples. Additionally, it is evident that the M-H algorithm with EL becomes increasingly unstable as r increases.

3.3 Comparison with the optimization methods

As we suggested in Section 2.3, the computation of the PEL estimator $\hat{\boldsymbol{\theta}}_n$ defined as (3) can be implemented using Algorithm 1 (referred to as M-H) and Algorithm 2 (referred to as MAMIS). In this part, we compare their performance with two optimization methods: (a) `optim`: A versatile R function for general-purpose optimization of objective functions, supporting various optimization algorithms like Nelder–Mead, quasi-Newton, and conjugate-gradient; and (b) `nlm`: An R function specialized in nonlinear optimization, particularly designed for finding minima of objective functions using Newton-type algorithms.

The choice of the proposal distribution plays a crucial role in achieving efficient sampling with BPEL. Within the context of the M-H algorithm, one commonly used scheme is the random walk M-H, where the proposal distribution takes the form of a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}_p)$ with the current state denoted as $\boldsymbol{\theta}$. It is essential to carefully select an appropriate value for σ^2. A small value for σ^2 can result in slow exploration of the state space, while a large value can lead to decreased acceptance rates, subsequently slowing down the algorithm. To strike a balance between exploration and acceptance rates, we can monitor the acceptance rate of the algorithm. In the simulation for M-H, we set $\sigma^2 = C(n \log r)^{-1}$ with some constant $C > 0$. We adjust the value of C until the acceptance rate closely matches the desired rate, typically aiming for ~ 0.234, as suggested in Gelman et al. (1997). It is known that the M-H algorithm requires some time to converge to its stationary distribution, especially when the initial point $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$ is situated in the tails of the

posterior distribution $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$. Considering this, we set a burn-in period of 500 iterations. For the MAMIS algorithm, we adhere to recommendations from Cornuet et al. (2012) and Mengersen et al. (2013) that advocate for the adoption of $\mathcal{T}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the proposal distribution. During each iteration k of MAMIS, we calculate the updated value $\hat{\boldsymbol{\zeta}}_{k+1} = \{\hat{\boldsymbol{\mu}}_{k+1}^{\top}, \text{vech}(\widehat{\boldsymbol{\Sigma}}_{k+1})^{\top}\}^{\top}$ for the parameter vector $\boldsymbol{\zeta} = \{\boldsymbol{\mu}^{\top}, \text{vech}(\boldsymbol{\Sigma})^{\top}\}^{\top}$ involved in the proposal distribution $\mathcal{T}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $\hat{\boldsymbol{\mu}}_{k+1} = N_k^{-1} \sum_{i=1}^{N_k} \omega_i^k \boldsymbol{\theta}_i^k$ and $\text{vech}(\widehat{\boldsymbol{\Sigma}}_{k+1}) = N_k^{-1} \sum_{i=1}^{N_k} \omega_i^k \text{vech}\{(\boldsymbol{\theta}_i^k - \hat{\boldsymbol{\mu}}_{k+1})(\boldsymbol{\theta}_i^k - \hat{\boldsymbol{\mu}}_{k+1})^{\top}\}$, where ω_i^k represents the corresponding importance weights, as outlined in Algorithm 2. In our simulations, we initialize $\widehat{\boldsymbol{\Sigma}}_1 = \mathbf{I}_p$, and the selection of $\hat{\boldsymbol{\mu}}_1$ is described in the next paragraph.

We conduct 200 replications following the DGP and explore various combinations of dimensionalities. Specifically, we consider $n \in \{120, 240\}$ and $r \in \{80, 160, 320, 640\}$. To assess the robustness of these methods with respect to initial points, we select 49 equally spaced grid points on the plane within the range of $[-3, 4] \times [-3, 4]$ as our chosen initial points. In the case of MAMIS, which is not an iterative algorithm, we set these initial points as the initial means $\hat{\boldsymbol{\mu}}_1$ for its proposal distribution $\mathcal{T}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to facilitate comparison. In our simulations, we identify the true global minima $\hat{\boldsymbol{\theta}}_n$ defined as (3) through exhaustive search. To achieve this, in each replication of the simulation (indexed by k), we generate a grid of 10,201 equally spaced points within the range $[-0.5, 1.5] \times [-0.5, 1.5]$. We then compute the posterior probabilities for these points and selected $\boldsymbol{\theta}_k^{\text{mode}}$ as the point with the highest probability. Since $\pi_0(\cdot)$ is selected as the improper uniform prior, $\boldsymbol{\theta}_k^{\text{mode}}$ is actually the required true global minima in the kth replication. We repeat this process for $k = 1$ to 200, and compare the outcomes obtained from both optimization and sampling methods by calculating the measure

$$\text{MSE}_1 = \frac{1}{200 \times 49} \sum_{k=1}^{200} \sum_{l=1}^{49} |\check{\boldsymbol{\theta}}_k(l) - \boldsymbol{\theta}_k^{\text{mode}}|_2^2.$$

Here, $\check{\boldsymbol{\theta}}_k(l)$ represents the related outcome in the kth replication initiated from the lth initial point.

In the context of BPEL sampling, we explore three scenarios with varying sample sizes of 1,500, 2,500, and 3,500, which we label as (M-H-1, M-H-2, and M-H-3) and (MAMIS-1, MAMIS-2, and MAMIS-3), respectively, for Algorithms 1 and 2. Additionally, we conduct an investigation into the influence of different values for the tuning parameter v. Table 1 presents the simulation results. The overall performance of the sampling approaches surpasses that of the optimization methods. Notably, for the nonlinear model, the optimization using the R function nlm is proven to be unreliable, resulting in highly unstable results. As the size of the generated samples increases, the performance of BPEL improves. Both M-H and MAMIS exhibit promising performance in both linear and nonlinear cases. For the nonlinear models, MAMIS significantly outperforms M-H, possibly owing to the advantages gained from employing importance weights for parameter estimation. The role of the tuning parameter v is pivotal, underscoring the merits from using the PEL approach in achieving more parsimonious models by effectively selecting most useful model conditions within the constraints of the available data information. When using very small values of v, such as 0.01, the performance of the methods becomes less satisfactory. Overall, the BPEL performs satisfactorily for a reasonable range of choices for v.

3.4 Comparison with competing methods

In this part, we compare the PEL estimator $\hat{\boldsymbol{\theta}}_n$ defined as (3) with two other estimators: the standard EL estimator $\tilde{\boldsymbol{\theta}}_n$ defined as (2) and the relaxed EL (REL) estimator introduced by Shi (2016). The REL is tailored for high-dimensional estimating equations, making it resilient to minor deviations from the equality constraints. Notice that the standard EL can only work for low-dimensional estimating equations. In line with our model specifications, where the two endogenous variables $u_{i,1}$ and $u_{i,2}$ are linked to IVs ($z_{i,1}$, $z_{i,2}$ and $z_{i,3}$, $z_{i,4}$, respectively) for each $i \in [n]$, we only use the first four moment conditions, that are related to the IVs $z_{i,1}$, $z_{i,2}$, $z_{i,3}$, and $z_{i,4}$, to produce the standard EL estimator $\tilde{\boldsymbol{\theta}}_n$. The computation of $\tilde{\boldsymbol{\theta}}_n$ can be implemented by the function `gel` in the R-package `gmm`. For both our PEL estimator $\hat{\boldsymbol{\theta}}_n$ and the REL estimator, we use all the r moment conditions.

Table 1. Comparison of Bayesian penalized empirical likelihood and optimization methods

		$h(v) = v, n = 120$				$h(v) = \sin v, n = 120$			
v	Methods	$r = 80$	$r = 160$	$r = 320$	$r = 640$	$r = 80$	$r = 160$	$r = 320$	$r = 640$
0.01	MAMIS-1	0.0606	0.0432	0.0371	0.0336	8.8598	7.0955	6.7810	6.7583
	MAMIS-2	0.0062	0.0020	0.0016	0.0015	8.2760	6.4363	6.0974	6.0576
	MAMIS-3	0.0023	0.0014	0.0014	0.0013	7.8634	6.0010	5.6087	5.6073
	M-H-1	0.0457	0.0015	0.0016	0.0015	12.7310	12.2970	12.2665	12.3473
	M-H-2	0.0411	0.0015	0.0015	0.0015	12.6824	12.2486	12.1999	12.3289
	M-H-3	0.0389	0.0014	0.0015	0.0014	12.6477	12.2179	12.1564	12.3042
	optim	0.2822	0.0720	0.0133	0.0132	12.2067	12.0219	11.9113	11.9760
	nlm	0.0956	0.0341	0.0222	0.0151	117934.2	115037.8	85081.4	83013.3
0.03	MAMIS-1	0.0465	0.0363	0.0317	0.0311	7.3502	6.2407	5.8879	6.2055
	MAMIS-2	0.0021	0.0011	0.0008	0.0009	6.5728	5.4791	5.0858	5.3622
	MAMIS-3	0.0008	0.0008	0.0007	0.0007	6.0018	4.9604	4.5459	4.7549
	M-H-1	0.0135	0.0009	0.0007	0.0008	12.5367	12.0546	12.1103	12.2611
	M-H-2	0.0116	0.0009	0.0007	0.0008	12.4304	11.9548	12.0239	12.1413
	M-H-3	0.0104	0.0008	0.0007	0.0007	12.3589	11.8816	11.9509	12.0655
	optim	0.0587	0.0109	0.0014	0.0014	12.8445	12.5361	12.3325	12.5015
	nlm	0.0385	0.0033	0.0015	0.0058	79382.0	73390.0	76096.0	94794.3
0.05	MAMIS-1	0.0451	0.0359	0.0331	0.0295	6.0750	5.5727	5.3221	5.3884
	MAMIS-2	0.0025	0.0018	0.0011	0.0008	5.0407	4.6182	4.3750	4.3938
	MAMIS-3	0.0017	0.0016	0.0008	0.0007	4.4061	3.9666	3.7920	3.7718
	M-H-1	0.0076	0.0015	0.0010	0.0008	12.0832	11.7687	11.8703	11.8866
	M-H-2	0.0074	0.0014	0.0009	0.0007	11.8869	11.5660	11.6859	11.7368
	M-H-3	0.0074	0.0013	0.0009	0.0007	11.7189	11.4244	11.5409	11.6131
	optim	0.0209	0.0010	0.0001	0.0000	13.2051	12.8147	12.6047	12.7401
	nlm	0.0167	0.0002	0.0004	0.0009	69229.7	58220.3	61362.5	63215.1

(*continued*)

Table 1. Continued

		$h(v) = v, n = 120$				$h(v) = \sin v, n = 120$			
v	Methods	$r = 80$	$r = 160$	$r = 320$	$r = 640$	$r = 80$	$r = 160$	$r = 320$	$r = 640$
0.07	MAMIS-1	0.0454	0.0381	0.0323	0.0293	4.6585	4.7192	4.4856	4.5785
	MAMIS-2	0.0049	0.0031	0.0023	0.0015	3.5217	3.6260	3.5231	3.5504
	MAMIS-3	0.0042	0.0029	0.0019	0.0014	2.8438	2.9326	2.9355	2.9075
	M-H-1	0.0065	0.0031	0.0020	0.0016	11.6003	11.3310	11.3530	11.5029
	M-H-2	0.0063	0.0030	0.0020	0.0015	11.2135	10.9535	11.1190	11.2268
	M-H-3	0.0062	0.0029	0.0019	0.0015	10.9140	10.7093	10.8781	11.0160
	optim	0.0117	0.0000	0.0007	0.0000	13.2631	13.0115	12.8281	12.8492
	nlm	0.0119	0.0000	0.0017	0.0010	49184.4	62755.1	63796.5	71037.4
		$h(v) = v, n = 240$				$h(v) = \sin v, n = 240$			
v	Methods	$r = 80$	$r = 160$	$r = 320$	$r = 640$	$r = 80$	$r = 160$	$r = 320$	$r = 640$
0.01	MAMIS-1	0.2600	0.0467	0.0297	0.0250	9.7398	7.7113	6.9972	6.7786
	MAMIS-2	0.1139	0.0047	0.0004	0.0004	9.1874	7.1095	6.4394	6.1002
	MAMIS-3	0.0911	0.0018	0.0002	0.0003	8.8130	6.6946	6.0186	5.6339
	M-H-1	0.4498	0.0371	0.0082	0.0004	13.2222	12.2080	12.0538	11.9004
	M-H-2	0.4279	0.0311	0.0067	0.0003	13.2205	12.1876	12.0544	11.8679
	M-H-3	0.4146	0.0277	0.0053	0.0003	13.2281	12.1745	12.0461	11.8451
	optim	1.0834	0.2247	0.0582	0.0160	12.8270	12.2083	12.2197	12.1037
	nlm	6.2781	0.0674	0.0163	0.0076	121272.7	113263.9	104125.3	87304.6
0.03	MAMIS-1	0.0707	0.0345	0.0279	0.0264	7.2303	6.6040	6.5472	6.3108
	MAMIS-2	0.0033	0.0012	0.0004	0.0004	6.3727	5.8076	5.8222	5.6557
	MAMIS-3	0.0007	0.0009	0.0002	0.0002	5.7517	5.2972	5.2865	5.1854
	M-H-1	0.0522	0.0034	0.0003	0.0002	12.8677	12.1031	12.1453	12.1239

(*continued*)

Table 1. Continued

		$h(v) = v, n = 240$				$h(v) = \sin v, n = 240$			
v	Methods	$r = 80$	$r = 160$	$r = 320$	$r = 640$	$r = 80$	$r = 160$	$r = 320$	$r = 640$
	M-H-2	0.0450	0.0034	0.0003	0.0002	12.7587	12.0285	12.0758	12.0806
	M-H-3	0.0388	0.0034	0.0002	0.0002	12.6799	11.9762	12.0297	12.0417
	optim	0.2977	0.0454	0.0125	0.0005	13.2494	12.7951	12.7153	12.8332
	nlm	0.1321	0.0127	0.0057	0.0047	90321.8	70994.2	90414.5	79523.2
0.05	MAMIS-1	0.0598	0.0387	0.0305	0.0261	5.8614	5.5769	5.6081	5.8757
	MAMIS-2	0.0043	0.0024	0.0010	0.0007	4.7997	4.6707	4.7641	4.9901
	MAMIS-3	0.0027	0.0019	0.0009	0.0005	4.0542	4.0482	4.1549	4.3903
	M-H-1	0.0063	0.0018	0.0010	0.0006	12.2347	11.8718	12.0358	12.0665
	M-H-2	0.0059	0.0016	0.0009	0.0005	12.0296	11.7100	11.9241	11.9669
	M-H-3	0.0056	0.0016	0.0009	0.0005	11.8578	11.5979	11.8305	11.8745
	optim	0.0827	0.0130	0.0073	0.0000	13.5813	13.0596	13.0502	13.0375
	nlm	0.0387	0.0023	0.0032	0.0052	48458.3	69014.4	62235.7	80570.4
0.07	MAMIS-1	0.0534	0.0381	0.0341	0.0278	4.5237	4.5635	4.9180	5.0988
	MAMIS-2	0.0071	0.0058	0.0036	0.0023	3.2789	3.5005	3.8945	4.1681
	MAMIS-3	0.0066	0.0054	0.0033	0.0021	2.5653	2.8253	3.2818	3.5420
	M-H-1	0.0069	0.0054	0.0036	0.0022	11.7248	11.5753	11.6807	12.0016
	M-H-2	0.0067	0.0053	0.0035	0.0021	11.3384	11.2954	11.4157	11.8458
	M-H-3	0.0067	0.0053	0.0035	0.0021	11.0287	11.0815	11.2360	11.7241
	optim	0.0169	0.0033	0.0007	0.0000	13.5938	13.3118	13.2545	13.2196
	nlm	0.0076	0.0015	0.0061	0.0006	72324.7	59709.3	80154.4	62439.4

Table 2. Comparison of Bayesian penalized empirical likelihood and other estimators

		$h(v) = v$				$h(v) = \sin v$			
n	Methods	$r = 80$	$r = 160$	$r = 320$	$r = 640$	$r = 80$	$r = 160$	$r = 320$	$r = 640$
120	MAMIS	0.0080	0.0096	0.0096	0.0114	0.0963	0.0829	0.0661	0.0620
	M-H	0.0086	0.0108	0.0118	0.0140	6.8664	7.4009	6.8831	7.2457
	EL	59.8762	58.7258	60.5130	60.0313	13.9942	14.0453	14.2512	14.4454
	REL	8.6108	8.8342	8.8781	9.2086	18.1845	18.5454	18.5858	18.9122
240	MAMIS	0.0036	0.0044	0.0047	0.0055	0.1417	0.1200	0.1069	0.1136
	M-H	0.0039	0.0048	0.0051	0.0061	13.1962	13.6146	12.9027	13.0548
	EL	57.9585	57.3146	57.5111	57.6992	14.0103	13.8869	14.0533	14.3296
	REL	8.2303	8.1680	8.1674	7.8542	19.3646	19.6221	19.9443	20.0887

For the selection of the tuning parameter in the REL estimator, we follow the recommendation in Shi (2016), using a consistent tuning parameter $0.5n^{-1/2}(\log r)^{1/2}$ throughout the simulations. Regarding the tuning parameter v in our BPEL, we employ the Bayesian information criterion (BIC) defined as

$$\mathrm{BIC}(v) = \log\left\{\frac{1}{r}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{x}_i;\hat{\boldsymbol{\theta}}_n^{(v)})\right|_2^2\right\} + |\mathcal{R}_n^{(v)}|n^{-1}\log n \tag{10}$$

for its selection, where $\hat{\boldsymbol{\theta}}_n^{(v)}$ denotes the associated PEL estimator with tuning parameter v calculated by our proposed sampling algorithm, and $\mathcal{R}_n^{(v)} = \mathrm{supp}\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n^{(v)})\}$ with $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n^{(v)}) = (\hat{\lambda}_1^{(v)}, \ldots, \hat{\lambda}_r^{(v)})^\top = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\hat{\boldsymbol{\theta}}_n^{(v)})} f_n(\boldsymbol{\lambda}; \hat{\boldsymbol{\theta}}_n^{(v)})$ with $f_n(\boldsymbol{\lambda};\boldsymbol{\theta})$ defined as (5). In practice, we set $\mathcal{R}_n^{(v)} = \{j \in [r] : |\hat{\lambda}_j^{(v)}| > 10^{-6}\}$ and restrict v in the interval $[0.05n^{-1/2}(\log r)^{1/2}, 0.75n^{-1/2}(\log r)^{1/2}]$.

For the same 49 initial points of the 200 replications mentioned in Section 3.3, we calculate the measure

$$\mathrm{MSE}_2 = \frac{1}{200\times 49}\sum_{k=1}^{200}\sum_{l=1}^{49}|\check{\boldsymbol{\theta}}_k(l) - \boldsymbol{\theta}_0|_2^2$$

to evaluate the performance of different estimators, where $\check{\boldsymbol{\theta}}_k(l)$ is the related estimator in the kth replication initiated from the lth initial point. Table 2 compares the measure MSE_2 for the three estimators: the PEL estimator (MAMIS, M-H), the standard EL estimator, and the REL estimator. The results for M-H and MAMIS are derived based on the generated samples of size 3,500. It becomes clear that BPEL demonstrates substantial performance improvements, clearly establishing its superiority over the other estimation methods. Particularly noteworthy is the effectiveness of MAMIS in addressing the challenges posed by nonlinear estimating equations, showing its promising performance.

3.5 Additional numerical studies

We provide additional simulation studies in the online supplementary material. Section A.1 examines the impact of prior specification. Section A.2 evaluates the performance of our method using an alternative DGP with data from a Student's t-distribution instead of a normal distribution. Section A.3 assesses the finite sample accuracy of the MCMC algorithms in approximating the posterior distribution. Section A.4 compares the posterior distributions resulting from different Bayesian EL formulations. Section A.5 presents the comparison between our method and two competing methods: approximate Bayesian computation and Bayesian synthetic likelihood. Overall, our findings confirm the highly competitive performance of the proposed BPEL with

the MCMC framework in terms of finite sample performance and accuracy in approximating the posterior distributions.

4 Real data analysis

International trade refers to the cross-border exchange of capital, commodities, and services between nations or regions. This type of trade typically constitutes a substantial portion of a country's gross domestic product (GDP). Eaton et al. (2011), hereafter referred to as EKK, combined an empirical model with microeconomic principles to analyse France's international trade patterns. Additionally, Shi (2016) utilized EKK's microeconomic model to derive parameter estimates for Chinese exporting companies. In this section, we reexamine the dataset previously examined in Shi (2016), employing the proposed BPEL approach.

The model proposed by EKK comprises five parameters denoted as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_5)^\top \in \boldsymbol{\Theta}$. The first component, θ_1, characterizes the distribution of production efficiency among firms, with a higher θ_1 indicating a larger proportion of manufacturers with lower efficiency. The second component, θ_2, quantifies the cost associated with accessing a fraction of potential buyers, where a higher θ_2 corresponds to lower costs. Parameters θ_3, θ_4, and θ_5 represent the standard deviation of the demand shock, the standard deviation of the entry cost shock, and the correlation coefficient between these two shocks, respectively. Each firm is identified by the index $i \in [n]$, while countries are represented by the index $j \in \{0\} \cup [r]$, with $j = 0$ denoting the home country.

According to the EKK's model, the sales of firm i in country j is specified as $Z_{i,j}(\boldsymbol{\theta}; e_{i,j}^{(1)}, e_{i,j}^{(2)}, e_i^{(3)}) = \kappa \bar{Z}_j (1 - \tau_{i,j})^{\theta_2/\theta_1} \tau_{i,j}^{-1/\theta_1} a_{i,j}^{(1)} / a_{i,j}^{(2)}$, where $a_{i,j}^{(1)} = \exp\{\theta_3 (1 - \theta_5^2)^{1/2} e_{i,j}^{(1)} + \theta_3 \theta_5 e_{i,j}^{(2)}\}$, $a_{i,j}^{(2)} = \exp\{\theta_4 e_{i,j}^{(2)}\}$, $\tau_{i,j} = \min\{1, e_i^{(3)} \bar{u}_i / \bar{u}_{i,j}\}$ and

$$\kappa = \left(\frac{\theta_1}{\theta_1 - 1} - \frac{\theta_1}{\theta_1 + \theta_2 - 1} \right) \exp\left\{ \frac{1}{2}\left(\theta_3 - \theta_1^2 \theta_4^2\right) + \theta_3 \theta_4 \theta_5 (\theta_1 - 1) + \frac{1}{2} \theta_4 (\theta_1 - 1)^2 \right\}$$

with $\bar{u}_{i,j} = (a_{i,j}^{(2)})^{\theta_1} N_j$ and $\bar{u}_i = \min\{\bar{u}_{i,0}, \max_{j \in [r]} \bar{u}_{i,j}\}$, and $(\bar{Z}_j, N_j)_{j \in \{0\} \cup [r]}$ are known constants. Here $e_{i,j}^{(1)} \sim \mathcal{N}(0, 1)$, $e_{i,j}^{(2)} \sim \mathcal{N}(0, 1)$ and $e_i^{(1)} \sim \mathcal{U}(0, 1)$ are mutually independent. Furthermore, $Z_{i,j}(\boldsymbol{\theta}; e_{i,j}^{(1)}, e_{i,j}^{(2)}, e_i^{(3)}) = 0$ means that the firm i is kept outside of the country j. As a pertinent economic indicator of our interest, the mean sale of all firms in country j is defined as $\mu_j(\boldsymbol{\theta}) = \mathbb{E}\{Z_{i,j}(\boldsymbol{\theta}; e_{i,j}^{(1)}, e_{i,j}^{(2)}, e_i^{(3)})\}$, where the expectation is taken respect to the random variables $\{e_{i,j}^{(1)}, e_{i,j}^{(2)}, e_i^{(3)}\}$. The dataset is sourced from the Chinese administrative databases, encompassing a total of $n = 6{,}754$ firms and their export data to $r = 126$ foreign destination countries in 2006. Leveraging this dataset, we can obtain the r-dimensional estimating function $\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta}) = \{g_1(\mathbf{x}_i; \boldsymbol{\theta}), \ldots, g_r(\mathbf{x}_i; \boldsymbol{\theta})\}^\top$, $i \in [n]$, with $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,r})^\top$, and $g_j(\mathbf{x}_i; \boldsymbol{\theta}) = x_{i,j} - \mu_j(\boldsymbol{\theta})$ for any $j \in [r]$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $x_{i,j}$ is the sale of firm i in country j from this dataset ($j = 0$ is not considered in this dataset).

Since the model is highly nonlinear with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, resulting in no closed-form expression for $\mu_j(\boldsymbol{\theta})$, we approximate it via numerical simulation (Eaton et al., 2011; Shi, 2016). Specifically, in the estimation, we utilize the 'artificial data' for another $5n = 33{,}770$ firms from the dataset. This involves simulating the entry decisions and sales across various countries for each of these artificial firms. Subsequently, we calculate sample means to approximate $\mu_j(\boldsymbol{\theta})$ for any $j \in \{0\} \cup [r]$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We generated samples of size 3,500 from the posterior distribution for the BPEL. To select the tuning parameter v, we employed the BIC as defined in (10). For the parameter space $\boldsymbol{\Theta}$, we adopted a compact range of values, specifically $\boldsymbol{\Theta} = [1.5, 10] \times [0.5, 5] \times [0.1, 5] \times [0.1, 5] \times [-0.9, 0.9]$, which is consistent with the economic context and aligns with the study of Shi (2016). To initiate the analysis, we selected 15 samples uniformly distributed within the parameter space $\boldsymbol{\Theta}$. Figure 2 presents the box-plots of the corresponding 15 estimates obtained by M-H and MAMIS from these initial values. The results for the REL with the same initial values are also included for comparative evaluation.

It is evident that for all five parameters, MAMIS exhibits the smallest variations in the resulting estimates, whereas the variations of M-H and REL are relatively similar. This consistency with the

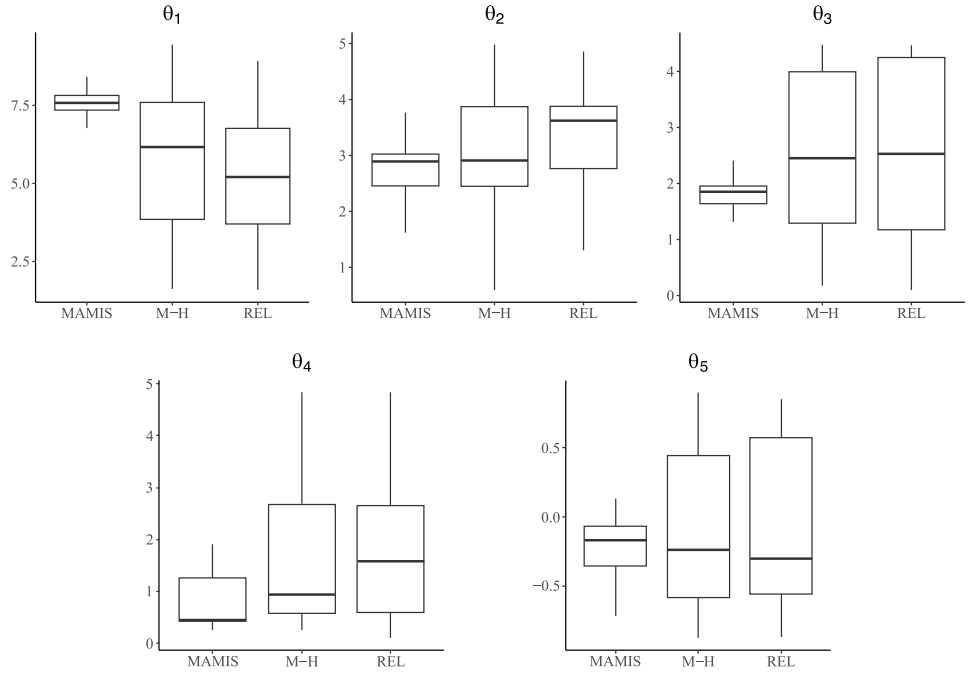

Figure 2. The box-plots of the estimated points.

findings in Sections 3.3 and 3.4 reaffirms the robustness of MAMIS when considering different initial points. Such robustness is desirable for conducting more in-depth analyses. For instance, let us take θ_5 into consideration which represents the correlation coefficient between the demand shock and the entry cost shock. The sign of its estimate carries the key implication. The 15 estimates of θ_5 obtained by REL and M-H, from different initial values, fall within the ranges of $(-0.8738, 0.8507)$ and $(-0.8774, 0.8996)$, respectively. In contrast, the estimates of θ_5 by MAMIS range in $(-0.7978, 0.1329)$, with the majority being negative, signalling a more assuring result.

We then proceed to examine the specific moments selected by the respective methods. For REL, we employ the greedy algorithm outlined in Section 3.2 of Shi (2016). To assess the effectiveness of moment selection, we validate whether or not the top 10 trading partners of China in terms of export volume in this dataset, including the USA, Japan, Germany, etc., are either selected or partially selected. We find that, although REL selects at least some of these countries for 10 out of the 15 initial values, the number of selected countries does not exceed 3. In contrast, for 13 out of the 15 initial values, M-H identifies at least some of these countries, with 9 of them including more than 3. In the case of MAMIS, 13 out of the 15 initial values result in the identification of some of these countries, and all of them include more than 3 countries. Additionally, the robustness of MAMIS with respect to the initial points provides enhanced reliability in this context.

5 Theoretical analysis

We introduce some additional notation first. For simplicity, write $\mathbb{E}_n(\cdot) = n^{-1}\sum_{i=1}^{n}\cdot$. For a $q \times q$ symmetric matrix $\mathbf{A}$, denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ the smallest and largest eigenvalues of $\mathbf{A}$, respectively. For a $q_1 \times q_2$ matrix $\mathbf{B} = (b_{i,j})_{q_1\times q_2}$, let $|\mathbf{B}|_\infty = \max_{i\in[q_1],j\in[q_2]}|b_{i,j}|$ be the super-norm. For the r-dimensional estimating function $\mathbf{g}(\cdot\,;\,\cdot) = \{g_1(\cdot\,;\,\cdot), \ldots, g_r(\cdot\,;\,\cdot)\}^\top$ and p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$, let $\nabla_{\boldsymbol{\theta}}\mathbf{g}(\cdot\,;\boldsymbol{\theta}) = \{\partial g_j(\cdot\,;\boldsymbol{\theta})/\partial\theta_k\}_{j\in[r],k\in[p]}$, an $r \times p$ matrix, be the first-order partial derivative of $\mathbf{g}(\cdot\,;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Let $\mathbf{V}(\boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}(\mathbf{x}_i;\boldsymbol{\theta})^{\otimes 2}\}$ and $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{x}_i;\boldsymbol{\theta})\}$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For a given index set $\mathcal{F}$, let $|\mathcal{F}|$ be its cardinality. Denote by $\mathbf{g}_{\mathcal{F}}(\cdot\,;\,\cdot)$ the subvector of $\mathbf{g}(\cdot\,;\,\cdot)$ collecting the components indexed by $\mathcal{F}$. Let $\mathbf{V}_{\mathcal{F}}(\boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}_{\mathcal{F}}(\mathbf{x}_i;\boldsymbol{\theta})^{\otimes 2}\}$ and $\boldsymbol{\Gamma}_{\mathcal{F}}(\boldsymbol{\theta}) = \mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathcal{F}}(\mathbf{x}_i;\boldsymbol{\theta})\}$. Analogously, we also write $\mathbf{a}_{\mathcal{F}}$ as the corresponding subvector of vector

a. For any two probability measures μ and ν, denote by $\mathcal{D}_{\mathrm{TV}}(\mu, \nu)$ the total variation distance between μ and ν.

5.1 Properties of the penalized empirical likelihood estimator

To investigate the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$ in (3), we assume some regularity conditions.

Condition 1 For any $\varepsilon > 0$, it holds that

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}:\, |\boldsymbol{\theta}-\boldsymbol{\theta}_0|_\infty > \varepsilon} |\mathbb{E}\{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}|_\infty \geq \Delta(\varepsilon),$$

where $\Delta(\cdot)$ is a nonnegative function satisfying $\liminf_{\varepsilon\to 0^+} \varepsilon^{-1}\Delta(\varepsilon) \geq K_1$ for some universal constant $K_1 > 0$.

Condition 2 (a) There exist universal constants $K_2 > 0$ and $\gamma > 4$ such that

$$\max_{j\in[r]} \mathbb{E}\left\{ \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |g_j(\mathbf{x}_i; \boldsymbol{\theta})|^\gamma \right\} \leq K_2$$

and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{j\in[r]} \mathbb{E}_n\{|g_j(\mathbf{x}_i; \boldsymbol{\theta})|^\gamma\} = O_{\mathrm{p}}(1)$. (b) There exist universal constants $0 < K_3 < K_4$ such that $K_3 < \lambda_{\min}\{\mathbf{V}(\boldsymbol{\theta}_0)\} \leq \lambda_{\max}\{\mathbf{V}(\boldsymbol{\theta}_0)\} < K_4$. (c) For any $\mathbf{x}$ and $j \in [r]$, $g_j(\mathbf{x}; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ satisfying

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{j\in[r],k\in[p]} \mathbb{E}_n\left\{ \left| \frac{\partial g_j(\mathbf{x}_i; \boldsymbol{\theta})}{\partial \theta_k} \right|^2 \right\} = O_{\mathrm{p}}(1)$$
$$= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{j\in[r],k_1,k_2\in[p]} \mathbb{E}_n\left\{ \left| \frac{\partial^2 g_j(\mathbf{x}_i; \boldsymbol{\theta})}{\partial \theta_{k_1} \partial \theta_{k_2}} \right|^2 \right\}.$$

Detailed discussion on Conditions 1 and 2 are given in Section B (online supplementary material). For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, define

$$\mathcal{M}^*_{\boldsymbol{\theta}} = \{j \in [r] : |\mathbb{E}_n\{g_j(\mathbf{x}_i; \boldsymbol{\theta})\}| \geq C_* \nu \rho'(0^+)\}$$

for some $C_* \in (0, 1)$. We assume the existence of a sequence $\ell_n \to \infty$ such that

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}:\, |\boldsymbol{\theta}-\boldsymbol{\theta}_0|_2 \leq c_n} |\mathcal{M}^*_{\boldsymbol{\theta}}| \leq \ell_n \right) \to 1$$

as $n \to \infty$, with some $c_n \to 0$ satisfying $\nu c_n^{-1} \to 0$. Proposition 1 shows that $\hat{\boldsymbol{\theta}}_n$ is consistent to the true parameter $\boldsymbol{\theta}_0$, allowing r growing exponentially with the sample size n.

Proposition 1 Let $P_\nu(\cdot) \in \mathscr{P}$ be a convex function for $\mathscr{P}$ defined as (4). Under Conditions 1, 2(a) and 2(b), if $\log r \ll n^{1/3}$ and $\ell_n n^{-1/2}(\log r)^{1/2} \ll \min\{\nu, n^{-1/\gamma}\}$, then the PEL estimator $\hat{\boldsymbol{\theta}}_n$ defined as (3) satisfies $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|_\infty = O_{\mathrm{p}}(\nu)$.

Proposition 1 establishes the consistency of the PEL estimator with diverging r, incorporating the impact of the penalty function. In particular, the convergence rate of $\hat{\boldsymbol{\theta}}_n$ is ν, provided that the tuning parameter ν in (3) satisfies $\nu \gg \ell_n n^{-1/2}(\log r)^{1/2}$. As a result, the convergence rate of $\hat{\boldsymbol{\theta}}_n$ is slower than $n^{-1/2}$, which can be viewed as the price paid for using the penalty in handling exponentially growing dimensionality r.

Recall $\rho(t; \nu) = \nu^{-1} P_\nu(t)$. For $P_\nu(\cdot) \in \mathscr{P}$ with $\mathscr{P}$ defined as (4), since $\rho'(0^+; \nu)$ is independent of ν, we write it as $\rho'(0^+)$ for simplicity. Let $\mathcal{R}_n = \mathrm{supp}\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)\}$ for the Lagrange multiplier $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n) =$

$(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)^\top = \arg\max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n(\hat{\boldsymbol{\theta}}_n)} f_n(\boldsymbol{\lambda}; \hat{\boldsymbol{\theta}}_n)$ with $f_n(\boldsymbol{\lambda}; \boldsymbol{\theta})$ defined as (5). Then $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)$ satisfy the score equation:

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{g}(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)}{1 + \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)^\top \mathbf{g}(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)} - \hat{\boldsymbol{\eta}}, \tag{11}$$

where $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_r)^\top$ with $\hat{\eta}_j = v\rho'(|\hat{\lambda}_j|; v)\mathrm{sgn}(\hat{\lambda}_j)$ for $\hat{\lambda}_j \neq 0$ and $\hat{\eta}_j \in [-v\rho'(0^+), v\rho'(0^+)]$ for $\hat{\lambda}_j = 0$. Here, an effective drastic dimension reduction is achieved with the associated sparse $\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n)$. The use of the penalty function $P_v(\cdot)$ leads to $\hat{\boldsymbol{\eta}}$ in (11), an extra term compared to that of the conventional EL. While $P_v(\cdot)$ ensures the consistency of $\hat{\boldsymbol{\theta}}_n$ as shown in Proposition 1, as we will show in Theorem 1 later, $\hat{\boldsymbol{\eta}}$ leads to a bias of the PEL estimator $\hat{\boldsymbol{\theta}}_n$.

We further remark that while penalizing the Lagrange multiplier in our PEL does effectively achieve the selection of moments, its properties in terms of the validity of the selected moments remain an interesting research question. On one hand, it is reasonable to expect that under appropriate conditions and with a suitably chosen tuning parameter, our PEL may correctly select the set of valid moments. On the other hand, the major challenge lies in the ambiguity of defining valid moments when the corresponding moment functions are evaluated at broad candidate values of the model parameters rather than the truth. This consideration opens the door to a research question of its own interest in the context of moment selection that we are interested in investigating in our future research.

To study the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$, we need the following regularity conditions.

Condition 3 Let $\mathbf{Q}_{\mathcal{F}} = \boldsymbol{\Gamma}_{\mathcal{F}}(\boldsymbol{\theta}_0)^{\top,\otimes 2}$ for any index set $\mathcal{F} \subset [r]$. There exist universal constants $0 < K_5 < K_6$ such that $K_5 < \lambda_{\min}(\mathbf{Q}_{\mathcal{F}}) \leq \lambda_{\max}(\mathbf{Q}_{\mathcal{F}}) < K_6$ for any index set $\mathcal{F} \subset [r]$ with $p \leq |\mathcal{F}| \leq \ell_n$.

Condition 4 (a) For the PEL estimator $\hat{\boldsymbol{\theta}}_n$ defined as (3), there exists a constant $\check{c} \in (C_*, 1)$ such that

$$\mathbb{P}\left[\bigcup_{j \in [r]} \{\check{c}v\rho'(0^+) \leq |\mathbb{E}_n\{g_j(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)\}| < v\rho'(0^+)\} \right] \to 0$$

as $n \to \infty$. (b) It holds that

$$\mathbb{P}\left[\bigcup_{j \in \mathcal{R}_n^c} \{|\hat{\eta}_j| = v\rho'(0^+)\} \right] \to 0$$

as $n \to \infty$.

Discussion of Conditions 3 and 4 are given in Section B (online supplementary material). Write $\widehat{\mathbf{V}}_{\mathcal{R}_n}(\hat{\boldsymbol{\theta}}_n) = \mathbb{E}_n\{\mathbf{g}_{\mathcal{R}_n}(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)^{\otimes 2}\}$ and $\widehat{\boldsymbol{\Gamma}}_{\mathcal{R}_n}(\hat{\boldsymbol{\theta}}_n) = \mathbb{E}_n\{\nabla_{\boldsymbol{\theta}} \mathbf{g}_{\mathcal{R}_n}(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_n)\}$. Define

$$\widehat{\mathbf{H}}_{\mathcal{R}_n} = \{\widehat{\boldsymbol{\Gamma}}_{\mathcal{R}_n}(\hat{\boldsymbol{\theta}}_n)^\top \widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1/2}(\hat{\boldsymbol{\theta}}_n)\}^{\otimes 2} \quad \text{and} \quad \hat{\boldsymbol{\psi}}_{\mathcal{R}_n} = \widehat{\mathbf{H}}_{\mathcal{R}_n}^{-1} \widehat{\boldsymbol{\Gamma}}_{\mathcal{R}_n}(\hat{\boldsymbol{\theta}}_n)^\top \widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\hat{\boldsymbol{\theta}}_n) \hat{\boldsymbol{\eta}}_{\mathcal{R}_n}, \tag{12}$$

where $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_r)^\top$ is specified in (11). We assume (r, ℓ_n, v) satisfy the following restrictions:

$$\begin{aligned} &\log r \ll \min\{n^{1/3}, n^{(\gamma-2)/(2\gamma)}\}, \quad \ell_n \ll \min\{n^{(\gamma-2)/(3\gamma)}(\log r)^{-2/3}, n^{1/5}(\log r)^{-2/5}\}, \\ &\text{and } \ell_n n^{-1/2}(\log r)^{1/2} \ll v \ll \ell_n^{-1/4} n^{-1/4}. \end{aligned} \tag{13}$$

The asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ is stated in Theorem 1, where the bias term $\hat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ comes from the penalty function $P_v(\cdot)$ imposed on the Lagrange multiplier $\boldsymbol{\lambda}$ in (3).

Theorem 1 Let $P_v(\cdot) \in \mathscr{P}$ be convex with bounded second-order derivative around 0, where $\mathscr{P}$ is defined as (4). Assume Conditions 1–4 hold with (r, ℓ_n, v) satisfying (13). For any $\mathbf{t} \in \mathbb{R}^p$ with $|\mathbf{t}|_2 = 1$, the PEL estimator $\hat{\boldsymbol{\theta}}_n$ defined as (3) satisfies $n^{1/2}\mathbf{t}^\top \widehat{\mathbf{H}}_{\mathcal{R}_n}^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 - \hat{\boldsymbol{\psi}}_{\mathcal{R}_n}) \to \mathcal{N}(0, 1)$ in distribution as $n \to \infty$, where $\widehat{\mathbf{H}}_{\mathcal{R}_n}$ and $\hat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ are defined in (12).

Here, the estimated bias $\hat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ can be easily calculated based on (12). Theorem 1 indicates that, upon correcting the bias by subtracting it from $\hat{\boldsymbol{\theta}}_n$, the resulting estimator $\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ will be $\sqrt{n}$ -consistent and asymptotically normal.

5.2 Properties of the posterior distribution and algorithms

For the proposed BPEL, we establish the Bernstein–von Mises theorem for the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$, as defined in (7). Furthermore, we provide theoretical assurances for the performance of Algorithms 1 and 2 in Section 2.3.

For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, write $\mathcal{R}(\boldsymbol{\theta}) = \text{supp}\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\}$ with

$$\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \{\hat{\lambda}_1(\boldsymbol{\theta}), \ldots, \hat{\lambda}_r(\boldsymbol{\theta})\}^\top = \arg\max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n(\boldsymbol{\theta})} f_n(\boldsymbol{\lambda}; \boldsymbol{\theta}),$$

where $f_n(\boldsymbol{\lambda}; \boldsymbol{\theta})$ is defined as (5). Then $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ satisfy the score equation:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})}{1 + \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})} - \hat{\boldsymbol{\eta}}(\boldsymbol{\theta}), \tag{14}$$

where $\hat{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \{\hat{\eta}_1(\boldsymbol{\theta}), \ldots, \hat{\eta}_r(\boldsymbol{\theta})\}^\top$ with $\hat{\eta}_j(\boldsymbol{\theta}) = v\rho'\{|\hat{\lambda}_j(\boldsymbol{\theta})|; v\}\text{sgn}\{\hat{\lambda}_j(\boldsymbol{\theta})\}$ for $\hat{\lambda}_j(\boldsymbol{\theta}) \neq 0$ and $\hat{\eta}_j(\boldsymbol{\theta}) \in [-v\rho'(0^+), v\rho'(0^+)]$ for $\hat{\lambda}_j(\boldsymbol{\theta}) = 0$. By the definition of the PEL estimator $\hat{\boldsymbol{\theta}}_n$, we have $f_n\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\} \geq f_n\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n); \hat{\boldsymbol{\theta}}_n\}$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. To investigate the asymptotic properties of the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$ defined as (7), we need to first study $\aleph_n(\boldsymbol{\theta}) = f_n\{\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\} - f_n\{\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_n); \hat{\boldsymbol{\theta}}_n\}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Given $\alpha_n = n^{-1/2}(\log r)^{1/2}$ and $\beta_n > 0$ satisfying $\ell_n^{1/2} v \ll \beta_n \ll \min\{\ell_n^{-1} n^{-1/\gamma}, v^{2/3}\ell_n^{-2/3} n^{-1/(3\gamma)}\}$, we split the whole parameter space $\boldsymbol{\Theta}$ into three regions: $\mathcal{C}_1 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n|_2 \leq \alpha_n\}$, $\mathcal{C}_2 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \alpha_n < |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n|_2 \leq \beta_n\}$ and $\mathcal{C}_3 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n|_2 > \beta_n\}$. Proposition 2 (online supplementary material) shows that the asymptotic behaviour of $\aleph_n(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ in these three regions are different.

Investigating the asymptotic behaviour of $\aleph_n(\boldsymbol{\theta})$ calls some new technical arguments. Write

$$\tilde{f}_n(\boldsymbol{\lambda}; \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\} \quad \text{and} \quad \tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\lambda} \in \hat{\Lambda}_n(\boldsymbol{\theta})} \tilde{f}_n(\boldsymbol{\lambda}; \boldsymbol{\theta}). \tag{15}$$

When r is a fixed constant, we know $2n\tilde{f}_n\{\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\}$ is the conventional log-EL ratio in the literature. The asymptotic behaviour of $2n\tilde{f}_n\{\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\}$ depends on the magnitude of $\mathbb{E}\{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}$. More specifically, under some mild conditions, it holds that (i) $2n\tilde{f}_n\{\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\}$ is asymptotically χ^2 distributed with degree of freedom r if $|\mathbb{E}\{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}|_2 \ll n^{-1/2}$, (ii) $2n\tilde{f}_n\{\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\}$ converges to a noncentral χ^2 distribution if $|\mathbb{E}\{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}|_2 \asymp n^{-1/2}$, and (iii) $2n\tilde{f}_n\{\tilde{\boldsymbol{\lambda}}(\boldsymbol{\theta}); \boldsymbol{\theta}\}$ diverges to ∞ in probability if $|\mathbb{E}\{\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta})\}|_2 \gg n^{-1/2}$. See, for example, Proposition 1 and Theorem 1 of Chang et al. (2013) for such results with $r = 1$. In comparison to $\tilde{f}_n(\boldsymbol{\lambda}; \boldsymbol{\theta})$ defined in (15), $f_n(\boldsymbol{\lambda}; \boldsymbol{\theta})$ involved in $\aleph_n(\boldsymbol{\theta})$ includes a penalty term imposed on the Lagrange multiplier $\boldsymbol{\lambda}$. This makes the standard technique for analysing the conventional log-EL ratio inapplicable. To further establish the Bernstein–von Mises theorem for the posterior distribution $\pi^\dagger(\boldsymbol{\theta} \mid \mathcal{X}_n)$ defined as (7), we assume the following regularity conditions.

Condition 5 (a) There exists a constant $\bar{c} \in (0, 1)$ such that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\theta} \in \mathcal{C}_1} \max_{j \in \mathcal{R}(\boldsymbol{\theta})^c} |\hat{\eta}_j(\boldsymbol{\theta})| \leq \bar{c}v\rho'(0^+)\right\} \to 1$$

as $n \to \infty$, where $\hat{\eta}_j(\boldsymbol{\theta})$ is specified in (14). (b) There exists $\kappa_n > 0$ satisfying $\max\{\ell_n^{1/2} n^{-1/2} (\log r)^{1/2},\ \ell_n \beta_n^{3/2} n^{1/(2\gamma)}\} \ll \kappa_n \ll v$ such that

$$\mathbb{P}\Big[\bigcup_{\boldsymbol{\theta}\in\mathcal{C}_2}\bigcup_{j\in\mathcal{R}_n}\{v\rho'(0^+) - \kappa_n < |\mathbb{E}_n\{g_j(\mathbf{x}_i;\boldsymbol{\theta})\}| < v\rho'(0^+) + \kappa_n\}\Big] \to 0$$

as $n \to \infty$. (c) There exist universal constants $K_7, K_8 > 0$ such that

$$\mathbb{P}\Big\{\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}([\mathbb{E}_n\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathcal{R}_n}(\mathbf{x}_i;\boldsymbol{\theta})\}]^{\otimes 2}) \geq K_7\Big\} \to 1 \quad \text{and}$$
$$\mathbb{P}\Big[\sup_{\boldsymbol{\theta}\in\mathcal{C}_3} \lambda_{\max}\{\widehat{\mathbf{V}}_{\mathcal{R}_n}(\boldsymbol{\theta})\} \leq K_8\Big] \to 1$$

as $n \to \infty$.

Condition 6 The prior density $\pi_0(\cdot)$ is continuously differentiable with bounded first-order derivatives and $\pi_0(\boldsymbol{\theta}_0) > 0$.

Detailed discussion of Conditions 5 and 6 are given in Section B (online supplementary material). Let $\Pi_n^{\dagger}(\cdot)$ be the measure which admits the posterior distribution $\pi^{\dagger}(\cdot \mid \mathcal{X}_n)$. Denote by $\mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\cdot)$ the Gaussian measure with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. To establish the Bernstein–von Mises theorem for the posterior distribution $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$ as in Theorem 2, we need to assume (r, ℓ_n, v) satisfy the following restrictions:

$$\begin{aligned}&\log r \ll n^{(\gamma-2)/(3\gamma)},\ \ell_n \ll \min\{n^{(\gamma-2)/(9\gamma)}(\log r)^{-1/9},\ n^{1/3}(\log r)^{-1},\ n^{(\gamma-2)/(2\gamma)}(\log r)^{-3/2}\},\\ &\quad \text{and } \ell_n n^{-1/2}(\log r)^{1/2} \ll v \ll \min\{\ell_n^{-7/2} n^{-1/\gamma},\ (\log r)^{-1}\}.\end{aligned} \tag{16}$$

Theorem 2 Let $P_v(\cdot) \in \mathscr{P}$ be convex and assume $\rho(t; v) = v^{-1}P_v(t)$ has bounded second-order derivative with respect to t around 0, where $\mathscr{P}$ is defined in (4). Assume Conditions 1–6 hold with (r, ℓ_n, v) satisfying (16). The posterior distribution $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$ converges in total variation toward a Gaussian distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}_n, n^{-1}\widehat{\mathbf{H}}_{\mathcal{R}_n}^{-1})$ in probability, that is, $\mathcal{D}_{\mathrm{TV}}(\Pi_n^{\dagger}, \mathcal{N}_{\hat{\boldsymbol{\theta}}_n, n^{-1}\widehat{\mathbf{H}}_{\mathcal{R}_n}^{-1}}) \to 0$ in probability as $n \to \infty$, where $\hat{\boldsymbol{\theta}}_n$ is the PEL estimator in (3), and $\widehat{\mathbf{H}}_{\mathcal{R}_n}$ is defined in (12).

Theorem 2 shows that $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$ has a Gaussian limiting distribution and it concentrates on a $n^{-1/2}$-ball centred at the PEL estimator $\hat{\boldsymbol{\theta}}_n$ of interest, which indicates that $\hat{\boldsymbol{\theta}}_n$ can be approximated by the mean of the posterior distribution $\pi^{\dagger}(\boldsymbol{\theta} \mid \mathcal{X}_n)$. More specifically, as shown in Corollary 1, the approximation error is of order smaller than $n^{-1/2}$.

Corollary 1 Under the conditions of Theorem 2, we have $|\mathbb{E}_{\boldsymbol{\theta}\sim\pi^{\dagger}}(\boldsymbol{\theta}) - \hat{\boldsymbol{\theta}}_n|_{\infty} = o_{\mathrm{p}}(n^{-1/2})$, where $\hat{\boldsymbol{\theta}}_n$ is the PEL estimator defined as (3), and $\mathbb{E}_{\boldsymbol{\theta}\sim\pi^{\dagger}}(\boldsymbol{\theta})$ is defined in (8).

Theorems 3 and 4 state the theoretical guarantees for Algorithms 1 and 2, respectively.

Theorem 3 For the density $\phi(\cdot \mid \cdot)$ of the proposal distribution in Algorithm 1, we assume $\phi(\boldsymbol{\vartheta} \mid \boldsymbol{\theta})$ is positive and continuous on $(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \in \boldsymbol{\Theta} \times \boldsymbol{\Theta}$. Conditional on $\mathcal{X}_n$, for any $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}$ such that $\pi^{\dagger}(\boldsymbol{\theta}^0 \mid \mathcal{X}_n) > 0$ with $\pi^{\dagger}(\cdot \mid \mathcal{X}_n)$ defined as (7), it holds that $\mathcal{D}_{\mathrm{TV}}(\mathcal{T}_{\boldsymbol{\theta}^0}^k, \Pi_n^{\dagger}) \to 0$ as $k \to \infty$, where $\mathcal{T}_{\boldsymbol{\theta}^0}^k(\cdot)$ is the measure which admits the distribution of the Markov chain determined by Algorithm 1 at kth step with initial point $\boldsymbol{\theta}^0$. Furthermore, conditional on $\mathcal{X}_n$, we have $|K^{-1}\sum_{k=1}^{K}\boldsymbol{\theta}^k -$

$\mathbb{E}_{\boldsymbol{\theta}\sim\pi^\dagger}(\boldsymbol{\theta})|_\infty \to 0$ almost surely as $K \to \infty$, where $\{\boldsymbol{\theta}^k\}_{k\geq 1}$ are generated via Algorithm 1 with the initial point $\boldsymbol{\theta}^0$ satisfying $\pi^\dagger(\boldsymbol{\theta}^0 \mid \mathcal{X}_n) > 0$.

Theorem 4 For the density $\varphi(\cdot\,;\,\cdot)$ of the proposal distribution and the function $\mathbf{h}: \mathbb{R}^p \mapsto \mathbb{R}^s$ in Algorithm 2, we assume $\varphi(\boldsymbol{\theta};\boldsymbol{\zeta})$ is positive and continuous on $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \boldsymbol{\Theta} \times \mathbb{R}^s$ and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\mathbf{h}(\boldsymbol{\theta})|_\infty \leq K_9$ for some universal constant $K_9 > 0$. Conditional on $\mathcal{X}_n$, if $\sum_{k=1}^\infty \exp(-CN_k) < \infty$ for any $C > 0$, then $|\widehat{\mathbb{E}}_{\pi\dagger,K}(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}\sim\pi^\dagger}(\boldsymbol{\theta})|_\infty \to 0$ almost surely as $K \to \infty$, where $\widehat{\mathbb{E}}_{\pi\dagger,K}(\boldsymbol{\theta})$ is the MAMIS estimator defined as (9).

6 Discussion

In this paper, we explore BPEL and demonstrate its promising performance using MCMC sampling as a competitive alternative to optimization in addressing EL problems. This framework has the potential for further advancements in several areas. To maintain focus and avoid digressions, we have confined our study to fixed-dimensional model parameters and exponentially growing moment conditions. However, there is significant interest in extending this approach to tackle variable and model selection using BPEL, which could accommodate high-dimensional sparse model parameters and potentially a continuum of moment conditions, as considered in Chaussé (2017). Incorporating specific priors in the context of concrete studies, particularly in high-dimensional problems, is another area of interest. Research in this direction presents additional challenges, especially in selecting appropriate priors, developing efficient sampling schemes, and conducting associated analyses.

In the broader context of Bayesian methodology, approximate Bayesian computation (ABC) and Bayesian synthetic likelihood (BSL) are two competitive methods for handling situations where the likelihood is difficult to evaluate or intractable. ABC and BSL have been extensively compared in the literature. We demonstrate that the rationale of ABC integrates well with our BPEL method, achieving both accuracy and computational efficiency. Our Algorithm 2, inspired by ABC, uses importance weights for samples drawn from an alternative distribution to address challenging sampling situations. Empirical evidence shows promising performance, particularly in difficult cases. BSL leverages the limiting distribution, such as the normal distribution, to handle intractable probability distributions, with the advantage of easy sampling from the normal distribution. We view our BPEL as a compelling alternative to BSL: EL uses a multinomial likelihood that incorporates model information without requiring a fully specified parametric model, making it a competitive option when the full likelihood is intractable.

Furthermore, we foresee the use of more sophisticated sampling schemes in conjunction with PEL as highly valuable for addressing complex problems with specific considerations. Examples include the Hamiltonian MCMC method examined in Chaudhuri et al. (2017) and the variational Bayesian approach explored in Yu and Bondell (2024). These avenues of research are part of our plans for future projects.

Acknowledgments

The authors are grateful to the Co-Editor Professor Daniela Witten, an Associate Editor and two referees for their helpful suggestions.

Conflicts of interest: None declared.

Funding

J.C. and Y.Z. were supported in part by the National Natural Science Foundation of China (grant nos. 72125008, 72495122, and 71991472). C.Y.T. was supported in part by the US National Science Foundation (grant no. DMS-2210687) and the National Institutes of Health (grant no. R01GM140476).

Data availability

The data that support the findings of this paper are openly available at the GitHub repository: https://github.com/JinyuanChang-Lab/BayesianPenalizedEL.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Bissiri P. G., Holmes C. C., & Walker S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(5), 1103–1130. https://doi.org/10.1111/rssb.12158

Brooks S., Gelman A., Jones G., & Meng X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Castillo I., Schmidt-Hieber J., & van der Vaart A. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics*, *43*(5), 1986–2018. https://doi.org/10.1214/15-AOS1334

Chang J., Chen S. X., & Chen X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, *185*(1), 283–304. https://doi.org/10.1016/j.jeconom.2014.10.011

Chang J., Chen S. X., Tang C. Y., & Wu T. T. (2021). High-dimensional empirical likelihood inference. *Biometrika*, *108*(1), 127–147. https://doi.org/10.1093/biomet/asaa051

Chang J., Shi Z., & Zhang J. (2023). Culling the herd of moments with penalized empirical likelihood. *Journal of Business & Economic Statistics*, *41*(3), 791–805. https://doi.org/10.1080/07350015.2022.2071903

Chang J., Tang C. Y., & Wu Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Annals of Statistics*, *41*(4), 2123–2148. https://doi.org/10.1214/13-AOS1139

Chang J., Tang C. Y., & Wu T. (2018). A new scope of penalized empirical likelihood with high-dimensional estimating equations. *Annals of Statistics*, *46*(6B), 3185–3216. https://doi.org/10.1214/17-AOS1655

Chaudhuri S., & Ghosh M. (2011). Empirical likelihood for small area estimation. *Biometrika*, *98*(2), 473–480. https://doi.org/10.1093/biomet/asr004

Chaudhuri S., Mondal D., & Yin T. (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *79*(1), 293–320. https://doi.org/10.1111/rssb.12164

Chaussé P. (2017). Generalized empirical likelihood for a continuum of moment conditions. https://api.semanticsholar.org/CorpusID:73607404

Chen S. X., Peng L., & Qin Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, *96*(3), 711–722. https://doi.org/10.1093/biomet/asp037

Cheng Y., & Zhao Y. (2019). Bayesian jackknife empirical likelihood. *Biometrika*, *106*(4), 981–988. https://doi.org/10.1093/biomet/asz031

Chib S., Shin M., & Simoni A. (2018). Bayesian estimation and comparison of moment condition models. *Journal of the American Statistical Association*, *113*(524), 1656–1668. https://doi.org/10.1080/01621459.2017.1358172

Cornuet J.-M., Marin J.-M., Mira A., & Robert C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, *39*(4), 798–812. https://doi.org/10.1111/sjos.2012.39.issue-4

Donald S. G., Imbens G. W., & Newey W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, *117*(1), 55–93. https://doi.org/10.1016/S0304-4076(03)00118-0

Eaton J., Kortum S., & Kramarz F. (2011). An anatomy of international trade: Evidence from French firms. *Econometrica*, *79*(5), 1453–1498. https://doi.org/10.3982/ECTA8318

Gelman A., Gilks W. R., & Roberts G. O. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120. https://doi.org/10.1214/aoap/1034625254

Godambe V. P., & Heyde C. C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review*, *55*(3), 231–244. https://doi.org/10.2307/1403403

Hesterberg T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, *37*(2), 185–194. https://doi.org/10.1080/00401706.1995.10484303

Hjort N. L., McKeague I., & Van Keilegom I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics*, *37*(3), 1079–1111. https://doi.org/10.1214/07-AOS555

Jain P., & Kar P. (2017). Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, *10*(3–4), 142–336. https://doi.org/10.1561/2200000058

Lazar N. A. (2003). Bayesian empirical likelihood. *Biometrika*, *90*(2), 319–326. https://doi.org/10.1093/biomet/90.2.319

Leng C., & Tang C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, *99*(3), 703–716. https://doi.org/10.1093/biomet/ass014

Ma Y.-A., Chen Y., Jin C., Flammarion N., & Jordan M. I. (2019). Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences USA*, *116*(42), 20881–20885. https://doi.org/10.1073/pnas.1820003116

Marin J.-M., Pudlo P., & Sedki M. (2019). Consistency of adaptive importance sampling and recycling schemes. *Bernoulli*, *25*(3), 1977–1998. https://doi.org/10.3150/18-BEJ1042

Mengersen K. L., Pudlo P., & Robert C. P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences USA*, *110*(4), 1321–1326. https://doi.org/10.1073/pnas.1208827110

Narisetty N., & He X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, *42*(2), 789–817. https://doi.org/10.1214/14-AOS1207

Ouyang J., & Bondell H (2023). Bayesian analysis of longitudinal data via empirical likelihood. *Computational Statistics & Data Analysis*, *187*, 107785. https://doi.org/10.1016/j.csda.2023.107785

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

Qin J., & Lawless J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, *22*(1), 300–325. https://doi.org/10.1214/aos/1176325370

Rao J. N. K., & Wu C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*(4), 533–544. https://doi.org/10.1111/j.1467-9868.2010.00747.x

Ripley B. D. (2006). *Stochastic simulation*. Wiley-Interscience.

Roberts G. O., & Rosenthal J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, *1*, 20–71. https://doi.org/10.1214/154957804100000024

Shi Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics*, *195*(1), 104–119. https://doi.org/10.1016/j.jeconom.2016.07.004

Tang C. Y., & Leng C. (2010). Penalized high dimensional empirical likelihood. *Biometrika*, *97*(4), 905–920. https://doi.org/10.1093/biomet/asq057

Tang R., & Yang Y. (2022). Bayesian inference for risk minimization via exponentially tilted empirical likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(4), 1257–1286. https://doi.org/10.1111/rssb.12510

Tsao M. (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *Annals of Statistics*, *32*(3), 1215–1221. https://doi.org/10.1214/009053604000000337

Vexler A., Tao G., & Hutson A. D. (2014). Posterior expectation based on empirical likelihoods. *Biometrika*, *101*(3), 711–718. https://doi.org/10.1093/biomet/asu018

Yang Y., & He X. (2012). Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, *40*(2), 1102–1131. https://doi.org/10.1214/12-AOS1005

Yu W., & Bondell H. D. (2024). Variational Bayes for fast and accurate empirical likelihood inference. *Journal of the American Statistical Association*, *119*(546), 1089–1101. https://doi.org/10.1080/01621459.2023.2169701

Zhao P., Ghosh M., Rao J. N. K., & Wu C. (2020). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 155–174. https://doi.org/10.1111/rssb.12342