Biometrics, 2024, 80(2), ujae035
https://doi.org/10.1093/biomtc/ujae035
Biometric Methodology

Efficient data integration under prior probability shift

Ming-Yueh Huang [1,*], Jing Qin [2], Chiung-Yu Huang [3]

[1]Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, [2]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, United States, [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, United States

[*]Corresponding author: Ming-Yueh Huang, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan (myh0728@uw.edu).

ABSTRACT

Conventional supervised learning usually operates under the premise that data are collected from the same underlying population. However, challenges may arise when integrating new data from different populations, resulting in a phenomenon known as dataset shift. This paper focuses on prior probability shift, where the distribution of the outcome varies across datasets but the conditional distribution of features given the outcome remains the same. To tackle the challenges posed by such shift, we propose an estimation algorithm that can efficiently combine information from multiple sources. Unlike existing methods that are restricted to discrete outcomes, the proposed approach accommodates both discrete and continuous outcomes. It also handles high-dimensional covariate vectors through variable selection using an adaptive least absolute shrinkage and selection operator penalty, producing efficient estimates that possess the oracle property. Moreover, a novel semiparametric likelihood ratio test is proposed to check the validity of prior probability shift assumptions by embedding the null conditional density function into Neyman's smooth alternatives (Neyman, 1937) and testing study-specific parameters. We demonstrate the effectiveness of our proposed method through extensive simulations and a real data example. The proposed methods serve as a useful addition to the repertoire of tools for dealing dataset shifts.

KEYWORDS: dataset shift; penalized likelihood; profile likelihood; semiparametric efficiency.

1 INTRODUCTION

Conventional supervised learning assumes that data are collected from the same underlying population. When new data from different populations are used, dataset shift can occur, compromising the performance of the algorithm. As an example, during the COVID-19 pandemic, the University of Michigan Hospital's early sepsis warning system, the epic sepsis model (ESM), experienced alert fatigue concerns and was paused in April 2020. According to Wong et al. (2021), despite a 35% decline in hospital admissions, the number of daily sepsis alerts increased by 43% after the start of the pandemic. The unexpected surge in alerts was believed to be due to changes in the patient case mix caused by the COVID-19 pandemic: patients who were hospitalized during the COVID pandemic were, in general, in more critical condition and therefore more likely to trigger alerts. This highlights the importance of accounting for dataset shift when developing robust machine learning models. In addition to shifts in patient characteristics, other common causes of dataset shift include change in policies (eg, introduction of a new clinical guideline), technologies (eg, adoption of a new assay), and patient behavior (eg, rise in cancer screening after widespread media coverage of a celebrity's illness), etc. To delve further into the various types of dataset shift and the potential real-world scenarios where they may occur, readers can refer to Storkey (2009) and Finlayson et al. (2021).

This paper focuses on prior probability shift, a type of dataset shift that is also known as label shift when dealing with categorical responses. Under prior probability shift, the marginal distribution of the outcome variable varies across datasets but the conditional distribution of features given outcome remains the same. A real-world example can be found in COVID-19 rapid antigen test, where the binary outcome is the true COVID infection status. The effectiveness of a diagnostic test is typically assessed based on its sensitivity and specificity, corresponding to the conditional distributions of test results (feature) given infection status (outcome). The sensitivity and specificity of the test are constant across regions and epidemic phases, while the prevalence of COVID-19 infection can vary significantly. It is well-recognized that shifts in prevalence rate can affect the test's predictive power, with a positive antigen test result being more likely to accurately identify a true infection during periods of heightened transmission.

Furthermore, the prior probability shift can arise due to selection bias. In particular, outcome-dependent sampling, where the probability of a study subject being included in the data depends on the outcome, may distort the true relationship between the outcome and covariates. A classic example is the case-control design, a powerful tool for studying rare outcomes or conditions. In this design, cases and controls are selected based on their binary outcome status, often resulting in cases being over-represented and controls under-represented compared to their actual

Received: June 27, 2023; Revised: January 25, 2024; Accepted: April 25, 2024

proportions in the underlying population. Although case-control studies do not produce samples that are representative of the underlying population of interest, it can be shown that the conditional distribution of covariates, given the outcome status, remains the same. Note that outcome-dependent sampling is not restricted to binary outcomes. In Section 4.3, we report an analysis of Osaka housing price datasets for the years 2018 and 2019. The goal is to evaluate how specific characteristics of the house affect its price, a continuous outcome. We note that the transaction records of sold properties do not constitute a representative sample of the entire housing stock. This is because houses with higher prices are less frequently sold, thereby introducing bias into the data collection process. In this analysis, we show that the outcome-dependent sampling, along with a shift in the marginal distribution of house prices, results in a prior probability shift between different years.

To account for prior probability shift, Saerens et al. (2002) employed a plug-in method, utilizing the relationship between $\mathbf{X}$ given Y and Y given $\mathbf{X}$, to adjust the prediction model. When dealing with categorical responses, an alternative approach involves leveraging the confusion matrix of a predefined classifier to rectify the prediction model (Saerens et al., 2002; Lipton et al., 2018; Garg et al., 2020). These methods estimate $f_{\mathbf{X}|Y}(\mathbf{x}\,|\,y)$ solely based on a single dataset during the training phase, thus leaves room for efficiency improvement through the integration of multiple datasets. Furthermore, these approaches primarily emphasize prediction accuracy, neglecting the consideration of statistical inference for the corrected prediction models. In this paper, we propose a maximum likelihood estimation procedure that effectively integrates information from different data sources while accounting for prior probability shift. We show that the proposed estimator is asymptotically unbiased and semiparametrically efficient, making it optimal in terms of minimizing the asymptotic variance. As demonstrated in our numerical analysis, the proposed method outperforms existing approaches by effectively reducing prediction errors. More importantly, our method is applicable to both discrete and continuous response variables.

In ensuring accurate and reliable predictions, the detection of prior probability shift plays a crucial role. Recognition of the presence of prior probability shift often relies on domain knowledge and careful assessment of new information. Incorporating a data-driven procedure into artificial intelligence systems can facilitate early detection of prior probability shift, allowing for timely adjustments to the prediction algorithm. Under the assumption of the same conditional density $f_{\mathbf{X}|Y}(\mathbf{x}\,|\,y)$, Saerens et al. (2002) proposed a likelihood ratio test for detecting differences in the outcome distribution. However, there remains a gap in checking whether $f_{\mathbf{X}|Y}(\mathbf{x}\,|\,y)$ remains the same across datasets. To bridge this gap, we introduce a novel test by embedding the null distribution $f_{\mathbf{X}|Y}(\mathbf{x}\,|\,y)$ into a smooth alternative (Neyman, 1937) and construct a semiparametric likelihood ratio test for testing study-specific parameters that characterize heterogeneity in the conditional distribution. The χ^2 distribution of the semiparametric likelihood ratio test statistic is established using the profile likelihood theory of Murphy and van der Vaart (2000). This innovative test enables the examination of the constancy of the conditional density $f_{\mathbf{X}|Y}(\mathbf{x}\,|\,y)$ across datasets and provides valuable insights into the presence of prior probability shift in prediction tasks.

The high-dimensional nature of covariates poses another challenge in developing machine learning algorithms, as it often leads to overfitting and reduces prediction performance. To address this issue, variable selection plays a crucial step in improving prediction accuracy and obtaining a more parsimonious model (Fan and Lv, 2010). Various penalization methods have been proposed to perform simultaneous model/variable selection and parameter estimation, including the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). The SCAD estimator possesses the oracle property, whereas the LASSO estimator is not consistent unless a nontrivial condition is met. Zou (2006) introduced an adaptive LASSO estimator to address this issue. Although initially developed for parametric models, we incorporate the adaptive LASSO penalty into the profile likelihood for handling high-dimensional covariate vectors within the semiparametric framework of the prior probability shift model. We show that the resulting maximum penalized likelihood estimator enjoys the oracle property.

The rest of this article is organized as follows. In Section 2, we propose a maximum likelihood estimation to estimate the prediction model under the prior probability shift and a testing procedure for the prior probability shift assumption. Section 3 extends the proposed method to deal with high-dimensional covariates by incorporating an adaptive LASSO penalty. We evaluate the performance of the proposed estimation, tests, and variable selection via Monte Carlo simulations and two real data examples in Section 4.

2 ESTIMATION AND TESTING UNDER PRIOR PROBABILITY SHIFT

2.1 Profile likelihood estimation

We begin by considering the case where two datasets were generated from populations with prior probability shift. Let Y_k be the outcome of interest and $\mathbf{X}_k$ be the vector of covariates in the kth population $(k = 1, 2)$. Here, we focus on the case of 2 datasets for clarity. However, our method can easily be extended to accommodate multiple datasets, as detailed in Section S1 of Supplementary Materials. The 2 datasets $\mathcal{D}_k = \{(Y_{ki}, \mathbf{X}_{ki}) : i = 1, \ldots, n_k\}$ $(k = 1, 2)$ consist of random samples from their respective populations and are assumed to be independent. Assume that the conditional density of Y_1 given $\mathbf{X}_1 = \mathbf{x}$ follows a parametric model $f(y\,|\,\mathbf{x};\boldsymbol{\theta})$, where f is a known function and $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter. The marginal distribution function of $\mathbf{X}_1$ is denoted by $G_1(\mathbf{x})$, and is left unspecified. The prior probability shift model assumes that the conditional density of $\mathbf{X}_k$ given $Y_k = y$, denoted by $f_k(\mathbf{x}\,|\,y)$, is identical for $k = 1, 2$. Applying the Bayes rule yields

$$f_k(\mathbf{x}\,|\,y)\mathrm{d}\mathbf{x} = \frac{f(y\,|\,\mathbf{x};\boldsymbol{\theta})\mathrm{d}G_1(\mathbf{x})}{\int f(y\,|\,\mathbf{u};\boldsymbol{\theta})\mathrm{d}G_1(\mathbf{u})} \quad (k = 1, 2). \qquad (1)$$

It further implies that the joint density of $(Y_2, \mathbf{X}_2)$ is

$$f_2(y, \mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}y = \frac{f(y \mid \mathbf{x}; \boldsymbol{\theta})\mathrm{d}G_1(\mathbf{x})}{\int f(y \mid \mathbf{u}; \boldsymbol{\theta})\mathrm{d}G_1(\mathbf{u})}\mathrm{d}F_2(y), \tag{2}$$

where $F_2(y)$ denotes the marginal distributions of Y_2, and is left unspecified. Consequently, we characterize the distributions subject to prior probability shift using a semiparametric model, wherein $\boldsymbol{\theta}$ is the finite-dimensional parameter and (G_1, F_2) are infinite-dimensional parameters. Here we focus on continuous outcome variables for simplicity, but the concepts and methods can be readily extended to discrete cases by replacing the probability density function with a probability mass function.

Saerens et al. (2002) proposed replacing $(\boldsymbol{\theta}, G_1)$ in (2) with estimators obtained using dataset $\mathcal{D}_1$ to construct prediction algorithm. However, this method only utilizes information from $\mathcal{D}_1$ and thus is expected to be inefficient. To address this limitation, we propose a maximum likelihood estimation approach that utilizes the full dataset $\mathcal{D}_1 \cup \mathcal{D}_2$, allowing us to achieve optimal efficiency. Note that the semiparametric log-likelihood function with full dataset is given by

$$\ell(\boldsymbol{\theta}, G_1, F_2) = \sum_{i=1}^{n_1} \log\{f(Y_{1i} \mid \mathbf{X}_{1i}; \boldsymbol{\theta})\Delta G_1(\mathbf{X}_{1i})\} + \sum_{i=1}^{n_2} \log\left\{\frac{f(Y_{2i} \mid \mathbf{X}_{2i}; \boldsymbol{\theta})\Delta G_1(\mathbf{X}_{2i})}{\int f(Y_{2i} \mid \mathbf{x}; \boldsymbol{\theta})\mathrm{d}G_1(\mathbf{x})}\Delta F_2(Y_{2i})\right\},$$

where $\Delta G_1(\mathbf{x})$ is the jump size of G_1 at the value $\mathbf{x}$ and $\Delta F_2(y)$ is the jump size of F_2 at the value of y. The log-likelihood function achieves its maximum if and only if G_1 has jumps on the distinct observed values of $\{\mathbf{X}_{1i}: i = 1, \ldots, n_1\} \cup \{\mathbf{X}_{2i}: i = 1, \ldots, n_2\}$ and F_2 has jumps on the distinct observed values of $\{Y_{2i}: i = 1, \ldots, n_2\}$; see Section S1 of Supplementary Materials. For ease of notation, we denote $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\} = \cup_{k=1}^{2} \mathcal{D}_k$, where $n = n_1 + n_2$. Let p_i be the jump size of G_1 at $\mathbf{X}_i$ $(i = 1, \ldots, n)$ and q_i be the jump size of F_2 at Y_{2i} $(i = 1, \ldots, n_2)$. In the presence of ties, the jump sizes of G_1 and F_2 are obtained by summing the corresponding p_i and q_i at the tied values. Denote $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_{n_2})$. With a slight abuse of notation, the log-likelihood function can be re-expressed as

$$\ell(\boldsymbol{\theta}, \mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \log f(Y_i \mid \mathbf{X}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log p_i - \sum_{i=1}^{n_2} \log \sum_{j=1}^{n} f(Y_{2i} \mid \mathbf{X}_j; \boldsymbol{\theta}) p_j + \sum_{i=1}^{n_2} \log q_i,$$

subject to the constraints $p_i, q_i \geq 0$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} p_i = 1$, and $\sum_{i=1}^{n_2} q_i = 1$. Clearly, the maximizer with respect to $\mathbf{q}$ is $\widehat{\mathbf{q}} = (\widehat{q}_1, \ldots, \widehat{q}_{n_2})$ with $\widehat{q}_i = n_2^{-1}$, $i = 1, \ldots, n_2$. That is, the nonparametric maximum likelihood estimator $\widehat{F}_2$ of F_2 is the empirical distribution of $\{Y_{2i}: i = 1, \ldots, n_2\}$. As the dimension of $\mathbf{p}$ increases with the sample size n, traditional optimization algorithms may not give a stable solution with respect to $\mathbf{p}$ when n is large. To tackle this issue, we utilize a profile likelihood approach, which offers an explicit and efficient estimation algorithm.

For each fixed $\boldsymbol{\theta}$, define $\widehat{\mathbf{p}}(\boldsymbol{\theta}) = \mathrm{argmax}_{\mathbf{p} \in \mathcal{P}_n} \ell(\boldsymbol{\theta}, \mathbf{p}, \widehat{\mathbf{q}})$. By applying the Lagrange method, $\widehat{\mathbf{p}}(\boldsymbol{\theta})$ is a maximizer of the Lagrange function

$$\mathcal{L}(\mathbf{p}, \eta) = \sum_{i=1}^{n} \log f(Y_i \mid \mathbf{X}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log p_i - \sum_{i=1}^{n_2} \log \sum_{j=1}^{n} f(Y_{2i} \mid \mathbf{X}_j; \boldsymbol{\theta}) p_j + \eta\left(\sum_{i=1}^{n} p_i - 1\right),$$

with η being a Lagrange multiplier. By differentiating $\mathcal{L}(\mathbf{p}, \eta)$ with respect to $(\mathbf{p}, \eta)$, we can show that $\eta = -n_1$ and $\widehat{\mathbf{p}}(\boldsymbol{\theta})$ solves the following equations:

$$p_i = \left\{n_1 + \sum_{j=1}^{n_2} \frac{f(Y_{2j} \mid \mathbf{X}_i; \boldsymbol{\theta})}{\sum_{k=1}^{n} f(Y_{2j} \mid \mathbf{X}_k; \boldsymbol{\theta}) p_k}\right\}^{-1}, \quad i = 1, \ldots, n. \tag{3}$$

The detailed derivation is given in Section S2 of Supplementary Materials. Let $\mathbf{T}(\mathbf{p}) = (T_1(\mathbf{p}), \ldots, T_n(\mathbf{p}))$ with $T_i(\mathbf{p})$ being the right-hand side of (3) for $i = 1, \ldots, n$. The equations can be rewritten as $\mathbf{p} = \mathbf{T}(\mathbf{p})$, which implies that the root is a fixed point of $\mathbf{T}(\mathbf{p})$. Following a standard argument of fixed-point theory, the root can be obtained by iteratively computing $\mathbf{T}(\mathbf{p})$ with a proper initial point. A pseudocode of this inner loop for the profile likelihood estimation is given by the function `Inner` in Algorithm 1. It is worth pointing out that $\ell(\boldsymbol{\theta}, \mathbf{p}, \widehat{\mathbf{q}})$ has a unique maximizer $\widehat{\mathbf{p}}(\boldsymbol{\theta})$ for each given $\boldsymbol{\theta}$, and the sequence $\widehat{\mathbf{p}}^{(m)}$ in Algorithm 1 converges to $\widehat{\mathbf{p}}(\boldsymbol{\theta})$ regardless of the initial point chosen. These results are summarized in the following theorem and their proofs are given in Section S3 of Supplementary Materials.

Theorem 1 *For any initial $\widehat{\mathbf{p}}^{(0)} \in \mathcal{P}_n$, the corresponding sequence $\{\widehat{\mathbf{p}}^{(m)} : m \geq 0\}$ in Algorithm 1 converges to the unique maximizer of $\ell(\boldsymbol{\theta}, \mathbf{p}, \widehat{\mathbf{q}})$ with respect to $\mathbf{p} \in \mathcal{P}_n$.*

The log-profile likelihood function of $\boldsymbol{\theta}$ is defined as $\ell_{\mathrm{pr}}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, \widehat{\mathbf{p}}(\boldsymbol{\theta}), \widehat{\mathbf{q}})$. To obtain the maximum profile likelihood estimator $\widehat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} \ell\{\boldsymbol{\theta}, \widehat{\mathbf{p}}(\boldsymbol{\theta}), \widehat{\mathbf{q}}\}$, we use the build-in R function `nlminb()`, which optimizes a multivariate function using a quasi-Newton method and a finite difference gradient. After obtaining $\widehat{\boldsymbol{\theta}}$, the marginal distribution of $\mathbf{X}_1$ can be estimated by $\widehat{G}_1(\mathbf{x}) = \sum_{i=1}^{n} \widehat{p}_i(\widehat{\boldsymbol{\theta}}) I(\mathbf{X}_i \leq \mathbf{x})$, where the inequality is applied component-wise. The pseudo code of the full estimation procedure is given by Algorithm 1.

Theoretically, the proposed method can treat either dataset as the starting point, provided that the model $f(y \mid \mathbf{x}; \boldsymbol{\theta})$ for the initial dataset is correctly specified. In practice, the choice of the starting dataset typically depends on practical considerations, often prioritizing the dataset collected earliest or originating from a primary study that motivates the analysis. If there is no specific dataset of primary interest, initiating the analysis with the dataset that has the largest sample size is recommended. This allows for a more robust check and verification of the initial model $f(y \mid \mathbf{x}; \boldsymbol{\theta})$.

Algorithm 1: The profile likelihood estimation

Input : Data $\mathcal{D}_1$ and $\mathcal{D}_2$, initial $\widehat{\mathbf{p}}^{(0)} = (p_1^{(0)}, \ldots, p_n^{(0)})$, and tolerance $\epsilon \rangle 0$

Output: $\widehat{\boldsymbol{\theta}}, \widehat{G}_1, \widehat{F}_2$

$\widehat{\mathbf{q}} \leftarrow (\Delta\widehat{F}_2(Y_{21}), \ldots, \Delta\widehat{F}_2(Y_{2n_2})) \leftarrow (n_2^{-1}, \ldots, n_2^{-1})$;

Function `Inner` $(\boldsymbol{\theta})$

$m \leftarrow 1$;

repeat

for $i \leftarrow 1$ **to** n **do**

$$\widehat{p}_i^{(m)} \leftarrow \left\{ n_1 + \sum_{j=1}^{n_2} \frac{f(Y_{2j} \mid \mathbf{X}_i; \boldsymbol{\theta})}{\sum_{k=1}^{n} f(Y_{2j} \mid \mathbf{X}_k; \boldsymbol{\theta}) \widehat{p}_k^{(m-1)}} \right\}^{-1};$$

$\widehat{\mathbf{p}}^{(m)} \leftarrow (p_1^{(m)}, \ldots, p_n^{(m)})$;

$m \leftarrow m + 1$;

until $\|\widehat{\mathbf{p}}^{(m)} - \widehat{\mathbf{p}}^{(m-1)}\| \langle \epsilon$;

Return $\widehat{\mathbf{p}}(\boldsymbol{\theta}) = \widehat{\mathbf{p}}^{(m)}$;

End Function;

$\widehat{\boldsymbol{\theta}} \leftarrow \text{argmax}_{\boldsymbol{\theta}} \ell\{\boldsymbol{\theta}, \texttt{Inner}(\boldsymbol{\theta}), \widehat{\mathbf{q}}\}$;

$(\Delta\widehat{G}_1(\mathbf{X}_1), \ldots, \Delta\widehat{G}_1(\mathbf{X}_n)) \leftarrow \texttt{Inner}(\widehat{\boldsymbol{\theta}})$;

2.2 Statistical inference and prediction

We now establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}$ by applying the profile likelihood theory (Murphy et al., 1997; Murphy and van der Vaart, 2000). Let $(\boldsymbol{\theta}_0, G_{10}, F_{20})$ be the true parameter values. With a slight abuse of notation, write $\mathbf{Z} = (Y_1, \mathbf{X}_1, Y_2, \mathbf{X}_2)$ and

$$\widetilde{\ell}(\mathbf{Z}; \boldsymbol{\theta}, G_1, F_2) = \kappa \log\{f(Y_1 \mid \mathbf{X}_1; \boldsymbol{\theta}) \mathrm{d}G_1(\mathbf{X}_1)\} + (1-\kappa) \log\left\{ \frac{f(Y_2 \mid \mathbf{X}_2; \boldsymbol{\theta}) \mathrm{d}G_1(\mathbf{X}_2)}{\int f(Y_2 \mid \mathbf{x}; \boldsymbol{\theta}) \mathrm{d}G_1(\mathbf{x})} \mathrm{d}F_2(Y_2) \right\},$$

where κ is the limit of $n_1/(n_1 + n_2)$ as $n_1, n_2 \to \infty$. Under Condition 3 and applying the law of large number, we can show that $n^{-1}\ell(\boldsymbol{\theta}, G_1, F_2)$ converges to $\widetilde{\ell}(\boldsymbol{\theta}, G_1, F_2) = \mathrm{E}\{\widetilde{\ell}(\mathbf{Z}; \boldsymbol{\theta}, G_1, F_2)\}$ as both n_1 and n_2 go to infinity. Coupled with the Glivenko-Cantelli result given in Section S5 of Supplementary Materials, this convergence is further shown to be uniform over $(\boldsymbol{\theta}, G_1, F_2)$. Combining with the identifiability of the parameter $(\boldsymbol{\theta}, G_1, F_2)$, whose proof is given in Section S6 of Supplementary Materials, we establish the consistency of the proposed estimator in the following theorem.

Theorem 2 (Consistency) *Suppose that Conditions 1-5 in Supplementary Materials are satisfied. Then,* $\sup_{\mathbf{x}} |\widehat{G}_1(\mathbf{x}) - G_{10}(\mathbf{x})|$ *and* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$ *converge almost surely to zero as* $n_1 \to \infty$ *and* $n_2 \to \infty$.

To establish the asymptotic normality and the semiparametric efficiency, we first derive the information operator. For a given direction $\mathbf{e} = (\mathbf{v}, u_1, u_2)$ with $\mathbf{v} \in \mathbb{R}^d$ and non-negative bounded functions $u_1(\mathbf{x})$ and $u_2(y)$, the information operator $\boldsymbol{\sigma}[\mathbf{e}] = (\boldsymbol{\sigma}_{\boldsymbol{\theta}}[\mathbf{e}], \sigma_{G_1}[\mathbf{e}], \sigma_{F_2}[\mathbf{e}])$ satisfies

$$\begin{aligned} &\mathrm{E}[\{\mathbf{v}^{\mathrm{T}} \partial_{\boldsymbol{\theta}} \widetilde{\ell}(\mathbf{Z}; \boldsymbol{\theta}_0, G_{10}, F_{20}) + \partial_{G_1} \widetilde{\ell}(\mathbf{Z}; \boldsymbol{\theta}_0, G_{10}, F_{20})[u_1] \\ &\quad + \partial_{F_2} \widetilde{\ell}(\mathbf{Z}; \boldsymbol{\theta}_0, G_{10}, F_{20})[u_2]\}^2] \\ &= \mathbf{v}^{\mathrm{T}} \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{e}) + \int \sigma_{G_1}[\mathbf{e}](\mathbf{x}) u_1(\mathbf{x}) \mathrm{d}G_{10}(\mathbf{x}) \\ &\quad + \int \sigma_{F_2}[\mathbf{e}](y) u_2(y) \mathrm{d}F_{20}(y), \end{aligned}$$

whose explicit form is given in Section S8 of Supplementary Materials. By observing that $\boldsymbol{\sigma}$ is a linear operator, Condition 2 ensures that $\boldsymbol{\sigma}$ is onto and continuously invertible. Let $\bar{\boldsymbol{\sigma}} = (\bar{\boldsymbol{\sigma}}_{\boldsymbol{\theta}}, \bar{\sigma}_{G_1}, \bar{\sigma}_{F_2})$ denote the inverse of $\boldsymbol{\sigma}$, $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ be the standard basis of $\mathbb{R}^d$, and $\boldsymbol{\Sigma}^{-1} = \{\bar{\boldsymbol{\sigma}}_{\boldsymbol{\theta}}[(\mathbf{e}_1, 0, 0)], \ldots, \bar{\sigma}_{\boldsymbol{\theta}}[(\mathbf{e}_d, 0, 0)]\}$. In Theorems 3, we show that $\boldsymbol{\Sigma}^{-1}$ is the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$. Moreover, Theorem 4 shows that $\widehat{\boldsymbol{\theta}}$ is semiparametric efficient for estimating $\boldsymbol{\theta}$.

Theorem 3 (Asymptotic normality) *Suppose that Conditions 1-5 in Supplementary Materials are satisfied. Then,* $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *converges in distribution to a d-variate normal distribution with mean zero and variance-covariance matrix* $\boldsymbol{\Sigma}^{-1}$.

Theorem 4 (Semiparametric efficiency) *Suppose that Conditions 1-5 in Supplementary Materials are satisfied. Then,* $\boldsymbol{\Sigma}^{-1}$ *is the semiparametric efficiency bound for estimating* $\boldsymbol{\theta}$.

We show in the proof of Theorem 3 that the information matrix $\boldsymbol{\Sigma}$ is identical to the negative second derivative of $\ell_{\mathrm{pr}}(\boldsymbol{\theta})$, which can be approximated by using finite differences. Specifically, we estimate $\boldsymbol{\Sigma} = (\Sigma_{ij})_{d\times d}$ by

$$\widehat{\Sigma}_{ii} = -\frac{\ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}} + \epsilon_i \mathbf{e}_i) - 2\ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}}) + \ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}} - \epsilon_i \mathbf{e}_i)}{n\epsilon_i^2},$$

$$\widehat{\Sigma}_{ij} = -\frac{\ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}} + \epsilon_i \mathbf{e}_i + \epsilon_j \mathbf{e}_j) - \ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}} + \epsilon_i \mathbf{e}_i) - \ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}} + \epsilon_j \mathbf{e}_j) + \ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}})}{n\epsilon_i\epsilon_j}, \quad i \neq j,$$

where ϵ_i and ϵ_j are pre-chosen small values. Let $\widehat{\mathbf{V}} = (V_{ij})_{d\times d} = \widehat{\boldsymbol{\Sigma}}^{-1}/n$. Then, a $(1-\alpha)$ confidence interval for the kth component of $\boldsymbol{\theta}$ is given by $\widehat{\boldsymbol{\theta}}_k \pm z_{1-\alpha/2}\widehat{V}_{kk}^{1/2}$ $(k = 1, \ldots, d)$, where z_p is the pth quantile of standard normal distribution. To test the null hypotheses H_0: $\theta_k = \theta_{k0}$ against the two-sided alternative H_A: $\theta_k \neq \theta_{k0}$, we consider the profile likelihood ratio statistic

$$\Lambda_{\mathrm{pr}}(\theta_{k0}) = 2\{\ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}}) - \ell_{\mathrm{pr}}(\widehat{\boldsymbol{\theta}}_0)\},$$

where $\widehat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ under H_0. By applying Corollary 2 of Murphy and van der Vaart (2000), it

can be shown that $\Lambda_{\text{pr}}(\theta_{k0})$ follows the χ^2_1 distribution under H_0 when n is sufficiently large.

Our proposed efficient estimator can further improve prediction of future outcomes. Let Y_k^{new} be the outcome corresponding to a new subject from the kth population ($k = 1, 2$) with covariate $\mathbf{X}_k^{\text{new}}$. Recall that Y_1^{new} given $\mathbf{X}_2^{\text{new}}$ follows the distribution $f(y \mid \mathbf{X}_1^{\text{new}}; \boldsymbol{\theta})$. Moreover, it follows from (2) that the density function of Y_2^{new} given $\mathbf{X}_2^{\text{new}}$ is

$$\frac{f(y \mid \mathbf{X}_2^{\text{new}}; \boldsymbol{\theta})\mathrm{d}R(y)}{\int f(s \mid \mathbf{X}_2^{\text{new}}; \boldsymbol{\theta})\mathrm{d}R(s)}, \tag{4}$$

where $\mathrm{d}R(y) = \mathrm{d}F_2(y)/\int f(y \mid \mathbf{x}; \boldsymbol{\theta})\mathrm{d}G_1(\mathbf{x})$ is the density ratio of Y_2 compared to Y_1. In the case of continuous outcomes, it is known that the optimal predictor for minimizing the mean squared prediction error is the conditional mean of the response given the covariates. As $\boldsymbol{\theta}$, G_1, and F_2 are unknown parameters, we plug-in the proposed efficient estimator and predict the new data Y_k^{new} with the conditional expectation of Y_k^{new} given $\mathbf{X}_k^{\text{new}}$ ($k = 1, 2$):

$$\widehat{Y}_1 = \int y f(y \mid \mathbf{X}_1^{\text{new}}; \widehat{\boldsymbol{\theta}})\mathrm{d}y \text{ and}$$

$$\widehat{Y}_2 = \frac{\int y f(y \mid \mathbf{X}_2^{\text{new}}; \widehat{\boldsymbol{\theta}})\mathrm{d}\widehat{R}(y)}{\int f(y \mid \mathbf{X}_2^{\text{new}}; \widehat{\boldsymbol{\theta}})\mathrm{d}\widehat{R}(y)}, \tag{5}$$

where $\mathrm{d}\widehat{R}(y) = \mathrm{d}\widehat{F}_2(y)/\int f(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})\mathrm{d}\widehat{G}_1(\mathbf{x})$. In the case of categorical outcome variables, say $Y_1, Y_2 \in \{y_1, \ldots, y_K\}$, the optimal predictor for minimizing the misclassification rate is to classify the subject to the category with highest posterior probabilities, which is also known as the Bayes optimal classifier. Thus, we predict Y_k^{new} ($k = 1, 2$) by

$$\widehat{Y}_1 = \underset{y \in \{y_1, \ldots, y_K\}}{\text{argmax}} \; f(y \mid \mathbf{X}_1^{\text{new}}; \widehat{\boldsymbol{\theta}}) \text{ and}$$

$$\widehat{Y}_2 = \underset{y \in \{y_1, \ldots, y_K\}}{\text{argmax}} \; \frac{f(y \mid \mathbf{X}_2^{\text{new}}; \widehat{\boldsymbol{\theta}})\Delta\widehat{R}(y)}{\sum_{k=1}^{K} f(y_k \mid \mathbf{X}_2^{\text{new}}; \widehat{\boldsymbol{\theta}})\Delta\widehat{R}(y_k)}, \tag{6}$$

where $\Delta\widehat{R}(y) = \Delta\widehat{F}_2(y)/\int f(y \mid \mathbf{x}; \widehat{\boldsymbol{\theta}})\mathrm{d}\widehat{G}_1(\mathbf{x})$.

2.3 Testing prior probability Shift

In this subsection, we propose a formal statistical test for assessing the prior probability shift assumption. Specifically, our goal is to test the null hypothesis that a common conditional density is shared across different datasets, that is, $H_0 : f_1(\mathbf{x} \mid y) \equiv f_2(\mathbf{x} \mid y)$. We employ a framework akin to the general methodology delineated in Thams et al. (2023) to elucidate our strategy. Let $\mathcal{F}_k$ be the collections of all possible joint density functions $f_k(y, \mathbf{x})$ with f_k^* representing the true models, $k = 1, 2$, for the 2 underlying populations. Recall that, under H_0, the joint density of $(Y_2, \mathbf{X}_2)$ is given by $f_2(y, \mathbf{x})\mathrm{d}y = f_1(y, \mathbf{x})\mathrm{d}F_2(y)/\int f_1(y, \mathbf{u})\mathrm{d}\mathbf{u}$. In other words, the prior probability shift assumption defines a map τ from $\mathcal{F}_1$ to $\mathcal{F}_2$, where $\tau(f_1) = f_1(y, \mathbf{x})\mathrm{d}F_2(y)/\int f_1(y, \mathbf{u})\mathrm{d}\mathbf{u}$. Now let $\mathcal{M} = \{f(y \mid \mathbf{x}; \boldsymbol{\theta})\mathrm{d}G_1(\mathbf{x})\mathrm{d}y\}$ be the family of distributions that satisfy our modeling assumption on $f_1(y, \mathbf{x})$, which is hence a subset of $\mathcal{F}_1$. This way, we have $f_1^* \in \mathcal{M}$, and the null hypothesis can then be reexpressed as $H_0 : \tau(f_1^*) = f_2^*$ or, equivalently, $H_0 : f_2^* \in \tau(\mathcal{M})$. Given that the null hypothesis is now solely a statement regarding f_2^*, constructing a test for H_0 solely necessitates the identification of suitable alternatives for f_2^*. While $\mathcal{F}_2$ may be an intuitive choice for the alternatives, its broad scope might result in a test statistic with excessive variability, thereby compromising the statistical power of the test. Instead, we suggest employing a proper subset H_A of $\mathcal{F}_2$ as the alternatives. This concept is illustrated in Figure 1.

Following the spirit of Neyman's smooth goodness-of-fit test, we propose to embed the null density $f_2(\mathbf{x} \mid y)$ within a larger parametric family of densities that differ smoothly from the null density. Specifically, we construct a smooth alternative of order L as

$$f_2(\mathbf{x} \mid y; \boldsymbol{\theta}, G_1, \boldsymbol{\gamma})\mathrm{d}\mathbf{x} = \frac{f(y \mid \mathbf{x}; \boldsymbol{\theta}) \exp\{\sum_{l=1}^{L} \gamma_l h_l(\mathbf{x}; \boldsymbol{\theta}, G_1)\}\mathrm{d}G_1(\mathbf{x})}{\int f(y \mid \mathbf{u}; \boldsymbol{\theta}) \exp\{\sum_{l=1}^{L} \gamma_l h_l(\mathbf{u}; \boldsymbol{\theta}, G_1)\}\mathrm{d}G_1(\mathbf{u})}. \tag{7}$$

Here, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$ is a L-dimensional vector of parameters and $\{h_l(\mathbf{x}; \boldsymbol{\theta}, G_1) : l = 1, \ldots, L\}$ is a set of orthonormal functions. In other words, we consider the alternative hypothesis $H_A : f_2(y, \mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}y = f_2(\mathbf{x} \mid y; \boldsymbol{\theta}, G_1, \boldsymbol{\gamma})\mathrm{d}F_2(y)\mathrm{d}\mathbf{x}$, in which the null is contained as a special case with $\boldsymbol{\gamma} = \mathbf{0}$. As pointed out in Rayner and Best (1990), the choice of $\{h_l(\mathbf{x}; \boldsymbol{\theta}, G_1) : l = 1, \ldots, L\}$ depends on the specific alternative density functions. For instance, when the null distribution is uniform and the alternative has a difference in mean relative to the null, Neyman utilized the first-order Legendre polynomial to detect the mean difference. Analogously, employing Legendre polynomials with a second, third, and the fourth order terms would allow the detection of differences in variance, skewness, and kurtosis, respectively. In the case of multivariate distribution, Macdonald polynomials, which are extensions of univariate Legendre polynomials, can be utilized. Notably, the first-order Macdonald polynomials are $x_1, \ldots, x_p$, each of which coincides with a first-order Legendre polynomial.

It is easy to see that, under (7), testing the null hypothesis is equivalent to testing $\boldsymbol{\gamma} = \mathbf{0}$. Under the smooth alternative, the log-likelihood is

$$\ell(\boldsymbol{\theta}, G_1, F_2, \boldsymbol{\gamma}) = \sum_{i=1}^{n_1} \log\{f(Y_{1i} \mid \mathbf{X}_{1i}; \boldsymbol{\theta})\Delta G_1(\mathbf{X}_{1i})\} + \sum_{i=1}^{n_2} \log\left\{\frac{f(Y_{2i} \mid \mathbf{X}_{2i}; \boldsymbol{\theta}) \exp\{\sum_{l=1}^{L} \gamma_l h_l(\mathbf{X}_{2i}; \boldsymbol{\theta}, G_1)\}\Delta G_1(\mathbf{X}_{2i})}{\int f(Y_{2i} \mid \mathbf{x}; \boldsymbol{\theta}) \exp\{\sum_{l=1}^{L} \gamma_l h_l(\mathbf{x}; \boldsymbol{\theta}, G_1)\}\mathrm{d}G_1(\mathbf{x})}\Delta F_2(Y_{2i})\right\}.$$

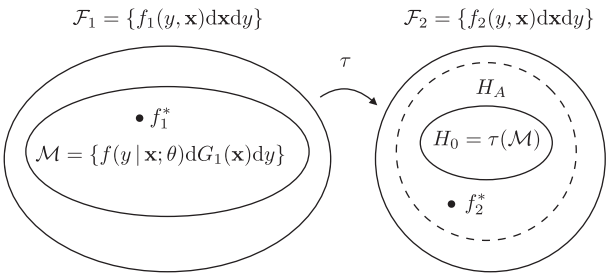

FIGURE 1 Illustration of distribution models $\mathcal{F}_1$ and $\mathcal{F}_2$ on heterogeneous populations, prior probability shift as a map τ, target hypothesis H_0, and the smooth alternatives H_A.

To test $\boldsymbol{\gamma} = \mathbf{0}$, we consider the likelihood ratio test statistic $R = -2\{\ell(\widehat{\boldsymbol{\theta}}, \widehat{G}_1, \widehat{F}_2, \mathbf{0}) - \ell(\widehat{\boldsymbol{\theta}}^A, \widehat{G}_1^A, \widehat{F}_2^A, \widehat{\boldsymbol{\gamma}}^A)\}$, where $(\widehat{\boldsymbol{\theta}}^A, \widehat{G}_1^A, \widehat{F}_2^A, \widehat{\boldsymbol{\gamma}}^A)$ is the maximum likelihood estimator that maximizes $\ell(\boldsymbol{\theta}, G_1, F_2, \boldsymbol{\gamma})$ under the alternative density. It is noted that the the maximizer $\widehat{F}_2^A$ coincides with $\widehat{F}_2$, and the maximization of $\ell(\boldsymbol{\theta}, G_1, \widehat{F}_2, \boldsymbol{\gamma})$ with respect to $(\boldsymbol{\theta}, G_1, \boldsymbol{\gamma})$ can be solved by a profile likelihood approach. More precisely, we modify the `Inner` function in Algorithm 1 to calculate $\widehat{\mathbf{p}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathrm{argmax}_{\mathbf{p} \in \mathcal{P}_n} \ell(\boldsymbol{\theta}, \mathbf{p}, \widehat{\mathbf{q}}, \boldsymbol{\gamma})$ for each given $(\boldsymbol{\theta}, \boldsymbol{\gamma})$; see the `AltInner` function in Algorithm A1 of Supplementary Materials. Subsequently, the maximum profile likelihood estimator under the alternatives is given by $(\widehat{\boldsymbol{\theta}}^A, \widehat{\boldsymbol{\gamma}}^A) = \mathrm{argmax}_{(\boldsymbol{\theta}, \boldsymbol{\gamma})} \ell(\boldsymbol{\theta}, \widehat{\mathbf{p}}(\boldsymbol{\theta}, \boldsymbol{\gamma}), \widehat{q}, \boldsymbol{\gamma})$. By applying Corollary 2 in Murphy and van der Vaart (2000), it can be shown that R follows an asymptotic χ^2_L distribution under H_0. Therefore, the prior probability shift assumption is rejected at a significance level of α if $R > \chi^2_{L,1-\alpha}$, where $\chi^2_{L,1-\alpha}$ represents the $(1-\alpha)$ quantile of the χ^2_L distribution. The pseudo code of the full testing procedure is given in Algorithm A1 of Supplementy Material.

3 SPARSE HIGH-DIMENSIONAL COVARIATE DATA

When dealing with a large number of covariates, the favorable asymptotic properties associated with maximum likelihood estimation may not hold true. As an example, Sur and Candès (2019) demonstrated that the maximum likelihood estimator of the logistic regression model is inconsistent and exhibits significantly larger variation when the ratio of the number of covariates to the sample size does not approach zero. A common practice for handling high-dimensional covariate data is to screen out irrelevant covariates using penalization techniques. In this section, we focus on sparse single-index models and incorporate an adaptive LASSO penalty (Zou, 2006) into the profile likelihood for simultaneous estimation of the model and selection of irrelevant covariates.

Suppose that $f(y \mid \mathbf{x}; \boldsymbol{\theta})$ is of the form $f(y \mid \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. In this model, a coefficient $\beta_k = 0$ implies that the corresponding covariate X_k is irrelevant. We are interested in a sparse high-dimensional setting where p is allowed to increase with n and the number of non-zero coefficients is relatively small. Write $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}, \boldsymbol{\beta}_{\mathcal{A}^c}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{\mathcal{A}}$ is a $p_{\mathcal{A}} \times 1$ vector containing the non-zero coefficients and $\boldsymbol{\beta}_{\mathcal{A}^c} = \mathbf{0}$ represents a collection of zero parameters. For simultaneous variable selection (estimating $\mathcal{A}$) and parameter estimation (estimating $\boldsymbol{\beta}_{\mathcal{A}}$), we propose the following adaptive LASSO penalized log-profile likelihood function:

$$\ell_{\mathrm{pe}}(\boldsymbol{\beta}) = \ell_{\mathrm{pr}}(\boldsymbol{\beta}, \widehat{\mathbf{p}}(\boldsymbol{\beta}), \widehat{\mathbf{q}}) - \lambda \sum_{k=1}^{p} \widehat{w}_k |\beta_k|,$$

where $\lambda > 0$ is a tuning parameter that controls the strength of penalty, and $\widehat{w}_k$, $k = 1, \ldots, p$, are adaptive weights. The proposed penalized estimator $\widehat{\boldsymbol{\beta}}_\lambda$ is defined as the maximizer of $\ell_{\mathrm{pe}}(\boldsymbol{\beta})$, and the indices of significant covariates are estimated by $\widehat{\mathcal{A}}_\lambda = \{k : \widehat{\beta}_{\lambda,k} \neq 0\}$.

Following Zou and Zhang (2009) and Zou and Hastie (2005), we propose to employ the adaptive weights $\widehat{w}_k = |\widetilde{\beta}_{\mathrm{EN},k}|^{-1}$, where $\widetilde{\boldsymbol{\beta}}_{\mathrm{EN}} = (\widetilde{\beta}_{\mathrm{EN},1}, \ldots, \widetilde{\beta}_{\mathrm{EN},p})^{\mathrm{T}}$ is the maximum penalized profile likelihood estimator obtained with elastic net penalty. Given that $\widetilde{\boldsymbol{\beta}}_{\mathrm{EN}}$ is a consistent estimator of $\boldsymbol{\beta}$, as demonstrated by Zou and Hastie (2005), the absolute value of $\widetilde{\beta}_{\mathrm{EN},k}$ reflects the importance of the kth covariate. This way, the adaptive LASSO method effectively selects important covariates by assigning smaller penalties to them compared to unimportant ones. Let $\boldsymbol{\beta}_0$ be the true parameter and $\mathcal{A}_0 = \{k : \beta_{0,k} \neq 0\}$. The asymptotic properties of $\widehat{\boldsymbol{\beta}}_\lambda$ are given in the following theorem.

Theorem 5 *Suppose that Conditions 1–5 in Supplementary Materials, $\lambda(p_{\mathcal{A}_0}/n)^{1/2} \min_{k \in \mathcal{A}_0} |\beta_k| \to 0$, and $(n/p)^{1/2} \min_{k \in \mathcal{A}_0} |\beta_k| \to \infty$ are satisfied. Then, $\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\| = O_p\{(p/n)^{1/2}\}$. Moreover, suppose that $\lambda/p \to \infty$, then, as $n \to \infty$, we have*

(i) $\widehat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}_0^c} = 0$ with probability tending to one;

(ii) $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}_0} - \boldsymbol{\beta}_{0,\mathcal{A}_0})$ converges in distribution to a $p_{\mathcal{A}_0}$-variate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}^{-1}_{\mathcal{A}_0}$, where $\boldsymbol{\Sigma}^{-1}_{\mathcal{A}_0}$ is the semiparametric efficiency bound under the model $f(y \mid \mathbf{x}; \boldsymbol{\theta}) = f(y \mid \mathbf{x}^{\mathrm{T}}_{\mathcal{A}_0}\boldsymbol{\beta}_{\mathcal{A}_0})$.

Under this penalized profile likelihood framework, both the dimensionality of predictors and the number of nonzero coefficients are allowed to increase with the sample size n. Additionally, the minimal signal strength $\min_{k \in \mathcal{A}_0} |\beta_k|$ is allowed to decrease with n. The properties (i) and (ii) imply that the proposed estimator enjoys the oracle property: the coefficients of irrelevant covariates equal zero with a probability tending to one, and $\widehat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}_0}$ is asymptotically as efficient as the maximum likelihood estimator, as if we knew $\boldsymbol{\beta}_{\mathcal{A}_0} = 0$.

In what follows, we describe the computational details for implementing the proposed penalized estimation. First, we utilize a computationally efficient coordinate descent method to solve the maximization problem. This method iteratively maximizes $\ell_{\mathrm{pe}}(\boldsymbol{\beta})$ along each coordinate direction β_k, $k = 1, \ldots, p$, while keeping other directions fixed during each iteration. Consequently, the original high-dimensional optimization problem is solved by a series of simpler, lower-dimensional ones. Second, we use the local quadratic approximation (Fan and Li, 2001) to

TABLE 1 The biases (Bias, × 100), SDs (SD, × 100), mean squared prediction errors (MSPE, × 100), and percentages of correct prediction (PCP) of different estimators under Scenarios (I) and (II) and different sample sizes n_1 and n_2.

Scenario (I)			Profile		Pooled		Plug-in			MSPE		
n_1	n_2		Bias	SD	Bias	SD	Bias	SD		Profile	Pooled	Plug-in
50	50	β_1	0	6	15	6	0	7	Y_1^{new}	26	28	26
		β_2	0	6	6	4	0	7	Y_2^{new}	15	29	16
		β_3		4	1	4		5				
200	50	β_1	0	3	6	3	0	3	Y_1^{new}	25	26	25
		β_2	0	3	6	3	0	3	Y_2^{new}	14	29	15
		β_3	0	2	1	2	0	3				
50	200	β_1	0	6	36	6	0	7	Y_1^{new}	24	37	24
		β_2	0	6	−3	4	0	7	Y_2^{new}	14	21	17
		β_3	0	3		2		5				
200	200	β_1	0	3	15	3	0	3	Y_1^{new}	26	28	26
		β_2	1	3	6	2	0	3	Y_2^{new}	15	25	15
		β_3	0	2	1	2	0	3				
Scenario (II)			Profile		Pooled		Plug-in			PCP		
n_1	n_2		Bias	SD	Bias	SD	Bias	SD		Profile	Pooled	Plug-in
50	50	β_1	−4	35	97	24	−13	47	Y_1^{new}	73.3	69.0	71.6
		β_2	4	30	4	31	14	56	Y_2^{new}	80.2	76.5	80.8
		β_3	−5	30	−5	30	−14	53				
200	50	β_1		18	41	15	−2	20	Y_1^{new}	77.1	74.1	76.5
		β_2	2	18	2	18	2	21	Y_2^{new}	82.1	66.6	81.9
		β_3	−2	18	−2	18	−3	22				
50	200	β_1	−3	33	161	16	−13	47	Y_1^{new}	76.9	60.1	76.6
		β_2	1	20	2	20	14	56	Y_2^{new}	82.9	81.7	82.2
		β_3		19		19	−14	53				
200	200	β_1	0	17	98	12	−2	20	Y_1^{new}	76.5	69.5	76.5
		β_2	1	14	1	14	2	21	Y_2^{new}	83.0	76.8	83.3
		β_3	−2	14	−2	14	−3	22				

The estimators include the profile, the pooled, and the partial estimators.

approximate the adaptive LASSO penalty:

$$\widehat{w}_k|\beta_k| \approx \widehat{w}_k|s| + \frac{\widehat{w}_k}{2|s|}(\beta_k^2 - s^2) \tag{8}$$

when $\beta_k \approx s$. The value s can be set to be any consistent estimate, such as $\widehat{\beta}_{\text{EN},k}$. Since the right-hand side of (8) is a smooth convex function, existing optimization algorithms, such as the R function `nlminb()`, can be used to maximize the log-penalized likelihood function. Finally, the tuning parameter λ is chosen as the minimizer of the BIC-type criterion (Wang et al., 2009):

$$\begin{aligned}\text{BIC}(\lambda) = &-2\ell_{\text{pr}}(\widehat{\boldsymbol{\beta}}_\lambda, \widehat{\mathbf{p}}(\widehat{\boldsymbol{\beta}}_\lambda), \widehat{\mathbf{q}}) \\ &+ \max\{\log\{\log(p)\}, 1\}|\widehat{\mathcal{A}}_\lambda| \log n,\end{aligned}$$

where $|\widehat{\mathcal{A}}_\lambda|$ is the number of elements in $\widehat{\mathcal{A}}_\lambda$.

4 NUMERICAL STUDIES

4.1 Monte carlo simulations

The finite-sample performance of the proposed profile likelihood estimator is compared with its competitors in two scenarios: one involving continuous outcomes and the other with binary outcomes. In Scenario (I), the covariate X_1 is generated from a standard normal distribution, and the outcome Y_1 is generated from a normal distribution with mean $\theta_1 + \theta_2 X_1$ and variance θ_3^2, where $(\theta_1, \theta_2, \theta_3) = (-1, 1, 0.5)$. Hence, given $Y_1 = y$, X_1 follows a normal distribution with mean $\theta_2(y - \theta_1)/(\theta_2^2 + \theta_3^2)$ and variance $\theta_3^2/(\theta_2^2 + \theta_3^2)$. Next, the outcome Y_2 is generated from a standard normal distribution, and the covariate X_2 is generated from the same conditional distribution as X_1 given Y_1. In Scenario (II), the covariate vector $\mathbf{X}_1 = (X_{1,1}, X_{1,2})$ is generated from a standard bivariate normal distribution, and the outcome Y_1 is generated from a logistic regression model $\text{logit}\text{P}(Y_1 = 1) = \theta_1 + \theta_2 X_{1,1} + \theta_3 X_{1,2}$ with $(\theta_1, \theta_2, \theta_3) = (-1, 1, -1)$. It implies that $\text{P}(Y_1 = 1) = (2\pi)^{-1/2} \int \exp\{-(u^2 + v^2)/2\}/\{1 + \exp(-\theta_1 - \theta_2 u - \theta_3 v\}\text{d}u\text{d}v \approx 0.294$. Moreover, the conditional density of $\mathbf{X}_1$ given $Y_1 = y$ is

$$\frac{(2\pi)^{-1/2}}{\text{P}(Y_1 = 1)^y \text{P}(Y_1 = 0)^{1-y}} \times \frac{\exp(y\mathbf{x}^\text{T}\boldsymbol{\theta})}{1 + \exp(\mathbf{x}^\text{T}\boldsymbol{\theta})} \exp\left(-\frac{\mathbf{x}^\text{T}\mathbf{x}}{2}\right). \tag{9}$$

Next, the outcome Y_2 is then generated from a Bernoulli distribution with $\text{P}(Y_2 = 1) = 0.8$, and the covariate X_2 is generated from the conditional density given in (9). The following simulation results are based on 1000 replications with different combinations of sample sizes $(n_1, n_2) = (50, 50), (50, 200), (200, 50), (200, 200)$.

We begin by comparing the performance of our proposed estimator for $\boldsymbol{\theta}$ with two existing methods: the pooled estimation that ignores the dataset shift and the plug-in estimation of Saerens et al. (2002). The empirical biases and SDs of the estimates for $\boldsymbol{\theta}$ are summarized in the first six columns of Table 1. As expected, the pooled estimator is biased because it does not account for the dataset shift. In contrast, the profile likelihood and plug-in methods are consistent and have negligible biases,

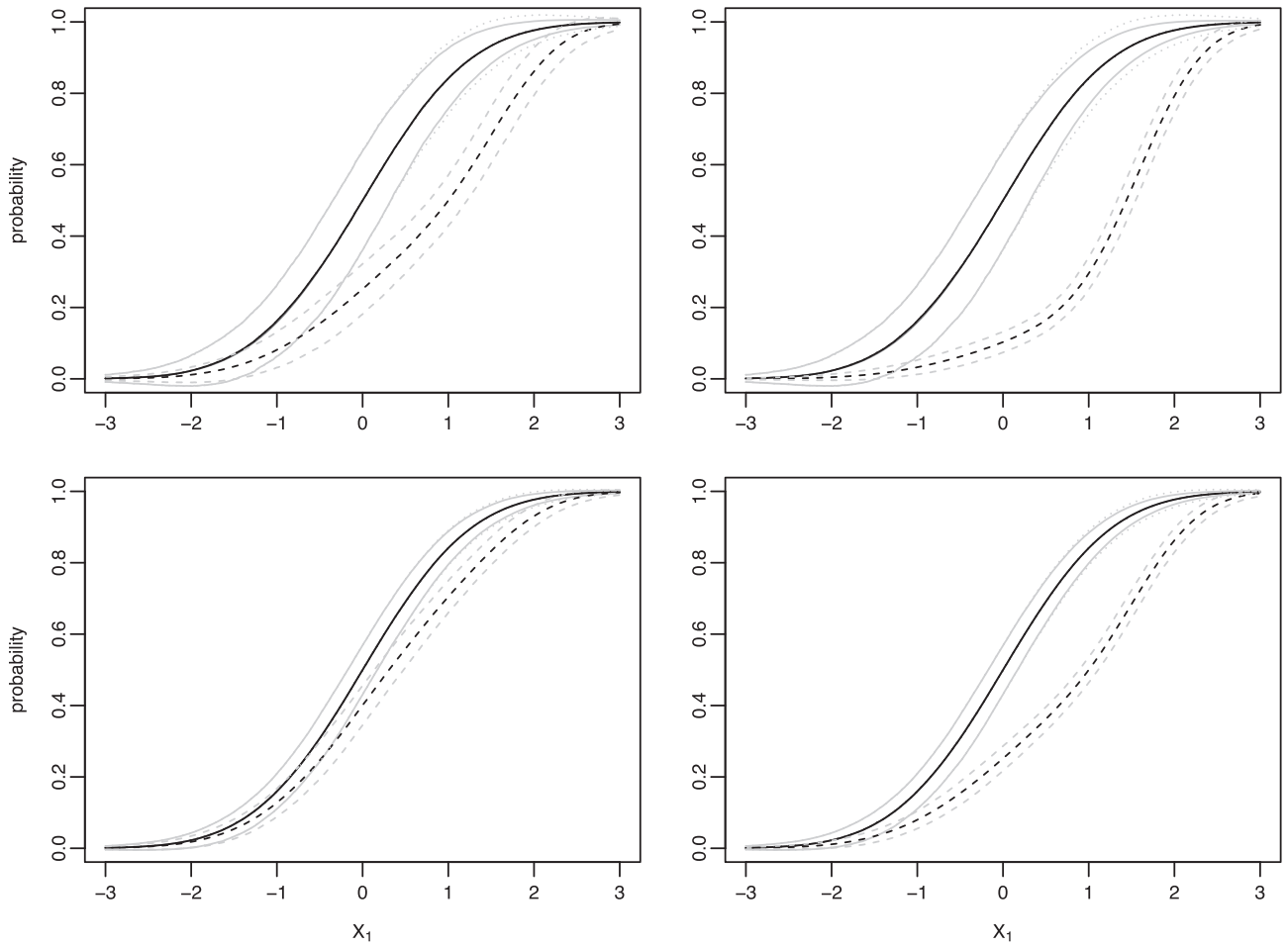

FIGURE 2 The empirical averages of different estimates for $G_1(\mathbf{x})$ (black lines) along with their 95% pointwise confidence intervals (gray lines) under Scenario (I) with different sample sizes. The estimates include the profile likelihood estimator $\widehat{G}_1$ (solid lines), the empirical distribution of $\{X_{1i}: i = 1, \ldots, n_1\} \cup \{X_{2i}: i = 1, \ldots, n_2\}$ (dashed lines), and the empirical distribution of $\{X_{1i}: i = 1, \ldots, n_1\}$ (dotted lines). Upper pannel: $n_1 = 50$; lower pannel: $n_1 = 200$; left pannel: $n_2 = 50$; right pannel: $n_2 = 200$.

with the profile likelihood method having smaller SDs. Next, we compare the performance of $\widehat{G}_1(\mathbf{x})$ with two other estimators under Scenario (I). The first estimator is the empirical distribution $\widetilde{G}_1$ based solely on $\{\mathbf{X}_{1i}: i = 1, \ldots, n_1\}$, while the second one is the empirical distribution $\check{G}_1$ derived from the pooled data $\{\mathbf{X}_{1i}: i = 1, \ldots, n_1\} \cup \{\mathbf{X}_{2i}: i = 1, \ldots, n_2\}$. Figure 2 depicts the Monte-Carlo averages over 1000 simulations for the three estimators along with their corresponding 95% pointwise confidence intervals. The biases of $\widehat{G}_1(\mathbf{x})$ and $\widetilde{G}_1$ are both close to zero. Moreover, $\widehat{G}_1(\mathbf{x})$ has narrower confidence intervals than $\widetilde{G}_1$, indicating an improvement in efficiency. In contrast, the pooled estimator $\check{G}_1$ that ignores dataset shift has noticeable biases.

Next, we compare the prediction performance of the 3 estimators using the mean squared prediction error $\text{MSPE} = \text{E}\{(\widehat{Y}_k - Y_k^{\text{new}})^2\}$ $(k = 1, 2)$ for continuous responses and the percentage of correct predictions $\text{PCP} = \text{P}(\widehat{Y}_k = Y_k^{\text{new}}) \times 100\%$ $(k = 1, 2)$ for binary responses. The Monte Carlo MSPEs and PCPs are summarized in the last three columns of Table 1. As expected, the predictions based on the pooled estimation demonstrate the highest MSPE and the lowest PCP, reinforcing that ignoring the dataset shift can lead to poor predictions. Moreover, the predictions based on the profile likelihood estimation exhibit lower MSPE and higher PCP compared with those based on the plug-in estimation. This finding supports the assertion that our efficient data integration algorithm can contribute to improved prediction performance.

We also assess the finite-sample performance of the proposed test for the prior probability shift assumption. We evaluate the size of the proposed test under Scenarios (I) and (II). The results, presented in the first 2 columns of Table 2, indicate that the test size is in close to the prespecified nominal level of 0.05. To assess the statistical power when the prior probability assumption is violated, we modify the setting in Scenario (I) by adding a constant δ to $\mathbf{X}_2$ so that the two conditional densities $f_1(\mathbf{x} \mid y)$ and $f_2(\mathbf{x} \mid y)$ differ. The power of our proposed test with different values of δ is summarized in the last five columns of Table 2. As expected, the power of the test increases as the sample size grows, with a more significant improvement observed when increasing n_1 compared to increasing n_2. This phenomenon can be attributed to the more precise estimation of $\boldsymbol{\theta}$ achieved by increasing n_1 as opposed to increasing n_2, as evident from the comparison of the SD of the profile likelihood estimator under $(n_1, n_2) = (200, 50)$ and $(n_1, n_2) = (50, 200)$ in Table 1.

TABLE 2 The sizes and the powers of testing the prior probability shift assumption under Scenarios (I) and (II), as well as modified Scenario 1 with different values of mean shift δ, which are evaluated using sample sizes of n_1 and n_2.

		size		power				
n_1	n_2	Scenario 1	Scenario 2	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0.6$	$\delta = 0.8$	$\delta = 1$
50	50	0.055	0.065	0.273	0.777	0.974	0.997	1.000
200	50	0.054	0.045	0.567	0.982	1.000	1.000	1.000
50	200	0.049	0.055	0.387	0.907	0.997	1.000	1.000
200	200	0.045	0.040	0.787	1.000	1.000	1.000	1.000

TABLE 3 The proportions of selecting $X_{1,1}$, $X_{1,2}$, $X_{1,3}$ and other covariates under the extended Scenario 1 with different sample sizes n_1 and n_2.

n_1	n_2	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	Others
50	50	1.000	1.000	0.955	0.000
200	50	1.000	1.000	0.998	0.000
50	200	1.000	1.000	0.995	0.000
200	200	1.000	1.000	1.000	0.000

Additionally, the power increases as the difference in $f(\mathbf{x} \mid y)$ between the two populations, quantified by δ, becomes larger. When the difference is small, a larger sample size is needed to detect the violation of the prior probability shift assumption, as seen in the case with $\delta = 0.2$ when $(n_1, n_2) = (200, 200)$.

Finally, to investigate the performance of the proposed penalized likelihood estimator, we extended Scenario (I) to include 100 covariates. Let $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,100})$, where the $X_{1,k}$'s are independent and follow standard normal distributions. The response Y_1 is generated from a normal distribution with mean $-1 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_1$ and variance of 0.5^2, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{100}) = (1, 0.5, 0.1, 0, \ldots, 0)^{\mathrm{T}}$. In this case, only $X_{1,1}$, $X_{1,2}$, and $X_{1,3}$ are relevant covariates. The data generating procedure for Y_2 and $\mathbf{X}_2$ follows Scenario (I). Table 3 summarizes the proportions of selecting each covariate based on our proposed estimator. The proposed estimator performs well, as it screens out all of the irrelevant covariates and selects the relevant covariates with high probability (ranging from 0.955 to 1.000). Moreover, false negatives occur more frequently when the signal is small ($\beta_3 = 0.1$) and the sample size is not large enough ($n_1 = n_2 = 50$).

4.2 Semisimulations using real data

In this subsection, we evaluate the performance of our proposed estimator using semisimulated data based on 2 well-known datasets: the airfoil dataset and the Pima Indian diabetes dataset. The former involves a continuous outcome and the latter deals with a binary outcome. The airfoil dataset consists of 1503 measurements of (log) airfoil sound pressure level, with 5 predictors including (log) frequency, angle of attack, chord length, free-stream velocity, and suction side (log) displacement thickness. On the other hand, the Pima Indian diabetes dataset includes 768 participants and the outcome of interest is the diagnosis of diabetes. The predictors include plasma glucose concentration in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age. Both datasets are available at the UC Irvine Machine Learning Repository.

Each simulation begins with a random split of data into two subsets: the training set $\mathcal{D}_{1,\text{train}}$ and its complement $\mathcal{D}^c_{1,\text{train}}$. Here, the training set contains r_1% of the total data. We then simulate a test set $\mathcal{D}_{1,\text{test}}$ by sampling 50% of the data from $\mathcal{D}^c_{1,\text{train}}$ without replacement. This implies that data points in $\mathcal{D}_{1,\text{test}}$ are exchangeable with those in $\mathcal{D}_{1,\text{train}}$, and therefore no prior probability shift occurs in this test set. Next, we simulate two sets $\mathcal{D}_{2,\text{train}}$ and $\mathcal{D}_{2,\text{test}}$ with prior probability shift. To achieve this, we sample an equal number of individuals from $\mathcal{D}_{1,\text{test}}$ with replacement, using a weight function proportional to the response values. Next, we sample r_2% of the individuals in $\mathcal{D}_2$ without replacement to form $\mathcal{D}_{2,\text{train}}$, and 50% of the individuals to form $\mathcal{D}_{2,\text{test}}$. To consider different sample sizes of test data, we set $r_1 = 25$, 50 and $r_2 = 25$, 50. Using the training datasets $\mathcal{D}_{1,\text{train}}$ and $\mathcal{D}_{2,\text{train}}$, we obtain the optimal predictor $\widehat{Y}_2$ given by (5) or (6). The prediction performance is then evaluated in terms of the mean squared prediction error $\text{MSPE} = \sum_{k=1}^{K}(\widehat{Y}_{2,k} - Y_{2,k})^2/375$ for the airfoil dataset, and in terms of the percentage of correct predictions $\text{PCP} = K^{-1}\sum_{k=1}^{K} I(\widehat{Y}_{2,k} = Y_{2,k})$ for the diabetes dataset, in the testing dataset $\mathcal{D}_{2,\text{test}} = \{(\mathbf{X}_{2,k}, Y_{2,k}) : k = 1, \ldots, K\}$.

Figure 3 shows the boxplots of the MSPEs over the 5000 simulations using the airfoil data. The pooled estimator exhibits significantly larger mean squared prediction errors compared to the other two estimators. This increased prediction error can be attributed to the biased estimation of the prediction model. Compared to the plug-in estimator, the proposed profile likelihood estimator demonstrates slightly smaller average and median MSPEs. As a result, our proposed method yields superior prediction performance in this semisimulation experiment. With the Pima Indian diabetes data, both the profile and the plug-in estimations yield a 100% PCP across 5000 simulations with varying combinations of $|\mathcal{D}_{1,\text{train}}|$ and $|\mathcal{D}_{2,\text{train}}|$. In contrast, the pooled estimation only achieves an average PCP ranging from 79.4% to 80.0%. These findings reinforce the argument that adjusting the prediction model to account for the prior probability shift is crucial for achieving superior classification results.

4.3 Analysis of Japan real estate price

In this section, we analyze the real estate transaction records in Osaka, compiled by the Ministry of Land, Infrastructure, Transport, and Tourism of Japan. Our objective is to assess how various key elements, such as time to the nearest station (X_1, minutes), land area (X_2, m^2), floor area (X_3, m^2), house age (X_4, years), breadth (X_5, m), coverage ratio (X_6), and floor area ratio (X_7), impact house prices (Y, million Yens) within the Osaka real estate market. This examination not only provides insights

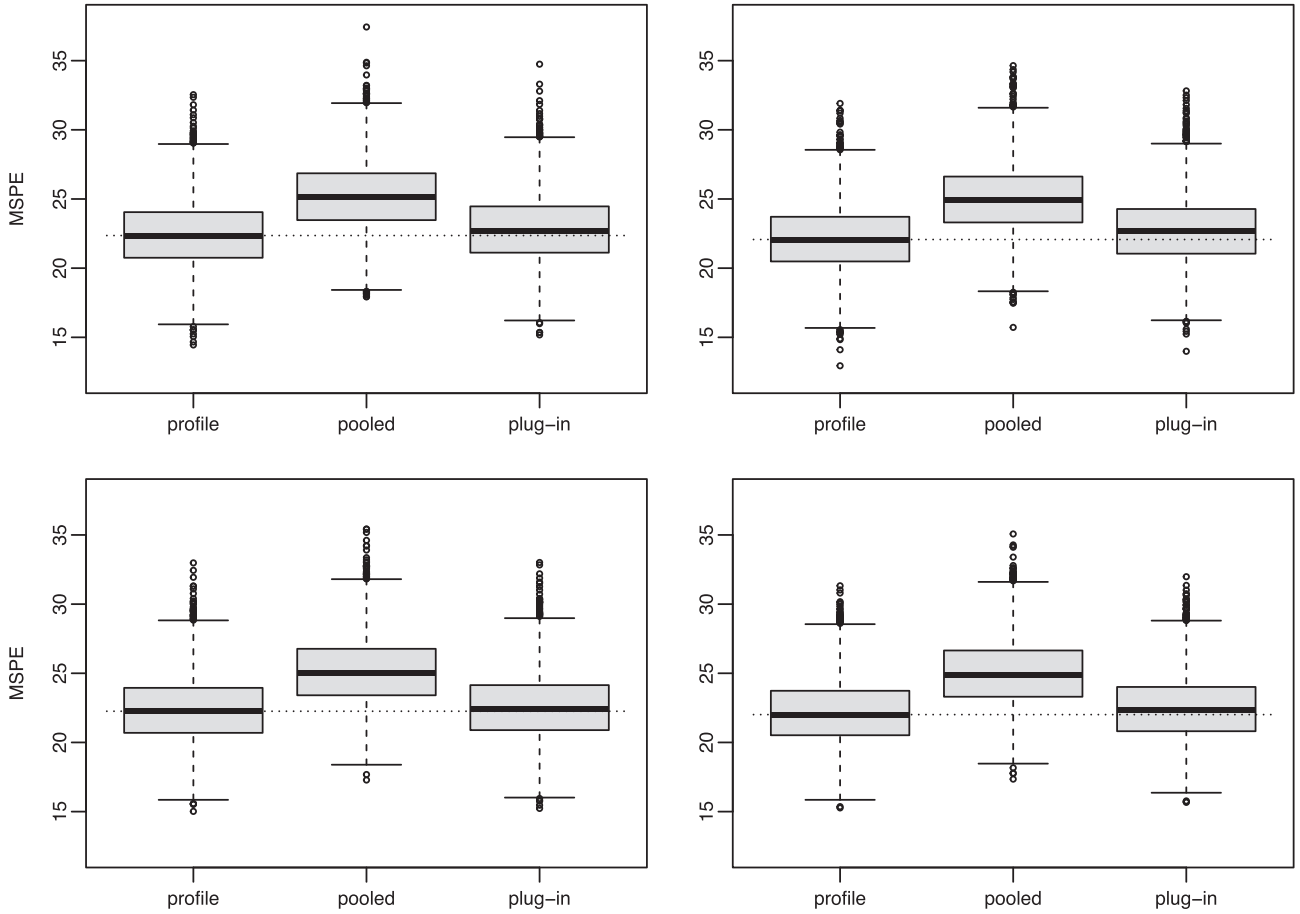

FIGURE 3 The boxplots of the mean squared prediction errors of the profile, pooled, and partial estimators obtained from 5000 times resampling in the airfoil data. Upper pannel: $|\mathcal{D}_{1,\text{train}}| = 375$; lower pannel: $|\mathcal{D}_{1,\text{train}}| = 750$; left pannel: $|\mathcal{D}_{2,\text{train}}| = 187$; right pannel: $|\mathcal{D}_{2,\text{train}}| = 375$.

for assessing the present value of properties but also serves as a foundation for predicting future property prices. It is important to note; however, that the transaction records are based on sold properties and do not constitute a random sample of the entire housing stock. In other words, the records are likely biased towards lower-priced sales, as higher-priced houses are less likely to be included due to fewer buyers in the market being able to afford them.

To verify the shift induced by the outcome-dependent sampling, we first derive the conditional distribution of the key factors given the house price. Let S denote an inclusion indicator, with inclusion probability $\mathrm{P}(S = 1 \mid Y = y, \mathbf{X} = \mathbf{x}) = \pi(y)$ for some function $\pi(y) \in [0, 1]$. Then, the conditional density of $\mathbf{X}$ given $Y = y$ and $S = 1$ is given by

$$\frac{\pi(y) f_{Y|\mathbf{X}}(y \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\pi(y) \int f_{Y|\mathbf{X}}(y \mid \mathbf{u}) f_{\mathbf{X}}(\mathbf{u}) \mathrm{d}\mathbf{u}} = \frac{f_{Y|\mathbf{X}}(y \mid \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\int f_{Y|\mathbf{X}}(y \mid \mathbf{u}) f_{\mathbf{X}}(\mathbf{u}) \mathrm{d}\mathbf{u}}. \quad (10)$$

This implies that the conditional density of $\mathbf{X}$ given $Y = y$ and $S = 1$ is identical to the underlying conditional density of $\mathbf{X}$ given $Y = y$. Moreover, the conditional density of Y given $S = 1$ is given by

$$\frac{\pi(y) \int f_{Y|\mathbf{X}}(y \mid \mathbf{u}) f_{\mathbf{X}}(\mathbf{u}) \mathrm{d}\mathbf{u}}{\int \pi(y) \int f_{Y|\mathbf{X}}(y \mid \mathbf{u}) f_{\mathbf{X}}(\mathbf{u}) \mathrm{d}\mathbf{u} \mathrm{d}y}, \quad (11)$$

which differs from the corresponding underlying quantity $f_Y(y) = \int f_{Y|\mathbf{X}}(y \mid \mathbf{u}) f_{\mathbf{X}}(\mathbf{u}) \mathrm{d}\mathbf{u}$ and leads to dataset shift. Equations (10)-(11) further demonstrate that when the datasets for 2018 and 2019 were collected with potentially different $\pi(y)$, the prior probability shift exists among these datasets.

There has been a tendency for the marginal distribution of house prices to shift towards lower prices in recent years. Upon initial examination of the 2506 records from 2018 and 1565 records from 2019, we observe a 4.84% decrease in the average house price, declining from 26.8 million Yens in 2018 to 25.5 million Yens in 2019. Additionally, the Kolmogorov-Smirnov test yields a P-value of 0.023, indicating a statistically significant shift in the distributions of transaction prices between the 2 years. These observations motivate us to apply the proposed methodology to deal with the prior probability shift in the transaction records in 2018 and 2019. We first fit a linear model to the centralized and standardized transaction prices in 2018

$$\widetilde{Y} = \theta_0 + \theta_1 X_1 + \cdots + \theta_7 X_7 + \varepsilon,$$

where $\widetilde{Y} = (Y - 26.8)/15.9$. Here, ε follows a normal distribution with mean zero and variance θ_8^2 and is assumed to be independent of $\mathbf{X}$. It implies that the common distribution of $\mathbf{X}$ given

TABLE 4 The estimated coefficients (Coef.), SEs, and *P*-values for testing zero-coefficients in the Japan real estate price data analysis.

	2018 & 2019			2018		
	Coef.	**SE**	***P*-value**	**Coef.**	**SE**	***P*-value**
Intercept	−0.7632	0.0120	<0.001	−1.0406	0.0135	<0.001
Time	−0.0178	0.0006	<0.001	−0.0164	0.0007	<0.001
Land area	0.3929	0.0083	<0.001	0.4211	0.0093	<0.001
Floor area	0.6792	0.0115	<0.001	0.7418	0.0129	<0.001
Age	−0.0284	0.0004	<0.001	−0.0264	0.0005	<0.001
Breadth	0.0269	0.0020	<0.001	0.0247	0.0023	<0.001
Coverage ratio	0.7492	0.0198	<0.001	0.7914	0.0223	<0.001
Floor area ratio	−0.1000	0.0062	<0.001	−0.0462	0.0070	<0.001

$\widetilde{Y} = y$ is given by

$$\frac{\phi\{(y-\theta_0-\theta_1x_1-\cdots-\theta_7x_7)/\theta_8\}\mathrm{d}G_1(\mathbf{x})}{\int \phi\{(y-\theta_0-\theta_1x_1-\cdots-\theta_7x_7)/\theta_8\}\mathrm{d}G_1(\mathbf{x})},$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. To test whether $f_{X|Y}(x|y)$ stays the same in 2019, we consider the smooth alternative

$$\frac{\phi\{(y-\theta_0-\theta_1x_1-\cdots-\theta_7x_7)/\theta_8\}\exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})\mathrm{d}G_1(\mathbf{x})}{\int \phi\{(y-\theta_0-\theta_1x_1-\cdots-\theta_7x_7)/\theta_8\}\exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})\mathrm{d}G_1(\mathbf{x})},$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_7)^{\mathrm{T}}$. Our proposed semiparametric likelihood ratio test for testing $\boldsymbol{\gamma} = \mathbf{0}$ resulted in a *P*-value of 0.325, suggesting that there is no strong evidence against the constancy of $f_{X|Y}(x|y)$ during the period 2018–2019.

The fitted linear models, obtained by using the proposed profile likelihood estimation method under the prior probability shift assumption and the maximum likelihood estimation based solely on the 2018 dataset, are summarized in Table 4. As anticipated, the proposed profile estimator, leveraging information from both the 2018 and 2019 datasets, exhibits a significant efficiency gain in estimating the effects of the covariates. The fitted models suggest that distance to the nearest station, age of the house, and floor-area ratio negatively impact house price. Notably, a one-minute increase in walking distance reduces the average price by 282.3 thousand Yen (95% confidence interval [CI]: 262.2-302.4), while each year of age decreases the price by 450.8 thousand Yen (95% CI: 437.7-463.9). In contrast, land area, floor area, breadth, and coverage ratio positively affect price. Interestingly, each additional square meter of floor area increases price by an average of 10.8 million Yen (95% CI: 10.4-11.1), whereas the same increase in land area leads to a lower price increase of 6.2 million Yen (95% CI: 6.0-6.5).

5 DISCUSSION

In this paper, we make 3 key contributions in addressing estimation and prediction problems under prior probability shift. First, we tackle the challenge of prior probability shift by developing an efficient estimation algorithm that can efficiently combine information from multiple data sources. The proposed method yields better prediction performance over existing ones. Second, our work fills a research gap in the literature by introducing a formal test for checking the assumption of prior probability shift, an aspect that has not been rigorously studied before. By embedding the null density within a larger parametric family of densities and testing deviations from the null, we propose a novel semiparametric likelihood ratio test for checking the assumption of prior probability shift. Finally, we extend our proposed approach to handle high-dimensional covariates by incorporating the adaptive LASSO penalty technique, thereby providing a flexible and effective solution for analyzing complex modern data scenarios.

Other than prior probability shift, there exist other types of dataset shifts. Interestingly, these shifts correspond to different sampling designs. For instance, we already demonstrated that outcome-dependent sampling aligns with an invariant conditional distribution $f_{\mathbf{X}|Y}(\mathbf{x}\,|\,y)$ across datasets, that is, prior probability shift. On the other hand, covariate-dependent sampling results in an invariant conditional distribution $f_{Y|\mathbf{X}}(y\,|\,\mathbf{x})$, that is, covariate shift. Recognizing the shared probability structure across datasets is a crucial step to ensure accurate and efficient data integration. This task, however, often demands a solid understanding of the data collection process. When such information is lacking, one can embed dataset shift in a general biased selection model that allows the shift (or sampling weight) to depend on the outcome and/or covariate in a prespecified functional form. Model selection can then be performed by testing parameter values in the selection bias model. Further investigation into the selection of different dataset shift assumptions and their implications is beyond the scope of this paper and will be explored in future research.

While the primary focus of this paper is on parametric models, our profile likelihood approach can be extended to accommodate more general semiparametric models, such as the single-index model of Delecroix et al. (2003) and the effective dimension reduction model of Li (1991). These models introduce additional nuisance parameters in addition to G_1 and F_2. Whether profile likelihood estimation can effectively handle these additional nuisances is a topic warranting further investigation. Another possible extension is the semisupervised learning problem. In some applications, responses for some individuals may be unavailable, resulting in a mix of labeled and unlabeled data. Semisupervised learning aims to develop prediction models using both labeled and unlabeled data. By treating the responses in the unlabeled data as missing variables, the proposed approach can handle the likelihood function of the observed data and tackle the semisupervised learning problem. These extensions will be explored in our future research.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices referenced in Sections 2-3, and codes (R package `EPPS`) are available with this paper at the *Biometrics* website on Oxford Academic.

FUNDING

None declared.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The airfoil dataset and the Pima Indian diabetes dataset are available in the University of California, Irvine, Machine Learning Repository at https://archive.ics.uci.edu. The Japan real estate price data can be downloaded from the online data science platform Kaggle at (https://www.kaggle.com/datasets/nishiodens/japan-real-estate-transaction-prices).

REFERENCES

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86, 213–226.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20, 101–148.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S. and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, 385, 283–286.

Garg, S., Wu, Y., Balakrishnan, S. and Lipton, Z. C. (2020). A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33, 3290–3300.

Li, K. -C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86, 316–342.

Lipton, Z., Wang, Y. -X. and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3122–3130. PMLRCambridge.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92, 968–976.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95, 449–485.

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20, 149–199.

Rayner, J. C. W. and Best, D. J. (1990). Smooth tests of goodness of fit: an overview. *International Statistical Review*, 58, 9–17.

Saerens, M., Latinne, P. and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14, 21–41.

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. *Dataset Shift in Machine Learning*, 30, 3–28.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 14516–14525.

Thams, N., Saengkyongam, S., Pfister, N. and Peters, J. (2023). Statistical testing under distributional shifts. *Journal of the Royal Statistial Society, Series B*, 85, 597–663.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B*, 71, 671–683.

Wong, A., Cao, J., Lyons, P. G., Dutta, S., Major, V. J., Ötleş, E. and Singh, K. (2021). Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. *JAMA Network Open*, 4, e2135286.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37, 1733–1751.

Received: June 27, 2023; Revised: January 25, 2024; Accepted: April 25, 2024