Journal of Machine Learning Research 24 (2023) 1-53 Submitted 1/23; Revised 8/23; Published 9/23

Optimal Parameter-Transfer Learning by Semiparametric Model Averaging

Xiaonan Hu XNHU@AMSS.AC.CN
School of Mathematical Sciences
Capital Normal University
Beijing, 100048, China
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190, China
Xinyu Zhang ∗ XINYU@AMSS.AC.CN
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190, China
International Institute of Finance, School of Management
University of Science and Technology of China
Hefei, 230026, Anhui, China
∗ *Corresponding author*

Editor: Amos Storkey

Abstract

In this article, we focus on prediction of a target model by transferring the information of source models. To be flexible, we use semiparametric additive frameworks for the target and source models. Inheriting the spirit of parameter-transfer learning, we assume that different models possibly share common knowledge across parametric components that is helpful for the target predictive task. Unlike existing parameter-transfer approaches, which need to construct auxiliary source models by parameter similarity with the target model and then adopt a regularization procedure, we propose a frequentist model averaging strategy with a J-fold cross-validation criterion so that auxiliary parameter information from different models can be adaptively transferred through data-driven weight assignments. The asymptotic optimality and weight convergence of our proposed method are built under some regularity conditions. Extensive numerical results demonstrate the superiority of the proposed method over competitive methods.

Keywords: asymptotic optimality, cross-validation, negative transfer, prediction, weighting

1. Introduction

Numerous machine learning techniques have been successfully applied in many fields under data-driven paradigms. As one of the most classical machine learning techniques, supervised learning frameworks commonly use given training data with labels to fit a model and then make predictions or inferences on unlabeled testing data based on the resulting model. The

performance is generally satisfactory when we have sufficient training data. However, in many real-world applications, it may be expensive or even unrealistic to collect such amount of data. For example, in some medical and biological studies, it is challenging to obtain a great deal of patient information from a medical institute due to ethical or cost issues, while there are many institutes owning related data sources that form the concept of "*data islands*". Considering the risk of policy, organizational interests, and individual privacy, it is difficult to cooperate by directly integrating all the data sets from multiple owners for unified analysis. Modeling the data sets separately may suffer from deteriorative generalization in cross-domain scenarios, and it also wastes resources and brings additional costs. Therefore, taking advantage of multi-source data reasonably to deal with these challenges motivates transfer learning, the aim of which is to improve the performance of a specific target task by transferring common knowledge shared across similar source domains (Pan and Yang, 2009; Blanchard et al., 2021). As a prevailing topic in computer science, transfer learning has been displaying great potential in modern applications, such as medical and biological studies (Shin et al., 2016), computer vision (Long et al., 2015), natural language processing (Raffel et al., 2020), and recommendation systems (Pan et al., 2010). Among the existing research, there are few studies on theoretical support for transfer learning frameworks prompting us to explore in this work.

Prediction is an important task in economic and statistical analysis, and statistical regression models are commonly adopted because of their convenience and interpretability. Semiparametric models, as a traditional class of statistical regression models, provide a flexible way to understand the complicated relationship between the response and the set of covariates by simultaneously considering parametric and nonparametric components. Although transfer learning has been widely studied for decades, little attention has been paid to it in statistical research. Specifically, the impact of transfer learning under semiparametric frameworks remains unclear, which motivates our study to try to fill this gap. In this article, we consider a partially linear model (PLM) with additive structures, which is a common type of semiparametric model in the literature, such as Stone (1985, 1986), Ma et al. (2006), Wang et al. (2011), and Ma and Zhu (2013). To accommodate the framework of transfer learning, inheriting the spirits of parameter-transfer learning (Pan and Yang, 2009), we further assume that different models on multiple data sets possibly share common knowledge across parametric components, meanwhile allowing heterogeneity in nonparametric components.

In recent years, a few studies on transfer learning under statistical models have sprung up. Bastani (2021) studies the single-source parameter-transfer approach under high-dimensional linear models and derives the estimation error bound, where the sample size of the source domain is larger than the dimension. Li et al. (2021) extend Bastani's work to the multi-source transfer learning framework with high-dimensional target and source models under some weaker assumptions, and the minimax optimality of the estimation error bound under ℓ_q-regularization ($q \in [0, 1]$) is proven. Tian and Feng (2022) further study the multi-source transfer learning framework in generalized linear model settings and develop a consistent procedure to detect transferable sources. In addition, Li et al. (2022b) extend the multi-source transfer learning framework to Gaussian graphical models and construct a multiple testing procedure for edge detection with false discovery rate control. Different from the current work, existing studies usually construct auxiliary source models based

on parameter similarity between the target and source models and adopt a regularization technique that requires properly selecting tuning parameters. These procedures also require equal dimensions for all the models and integrate data sets in the estimation process, which may bring limitations in practical application. Specifically, when heterogeneous data are collected from different owners, it is often difficult to aggregate the complete data sets from all parties without boundaries, and generally, only summary statistics can be obtained to protect data privacy. In addition, these works mainly focus on parameter estimation while rare attention is paid to out-of-sample prediction. There are some other topics closely related to our work, such as multi-task learning (Evgeniou and Pontil, 2004; Ando et al., 2005) and integrative analysis (Ma et al., 2011; Liu et al., 2014). However, their goals are mainly to simultaneously estimate parameters for all the models considering the similarity and heterogeneity among model parameters. In addition, they generally assume that all the models are correctly specified, whereas our framework does not require this assumption.

To make better predictions of the target model, model averaging is an effective strategy combining information from multiple candidate models. There exists rich literature on model averaging for several decades (Clarke, 2003), and we consider the asymptotically optimal methods in the frequentist model averaging framework in this work. For model averaging studies on PLM, Zhang and Wang (2019) study the optimal model averaging method for PLM with heteroscedasticity and propose a Mallows-type weight choice criterion in a kernel smoothing framework. Furthermore, frequentist model averaging estimators for other variants of partially linear models have also been studied, such as varying-coefficient partially linear models (Li et al., 2018; Zhu et al., 2019; Li et al., 2022a), partially linear functional additive models (Liu and Zhang, 2021) and semi-functional partially linear models (Jiang et al., 2021). In this article, we contribute to taking the model averaging methodology as a bridge for knowledge transfer between the possibly shared parameter information and the predictive task for the target model. Different from traditional model averaging procedures, we combine parameter information from multiple semiparametric models using samples from heterogeneous populations. To estimate multiple semiparametric models, a polynomial spline-based estimator is adopted to approximate nonparametric functions, which can be implemented more cheaply than kernel-based smoothing approaches.

Our approach has some appealing advantages. First, when the target model is misspecified, our procedure can asymptotically obtain optimal prediction in the sense of achieving the lowest possible out-of-sample prediction risk. Second, when the target model is correctly specified, those models possibly sharing auxiliary information are automatically distinguished by weight assignments. In the studies of transfer learning, sometimes knowledge transfer may even hurt the learning performance of the target task when source models are not related to the target model, the phenomenon of which is referred to as "*negative transfer*" in the literature (Pan and Yang, 2009). Some recent works also concern this problem, such as Li et al. (2021) and Tian and Feng (2022), where they introduce algorithms to construct auxiliary source models based on parameter similarity. Further, theoretical properties support that transferring knowledge among such auxiliary source models can improve the performance of the target model under certain conditions. Our proposed method does not require knowing auxiliary source models in advance and theoretically ensures that the parameter transfer asymptotically occurs in potential auxiliary models, which attempts to address the negative transfer problem from a new perspective. It can be seen that our

method has theoretical guarantees in both correct and incorrect target model settings, which is a desirable feature in applications. Finally, our procedure offers an alternative strategy for massive data analysis. Specifically, we can split the full data set into many batches and estimate each batch of data in parallel. We can then aggregate estimators through our transfer learning mechanism to achieve predictions. This approach is similar to the divide and conquer technique in the literature of distributed learning (Zhang et al., 2013; Battey et al., 2018). By transmitting only summary statistics information instead of pooling multiple data sets together, our framework provides a feasible strategy to effectively protect the privacy of individual data. Relevant studies can be found in the literature of meta-analysis (Xie et al., 2011; Kundu et al., 2019).

In conclusion, the primary contributions of this work can be summarized as follows.

- In contrast to traditional frequentist model averaging approaches for semiparametric models, we adopt a spline-based estimator and propose a data-driven weight choice criterion in the scenario of multiple populations, which provides more insights for model averaging research.

- For transfer learning frameworks, we develop a parameter-transfer approach aimed at the predictive task under statistical models. We contribute to taking a model averaging strategy to adaptively transfer possibly shared parameter information instead of auxiliary model selection. Some appealing properties for parameter-transfer learning are established from a statistical view.

The rest of the paper is organized as follows. In Section 2, we introduce our model framework and weight choice criterion. Section 3 provides the theoretical properties of our approach, including asymptotic optimality and weight convergence under certain regularity conditions. Extensive simulation studies and a real data example are conducted in Section 4 and Section 5, respectively. Concluding remarks are summarized in Section 6. All the technical details and additional numerical results are presented in the Appendix.

2. Optimal Parameter-Transfer Approach

In this section, we first introduce our semiparametric model setting under transfer learning framework and then propose a parameter-transfer approach based on frequentist model averaging. We further provide a cross-validation based procedure to choose proper weights.

2.1 Model Framework

Assume that target data $\{y_i^{(0)}, \boldsymbol{x}_i^{(0)}, \boldsymbol{z}_i^{(0)}\}$ for $i = 1, \ldots, n_0$ and source data $\{y_i^{(m)}, \boldsymbol{x}_i^{(m)}, \boldsymbol{z}_i^{(m)}\}$ for $m = 1, \ldots, M$, $i = 1, \ldots, n_m$ are independent samples from $M + 1$ heterogeneous populations. For the mth data set, $m = 0, \ldots, M$, $y_i^{(m)}$ are continuous scalar responses, $\boldsymbol{x}_i^{(m)} = (x_{i1}^{(m)}, \ldots, x_{ip}^{(m)})^T$ are p-dimensional i.i.d observations, and $\boldsymbol{z}_i^{(m)} = (z_{i1}^{(m)}, \ldots, z_{iq_m}^{(m)})^T$ are q_m-dimensional i.i.d observations. Here, different $\boldsymbol{z}_i^{(m)}$ are allowed in different data sets. Suppose that the target and source samples follow $M + 1$ semiparametric additive linear models, which are referred to as the target model and source models as follows. For

$m = 0, \ldots, M$, $i = 1, \ldots, n_m$,

$$y_i^{(m)} = \mu_i^{(m)} + \varepsilon_i^{(m)} = (\boldsymbol{x}_i^{(m)})^T \boldsymbol{\beta}^{(m)} + g^{(m)}(\boldsymbol{z}_i^{(m)}) + \varepsilon_i^{(m)}, \tag{1}$$

where $\mu_i^{(m)}$ contains both a parametric component of $(\boldsymbol{x}_i^{(m)})^T \boldsymbol{\beta}^{(m)}$ with an unknown parameter $\boldsymbol{\beta}^{(m)} \in \mathbb{R}^p$ and nonparametric components of $g^{(m)}(\boldsymbol{z}_i^{(m)}) = \sum_{l=1}^{q_m} g_l^{(m)}(z_{il}^{(m)})$ with a commonly adopted additive structure, $g_l^{(m)}$ is a one-dimensional unknown smooth function, and $\varepsilon_i^{(m)}$ are independent random errors with $E(\varepsilon_i^{(m)} | \boldsymbol{x}_i^{(m)}, \boldsymbol{z}_i^{(m)}) = 0$ and $E\{(\varepsilon_i^{(m)})^2 | \boldsymbol{x}_i^{(m)}, \boldsymbol{z}_i^{(m)}\} = \sigma_{i,m}^2$. Note that $\boldsymbol{\beta}^{(m)}$ in different source models are allowed to be identical or different from the target model. Here, the dimension p of the parametric component for each model is allowed to go to infinity, and q_m is fixed for $m = 0, \ldots, M$.

In the context of transfer learning, we expect to improve the prediction of the new response $y_{n_0+1}^{(0)}$ given the corresponding set of covariates $\{\boldsymbol{x}_{n_0+1}^{(0)}, \boldsymbol{z}_{n_0+1}^{(0)}\}$ from the target population through transferring common knowledge from source models. To accommodate the transfer learning framework under semiparametric models, we specifically borrow the idea of parameter-transfer approaches and further assume that source models possibly share parameter information with the target model. Unlike some recent works on transfer learning under linear models (Li et al., 2021) and generalized linear models (Tian and Feng, 2022), it is not necessary for our proposed method to construct auxiliary source models based on parameter similarity and integrate data sets by a regularization procedure.

To estimate our models (1), we consider a polynomial spline-based method to approximate nonparametric parts. The corresponding theoretical properties have been well established in the literature (De Boor, 2001). Let $\Psi_l^{(m)}$ be the polynomial spline space consisting of functions of degree $r_l^{(m)} \geq 1$ and $S_l^{(m)}$ be the number of interior knots in the interval $[0, 1]$ for the mth model. Here, the number of interior knots can vary from different models and is allowed to be divergent as the sample size increases. Assume that there exists a normalized B-spline basis $B_l^{(m)}(z) = \{b_{l1}(z), \ldots, b_{lv_l^{(m)}}(z)\}^T$ of the spline space, where $v_l^{(m)} = r_l^{(m)} + S_l^{(m)}$ and $v_l^{(m)}$ is allowed to increase as the sample size increases. As discussed in De Boor (2001), nonparametric functions can be well approximated by the linear combination of B-spline basis functions under certain conditions. Therefore, the estimator can be transformed into a least squares formula as follows.

$$\widehat{\boldsymbol{\theta}}^{(m)} = \arg\min_{\boldsymbol{\theta}^{(m)}} \sum_{i=1}^{n_m} \left\{ y_i^{(m)} - (\boldsymbol{d}_i^{(m)})^T \boldsymbol{\theta}^{(m)} \right\}^2, \quad m = 0, \ldots, M, \tag{2}$$

where $\boldsymbol{d}_i^{(m)} = [(\boldsymbol{x}_i^{(m)})^T, \{B_1^{(m)}(z_{i1}^{(m)})\}^T, \ldots, \{B_{q_m}^{(m)}(z_{iq_m}^{(m)})\}^T]^T$, $\boldsymbol{\theta}^{(m)} = \{(\boldsymbol{\beta}^{(m)})^T, (\boldsymbol{\gamma}_1^{(m)})^T, \ldots, (\boldsymbol{\gamma}_{q_m}^{(m)})^T\}^T$, and $\boldsymbol{\gamma}_l^{(m)} = (\gamma_{l1}^{(m)}, \ldots, \gamma_{lv_l^{(m)}}^{(m)})^T$ for $m = 0, \ldots, M$ and $l = 1, \ldots, q_m$. Let the total dimension of the mth model be $p_m = \sum_{l=1}^{q_m} v_l^{(m)} + p$.

2.2 Model Averaging Prediction Procedure

In this section, we introduce a frequentist model averaging strategy to transfer possibly shared parameter information from multiple models for our prediction of the target model,

which has the advantages of achieving asymptotically optimal prediction and adaptively using potential auxiliary models by reasonable weight assignments. Since we have little prior knowledge about auxiliary parameter information from source models in practice, we simply put all the M source models into our transfer learning framework. To develop our transfer learning framework, we first construct $M+1$ models based on B-spline basis approximations. Specifically, the estimators of $\mu_i^{(0)}$ corresponding to the $M+1$ models, including the target model ($m=0$) and M source models ($m=1,\ldots,M$) with possibly shared parameters, are defined as

$$\widehat{\mu}_{i,m}^{(0)} = (\boldsymbol{d}_i^{(0)})^T \widehat{\boldsymbol{\theta}}_m^{(0)} = \begin{cases} (\boldsymbol{x}_i^{(0)})^T \widehat{\boldsymbol{\beta}}^{(0)} + \sum_{l=1}^{q_0} \{B_l^{(0)}(z_{il}^{(0)})\}^T \widehat{\boldsymbol{\gamma}}_l^{(0)}, & m=0, \\ (\boldsymbol{x}_i^{(0)})^T \widehat{\boldsymbol{\beta}}^{(m)} + \sum_{l=1}^{q_0} \{B_l^{(0)}(z_{il}^{(0)})\}^T \widehat{\boldsymbol{\gamma}}_l^{(0)}, & m=1,\ldots,M, \end{cases} \tag{3}$$

where $\widehat{\boldsymbol{\theta}}_m^{(0)} = \{(\widehat{\boldsymbol{\beta}}^{(m)})^T, (\widehat{\boldsymbol{\gamma}}_1^{(0)})^T, \ldots, (\widehat{\boldsymbol{\gamma}}_{q_0}^{(0)})^T\}^T$. Slightly different from the construction of candidate models in previous model averaging literature, there exists uncertainty about the parameter information of which model can be transferred to the target model. In other words, the informative level of different models is not clear, and only information of $\widehat{\boldsymbol{\beta}}^{(m)}$ is allowed to be transferred between models.

Algorithm 1: Trans-SMAP

Input: Training samples, including the target and source data, $\{(\boldsymbol{x}_i^{(m)}, \boldsymbol{z}_i^{(m)}, y_i^{(m)}); i=1,\ldots,n_m, m=0,\ldots,M\}$ from the target and source models (1) and the new sample $\{\boldsymbol{x}_{n_0+1}^{(0)}, \boldsymbol{z}_{n_0+1}^{(0)}\}$ from the target model.

Output: Prediction of $y_{n_0+1}^{(0)}$ associated with the new sample $\{\boldsymbol{x}_{n_0+1}^{(0)}, \boldsymbol{z}_{n_0+1}^{(0)}\}$.

Step 1. Estimate parameter $\boldsymbol{\theta}^{(m)}$ for the target model using training samples by (2), and denote the estimator by $\widehat{\boldsymbol{\theta}}^{(m)}$.

Step 2. Split training samples from the target model into J subgroups with $2 \le J \le n_0$.

Step 3. **foreach** $j \in \{1,\ldots,J\}$ **do**

- Step 3.1. For $m=1,\ldots,M$, estimate $\boldsymbol{\theta}^{(m)}$ separately by (2) with all the training samples, and estimate $\boldsymbol{\theta}^{(0)}$ only with the training samples in subgroup $\mathcal{G}_j^c$.
- Step 3.2. For $i \in \mathcal{G}_j, m=0,\ldots,M$, perform the prediction of $y_i^{(0)}$ based on (3) as $\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}$.
- Step 3.3. Construct the weighted combination as $\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w}) = \sum_{m=0}^{M} w_m \widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}$.

Step 4. Select the weight vector $\widehat{\boldsymbol{w}}$ by minimizing the J-fold cross-validation criterion (5).

Step 5. Given the new sample $\{\boldsymbol{x}_{n_0+1}^{(0)}, \boldsymbol{z}_{n_0+1}^{(0)}\}$ from the target model, obtain the model averaging prediction by plugging in $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{\theta}}^{(m)}$, that is $\widehat{\mu}_{n_0+1}^{(0)}(\widehat{\boldsymbol{w}}) = \sum_{m=0}^{M} \widehat{w}_m \widehat{\mu}_{n_0+1,m}^{(0)}$.

Following the previous studies on model averaging, the final prediction can be defined as a weighted average of $\widehat{\mu}_{i,m}^{(0)}$ expressed as $\widehat{\mu}_i^{(0)}(\boldsymbol{w}) = \sum_{m=0}^{M} w_m \widehat{\mu}_{i,m}^{(0)}$, where $\boldsymbol{w} = (w_0,\ldots,w_M)^T$ is the weight vector in the space $\mathcal{W} = \{\boldsymbol{w} \in [0,1]^{M+1} : \sum_{m=0}^{M} w_m = 1\}$. To determine a proper choice of weights, we adopt a J-fold ($J>1$) cross-validation criterion. Specifically,

we randomly divide the target samples into J mutually exclusive subgroups $\mathcal{G}_1, \ldots, \mathcal{G}_J$. For simplicity, we assume that all subgroups have equal size n_0/J, which is a positive integer. For $j = 1, \ldots, J$, let $\mathcal{G}_j^c = \{1, \ldots, n_0\} \backslash \mathcal{G}_j$ denote the set $\{1, \ldots, n_0\}$ excluding the elements in $\mathcal{G}_j$. Then the J-fold cross-validation based weight choice criterion is defined as

$$CV(\boldsymbol{w}) = \frac{1}{n_0} \sum_{j=1}^{J} \sum_{i \in \mathcal{G}_j} \left\{ y_i^{(0)} - \widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w}) \right\}^2, \tag{4}$$

where $\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})$ is the weighted average of $\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}$, and the definition of $\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}$ is similar to $\widehat{\mu}_{i,m}^{(0)}$ except that the estimator is based on data corresponding to the subgroup $\mathcal{G}_j^c$. Note that criterion (4) will reduce to the leave-one-out cross-validation criterion when $J = n_0$. The weight vector can be obtained by solving the following constrained optimization problem

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} CV(\boldsymbol{w}). \tag{5}$$

Then the resulting model averaging prediction of $y_{n_0+1}^{(0)}$ associated with the new sample from the target model is given by $\widehat{\mu}_{n_0+1}^{(0)}(\widehat{\boldsymbol{w}})$. We summarize our procedure in Algorithm 1, and we term it **Trans**fer learning for **S**emiparametric **M**odel **A**veraging **P**rediction (Trans-SMAP). As suggested by an anonymous referee, we provide a discussion about the computational complexity of Algorithm 1 in Appendix B.2.

Remark 1 *The rationale of the proposed criterion is to find a proper weighted averaging estimator by minimizing the expected squared loss* $E[\{y_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2]$*, and we consider a data-driven approach to approximate it through the* J*-fold cross-validation criterion* (4)*. Formally, we prove that* $CV(\boldsymbol{w})$ *is an asymptotically unbiased estimator of* $E[\{y_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2]$ *that is expected to minimize. The details are provided in Appendix A.2.*

Remark 2 *The choice of* J *in criterion* (4) *is usually uncertain in practice. Since there is no theoretically optimal value, we manually use the 5-fold CV criterion in terms of computational efficiency in this paper. To be more convincing, we design additional experiments to compare different choices of* J *together with the leave-one-out procedure in terms of prediction error and time consumption. In conclusion, we find that the prediction performance is not sensitive to the choice of* J*, but the time consumption of the* J*-fold CV criteria can be greatly reduced compared to the leave-one-out procedure as the sample size increases. More details can be referred to Appendix C.5.*

3. Theoretical Properties

In this section, we establish some statistical properties of our method. Specifically, we derive the properties under the misspecified target model and correctly specified target model, respectively. Note that a correctly specified target model in our setting means that all the variables in the true target model (1) are included when fitting the model.

Define the risk function as $R(\boldsymbol{w}) = E[\{\mu_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2]$ and the prediction risk function as $\mathrm{PR}(\boldsymbol{w}) = E[\{y_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2]$. It is easy to see the decomposition $\mathrm{PR}(\boldsymbol{w}) = E\{(y_{n_0+1}^{(0)} - \mu_{n_0+1}^{(0)})^2\} + R(\boldsymbol{w})$, where the first term is unrelated to $\boldsymbol{w}$. This decomposition

implies that our objective is equivalent to minimizing $R(\boldsymbol{w})$ based on Remark 1. Therefore, we will first establish the asymptotic optimality of our model averaging procedure with respect to minimizing $R(\boldsymbol{w})$.

Note that all the limiting processes throughout this paper correspond to $\underline{n} \to \infty$, where $\underline{n} = \min_{0\le m\le M} n_m$. Let $\overline{p} = \max_{0\le m\le M} p_m$ and $\overline{v}^{(0)} = \max_{1\le l\le q_0} v_l^{(0)}$. In addition, we allow the number of source models M to go to infinity. Let $a \vee b$ denote $\max\{a, b\}$ and $a \wedge b$ denote $\min\{a, b\}$.

3.1 Asymptotic Optimality under Misspecified Target Model

For convenience, before we provide the theoretical properties formally, some notations need to be defined. Suppose that the pseudo-true values of parameters exist for each model, and let $\widetilde{\boldsymbol{\theta}}_m^{(0)} = \{(\widetilde{\boldsymbol{\beta}}^{(m)})^T, (\widetilde{\boldsymbol{\gamma}}_1^{(0)})^T, \ldots, (\widetilde{\boldsymbol{\gamma}}_{q_0}^{(0)})^T\}^T$ denote the corresponding values for $m = 1, \ldots, M$. For $j = 1, \ldots, J$, $i \in \mathcal{G}_j$, denote the in-sample prediction of the mth model using the subgroup samples in $\mathcal{G}_j^c$ by $\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)} = (\boldsymbol{d}_i^{(0)})^T \widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}$, where $\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)} = \{(\widehat{\boldsymbol{\beta}}_{[\mathcal{G}_j^c]}^{(0)})^T, (\widehat{\boldsymbol{\gamma}}_{1,[\mathcal{G}_j^c]}^{(0)})^T, \ldots, (\widehat{\boldsymbol{\gamma}}_{q_0,[\mathcal{G}_j^c]}^{(0)})^T\}^T$ for $m = 0$ and $\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)} = \{(\widehat{\boldsymbol{\beta}}^{(m)})^T, (\widehat{\boldsymbol{\gamma}}_{1,[\mathcal{G}_j^c]}^{(0)})^T, \ldots, (\widehat{\boldsymbol{\gamma}}_{q_0,[\mathcal{G}_j^c]}^{(0)})^T\}^T$ for $m = 1, \ldots, M$. Then the weighted averaging estimator can be written as $\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w}) = \sum_{m=0}^M w_m \widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}$. Note that $\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}$ and $\widehat{\boldsymbol{\theta}}_m^{(0)}$ have identical limiting values $\widetilde{\boldsymbol{\theta}}_m^{(0)}$ under large samples. In addition, the in-sample prediction of the mth model based on the pseudo-true values is defined as $\widetilde{\mu}_{i,m}^{(0)} = (\boldsymbol{d}_i^{(0)})^T \widetilde{\boldsymbol{\theta}}_m^{(0)}$, and the corresponding averaging prediction is $\widetilde{\mu}_i^{(0)}(\boldsymbol{w}) = \sum_{m=0}^M w_m \widetilde{\mu}_{i,m}^{(0)}$.

Now, we introduce some notations for the prediction associated with the new sample. Define the prediction of $y_{n_0+1}^{(0)}$ under the mth model based on $\widetilde{\boldsymbol{\theta}}_m^{(0)}$ and $\widehat{\boldsymbol{\theta}}_m^{(0)}$ as $\widetilde{\mu}_{n_0+1,m}^{(0)} = (\boldsymbol{d}_{n_0+1}^{(0)})^T \widetilde{\boldsymbol{\theta}}_m^{(0)}$ and $\widehat{\mu}_{n_0+1,m}^{(0)} = (\boldsymbol{d}_{n_0+1}^{(0)})^T \widehat{\boldsymbol{\theta}}_m^{(0)}$. Then the corresponding averaging predictions are $\widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) = \sum_{m=0}^M w_m \widetilde{\mu}_{n_0+1,m}^{(0)}$ and $\widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) = \sum_{m=0}^M w_m \widehat{\mu}_{n_0+1,m}^{(0)}$.

Further, denote the risk function calculated based on the pseudo-true values by $\widetilde{R}(\boldsymbol{w}) = E[\{\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2]$. Let $\xi_n = \inf_{\boldsymbol{w}\in\mathcal{W}} \widetilde{R}(\boldsymbol{w})$ be the minimum risk over the class of averaging estimators. Let $\mathcal{O}(\widetilde{\boldsymbol{\theta}}_m^{(0)}, c)$ denote a neighborhood of $\widetilde{\boldsymbol{\theta}}_m^{(0)}$ for some constant c such that $\|\widetilde{\boldsymbol{\theta}}_m^{(0)} - \boldsymbol{\theta}\| \le c$ for any $\boldsymbol{\theta} \in \mathcal{O}(\widetilde{\boldsymbol{\theta}}_m^{(0)}, c)$. To establish the asymptotic optimality for our method, some regularity conditions are necessarily stated as follows.

Condition 1 *Let s be a positive integer and $t \in (0, 1]$ such that $\kappa = s + t > 1.5$, and let $\mathcal{F}$ denote the collection of functions f on $[0, 1]$ whose sth derivative, $f^{[s]}$, exists and satisfies the Lipschitz condition of order t; that is $|f^{[s]}(x') - f^{[s]}(x)| \le C_0|x' - x|^t$, $0 \le x', x \le 1$, for some positive C_0. Then, (i) the nonparametric functions $g_l^{(m)}$ for $l = 1, \ldots, q_m, m = 0, \ldots, M$ in model (1) belong to $\mathcal{F}$; (ii) the number of interior knots for each spline approximation $S_l^{(m)}$ satisfies $n_m^{1/2\kappa} \ll S_l^{(m)} \ll n_m^{1/3}$.*

Condition 2 *For $m = 0, \ldots, M$, the distribution of $z_{il}^{(m)}$ for $l = 1, \ldots, q_m$ is absolutely continuous, and its density is bounded away from zero and infinity uniformly over l.*

Condition 3 *Suppose that* $M \leq \underline{n}$. *Uniformly for* $m = 0, \ldots, M$, *there exist limiting values* $\widetilde{\boldsymbol{\theta}}^{(m)}$ *such that* $\|\widehat{\boldsymbol{\theta}}^{(m)} - \widetilde{\boldsymbol{\theta}}^{(m)}\| = O_p(p_m^{1/2} n_m^{-1/2} M^{1/2})$, *where* $\widehat{\boldsymbol{\theta}}^{(m)} = \{(\widehat{\boldsymbol{\beta}}^{(m)})^T, (\widehat{\boldsymbol{\gamma}}_1^{(m)})^T, \ldots, (\widehat{\boldsymbol{\gamma}}_{q_m}^{(m)})^T\}^T$ *and* $\widetilde{\boldsymbol{\theta}}^{(m)} = \{(\widetilde{\boldsymbol{\beta}}^{(m)})^T, (\widetilde{\boldsymbol{\gamma}}_1^{(m)})^T, \ldots, (\widetilde{\boldsymbol{\gamma}}_{q_m}^{(m)})^T\}^T$.

Condition 4 *For* $m = 0, \ldots, M$, $\overline{p} = o(n_m^{1/2})$.

Condition 5 *The expectations* $E\{(\mu_i^{(0)})^4\}$, $E\{(\widetilde{\mu}_{i,m}^{(0)})^4\}$, *and* $E\{(\varepsilon_i^{(0)})^4\}$ *exist for* $m = 0, \ldots, M$.

Condition 6 *For* $i = 1, \ldots, n_0$, $j = 1, \ldots, J$, *(i)* $\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}$ *is differentiable with respect to* $\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}$; *(ii) there exists a positive constant* c *such that*

$$E\left(\sup_{\bar{\boldsymbol{\theta}} \in \mathcal{O}(\widetilde{\boldsymbol{\theta}}_m^{(0)}, c)} \left\| \frac{\partial \widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}}{\partial \widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}} \Big|_{\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)} = \bar{\boldsymbol{\theta}}} \right\|^2\right) = O(p_m)$$

uniformly for $m = 0, \ldots, M$.

Condition 7 $\xi_n^{-1} \sup_{\boldsymbol{w} \in \mathcal{W}} |\{\mu_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2 - \{\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\}^2|$ *is uniformly integrable.*

Condition 8 $\{(\overline{p}\underline{n}^{-1/2} M^{1/2}) \vee (\overline{p}^2 \underline{n}^{-1} M)\} \xi_n^{-1} = o(1)$.

Conditions 1 and 2 are mild and commonly assumed for nonparametric models with spline-based approximation in the literature, such as Stone (1985), Stone (1986), Wang et al. (2011) and Zhang and Liang (2011). Condition 3 ensures that the estimator $\widehat{\boldsymbol{\theta}}^{(m)}$ in each model has the limiting value $\widetilde{\boldsymbol{\theta}}^{(m)}$. This can be regarded as a variant of the common condition to establish the asymptotic properties in literature; see, for example, Zhang et al. (2016), Ando and Li (2017), and Zhang and Liu (2023). When the number of source models M is fixed, the required convergence rates are slower than the rates for parametric models established in White (1982) due to the semiparametric model settings. We further weaken the convergence rates to accommodate uniform convergence under a possibly diverging M. Note that the dimension of the parametric component in each model allows to be divergent in our setting, and the extension to high-dimensional settings of $p > n_m$ is left for future study.

Condition 4 restricts the divergence rates of dimensions of the target and source models based on B-spline approximations as the sample size increases, which is commonly assumed in the literature (Liao et al., 2021). Suppose that the polynomial degree of all spline basis functions is fixed in our setting. Theorem 2 in Stone (1986) shows that the cubic spline estimator of the nonparametric function achieves the optimal convergence rate if the number of interior knots has the order of $n_m^{1/5}$. When the spline estimators for each model achieve the optimal convergence rate in our framework, Condition 4 still holds.

Conditions 5-7 are some mild technical conditions. Specifically, Condition 5 includes some general constraints of moments on the conditional expectation, prediction, and error, which are commonly assumed in the literature (Wan et al., 2010; Ando and Li, 2017; Zhang and Wang, 2019). Condition 6 concerns boundedness and differentiability; it is also adopted

in Zhang and Liu (2023). Condition 7 is mainly imposed for technical needs to obtain expectation in the proof.

Condition 8 plays an important role in the proof of our theorem. It restricts the divergence rate of the number of source models and requires that the target model is sufficiently misspecified. Similar conditions in the literature are Condition 7 in Ando and Li (2014), Condition C.6 in Zhang et al. (2016), and Assumption 5 in Zhang and Liu (2023). A detailed explanation of the misspecification of the target model under Condition 8 is given in Appendix A.1. Note that Condition 8 also implies an upper bound on the number of source models M. Specifically, if we assume that the target model is misspecified and $\overline{p}\underline{n}^{-1/2}M^{1/2} = o(1)$, then we have $M^{1/2} = o(\overline{p}^{-1}\underline{n}^{1/2}\xi_n)$ based on Condition 8. If we further assume $\xi_n = O(1)$ and $\overline{p} = O(\underline{n}^{1/2-\zeta})$ for $0 < \zeta \leq 1/2$ based on Condition 4, then the order of M is $o(\underline{n}^{2\zeta})$.

Next, we formally present the theoretical property under the above conditions in Theorem 1. The proof is provided in Appendix A.3.

Theorem 1 *Under Conditions 1-8, we have*

$$\frac{R(\widehat{\boldsymbol{w}})}{\inf\limits_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w})} \to 1$$

in probability.

Theorem 1 indicates that the proposed model averaging prediction has the asymptotic optimality in the sense of achieving the lowest possible out-of-sample prediction risk, which is a fundamental but important property in the frequentist model averaging literature. Unlike most previous works, the proposed procedure is constructed using multiple data sets, and the corresponding asymptotic optimality is established based on the out-of-sample prediction risk, which is more practical for predictive task in our context. It is worth noting that our result of asymptotic optimality always holds regardless of whether the source models are correct, which is not surprising since the target model is our main concern and naturally dominates the performance.

3.2 Weight Convergence under Correct Target Model

We now turn to the case of the correct target model. Since there are no requirements for the model specification of source models, it may contain both correct and misspecified models. Define the informative models as the models having the same pseudo-true values of $\widetilde{\boldsymbol{\beta}}^{(m)}$ as the target model, and let $\mathcal{I} \subseteq \{0, \ldots, M\}$ be the corresponding set of indices. Obviously, $0 \in \mathcal{I}$. Let $\mathcal{I}^c$ be the complement of $\mathcal{I}$. Note that the informative models are characterized by the parameter effects in the limit, which also reflects the similarity between the estimators for the target model and source models in a sense. We further define the sum of the weight estimators for the informative models as $\widehat{\tau} = \sum_{m\in\mathcal{I}} \widehat{w}_m$. To study the theoretical property of the weights, we rely on an alternative for Condition 8 and some other technical conditions. Let $\overline{\mathcal{W}} = \{\boldsymbol{w} \in \mathcal{W} : \sum_{m\in\mathcal{I}} w_m = 0\}$ be the subset of $\mathcal{W}$ that assigns all the weights to the models belonging to $\mathcal{I}^c$. Then the conditions are presented as follows.

Condition 9 $(\overline{v}^{(0)})^{1/2-\kappa} = o\{(\overline{p}\underline{n}^{-1/2}M^{1/2}) \wedge (\overline{p}^2\underline{n}^{-1}M)\}$, *where* κ *is defined in Condition 1.*

Condition 10 $E\left(\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1,m}\right) = O(1)$ *uniformly for* $m \in \mathcal{I}^c$.

Condition 11 $\{(\overline{p}\underline{n}^{-1/2}M^{1/2}) \vee (\overline{p}^2\underline{n}^{-1}M)\}\{\inf_{\boldsymbol{w}\in\overline{\mathcal{W}}} \widetilde{R}(\boldsymbol{w})\}^{-1} = o(1)$.

Condition 12 $\{\inf_{\boldsymbol{w}\in\overline{\mathcal{W}}} \widetilde{R}(\boldsymbol{w})\}^{-1} = O(1)$.

Condition 9 specifically constrains the divergence rate of dimension of the spline basis in the target model. Condition 10 restricts the boundedness of approximation between $\mu^{(0)}_{n_0+1}$ and $\widetilde{\mu}^{(0)}_{n_0+1,m}$ based on the models belonging to $\mathcal{I}^c$; this restriction is common and reasonable. Condition 11 imposes a restriction on the growth rate of the minimum risk based on the averaging prediction over models in $\mathcal{I}^c$, which is an extension to Assumption 6 in Zhang and Liu (2023) with semiparametric model settings. Note that Condition 11 implies that the risk of the target model is sufficiently large when the target model is misspecified, so we do not need the constraint in Condition 8. Based on this condition, the resulting risk will converge to zero when combining any informative source models with the correct target model. On the contrary, the risk will be much larger than $(\overline{p}\underline{n}^{-1/2}M^{1/2}) \vee (\overline{p}^2\underline{n}^{-1}M)$ asymptotically when combining any source model belonging to $\mathcal{I}^c$ even with the correct target model. Therefore, it is intuitively clear that none of the models in $\mathcal{I}^c$ will be asymptotically assigned nonzero weights based on our criterion. Condition 12 restricts the risk of the model averaging estimator under misspecified models. It can be seen that Condition 11 is satisfied if we have $\overline{p}^2M = o(n)$ combined with Condition 12. We summarize the property of weight convergence in the following Theorem 2, and the proof is provided in Appendix A.4.

Theorem 2 *If Conditions 1-6 and 9-11 are satisfied, then* $\widehat{\tau} \to 1$ *in probability.*

Theorem 2 demonstrates that our procedure asymptotically assigns all the weights to informative models when the target model is correctly specified, which can be regarded as a type of consistency property in model selection. In other words, the proposed model averaging criterion can consistently select informative models by weight assignments, which is an important distinction compared to existing parameter-transfer learning frameworks. From the definition of the informative models, both correct and incorrect source models may contribute to the prediction if the parameter information is similar enough. In addition, our method yields robust performance even when several models have strong dissimilarity because those models are assigned small or zero weights asymptotically.

As discussed in previous sections, we hope to avoid the negative transfer problem in practice. Next, we further discuss the difference between the upper bound of the risk of our method and that of the least squares estimator on target data only, which is summarized in the following corollary. Let $\bar{R}(\widehat{\boldsymbol{w}})$ and $\bar{R}_0$ denote the upper bounds of the risk of our Trans-SMAP and the least squares estimator on target data, respectively. The technical details are provided in Appendix A.5.

Corollary 3 *Assume that the dimensions and sample sizes of the target and source data satisfy* $\bar{p}^2\underline{n}^{-1} = O(p_0^2 n_0^{-1})$. *If Conditions 1-8 hold or Conditions 1-7, Conditions 9-10 and Condition 12 hold, then* $\bar{R}(\widehat{\boldsymbol{w}}) = O_p(\bar{R}_0)$.

Corollary 1 demonstrates that the upper bound of the risk of our Trans-SMAP has no larger order than that of the least squares estimator on target data regardless of whether the target model is correct. Hence, in a sense, it also verifies that our procedure provides a reasonable strategy to mitigate the potential negative transfer problem.

4. Simulation Studies

In this section, we evaluate the finite sample performance of our procedure in various numerical experiments. For comparison, we consider the following seven competitive procedures: transfer learning by the simple averaging procedure (termed as "Trans-SimpMA"), transfer learning by the smoothed AIC and BIC (Buckland et al., 1997) based model averaging procedures (termed as "Trans-SAIC" and "Trans-SBIC"), least squares estimator using the target data only (termed as "LSE-Tar"), least squares estimator using all the data (termed as "LSE-All"), Trans-Lasso (Li et al., 2021), and Trans-GLM (Tian and Feng, 2022). Specifically, Trans-SimpMA, Trans-SAIC, and Trans-SBIC construct the corresponding weighted averaging estimators with equal weight $1/(M+1)$, $\exp(-\mathrm{AIC}_m/2)/\sum_{m=0}^{M}\exp(-\mathrm{AIC}_m/2)$, and $\exp(-\mathrm{BIC}_m/2)/\sum_{m=0}^{M}\exp(-\mathrm{BIC}_m/2)$, respectively, where $\mathrm{AIC}_m = \log\{n_m^{-1}\sum_{i=1}^{n_m}(y_i^{(m)} - \widehat{\mu}_i^{(m)})^2\} + 2p_m/n_m$ and $\mathrm{BIC}_m = \log\{n_m^{-1}\sum_{i=1}^{n_m}(y_i^{(m)} - \widehat{\mu}_i^{(m)})^2\} + p_m\log n_m/n_m$. The purpose of comparing these methods is to verify the superiority of our proposed method. LSE-Tar performs the prediction with the least squares estimator (2) using the target data. LSE-All performs the prediction with the least squares estimator based on all the target and source data by minimizing the following integrative loss function $L(\boldsymbol{\beta}, \boldsymbol{\gamma}_1^{(0)}, ..., \boldsymbol{\gamma}_{q_0}^{(0)}, \ldots, \boldsymbol{\gamma}_{q_m}^{(m)}) = (2\sum_{m=0}^{M} n_m)^{-1}\sum_{m=0}^{M}\sum_{i=1}^{n_m}[y_i^{(m)} - (\boldsymbol{x}_i^{(m)})^T\boldsymbol{\beta} - \sum_{l=1}^{q_m}\{B_l^{(m)}(z_{il}^{(m)})\}^T\boldsymbol{\gamma}_l^{(m)}]^2$. The purpose of considering LSE-Tar and LSE-All is to understand the effect of reasonable knowledge transfer. To comprehensively demonstrate the superiority of our procedure, we also consider two recent transfer learning approaches, Trans-Lasso and Trans-GLM, related to our framework. All experiments are implemented in R software, and more details can be seen in Appendix B.1.

4.1 Simulation Design

Set the target sample size $n_0 = 150$, and source sample sizes $(n_1, n_2, n_3) = (200, 200, 150)$. For the parametric components, $\boldsymbol{x}_i^{(m)}$ from the target and source models are generated from a 6-dimensional multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with $\Sigma = [\Sigma_{aa'}]_{6\times 6}$, where $\Sigma_{aa'} = 0.5^{|a-a'|}$. Set the parametric coefficient vectors of the target and source models as $\boldsymbol{\beta}^{(0)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T$, $\boldsymbol{\beta}^{(1)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3, 1.8)^T + \delta_1$, $\boldsymbol{\beta}^{(2)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T + \delta_2$, and $\boldsymbol{\beta}^{(3)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T + \delta_3$, where δ_1, δ_2, δ_3 are the parametric coefficient differences relative to the target model. Further, we set $\delta_1 = 0.02$, $\delta_2 = 0.3$, and $\delta_3 = 0$, so the parameters of the first and second source models are different from the target model, and the third source model is informative because its coefficient is exactly same as that of the target model. Note that the parametric coefficient

of the first source model is a 7-dimensional vector, whereas the others are 6-dimensional vectors. Here, we always omit the last component of $\boldsymbol{x}_i^{(1)}$ when fitting the first source model, so the model is misspecified. For the other models, we do not ignore any components, and then they are all correctly specified. The above setting is the case of the correct target model. When the target model is misspecified, we keep other settings unchanged except setting $\boldsymbol{\beta}^{(0)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3, 0.1)^T$, and we similarly omit the last component of $\boldsymbol{x}_i^{(0)}$.

We set the dimensions of the nonparametric component for each model as $q_m = 3$ for $m = 0, \ldots, M$, and we generate $z_{il}^{(m)}$ from a uniform distribution $U(0, 1)$. The following nonlinear functions for different models are considered: $g^{(0)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2) + u_3$, $g^{(1)}(\boldsymbol{u}) = 2(u_1 + 0.5)^3 + \cos(\pi u_2) + u_3$, $g^{(2)}(\boldsymbol{u}) = 2.5(u_1 + 0.3)^3 + \sin(\pi u_2) + 1.5u_3$, and $g^{(3)}(\boldsymbol{u}) = 1.8(u_1 + 0.3)^3 + \cos(\pi u_2) + u_3$. In order to accommodate the settings of Trans-Lasso and Trans-GLM for convenient comparison, we consider the scenairo of multiple data with equal dimensions, but our framework is not limited to this setting. Hence, we conduct additional simulation studies in heterogeneous dimension settings in Appendix C.2. The random error term $\varepsilon_i^{(m)}$ for the $M + 1$ models follows a normal distribution $N(0, 0.5^2)$ with $\sigma_{i,m} = 0.5$ for $m = 0, \ldots, M, i = 1, \ldots, n_m$. Here, we mainly consider the homoscedastic setting as an example. Since our framework is compatible with heteroscedasticity, we further conduct the heteroscedastic design in Appendix C.4. To evaluate the prediction performance, we generate new samples from the target model with sample size $n^* = 500$. Furthermore, we design alternative settings to increase the sample sizes to $(300, 350, 350, 250)$ and $(500, 550, 500, 450)$, while keeping the other settings invariant. All the experiments are replicated 500 times. Following a referee's suggestion, we conduct an additional simulation study in high-dimensional settings that matches the assumptions of competing methods like Trans-Lasso to provide a relatively fair comparison. Further details regarding this simulation study can be found in Appendix C.8.

Next, we consider increasing the number of source models. Let the sample sizes be $(n_0, \ldots, n_6) = (150, 200, 150, 200, 150, 150, 200)$. The parametric coefficients of the target and source models are set as $\boldsymbol{\beta}^{(0)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T$, $\boldsymbol{\beta}^{(1)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3, 1.8)^T + \delta_1$, $\boldsymbol{\beta}^{(2)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T + \delta_2$, $\boldsymbol{\beta}^{(3)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T + \delta_3$, $\boldsymbol{\beta}^{(4)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3, 1.8)^T + \delta_4$, $\boldsymbol{\beta}^{(5)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T + \delta_5$, and $\boldsymbol{\beta}^{(6)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3)^T + \delta_6$, where $\delta_1 = 0.02, \delta_2 = 0.02, \delta_3 = 0.3, \delta_4 = 0, \delta_5 = 0.02, \delta_6 = 0.3$. Let the nonlinear functions for different models be $g^{(0)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2) + u_3$, $g^{(1)}(\boldsymbol{u}) = 2(u_1 + 0.5)^3 + \cos(\pi u_2) + u_3$, $g^{(2)}(\boldsymbol{u}) = 2.5(u_1 + 0.3)^3 + \sin(\pi u_2) + 1.5u_3$, $g^{(3)}(\boldsymbol{u}) = 1.8(u_1 + 0.3)^3 + \cos(\pi u_2) + u_3$, $g^{(4)}(\boldsymbol{u}) = 1.5(u_1 + 0.5)^3 + \cos(2\pi u_2) + u_3^2$, $g^{(5)}(\boldsymbol{u}) = (u_1 + 0.6)^2 + \cos(\pi u_2) + 1.3u_3^2$, and $g^{(6)}(\boldsymbol{u}) = 1.3(u_1 + 0.5)^2 + \cos(2\pi u_2) + 1.6u_3$. Similarly, we design additional settings with large sample sizes $(300, 350, 300, 350, 300, 300, 350)$ and $(500, 550, 500, 550, 500, 500, 550)$.

4.2 Simulation Results

We evaluate all the methods by the mean squared error (MSE) based on the new samples with sample size n^*, which is expressed as $\text{MSE} = \sum_{i=1}^{n^*} (\widehat{\mu}_i^{(0)} - \mu_i^{(0)})^2 / n^*$. The results of the averaged MSE based on 500 replications are reported in Table 1. The rows labeled "Uplift Rate" in each table represent the percentage improvement of the averaged MSE for the

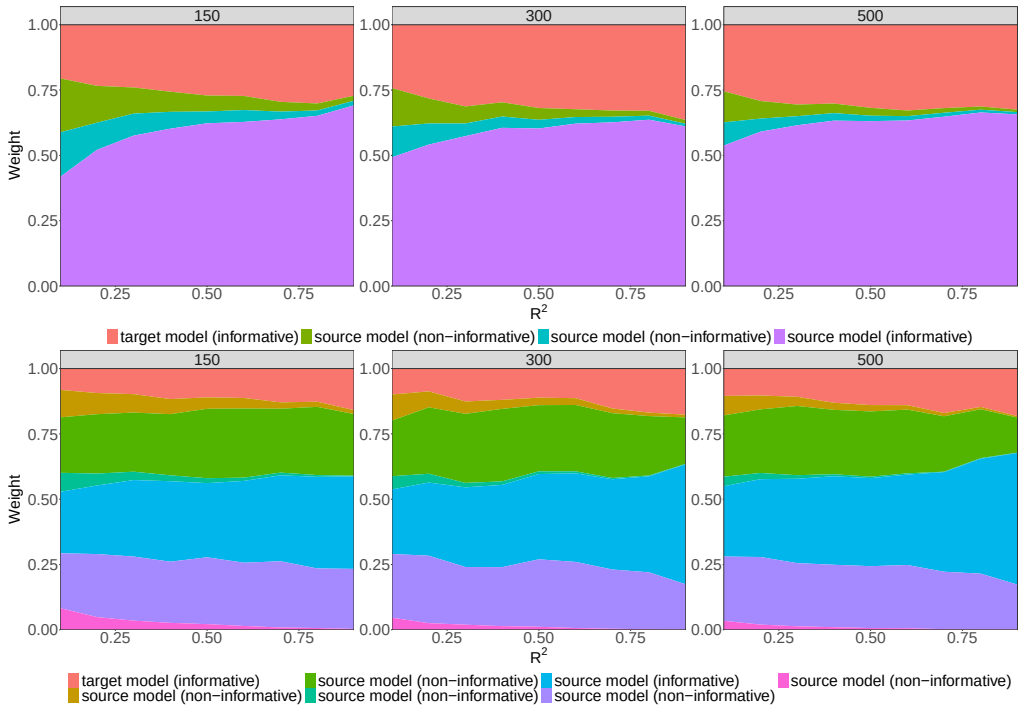

Figure 1: Stacked area plots of the averaged weight assignments based on our method when the target model is correctly specified with $M = 3$ (first line) and $M = 6$ (second line).

proposed method relative to the smallest result among alternative methods. According to the results in Table 1, we see that the proposed Trans-SMAP outperforms all the competitive methods. The improvement is particularly significant as the number of models increases, especially in the correct target model scenario. When the target model is misspecified, the superiority is not surprising because Theorem 1 provides the theoretical guarantee for the asymptotic optimality. The proposed method still has advantages when the target model is correctly specified due to that our procedure asymptotically uses the informative models based on Theorem 2, whereas LSE-Tar does not borrow any auxiliary information, and other approaches may not use auxiliary information effectively. It is worth noting that the recent transfer learning approaches, Trans-Lasso and Trans-GLM, are also not comparable to our method. Except for the inherent limitations discussed in the introduction, the most important reason is that these procedures are aimed at parametric models. To evaluate the performance under various levels of noise, we let the variance of random error vary such that $R^2 = \text{var}(\mu_i^{(0)})/\text{var}(y_i^{(0)})$ ranges from 0.1 to 0.9 with increments of 0.1, and the detailed results are provided in Appendix C.1. Additionally, we design more general experiments to further demonstrate the stability of our procedure under potential negative transfer scenarios in Appendix C.3. The details are reported in Tables 4-7, and the corresponding results still support our method.

		Correct Target Model			Misspecified Target Model		
	Method	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$
$M = 3$	Trans-SMAP	0.027	0.013	0.008	0.034	0.021	0.015
		(0.010)	(0.005)	(0.003)	(0.010)	(0.005)	(0.003)
	Trans-SimpMA	0.239	0.226	0.220	0.218	0.205	0.198
		(0.038)	(0.028)	(0.023)	(0.034)	(0.024)	(0.021)
	Trans-SBIC	0.188	0.173	0.165	0.180	0.165	0.156
		(0.028)	(0.020)	(0.016)	(0.027)	(0.019)	(0.016)
	Trans-SAIC	0.183	0.170	0.165	0.174	0.161	0.155
		(0.027)	(0.020)	(0.016)	(0.026)	(0.019)	(0.016)
	LSE-Tar	0.030	0.015	0.009	0.038	0.022	0.016
		(0.011)	(0.005)	(0.003)	(0.011)	(0.005)	(0.004)
	LSE-All	0.347	0.309	0.260	0.316	0.279	0.234
		(0.065)	(0.046)	(0.032)	(0.062)	(0.040)	(0.029)
	Trans-Lasso	0.123	0.110	0.104	0.130	0.117	0.112
		(0.013)	(0.008)	(0.006)	(0.014)	(0.009)	(0.007)
	Trans-GLM	0.124	0.109	0.103	0.131	0.116	0.112
		(0.017)	(0.007)	(0.006)	(0.017)	(0.008)	(0.007)
	Uplift Rate	11.11%	15.38%	12.50%	11.76%	4.76%	6.67%
$M = 6$	Trans-SMAP	0.026	0.013	0.008	0.034	0.020	0.015
		(0.010)	(0.004)	(0.003)	(0.010)	(0.005)	(0.003)
	Trans-SimpMA	0.211	0.199	0.197	0.194	0.181	0.178
		(0.026)	(0.020)	(0.018)	(0.024)	(0.019)	(0.016)
	Trans-SBIC	0.195	0.176	0.173	0.182	0.166	0.161
		(0.023)	(0.017)	(0.015)	(0.021)	(0.016)	(0.014)
	Trans-SAIC	0.189	0.174	0.172	0.177	0.164	0.160
		(0.023)	(0.017)	(0.015)	(0.021)	(0.016)	(0.014)
	LSE-Tar	0.031	0.015	0.009	0.039	0.022	0.016
		(0.012)	(0.005)	(0.003)	(0.013)	(0.006)	(0.003)
	LSE-All	0.336	0.278	0.259	0.311	0.256	0.238
		(0.049)	(0.031)	(0.025)	(0.043)	(0.030)	(0.024)
	Trans-Lasso	0.124	0.109	0.104	0.133	0.117	0.111
		(0.016)	(0.008)	(0.006)	(0.015)	(0.009)	(0.007)
	Trans-GLM	0.127	0.108	0.104	0.136	0.116	0.111
		(0.019)	(0.008)	(0.006)	(0.019)	(0.008)	(0.006)
	Uplift Rate	19.23%	15.38%	12.50%	14.71%	10.00%	6.67%

Table 1: The averaged MSE of out-of-sample prediction for different methods. The standard errors are given in parenthesis.

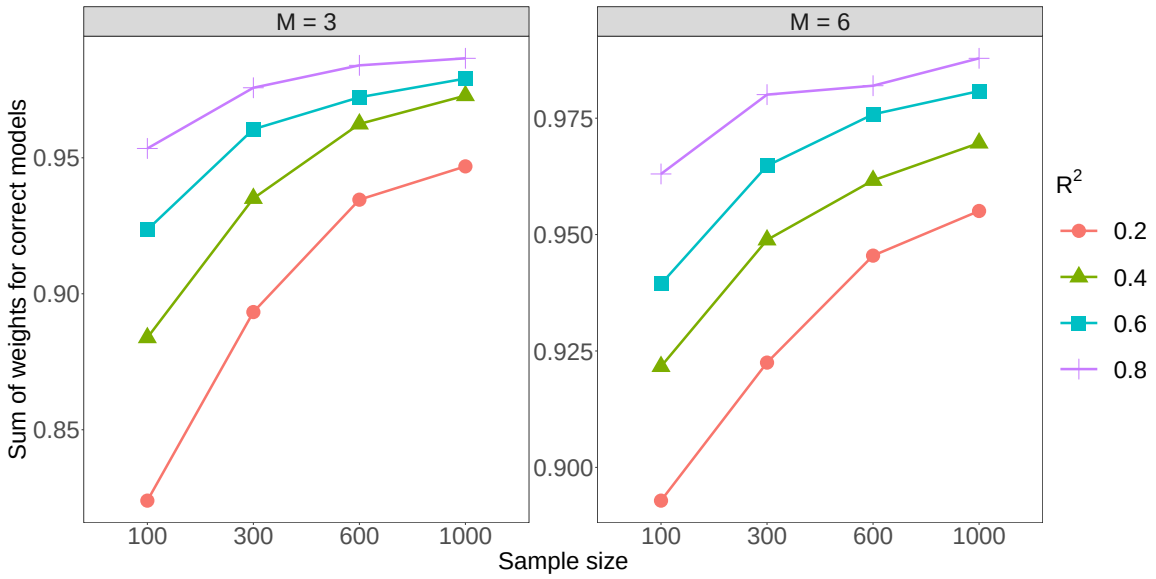

Figure 2: The averaged sum of weights for informative models.

We further evaluate the performance of weight estimation. Figure 1 shows the results of averaged weight assignments when the target model is correctly specified, where areas with different colors denote averaging weights under various R^2. It can be observed that large weights tend to be assigned to the target model as well as source models with small differences in parameter effects, and the trend becomes more significant as R^2 grows. This is not unexpected because the target model is of interest, and smaller differences indicate possibly more informative models for knowledge transfer. Notice that the weight for the informative source model is even larger than the weight for the target model, a possible reason of which is that the informative source model has a larger sample size than the target model. To exclude the influence of confounding factors on the weight estimation, we simply adjust the settings to let the target and source sample sizes be equal but vary from $\{100, 300, 600, 1000\}$, and we set only one source model as the misspecified model with the remaining models being informative. The relationship between the sum of weights for informative models and the sample size is illustrated in Figure 2, which clearly verifies the property in Theorem 2 that the sum of weights for informative models will be close to 1 as the sample size increases.

5. Empirical Data Analysis

In this section, we apply our approach to analyze housing rental information data in Beijing, which is drawn from a publicly available data set on `http://www.idatascience.cn/dataset`. Our primary goal is to predict the monthly rent that is conducive to better understanding and following up the housing rental market. Considering the similarity of geographical location, population structure, and rental demand, we choose five adjacent districts distributed in southwestern Beijing for our analysis, and the specific locations of rental houses

are marked in Figure 7. Overall, the data set for our analysis contains 1409 observations with 33 variables distributed in five districts (Daxing, Fangshan, Fengtai, Mentougou, and Shijingshan). To accommodate the transfer learning framework, we take the data from different districts as multi-source data sets. The sample sizes for the source domains of Daxing, Fangshan, Fengtai, Mentougou, and Shijingshan are $(n_1, ..., n_5) = (291, 247, 339, 263, 269)$. The response variable, denoted by $Y^{(m)}$ for $m = 1, 2, 3, 4, 5$, indicates the natural logarithm of the monthly rent. After excluding irrelevant variables by the preliminary variable selection, only ten covariates remain in our models, including the number of rooms ($X_1^{(m)}$), the number of restrooms ($X_2^{(m)}$), the number of living rooms ($X_3^{(m)}$), total area ($X_4^{(m)}$), have or not a bed ($X_5^{(m)}$), have or not a wardrobe ($X_6^{(m)}$), have or not a air conditioner ($X_7^{(m)}$), have or not a fuel gas ($X_8^{(m)}$), total floor ($Z_1^{(m)}$), the number of schools within 3 km ($Z_2^{(m)}$). More details can be seen in Table 10 in Appendix C.6. All the covariates have been properly transformed and scaled. Since we have little prior knowledge of whether the relationship between the response and each predictor is linear or nonlinear, we further conduct the marginal visualization for each district to determine our model specification. The marginal relationships between the natural logarithm of the monthly rent and ten predictors are plotted in Figure 8-Figure 12 in Appendix C.6. Since our framework theoretically allows transferring among parametric regression models, we construct an ordinary linear regression model for Fengtai. Therefore, we adopt the following models in this analysis

$$\mathrm{Y}^{(m)} = \begin{cases} \beta_0^{(m)} + \sum_{j=1}^{8} \beta_j^{(m)} X_j^{(m)} + \sum_{j'=1}^{2} g_{j'}^{(m)}(Z_{j'}^{(m)}) + \varepsilon^{(m)} & (m = 1, 2, 4, 5), \\ \beta_0^{(m)} + \sum_{j=1}^{8} \beta_j^{(m)} X_j^{(m)} + \sum_{j'=1}^{2} \gamma_{j'}^{(m)} Z_{j'}^{(m)} + \varepsilon^{(m)} & (m = 3). \end{cases}$$

To further demonstrate the replicability of our proposal, each data set will be regarded as the target domain, and the others will serve as source domains. Then we consider multiple combinations of the target model and source models and carry out our procedure one by one.

Next, we fit semiparametric models introduced in Section 2.1. To evaluate the out-of-sample prediction risk, we randomly split the target samples into two subgroups with equal size as the training and testing data. Then we calculate the mean squared prediction error MSPE $= 2\|Y_{[k]}^{(m)} - \widehat{Y}_{[k]}^{(m)}\|^2/n_m, m = 1, \ldots, 5$, where subscript $[k]$ denotes the kth replication, and n_m refers to the sample size of the mth data set. We repeat the above process 500 times, and the corresponding results are illustrated in Figure 3. Following a referee's advice, we further conduct an additional simulation study similar to the real data structure in Appendix C.7. The detailed results are summarized in Table 11.

According to Figure 3, it can be seen that all the CV criteria based Trans-SMAP perform similarly, and they outperform alternative methods for most of the target domains. Specifically, Trans-SMAP (5-fold CV) yields a smaller median of the MSPE than Trans-SimpMA, Trans-SBIC, and Trans-SAIC for Daxing, Fangshan, Fengtai, and Mentougou, which demonstrates the superiority of our weight choice criterion with theoretical support. LSE-Tar performs worse than Trans-SMAP (5-fold CV) for all target domains except Fengtai, but the advantage of LSE-Tar is relatively small. LSE-All performs much worse than our method in all scenarios. The poor performance of LSE-Tar and LSE-All results from their ineffective use of potential auxiliary information. Other parameter-transfer approaches,

	Daxing	Fangshan	Fengtai	Mentougou	Shijingshan
Daxing	0.366	0.114	0.163	0.010	0.347
Fangshan	0.189	0.297	0.002	0.397	0.116
Fengtai	0.125	0.003	0.656	0.000	0.216
Mentougou	0.000	0.199	0.000	0.780	0.020
Shijingshan	0.266	0.131	0.139	0.281	0.183

Table 2: The averaged weight assignments for different target domains. The rows represent different target domains, and columns are the corresponding models.

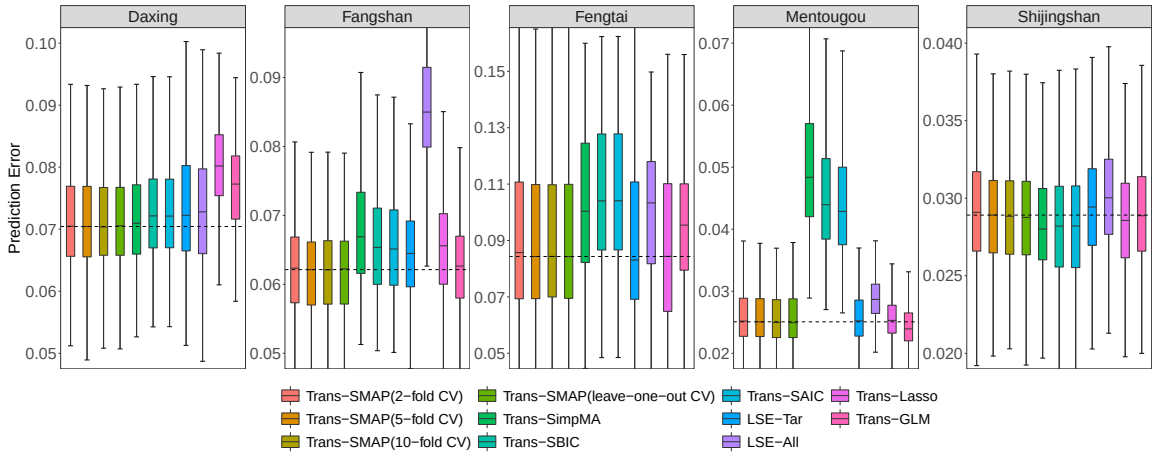

Figure 3: Boxplots of the MSPE for different target domains in the housing rental information data analysis.

Trans-Lasso and Trans-GLM, also perform much or slightly worse than Trans-SMAP (5-fold CV) for all of these target domains except Mentougou. This result shows that the prediction performance of our method is not sensitive to the choice of J. This conclusion is consistent with the results of our simulation studies, which are detailed in Figure 6 in Appendix C.5. However, it is important to note that our analysis of the influence of J is based on a numerical perspective, and a rigorous theoretical analysis is left for future research.

We further examine the weight assignments for the different models achieved by our criterion in Table 2. To be more intuitive, we plot a directed graph in Appendix C.6 to visualize the transfer network. Some interesting findings can be observed from the results. First, the weights are adaptively assigned to different models for different target domains with distinctiveness. Specifically, the weights for some models are very small or even exactly zero, which partly reflects the weak transferability. Second, from the weight assignments for the target domains of Fengtai and Mentougou, the target model plays a more important role

than the source models. This indicates that limited knowledge may be transferred to target tasks. On the contrary, for the target domains of Fangshan and Shijingshan, we can observe that the source models indeed improve the predictive job from the weight assignments and MSPE, among which Mentougou is commonly informative for both target models. Third, we find that Fangshan and Mentougou mutually serve as their most informative source models, so they may help each other to improve their performance.

In summary, the empirical data example demonstrates the effectiveness of our proposed Trans-SMAP in terms of the MSPE compared to competitive approaches and the proper weight assignments for different models, which suggests a promising strategy for predictive tasks in future applications.

6. Concluding Remarks

In the context of transfer learning, we propose an optimal parameter-transfer approach for prediction under a flexible semiparametric additive framework. We develop a model averaging approach to transfer possibly shared parameter information from source models to the target model. The asymptotic optimality of the out-of-sample prediction risk under the misspecified target model and weight convergence under the correct target model are derived. Extensive numerical results demonstrate our effectiveness compared to alternative methods and further support the theoretical findings. Note that our framework allows adopting parametric frameworks in different models, which demonstrates its flexibility in applications. Even though equal dimensions of parametric components are assumed in our setting, the proposed method theoretically allows for a more general scenario with different dimensions. In addition, our procedure provides a feasible strategy to deal with massive data. Specifically, we can split the data first and carry out estimating processes for each data set in parallel, after which we aggregate the corresponding estimators to construct the prediction by our strategy. Since only parameter estimators are exposed across multiple data sets in our procedure, in a sense, our approach can effectively protect the privacy of original data.

Several promising future attempts are worth further research. First, it would be interesting to further study the statistical inference for the resulting model averaging prediction. In this regard, some asymptotic distribution theories for the frequentist model averaging estimator have been established in the literature; see related works from Hjort and Claeskens (2003), Liu (2015), and Zhang and Liu (2019). Second, optimal parameter-transfer approaches under some variants of the semiparametric framework are also appealing, such as varying-coefficient models, single-index models, and their generalized versions with extensions to high-dimensional scenarios. Third, it is intuitive to combine multiple models by the traditional model averaging approaches instead of using a single model for each data set. Fourth, it is necessary to consider a data-driven procedure to select J in our criterion instead of just using some given values, and the theoretical investigation warrants further research. Last, transferring shared information of the nonparametric components is a very interesting topic. One possible strategy is to directly transfer the estimates of nonparametric functions in different models. Alternatively, we could consider transferring the hyperparameters in the corresponding nonparametric estimation methods, such as the number of internal knots or degree of the piecewise polynomial in spline-based approaches, and kernel function or

bandwidth in kernel-based methods. The specific methodology needs in-depth research in the future.

Acknowledgments

The authors are very grateful to the action editor and two anonymous referees for their constructive comments and suggestions that substantially improve the original manuscript. Zhang's work was supported by the National Natural Science Foundation of China under Grants 71925007, 72091212, 71988101 and 12288201, and the CAS Project for Young Scientists in Basic Research under Grant YSBR-008. Hu's work was supported by the China Postdoctoral Science Foundation under Grant 2021M703428. No potential conflict of interest is reported by the authors.

Appendix A. Technical Details

In this section, we provide some technical details in Section 3 and proofs of Theorem 1, Theorem 2, and Corollary 3.

A.1 Verification of Condition 8

Let $\|f\|^2 = E\{f^2(X)\} = \int_0^1 f^2(x)p(x)dx$ denote the L^2 norm of the function f on $[0,1]$, where $p(x)$ is the density of X. Denote the additive function of the target model by $g^{(0)}$ that can be represented in the form $g^{(0)}(\boldsymbol{z}^{(0)}_{n_0+1}) = \sum_{l=1}^{q_0} g_l^{(0)}(z^{(0)}_{n_0+1,l})$, where the component $g_l^{(0)}$ belongs to the space $\mathcal{F}$ introduced in Condition 1. Let $\widetilde{g}^{(0)}$ be the additive spline of the target model with the form $\widetilde{g}^{(0)}(\boldsymbol{z}^{(0)}_{n_0+1}) = \sum_{l=1}^{q_0} \widetilde{g}_l^{(0)}(z^{(0)}_{n_0+1,l}) = \sum_{l=1}^{q_0}(B_l^{(0)}(z^{(0)}_{n_0+1,l}))^T\widetilde{\boldsymbol{\gamma}}_l^{(0)}$, where the component $\widetilde{g}_l^{(0)}$ belongs to the space $\Psi^{(0)}$ introduced in Section 2.1.

In order to better understand Condition 8, suppose that the target model is correctly specified, according to Lemma 8 of Stone (1986), we have $\|g^{(0)} - \widetilde{g}^{(0)}\|^2 = O\{(\overline{v}^{(0)})^{-2\kappa}\}$, where $\overline{v}^{(0)} = \max_{1\le l\le q_0} v_l^{(0)} = \max_{1\le l\le q_0}\{r_l^{(0)} + S_l^{(0)}\}$ is the maximal dimension of the B-spline basis for the target model, and κ is defined in Condition 1. Then

$$\begin{aligned}
\xi_n &= \inf_{\boldsymbol{w}\in\mathcal{W}} \widetilde{R}(\boldsymbol{w}) \\
&= \inf_{\boldsymbol{w}\in\mathcal{W}} E\left[\left\{\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\right\}^2\right] \\
&\le E\left[\left\{\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1,0}\right\}^2\right] \\
&= E\left\{\left[(\boldsymbol{x}^{(0)}_{n_0+1})^T(\boldsymbol{\beta}^{(0)} - \widetilde{\boldsymbol{\beta}}^{(0)}) + \left\{\sum_{l=1}^{q_0} g_l^{(0)}(z^{(0)}_{n_0+1,l}) - \sum_{l=1}^{q_0}(B_l^{(0)}(z^{(0)}_{n_0+1,l}))^T\widetilde{\boldsymbol{\gamma}}_l^{(0)}\right\}\right]^2\right\} \\
&= E\left[\left\{g^{(0)}(\boldsymbol{z}^{(0)}_{n_0+1}) - \widetilde{g}^{(0)}(\boldsymbol{z}^{(0)}_{n_0+1})\right\}^2\right] \\
&= O\{(\overline{v}^{(0)})^{-2\kappa}\}. \qquad (6)
\end{aligned}$$

Hence, Condition 8 is violated based on Condition 1.

A.2 Proof of Asymptotic Unbiasedness

Proof Recall the previous notations as follows.

$$\widehat{\mu}^{(0)}_{i,[\mathcal{G}_j^c]}(\boldsymbol{w}) = \sum_{m=0}^{M} w_m \widehat{\mu}^{(0)}_{i,m,[\mathcal{G}_j^c]},\ i\in\mathcal{G}_j,\ j=1,\ldots,J,$$

$$\widehat{\mu}^{(0)}_{n_0+1}(\boldsymbol{w}) = \sum_{m=0}^{M} w_m \widehat{\mu}^{(0)}_{n_0+1,m}.$$

Based on the definition of $CV(\boldsymbol{w})$, we have the following decompositions

$$E\{CV(\boldsymbol{w})\} = E\left[\left\{y_i^{(0)} - \widehat{\mu}^{(0)}_{i,[\mathcal{G}_j^c]}(\boldsymbol{w})\right\}^2\right]$$

$$= E\left[\{y_i^{(0)} - \mu_i^{(0)}\}^2\right] + E\left[\left\{\mu_i^{(0)} - \widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})\right\}^2\right]$$

and

$$\begin{aligned}\mathrm{PR}(\boldsymbol{w}) &= E\left[\left\{y_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2\right] \\ &= E\left[\left\{y_{n_0+1}^{(0)} - \mu_{n_0+1}^{(0)}\right\}^2\right] + E\left[\left\{\mu_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2\right].\end{aligned}$$

Comparing the above formulas, since the samples from the target population are independent with identical distribution, we have $E[\{y_i^{(0)} - \mu_i^{(0)}\}^2] = E[\{y_{n_0+1}^{(0)} - \mu_{n_0+1}^{(0)}\}^2]$ and $E[\{\mu_i^{(0)} - \widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})\}^2] = E[\{\mu_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})\}^2]$. Note that $\widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})$ and $\widehat{\mu}_{n_0+1,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})$ are weighted averaging predictions with the same definition except that the latter uses $n_0 - n_0/J$ observations to estimate parameters $\boldsymbol{\theta}_m^{(0)}$ instead of all samples. Since n_0 and $n_0 - n_0/J$ have the same order for any $J \in \{2, \ldots, n_0\}$ as $\underline{n} \to \infty$, $\widehat{\boldsymbol{\theta}}_m^{(0)}$ and $\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}$ have the same limiting values. Therefore, we have $\widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) - \widehat{\mu}_{n_0+1,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w}) \to 0$ in probability, and then $E(CV(\boldsymbol{w})) = \mathrm{PR}(\boldsymbol{w}) + o(1)$ for any $\boldsymbol{w}$. This completes the proof.

■

A.3 Proof of Theorem 1

First we list the following lemma without proof, which is also provided in Lemma 1 in Zhang (2010), Lemma 1 in Gao et al. (2019), and Lemma 1 in Zhang and Liu (2023), to derive the asymptotic optimality theory for completeness.

Lemma 4 *Let*

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}}\{R(\boldsymbol{w}) + a_n(\boldsymbol{w}) + b_n\}.$$

If

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|a_n(\boldsymbol{w})|}{\widetilde{R}(\boldsymbol{w})} = o_p(1)$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|R(\boldsymbol{w}) - \widetilde{R}(\boldsymbol{w})|}{\widetilde{R}(\boldsymbol{w})} = o_p(1),$$

and there exists a positive constant c so that $\lim_{\underline{n}\to\infty} \inf_{\boldsymbol{w}\in\mathcal{W}} \widetilde{R}(\boldsymbol{w}) \geq c$ almost surely, then we have

$$\frac{R(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w})} \to 1$$

in probability.

Now, we formally prove the Theorem 1.

Proof Let

$$CV^*(\boldsymbol{w}) = CV(\boldsymbol{w}) - \frac{1}{n_0}\sum_{i=1}^{n_0}(y_i^{(0)} - \mu_i^{(0)})(y_i^{(0)} + \mu_i^{(0)}).$$

Since $n_0^{-1}\sum_{i=1}^{n_0}(y_i^{(0)} - \mu_i^{(0)})(y_i^{(0)} + \mu_i^{(0)})$ is unrelated to $\boldsymbol{w}$, our weight choice criterion is equivalent to

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} CV^*(\boldsymbol{w}).$$

According to Lemma 4, Theorem 1 is valid if the following equalities hold

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|R(\boldsymbol{w}) - \widetilde{R}(\boldsymbol{w})|}{\widetilde{R}(\boldsymbol{w})} = o_p(1) \tag{7}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|CV^*(\boldsymbol{w}) - \widetilde{R}(\boldsymbol{w})|}{\widetilde{R}(\boldsymbol{w})} = o_p(1). \tag{8}$$

We first consider (7). Observe that

$$\begin{aligned}
&\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\left\{\mu_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2 - \left\{\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2\right| \\
&= \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\left\{\widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}\left\{\widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) + 2\widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) + 2\mu_{n_0+1}^{(0)}\right\}\right| \\
&= \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\left\{\sum_{m=0}^{M} w_m(\widehat{\boldsymbol{\theta}}_m^{(0)} - \widetilde{\boldsymbol{\theta}}_m^{(0)})^T\frac{\partial\widehat{\mu}_{n_0+1,m}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_m^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_m^{(0)}=\bar{\boldsymbol{\theta}}}\right\}\left\{\sum_{m=0}^{M} w_m(\widehat{\boldsymbol{\theta}}_m^{(0)} - \widetilde{\boldsymbol{\theta}}_m^{(0)})^T\frac{\partial\widehat{\mu}_{n_0+1,m}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_m^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_m^{(0)}=\bar{\boldsymbol{\theta}}}\right.\right. \\
&\qquad\left.\left. + 2\sum_{m=0}^{M} w_m\widetilde{\mu}_{n_0+1,m}^{(0)} + 2\mu_{n_0+1}^{(0)}\right\}\right| \\
&\le \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left\{\sum_{m=0}^{M} w_m\left\|\widehat{\boldsymbol{\theta}}_m^{(0)} - \widetilde{\boldsymbol{\theta}}_m^{(0)}\right\|\left\|\frac{\partial\widehat{\mu}_{n_0+1,m}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_m^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_m^{(0)}=\bar{\boldsymbol{\theta}}}\right\|\right\}\left\{\sum_{m=0}^{M} w_m\left\|\widehat{\boldsymbol{\theta}}_m^{(0)} - \widetilde{\boldsymbol{\theta}}_m^{(0)}\right\|\right. \\
&\qquad\left.\times\left\|\frac{\partial\widehat{\mu}_{n_0+1,m}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_m^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_m^{(0)}=\bar{\boldsymbol{\theta}}}\right\| + 2\sum_{m=0}^{M} w_m|\widetilde{\mu}_{n_0+1,m}^{(0)}| + 2|\mu_{n_0+1}^{(0)}|\right\} \\
&= \xi_n^{-1}O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M) \\
&= o_p(1),
\end{aligned} \tag{9}$$

where the second equality uses Condition 6, the third equality uses Conditions 5 and 6, and the last equality uses Condition 8. Therefore, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{|R(\boldsymbol{w}) - \widetilde{R}(\boldsymbol{w})|}{\widetilde{R}(\boldsymbol{w})} \\
&\le \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}|R(\boldsymbol{w}) - \widetilde{R}(\boldsymbol{w})| \\
&= \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|E\left[\left\{\mu_{n_0+1}^{(0)} - \widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2 - \left\{\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2\right]\right|
\end{aligned}$$

$$\leq E\left[\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\left\{\mu_{n_0+1}^{(0)}-\widehat{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2-\left\{\mu_{n_0+1}^{(0)}-\widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2\right|\right]$$
$$=o(1),$$

where the last equality is based on Condition 7 and (9). Hence, we obtain (7).

Next, we consider (8). Similar to the derivation of (9), we have

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left[\left\{y_i^{(0)}-\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})\right\}^2-\left\{y_i^{(0)}-\widetilde{\mu}_i^{(0)}(\boldsymbol{w})\right\}^2\right]\right|$$
$$=\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left\{\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})-\widetilde{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})\right\}\left\{\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})-\widetilde{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})+2\widetilde{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})+2y_i^{(0)}\right\}\right|$$
$$=\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left\{\sum_{m=0}^{M}w_m(\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}-\widetilde{\boldsymbol{\theta}}_m^{(0)})^T\frac{\partial\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}=\bar{\boldsymbol{\theta}}}\right\}\left\{\sum_{m=0}^{M}w_m\right.\right.$$
$$\left.\left.\times(\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}-\widetilde{\boldsymbol{\theta}}_m^{(0)})^T\frac{\partial\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}=\bar{\boldsymbol{\theta}}}+2\sum_{m=0}^{M}w_m\widetilde{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}+2\mu_i^{(0)}+2\varepsilon_i^{(0)}\right\}\right|$$
$$\leq\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left\{\sum_{m=0}^{M}w_m\left\|\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}-\widetilde{\boldsymbol{\theta}}_m^{(0)}\right\|\left\|\frac{\partial\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}=\bar{\boldsymbol{\theta}}}\right\|\right\}\left\{\sum_{m=0}^{M}w_m\right.$$
$$\left.\times\left\|\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}-\widetilde{\boldsymbol{\theta}}_m^{(0)}\right\|\left\|\frac{\partial\widehat{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}}{\partial\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}}\Big|_{\widehat{\boldsymbol{\theta}}_{m,[\mathcal{G}_j^c]}^{(0)}=\bar{\boldsymbol{\theta}}}\right\|+2\sum_{m=0}^{M}w_m|\widetilde{\mu}_{i,m,[\mathcal{G}_j^c]}^{(0)}|+2|\mu_i^{(0)}|+2|\varepsilon_i^{(0)}|\right\}$$
$$=O_p(\overline{p}\underline{n}^{-1/2}M^{1/2}+\overline{p}^2\underline{n}^{-1}M). \tag{10}$$

Observe that

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{\left|CV^*(\boldsymbol{w})-\widetilde{R}(\boldsymbol{w})\right|}{\widetilde{R}(\boldsymbol{w})}$$
$$\leq\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|CV^*(\boldsymbol{w})-\widetilde{R}(\boldsymbol{w})\right|$$
$$=\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left[\left\{y_i^{(0)}-\widehat{\mu}_{i,[\mathcal{G}_j^c]}^{(0)}(\boldsymbol{w})\right\}^2-(y_i^{(0)}-\mu_i^{(0)})(y_i^{(0)}+\mu_i^{(0)})\right]\right.$$
$$\left.-E\left[\left\{\mu_{n_0+1}^{(0)}-\widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w})\right\}^2\right]\right|$$
$$\leq\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\left[\left\{y_i^{(0)}-\widetilde{\mu}_i^{(0)}(\boldsymbol{w})\right\}^2-(y_i^{(0)}-\mu_i^{(0)})(y_i^{(0)}+\mu_i^{(0)})\right]-E\left[\left\{\mu_{n_0+1}^{(0)}\right.\right.\right.$$

$$
\begin{aligned}
&- \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\Big\}^2\Big]\Big| + \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left[\Big\{y_i^{(0)} - \widehat{\mu}^{(0)}_{i,[\mathcal{G}_j^c]}(\boldsymbol{w})\Big\}^2 - \Big\{y_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2\right]\right| \\
&= \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\left[\Big\{y_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2 - (y_i^{(0)} - \mu_i^{(0)})(y_i^{(0)} + \mu_i^{(0)})\right] - E\Big[\Big\{\mu^{(0)}_{n_0+1}\right. \\
&\qquad\qquad \left.- \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\Big\}^2\Big]\right| + \xi_n^{-1}O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M) \\
&\le \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\left[\Big\{y_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2 - (y_i^{(0)} - \mu_i^{(0)})(y_i^{(0)} + \mu_i^{(0)}) - \Big\{\mu_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2\right]\right| \\
&\quad + \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\Big\{\mu_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2 - E\left[\Big\{\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\Big\}^2\right]\right| \\
&\quad + \xi_n^{-1}O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M) \\
&= \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\Big\{\mu_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2 - E\left[\Big\{\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\Big\}^2\right]\right| \\
&\quad + \xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\Big\{2\widetilde{\mu}^{(0)}_i(\boldsymbol{w})(y_i^{(0)} - \mu_i^{(0)})\Big\}\right| \\
&\quad + \xi_n^{-1}O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M), \qquad (11)
\end{aligned}
$$

where the second equality is based on (10). Hence, to obtain (8), it suffices to prove

$$
\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\Big\{\mu_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2 - E\left[\Big\{\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\Big\}^2\right]\right| = o_p(1) \qquad (12)
$$

and

$$
\xi_n^{-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\Big\{\widetilde{\mu}^{(0)}_i(\boldsymbol{w})(y_i^{(0)} - \mu_i^{(0)})\Big\}\right| = o_p(1). \qquad (13)
$$

To prove (12), observe that

$$
\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\Big\{\mu_i^{(0)} - \widetilde{\mu}^{(0)}_i(\boldsymbol{w})\Big\}^2 - E\left[\Big\{\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1}(\boldsymbol{w})\Big\}^2\right]\right| \\
&= \sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\left\{\mu_i^{(0)} - \sum_{m=0}^{M} w_m\widetilde{\mu}^{(0)}_{i,m}\right\}^2 - E\left[\left\{\mu^{(0)}_{n_0+1} - \sum_{m=0}^{M} w_m\widetilde{\mu}^{(0)}_{n_0+1,m}\right\}^2\right]\right| \\
&= \sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\left\{\sum_{m=0}^{M} w_m(\mu_i^{(0)} - \widetilde{\mu}^{(0)}_{i,m})\right\}^2 - E\left[\left\{\sum_{m=0}^{M} w_m(\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1,m})\right\}^2\right]\right| \\
&\le \sup_{\boldsymbol{w}\in\mathcal{W}}\sum_{m=0}^{M}\sum_{m'=0}^{M} w_m w_{m'}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}(\mu_i^{(0)} - \widetilde{\mu}^{(0)}_{i,m})(\mu_i^{(0)} - \widetilde{\mu}^{(0)}_{i,m'}) - E\Big\{(\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1,m})\right.
\end{aligned}
$$

$$\times (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m'}^{(0)})\Big\}\Big|$$

$$= n_0^{-1/2} \sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{m=0}^{M}\sum_{m'=0}^{M} w_m w_{m'} \widehat{\eta}_{m,m'}$$

$$\leq n_0^{-1/2} \sup_{m,m'} \widehat{\eta}_{m,m'}$$

$$= n_0^{-1/2} o_p(\xi_n n_0^{1/2}), \tag{14}$$

where

$$\widehat{\eta}_{m,m'} = \left| n_0^{-1/2} \sum_{i=1}^{n_0} \left[(\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)})(\mu_i^{(0)} - \widetilde{\mu}_{i,m'}^{(0)}) - E\left\{ (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)})(\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m'}^{(0)}) \right\} \right] \right|,$$

and the last equality in (14) is due to the following (15) and (16).

From Condition 5, we have

$$\mathrm{var}\left\{ (\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)})(\mu_i^{(0)} - \widetilde{\mu}_{i,m'}^{(0)}) \right\} = O(1) \tag{15}$$

uniformly for $m, m' \in \{0, \ldots, M\}$. By Chebyshev inequality, we derive that for any $\nu > 0$,

$$\begin{aligned}
&\Pr\left\{ \xi_n^{-1} n_0^{-1/2} \sup_{m,m'} \widehat{\eta}_{m,m'} > \nu \right\} \\
&\leq \sum_{m=0}^{M}\sum_{m'=0}^{M} \Pr\left\{ \xi_n^{-1} n_0^{-1/2} \widehat{\eta}_{m,m'} > \nu \right\} \\
&\leq \xi_n^{-2} n_0^{-1} \nu^{-2} \sum_{m=0}^{M}\sum_{m'=0}^{M} \mathrm{var}\left\{ (\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)})(\mu_i^{(0)} - \widetilde{\mu}_{i,m'}^{(0)}) \right\} \\
&= O(\xi_n^{-2} n_0^{-1} M^2) = o(1).
\end{aligned} \tag{16}$$

Together with Condition 8, we have $\sup_{m,m'} \widehat{\eta}_{m,m'} = o_p(\xi_n n_0^{1/2})$, and then obtain (12).

Similar to the derivation of (12), from Condition 5, we have

$$\mathrm{var}\left\{ \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\} = O(1) \tag{17}$$

uniformly for $m = 0, \ldots, M$. Further, for any $\nu > 0$,

$$\begin{aligned}
&\Pr\left\{ \xi_n^{-1} \sup_m \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| > \nu \right\} \\
&\leq \sum_{m=0}^{M} \Pr\left\{ \xi_n^{-1} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| > \nu \right\} \\
&\leq \xi_n^{-2} n_0^{-1} \nu^{-2} \sum_{m=0}^{M} \mathrm{var}\left\{ \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\}
\end{aligned}$$

$$= O(\xi_n^{-2} n_0^{-1} M) = o(1). \tag{18}$$

Therefore, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_i^{(0)}(\boldsymbol{w})(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| \\
&= \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \sum_{m=0}^{M} w_m \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| \\
&= \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \sum_{m=0}^{M} w_m \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| \\
&\leq \sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{m=0}^{M} w_m \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_{i,m}^{(0)}(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| \\
&\leq \sup_{m} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_{i,m}^{(0)} y_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)} \mu_i^{(0)} \right\} \right| \\
&= o_p(\xi_n),
\end{aligned}$$

where the last equality is based on (17) and (18). Then (13) is obtained. This completes the proof of Theorem 1. ■

A.4 Proof of Theorem 2

Proof Consider that Theorem 2 trivially holds if $\mathcal{I}^c$ is empty, thus we just need to discuss the case that $\mathcal{I}^c$ is not empty. Similar to the derivation of (14), for any constant $\nu > 0$,

$$\begin{aligned}
&\Pr\left\{ \frac{\sqrt{n_0}}{M} \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \mu_i^{(0)} - \widetilde{\mu}_i^{(0)}(\boldsymbol{w}) \right\}^2 - E\left[\left\{ \mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) \right\}^2 \right] \right| > \nu \right\} \\
&= \Pr\left\{ \frac{\sqrt{n_0}}{M} \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \mu_i^{(0)} - \sum_{m=0}^{M} w_m \widetilde{\mu}_{i,m}^{(0)} \right\}^2 - E\left[\left\{ \mu_{n_0+1}^{(0)} - \sum_{m=0}^{M} w_m \widetilde{\mu}_{n_0+1,m}^{(0)} \right\}^2 \right] \right| > \nu \right\} \\
&= \Pr\left\{ \frac{\sqrt{n_0}}{M} \sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \sum_{m=0}^{M} w_m (\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)}) \right\}^2 - E\left[\left\{ \sum_{m=0}^{M} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\}^2 \right] \right| > \nu \right\} \\
&\leq \Pr\left\{ \frac{\sqrt{n_0}}{M} \sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{m=0}^{M} \sum_{m'=0}^{M} w_m w_{m'} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} (\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)})(\mu_i^{(0)} - \widetilde{\mu}_{i,m'}^{(0)}) - E\left\{ (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right.\right.\right. \\
&\qquad\qquad \left.\left.\left. \times (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m'}^{(0)}) \right\} \right| > \nu \right\} \\
&\leq \sum_{m=0}^{M} \sum_{m'=0}^{M} \Pr\left\{ \left| \frac{1}{n_0} \sum_{i=1}^{n_0} (\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)})(\mu_i^{(0)} - \widetilde{\mu}_{i,m'}^{(0)}) - E\left\{ (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right.\right.\right.
\end{aligned}$$

$$\times (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m'}^{(0)}) \Big\} \Big| > M n_0^{-1/2} \nu \Big\}$$

$$\leq M^{-2}\nu^{-2} \sum_{m=0}^{M} \sum_{m'=0}^{M} \mathrm{var} \left\{ (\mu_i^{(0)} - \widetilde{\mu}_{i,m}^{(0)})(\mu_i^{(0)} - \widetilde{\mu}_{i,m'}^{(0)}) \right\}$$

$$= O(\nu^{-2}),$$

where the second inequality uses Boole's inequality, the third inequality uses Chebyshev's inequality, and the last equality is based on (15). Therefore, it follows that

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \mu_i^{(0)} - \widetilde{\mu}_i^{(0)}(\boldsymbol{w}) \right\}^2 - E\left[\left\{ \mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) \right\}^2 \right] \right| = O_p(n_0^{-1/2} M). \tag{19}$$

Similarly,

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \widetilde{\mu}_i^{(0)}(\boldsymbol{w})(y_i^{(0)} - \mu_i^{(0)}) \right\} \right| = O_p(n_0^{-1/2} M^{1/2}). \tag{20}$$

Hence, combining (19), (20), and (11), we have

$$CV^*(\boldsymbol{w}) = \widetilde{R}(\boldsymbol{w}) + O_p(\overline{p}\underline{n}^{-1/2} M^{1/2} + \overline{p}^2 \underline{n}^{-1} M). \tag{21}$$

Let $\boldsymbol{\vartheta}$ be a weight vector with $\vartheta_m = 0$ for $m \in \mathcal{I}$ and $\vartheta_m = w_m/(1-\tau)$ for $m \in \mathcal{I}^c$, where $\tau = \sum_{m\in\mathcal{I}} w_m$. According to Lemma 7 and Lemma 8 in Stone (1986), we have $\|g^{(0)} - \widetilde{g}^{(0)}\|_\infty = O\{(\overline{v}^{(0)})^{1/2-\kappa}\}$, where $\|f\|_\infty = \sup_{0\leq x\leq 1}|f(x)|$ denotes the supnorm of the function f on $[0,1]$. Then using $\boldsymbol{\vartheta}$, we have

$$\begin{aligned}
\widetilde{R}(\boldsymbol{w}) &= E\left[\left\{ \mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1}^{(0)}(\boldsymbol{w}) \right\}^2 \right] \\
&= E\left[\left\{ \sum_{m=0}^{M} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\}^2 \right] \\
&= E\left[\left\{ \sum_{m\in\mathcal{I}^c} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) + \sum_{m\in\mathcal{I}} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\}^2 \right] \\
&= E\left[\left\{ \sum_{m\in\mathcal{I}^c} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\}^2 \right] + E\left[\left\{ \sum_{m\in\mathcal{I}} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\}^2 \right] \\
&\quad + E\left[2 \left\{ \sum_{m\in\mathcal{I}^c} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\} \left\{ \sum_{m\in\mathcal{I}} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\} \right] \\
&= (1-\tau)^2 E\left[\left\{ \sum_{m\in\mathcal{I}^c} (1-\tau)^{-1} w_m (\mu_{n_0+1}^{(0)} - \widetilde{\mu}_{n_0+1,m}^{(0)}) \right\}^2 \right] \\
&\quad + \tau^2 E\left[\left\{ g^{(0)}(\boldsymbol{z}_{n_0+1}^{(0)}) - \widetilde{g}^{(0)}(\boldsymbol{z}_{n_0+1}^{(0)}) \right\}^2 \right]
\end{aligned}$$

$$
\begin{aligned}
&+ E\left(2\tau\left\{\sum_{m\in\mathcal{I}^c} w_m(\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1,m})\right\}\left\{g^{(0)}(\boldsymbol{z}^{(0)}_{n_0+1}) - \widetilde{g}^{(0)}(\boldsymbol{z}^{(0)}_{n_0+1})\right\}\right)\\
&= (1-\tau)^2\widetilde{R}(\boldsymbol{\vartheta}) + O\{(\overline{v}^{(0)})^{-2\kappa}\} + O\{(\overline{v}^{(0)})^{1/2-\kappa}\}E\left\{\sum_{m\in\mathcal{I}^c} w_m(\mu^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1,m})\right\}\\
&= (1-\tau)^2\widetilde{R}(\boldsymbol{\vartheta}) + O\{(\overline{v}^{(0)})^{-2\kappa}\} + O\{(\overline{v}^{(0)})^{1/2-\kappa}\}\\
&= (1-\tau)^2\widetilde{R}(\boldsymbol{\vartheta}) + O\{(\overline{v}^{(0)})^{1/2-\kappa}\}, \qquad (22)
\end{aligned}
$$

where the fifth equality is based on the definition of informative models that have the same pseudo-true values with the target model, the sixth equality is similar to the derivation of (6), and the last but one equality uses Condition 10. Here, the notation $\widetilde{R}$ denotes the function of $\boldsymbol{\vartheta}$ with the same definition as previous. Then based on (21), (22), and Condition 9, replacing $\boldsymbol{w}$ with $\widehat{\boldsymbol{w}}$, we have

$$
CV^*(\widehat{\boldsymbol{w}}) = (1-\widehat{\tau})^2\widetilde{R}(\widehat{\boldsymbol{\vartheta}}) + O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M). \qquad (23)
$$

Note that in all the functions of $\boldsymbol{w}$ such as $\widehat{R}(\boldsymbol{w})$, we calculate expectations firstly and then plug in $\widehat{\boldsymbol{w}}$. Let $\breve{\boldsymbol{w}}$ be the weight vector with the first component one and the others zero. Then we have

$$
\begin{aligned}
CV^*(\breve{\boldsymbol{w}}) &= \widetilde{R}(\breve{\boldsymbol{w}}) + O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M)\\
&= O((v^{(0)})^{-2\kappa}) + O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M)\\
&= O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M). \qquad (24)
\end{aligned}
$$

Next, from (23), (24), and the fact that $\widehat{\boldsymbol{w}}$ minimizes $CV^*(\boldsymbol{w})$, we have

$$
\begin{aligned}
&(1-\widehat{\tau})^2\widetilde{R}(\widehat{\boldsymbol{\vartheta}}) + O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M)\\
&\le CV^*(\breve{\boldsymbol{w}})\\
&= O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M).
\end{aligned}
$$

Hence,

$$
(1-\widehat{\tau})^2 \inf_{\boldsymbol{w}\in\overline{\mathcal{W}}} \widetilde{R}(\boldsymbol{w}) \le O_p(\overline{p}\underline{n}^{-1/2}M^{1/2} + \overline{p}^2\underline{n}^{-1}M), \qquad (25)
$$

and it implies $\widehat{\tau}\to 1$ in probability based on Condition 11. This completes the proof. ■

A.5 Proof of Corollary 3

Proof To prove the corollary, we require considering the cases of the correctly specified target model and the misspecified target model. When the target model is misspecified, the result of Theorem 1 ensures that our method yields the minimum risk, and it obviously implies the conclusion.

Next, we mainly discuss the case of the correctly specified target model. When the target model is correct, the prediction of least squares estimator on target data can be written as

$$
\widehat{\mu}^{(0)}_{n_0+1} - \mu^{(0)}_{n_0+1} = (\widehat{\mu}^{(0)}_{n_0+1} - \widetilde{\mu}^{(0)}_{n_0+1}) + (\widetilde{\mu}^{(0)}_{n_0+1} - \mu^{(0)}_{n_0+1})
$$

$$= O_p(p_0 n_0^{-1/2} M^{1/2}) + O((\overline{v}^{(0)})^{-\kappa})$$
$$= O_p(p_0 n_0^{-1/2} M^{1/2}),$$

where the last equality is based on the definition of $\overline{v}^{(0)}$ and Condition 1. Then the risk of least squares estimator on target data is $O(p_0^2 n_0^{-1} M)$. In addition, the prediction of Trans-SMAP satisfies

$$\begin{aligned}
\widehat{\mu}^{(0)}_{n_0+1}(\widehat{\boldsymbol{w}}) &= \sum_{m=0}^{M} \widehat{w}_m \widehat{\mu}^{(0)}_{n_0+1,m} \\
&= \mu^{(0)}_{n_0+1} + \sum_{m=0}^{M} \widehat{w}_m (\widehat{\mu}^{(0)}_{n_0+1,m} - \mu^{(0)}_{n_0+1}) \\
&= \mu^{(0)}_{n_0+1} + \sum_{m\in\mathcal{I}} \widehat{w}_m (\widehat{\mu}^{(0)}_{n_0+1,m} - \mu^{(0)}_{n_0+1}) + \sum_{m\in\mathcal{I}^c} \widehat{w}_m (\widehat{\mu}^{(0)}_{n_0+1,m} - \mu^{(0)}_{n_0+1}) \\
&= \mu^{(0)}_{n_0+1} + \sum_{m\in\mathcal{I}} \widehat{w}_m (\widehat{\mu}^{(0)}_{n_0+1,m} - \widetilde{\mu}^{(0)}_{n_0+1,m}) + \sum_{m\in\mathcal{I}} \widehat{w}_m (\widetilde{\mu}^{(0)}_{n_0+1,m} - \mu^{(0)}_{n_0+1}) \\
&\quad + \sum_{m\in\mathcal{I}^c} \widehat{w}_m (\widehat{\mu}^{(0)}_{n_0+1,m} - \mu^{(0)}_{n_0+1}) \\
&= \mu^{(0)}_{n_0+1} + O_p(\overline{p}\underline{n}^{-1/2} M^{1/2}) + O\{(\overline{v}^{(0)})^{1/2-\kappa}\} + O_p(1-\widehat{\tau}) \\
&= \mu^{(0)}_{n_0+1} + O_p(\overline{p}\underline{n}^{-1/2} M^{1/2}) + O\{(\overline{v}^{(0)})^{1/2-\kappa}\} + O_p(\overline{p}^{1/2}\underline{n}^{-1/4} M^{1/4} + \overline{p}\underline{n}^{-1/2} M^{1/2}) \\
&\quad \times \left\{ \inf_{\boldsymbol{w}\in\overline{\mathcal{W}}} \widetilde{R}(\boldsymbol{w}) \right\}^{-1/2} \\
&= \mu^{(0)}_{n_0+1} + O_p\{\overline{p}\underline{n}^{-1/2} M^{1/2}\},
\end{aligned}$$

where the last but one equality is based on (25) and the last equality is based on Conditions 1, 9 and 12. Therefore, the risk of Trans-SMAP is $O_p(\overline{p}^2 \underline{n}^{-1} M)$. Since it obviously has $\overline{p}^2 \underline{n}^{-1} > p_0^2 n_0^{-1}$, we have $\bar{R}(\widehat{\boldsymbol{w}}) = O_p(\bar{R}_0)$ as long as $\overline{p}^2 \underline{n}^{-1} = O(p_0^2 n_0^{-1})$. This completes the proof. ■

Appendix B. Implementation Details in Numerical Experiments

B.1 Implementation Details of Different Methods

In our simulation study, we implement all the numerical experiments with R software. To implement our Trans-SMAP procedure, we apply the cubic B-splines to approximate additive functions, set $r_l^{(m)} = 3$ for all spline estimators, and specify the number of knots through the argument "`df`" in the R function "`bs`". Here, we set `df` $= 3$ for each spline estimator in $M+1$ models for simplicity and efficiency. Note that the number of knots can also be properly determined by criteria such as cross-validation. Since the estimation accuracy of nonparametric components is not our goal, to reduce the computational complexity, we do not focus on selecting the number of knots and simply adopt a fixed setting in the simulation study.

The optimization of our weight criterion can be formulated as a constrained quadratic programming problem, which can be efficiently solved by the existing function "`solve.QP`" in the R software package "`quadprog`". Specifically, let $\mathbf{Q} = n_0^{-1}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\{(y_i^{(0)}\mathbf{1} - \widehat{\boldsymbol{y}}_{i,[\mathcal{G}_j^c]})(y_i^{(0)}\mathbf{1} - \widehat{\boldsymbol{y}}_{i,[\mathcal{G}_j^c]})^T\}$, where $\widehat{\boldsymbol{y}}_{i,[\mathcal{G}_j^c]} = (\widehat{\mu}^{(0)}_{i,0,[\mathcal{G}_j^c]},\ldots,\widehat{\mu}^{(0)}_{i,M,[\mathcal{G}_j^c]})^T$ and $\mathbf{1}$ is an $(M+1)\times 1$ column vector with ones. Further, with $\sum_{m=0}^{M} w_m = 1$, we have

$$\begin{aligned}
CV(\boldsymbol{w}) &= \frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left\{y_i^{(0)} - \widehat{\mu}^{(0)}_{i,[\mathcal{G}_j^c]}(\boldsymbol{w})\right\}^2 \\
&= \frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left\{\boldsymbol{w}^T(y_i^{(0)}\mathbf{1} - \widehat{\boldsymbol{y}}_{i,[\mathcal{G}_j^c]})\right\}^2 \\
&= \boldsymbol{w}^T\left[\frac{1}{n_0}\sum_{j=1}^{J}\sum_{i\in\mathcal{G}_j}\left\{(y_i^{(0)}\mathbf{1} - \widehat{\boldsymbol{y}}_{i,[\mathcal{G}_j^c]})(y_i^{(0)}\mathbf{1} - \widehat{\boldsymbol{y}}_{i,[\mathcal{G}_j^c]})^T\right\}\right]\boldsymbol{w} \\
&= \boldsymbol{w}^T\mathbf{Q}\boldsymbol{w}.
\end{aligned}$$

The recent transfer learning methods, Trans-Lasso and Trans-GLM, can be easily implemented via open-source programs and the R package "`glmtrans`". In addition, we have also attempted to compare our method with the integrative analysis method for semiparametric models (Li et al., 2019) in a small experiment. We find that the integrative analysis method is less effective than any other method used in this article, which possibly due to the following reasons. First, it adopts a group lasso type penalization that needs a suitable tuning parameter selection, and the resultant biased estimators may lead to an unsatisfactory prediction. Second, the goal and framework of integrative analysis, which can be regarded as multi-task learning, differ from transfer learning. It aims to identify important predictors and estimate parameters in high-dimensional settings and has no theoretical guarantee for out-of-sample prediction, meanwhile all the models are assumed to be correctly specified and of equal concern.

B.2 Computational Complexity Analysis of the Trans-SMAP Procedure

The calculation of our algorithm is mainly concentrated in the following two stages: cross-validation (Step 3.1) and the optimization of weight criterion (Step 4). In the cross-validation step, we need to solve the parameter estimation of each model in each iteration. The computational burden of this step mainly comes from the B-spline expansion of nonparametric components and least squares estimation of equation (2). Specifically, the B-splines of degree $r_l^{(m)}$ for the lth covariate in function $g_l^{(m)}(\cdot)$ require $O((r_l^{(m)})^2)$ computation using De Boor's algorithm (De Boor, 2001; Toraichi et al., 1987). Assuming that J is a positive constant, the computational complexity of Step 3.1 is $\sum_{m=0}^{M} O(q_m(r_l^{(m)})^2 + p_m^2 n_m)$. Moreover, we can estimate the $M+1$ models in this step in parallel.

In the optimization of weight criterion step, Appendix B.1 shows that $CV(\boldsymbol{w})$ is a quadratic function of weights. We can formulate the optimization problem as a constrained quadratic programming problem. Under some weak conditions, we can use the ellipsoid or interior point method to solve the quadratic programming problem in polynomial time

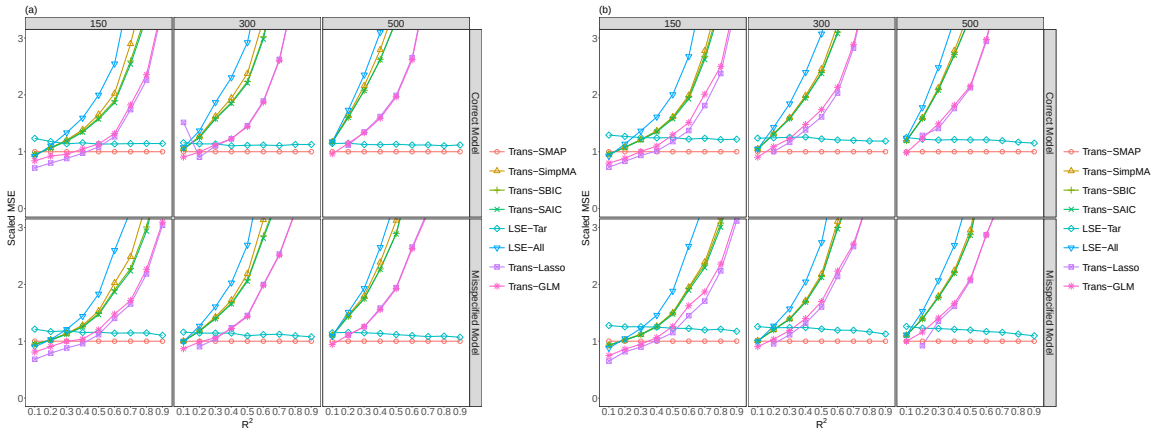

Figure 4: The scaled averaged MSE of out-of-sample prediction in homogeneous dimension settings with (a) $M = 3$ and (b) $M = 6$. Since the numerical experiments of Trans-Lasso for $n_0 \in \{300, 500\}$ and $R^2 = 0.1$ are infeasible, the corresponding results are not plotted.

(Kozlov et al., 1979). The computational complexity of solving weights in (5) is $O((M + 1)n_0^2)$. Hence, the total computational complexity of our algorithm is $\sum_{m=0}^{M} O(q_m(r_l^{(m)})^2 + p_m^2 n_m) + O((M + 1)n_0^2)$.

In summary, although our algorithm may seem complicated, the computational burden is acceptable in theory. Moreover, the computational efficiency has been validated in our numerical simulation studies. For instance, we compare several cross-validation procedures with different choices of J in Appendix C.5. The results show that even the most time-consuming procedure, leave-one-out cross-validation, takes only 3.865 seconds in a single replicate under the settings of $M = 3$ and $n_0 = 500$. Therefore, we believe that our proposed algorithm is computationally feasible for practical applications.

Appendix C. Additional Numerical Results

C.1 Supplemental Results in Homogeneous Dimension Settings

Figure 4 presents the relationship between the R^2 and the scaled MSE with respect to Trans-SMAP. It can be seen from Figure 4 that Trans-SMAP yields the smallest MSE over most of the range of R^2. Specifically, the advantage of our method over Trans-SimpMA, Trans-SBIC, Trans-SAIC, LSE-All, Trans-Lasso, and Trans-GLM becomes apparent as R^2 increases gradually. For example, when the target model is correctly specified, the gain of our method is possibly due to large weights being assigned to informative models, which can also be validated from Figure 1 in Section 4.2. Note that Trans-SMAP always performs slightly better than LSE-Tar in all of our scenarios. As the sample size increases, Trans-SMAP still dominates alternative methods for a wide range of R^2.

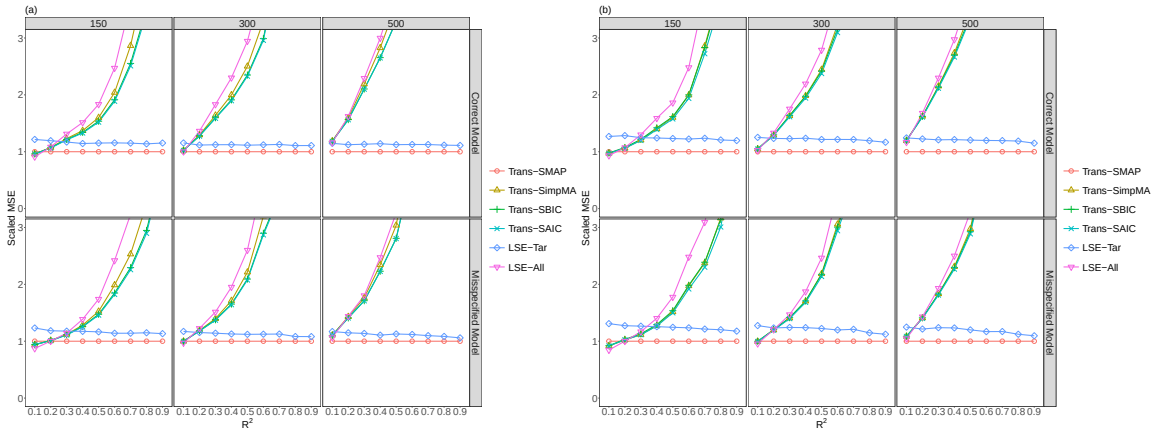

Figure 5: The scaled averaged MSE of out-of-sample prediction in heterogeneous dimension settings with (a) $M = 3$ and (b) $M = 6$. Since the numerical experiments of Trans-Lasso for $n_0 \in \{300, 500\}$ and $R^2 = 0.1$ are infeasible, the corresponding results are not plotted.

C.2 Simulation Study in Heterogeneous Dimension Settings

In this section, we design additional settings similar to that in Section 4.1 except for generating multiple data sets with heterogeneous dimensions of nonparametric parts. For $M = 3$, we set the dimensions of the nonparametric component for each model as $(q_0, q_1, q_2, q_3) = (3, 2, 2, 1)$, and the following nonlinear functions for different models are considered: $g^{(0)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2) + u_3$, $g^{(1)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_1) + u_2$, $g^{(2)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2) + u_1$, and $g^{(3)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_1) + u_1$. For $M = 6$, let the dimensions of the nonparametric variables be $(q_0, ..., q_6) = (3, 2, 2, 1, 3, 2, 2)$, and let the corresponding nonlinear functions for different models be $g^{(0)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2) + u_3$, $g^{(1)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_1) + u_2$, $g^{(2)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2) + u_1$, $g^{(3)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_1) + u_1$, $g^{(4)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \cos(\pi u_2) + u_3$, $g^{(5)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \cos(\pi u_1) + u_2$, and $g^{(6)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \cos(\pi u_2) + u_1$. All the other settings are consistent with the design in Section 4.1.

Since the frameworks of Trans-Lasso and Trans-GLM require equal dimensions of covariates, we omit these two methods in this simulation study. The MSE of our proposed Trans-SMAP and alternative methods are shown in Table 3 and Figure 5. The corresponding results are similar to that in Figure 4 and Table 1 in Section 4.2. It can be seen that our method still outperforms all the competitive methods under various simulation settings. Therefore, it reflects the flexibility and effectiveness of our approach in more practical scenarios.

C.3 Stability Analysis under Various Dissimilarities of Parameter Effects

In Section 4.1, we only design the comparison studies under fixed differences of parametric coefficients. To further demonstrate the stability of our procedure in potential negative

		Correct Target Model			Misspecified Target Model		
	Method	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$
$M = 3$	Trans-SMAP	0.026	0.013	0.008	0.035	0.021	0.016
		(0.009)	(0.005)	(0.003)	(0.011)	(0.005)	(0.003)
	Trans-SimpMA	0.238	0.224	0.219	0.218	0.206	0.199
		(0.037)	(0.029)	(0.024)	(0.034)	(0.027)	(0.021)
	Trans-SBIC	0.182	0.169	0.166	0.173	0.163	0.156
		(0.025)	(0.021)	(0.017)	(0.026)	(0.020)	(0.016)
	Trans-SAIC	0.180	0.167	0.165	0.171	0.162	0.156
		(0.025)	(0.021)	(0.017)	(0.026)	(0.019)	(0.016)
	LSE-tar	0.029	0.014	0.009	0.039	0.022	0.016
		(0.011)	(0.005)	(0.003)	(0.013)	(0.006)	(0.003)
	LSE-All	0.329	0.289	0.244	0.300	0.266	0.220
		(0.060)	(0.045)	(0.032)	(0.057)	(0.040)	(0.028)
	Uplift Rate	11.54%	7.69%	12.50%	11.43%	4.76%	0.00%
$M = 6$	Trans-SMAP	0.025	0.013	0.007	0.033	0.021	0.015
		(0.009)	(0.005)	(0.003)	(0.010)	(0.005)	(0.003)
	Trans-SimpMA	0.210	0.200	0.196	0.192	0.184	0.179
		(0.024)	(0.020)	(0.016)	(0.024)	(0.017)	(0.015)
	Trans-SBIC	0.187	0.177	0.173	0.176	0.168	0.162
		(0.021)	(0.017)	(0.015)	(0.022)	(0.015)	(0.013)
	Trans-SAIC	0.185	0.176	0.173	0.174	0.167	0.162
		(0.021)	(0.017)	(0.015)	(0.022)	(0.015)	(0.013)
	LSE-tar	0.030	0.015	0.009	0.038	0.023	0.017
		(0.011)	(0.006)	(0.003)	(0.011)	(0.006)	(0.003)
	LSE-All	0.301	0.249	0.227	0.276	0.228	0.208
		(0.043)	(0.030)	(0.022)	(0.041)	(0.026)	(0.020)
	Uplift Rate	20.00%	15.38%	28.57%	15.15%	9.52%	13.33%

Table 3: The averaged MSE of out-of-sample prediction in heterogeneous dimension settings. The standard errors are given in parenthesis.

transfer scenarios, we conduct the following simulation by varying the difference of parametric coefficients for the target model. Specifically, let the values of δ_2 for $M = 3$ and δ_3, δ_6 for $M = 6$ in coefficient vectors vary from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and all the other settings are consistent with the settings in Section 4.1. Here, we display the results of averaged MSE based on 500 replications in homoscedastic settings with the fixed level of noise, where Tables 4-5 are corresponding results associated with heterogeneous dimension settings and Tables 6-7 are for homogeneous dimension settings.

Similarly, it is clearly observed from all the tables that our proposed Trans-SMAP still yields significant improvement compared to competitive methods. In addition, we can find that the prediction accuracy of Trans-SMAP, LSE-Tar, Trans-Lasso, and Trans-GLM is insensitive to the levels of dissimilarity, while the performance of other methods gets worse as the difference increases. Specifically, the stable performance shown by Trans-SMAP, Trans-Lasso, and Trans-GLM is due to their suitable strategies of knowledge transfer, among which our method brings additional predictive benefits compared to LSE-Tar. Note that the inferior performance of Trans-Lasso and Trans-GLM compared to LSE-Tar mainly results from the semiparametric model settings. For the dissimilarity sensitive methods, such as Trans-SimpMA, Trans-SBIC, Trans-SAIC, and LSE-All, the results demonstrate that transferring information from certain sources even leads to unsatisfactory performance compared to LSE-Tar. Therefore, in a sense, the proposed Trans-SMAP has ability to avoid the negative transfer problem.

C.4 Simulation Study in Heteroscedastic Settings

In this section, we supplement additional simulation studies in heteroscedastic settings to evaluate our method comprehensively. Based on the model settings in Section 2.1, our method allows for heteroscedastic cases. For simplicity, we only consider generating data following the homogeneous settings except that the random errors of the m-th model are normally distributed with heteroscedasticity as $\varepsilon_i^{(m)} \sim N(0, 0.5(x_{i1}^{(m)})^2)$ for $m = 0, \ldots, M$.

The corresponding results of MSE are presented in Tables 8 and 9. According to the results, Trans-SMAP similarly performs the best in both the correct and misspecified target model settings. It is worth noting that the improvement of Trans-SMAP is more significant than those in homoscedastic settings based on the uplift rate in the tables.

C.5 Comparison of Various CV Criteria

To examine the impact of the choice of J in our cross-validation criterion, we analyze the performance of the 2-fold CV, 5-fold CV, 10-fold CV, and leave-one-out CV based Trans-SMAP in terms of the MSE and time consumption. For simplicity, we generate data following the homogeneous settings for $M = 3$ as an example, and the corresponding results are illustrated in Figure 6.

From Figure 6 (a) and (c), it can be seen that all the CV criteria based Trans-SMAP perform similarly better than alternative methods. However, the leave-one-out procedure takes a larger amount of time than alternative criteria as the sample size increases based on Figure 6 (b). For instance, the 5-fold CV takes 0.039 seconds that is approximately 100 times faster than 3.865 seconds of the leave-one-out CV when the target sample size $n_0 = 500$. Hence, we advocate using the J-fold CV criterion in this article instead of the

		Correct Target Model					Misspecified Target Model				
	δ_2	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
$n_0 = 150$	Trans-SMAP	0.026	0.027	0.027	0.026	0.026	0.035	0.035	0.034	0.034	0.035
		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.010)	(0.011)
	Trans-SimpMA	0.115	0.237	0.431	0.693	1.030	0.106	0.221	0.405	0.654	0.976
		(0.024)	(0.038)	(0.051)	(0.074)	(0.097)	(0.023)	(0.035)	(0.052)	(0.066)	(0.086)
	Trans-SBIC	0.054	0.183	0.414	0.744	1.191	0.054	0.175	0.398	0.731	1.152
		(0.013)	(0.028)	(0.051)	(0.089)	(0.137)	(0.013)	(0.027)	(0.052)	(0.087)	(0.129)
	Trans-SAIC	0.054	0.181	0.409	0.735	1.176	0.054	0.173	0.393	0.722	1.138
		(0.013)	(0.028)	(0.051)	(0.088)	(0.136)	(0.013)	(0.027)	(0.051)	(0.086)	(0.128)
	LSE-tar	0.029	0.031	0.030	0.031	0.030	0.039	0.039	0.038	0.039	0.039
		(0.011)	(0.012)	(0.011)	(0.012)	(0.011)	(0.012)	(0.013)	(0.012)	(0.011)	(0.012)
	LSE-All	0.153	0.326	0.601	0.976	1.445	0.140	0.303	0.564	0.914	1.386
		(0.037)	(0.059)	(0.099)	(0.167)	(0.242)	(0.035)	(0.054)	(0.100)	(0.153)	(0.237)
	Uplift Rate	11.54%	14.81%	11.11%	19.23%	15.38%	11.43%	11.43%	11.76%	14.71%	11.43%
$n_0 = 300$	Trans-SMAP	0.013	0.013	0.013	0.013	0.013	0.021	0.021	0.021	0.020	0.021
		(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
	Trans-SimpMA	0.102	0.226	0.416	0.680	1.009	0.091	0.206	0.394	0.648	0.971
		(0.017)	(0.028)	(0.043)	(0.062)	(0.080)	(0.016)	(0.026)	(0.041)	(0.059)	(0.076)
	Trans-SBIC	0.043	0.171	0.399	0.731	1.168	0.042	0.162	0.388	0.715	1.149
		(0.008)	(0.021)	(0.042)	(0.072)	(0.109)	(0.008)	(0.018)	(0.041)	(0.069)	(0.103)
	Trans-SAIC	0.042	0.170	0.396	0.725	1.159	0.042	0.161	0.385	0.709	1.140
		(0.008)	(0.021)	(0.041)	(0.071)	(0.109)	(0.008)	(0.018)	(0.041)	(0.069)	(0.103)
	LSE-tar	0.014	0.014	0.014	0.015	0.014	0.022	0.022	0.022	0.022	0.022
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
	LSE-All	0.129	0.292	0.542	0.889	1.324	0.115	0.267	0.517	0.858	1.265
		(0.024)	(0.042)	(0.072)	(0.119)	(0.174)	(0.022)	(0.041)	(0.072)	(0.114)	(0.169)
	Uplift Rate	7.69%	7.69%	7.69%	15.38%	7.69%	4.76%	4.76%	4.76%	10.00%	4.76%
$n_0 = 500$	Trans-SMAP	0.008	0.008	0.008	0.008	0.007	0.016	0.016	0.015	0.016	0.015
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	Trans-SimpMA	0.099	0.219	0.412	0.675	1.003	0.086	0.200	0.385	0.641	0.963
		(0.014)	(0.024)	(0.036)	(0.053)	(0.078)	(0.013)	(0.023)	(0.035)	(0.052)	(0.067)
	Trans-SBIC	0.039	0.165	0.394	0.727	1.156	0.038	0.157	0.380	0.709	1.136
		(0.007)	(0.018)	(0.036)	(0.064)	(0.106)	(0.006)	(0.016)	(0.037)	(0.061)	(0.090)
	Trans-SAIC	0.039	0.165	0.394	0.726	1.155	0.038	0.157	0.380	0.708	1.135
		(0.007)	(0.018)	(0.036)	(0.064)	(0.106)	(0.006)	(0.016)	(0.036)	(0.061)	(0.090)
	LSE-tar	0.009	0.009	0.009	0.009	0.008	0.017	0.017	0.016	0.016	0.016
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	LSE-All	0.116	0.244	0.446	0.723	1.055	0.101	0.224	0.417	0.684	1.024
		(0.019)	(0.032)	(0.053)	(0.083)	(0.123)	(0.017)	(0.03)	(0.052)	(0.083)	(0.121)
	Uplift Rate	12.50%	12.50%	12.50%	12.50%	14.29%	6.25%	6.25%	6.67%	0.00%	6.67%

Table 4: The averaged MSE of out-of-sample prediction in heterogeneous dimension settings for $M = 3$. The standard errors are given in parenthesis.

		Correct Target Model					Misspecified Target Model				
	δ_3, δ_6	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
$n_0 = 150$	Trans-SMAP	0.025	0.025	0.025	0.025	0.025	0.033	0.033	0.033	0.033	0.033
		(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
	Trans-SimpMA	0.076	0.209	0.439	0.757	1.171	0.069	0.195	0.410	0.721	1.126
		(0.014)	(0.026)	(0.041)	(0.063)	(0.092)	(0.014)	(0.022)	(0.040)	(0.060)	(0.084)
	Trans-SBIC	0.050	0.186	0.440	0.808	1.278	0.050	0.177	0.415	0.775	1.247
		(0.011)	(0.022)	(0.044)	(0.078)	(0.118)	(0.011)	(0.020)	(0.042)	(0.072)	(0.112)
	Trans-SAIC	0.050	0.184	0.434	0.797	1.261	0.050	0.175	0.410	0.765	1.231
		(0.011)	(0.022)	(0.043)	(0.077)	(0.117)	(0.011)	(0.020)	(0.042)	(0.071)	(0.111)
	LSE-tar	0.030	0.030	0.030	0.030	0.029	0.038	0.038	0.038	0.039	0.038
		(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
	LSE-All	0.104	0.303	0.643	1.126	1.727	0.093	0.279	0.599	1.067	1.673
		(0.020)	(0.045)	(0.083)	(0.143)	(0.227)	(0.019)	(0.040)	(0.079)	(0.146)	(0.220)
	Uplift Rate	20.00%	20.00%	20.00%	20.00%	16.00%	15.15%	15.15%	15.15%	18.18%	15.15%
$n_0 = 300$	Trans-SMAP	0.012	0.012	0.012	0.012	0.013	0.020	0.020	0.020	0.021	0.020
		(0.005)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)
	Trans-SimpMA	0.065	0.200	0.426	0.747	1.151	0.058	0.183	0.400	0.710	1.111
		(0.009)	(0.020)	(0.034)	(0.053)	(0.081)	(0.009)	(0.018)	(0.033)	(0.054)	(0.078)
	Trans-SBIC	0.040	0.177	0.427	0.795	1.266	0.040	0.167	0.408	0.762	1.237
		(0.006)	(0.017)	(0.035)	(0.063)	(0.101)	(0.006)	(0.016)	(0.036)	(0.064)	(0.100)
	Trans-SAIC	0.040	0.176	0.425	0.791	1.260	0.040	0.166	0.406	0.759	1.231
		(0.006)	(0.017)	(0.035)	(0.063)	(0.101)	(0.006)	(0.016)	(0.036)	(0.063)	(0.100)
	LSE-tar	0.014	0.014	0.014	0.014	0.015	0.022	0.023	0.022	0.023	0.023
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.005)
	LSE-All	0.082	0.249	0.529	0.921	1.431	0.073	0.228	0.494	0.881	1.386
		(0.013)	(0.029)	(0.057)	(0.096)	(0.155)	(0.012)	(0.026)	(0.056)	(0.091)	(0.148)
	Uplift Rate	16.67%	16.67%	16.67%	16.67%	15.38%	10.00%	15.00%	10.00%	9.52%	15.00%
$n_0 = 500$	Trans-SMAP	0.008	0.007	0.008	0.007	0.008	0.015	0.015	0.015	0.015	0.015
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	Trans-SimpMA	0.061	0.195	0.420	0.739	1.148	0.053	0.177	0.393	0.706	1.099
		(0.008)	(0.017)	(0.031)	(0.054)	(0.075)	(0.006)	(0.016)	(0.029)	(0.051)	(0.075)
	Trans-SBIC	0.036	0.173	0.421	0.784	1.258	0.035	0.162	0.401	0.760	1.223
		(0.005)	(0.015)	(0.032)	(0.061)	(0.090)	(0.004)	(0.014)	(0.031)	(0.058)	(0.093)
	Trans-SAIC	0.036	0.172	0.421	0.783	1.255	0.035	0.162	0.400	0.758	1.220
		(0.005)	(0.015)	(0.032)	(0.061)	(0.090)	(0.004)	(0.014)	(0.031)	(0.058)	(0.092)
	LSE-tar	0.009	0.008	0.009	0.009	0.009	0.016	0.016	0.016	0.016	0.016
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	LSE-All	0.074	0.227	0.484	0.843	1.319	0.064	0.207	0.453	0.810	1.264
		(0.010)	(0.022)	(0.043)	(0.079)	(0.123)	(0.008)	(0.021)	(0.040)	(0.073)	(0.120)
	Uplift Rate	12.50%	14.29%	12.50%	28.57%	12.50%	6.67%	6.67%	6.67%	6.67%	6.67%

Table 5: The averaged MSE of out-of-sample prediction in heterogeneous dimension settings for $M = 6$. Set the values of δ_3 and δ_6 be equal but vary from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The standard errors are given in parenthesis.

		Correct Target Model					Misspecified Target Model				
	δ_2	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
$n_0 = 150$	Trans-SMAP	0.026	0.026	0.026	0.027	0.026	0.034	0.034	0.034	0.035	0.035
		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
	Trans-SimpMA	0.116	0.238	0.427	0.694	1.026	0.104	0.218	0.404	0.660	0.976
		(0.024)	(0.037)	(0.055)	(0.073)	(0.095)	(0.022)	(0.036)	(0.051)	(0.071)	(0.094)
	Trans-SBIC	0.056	0.188	0.424	0.777	1.228	0.055	0.179	0.414	0.758	1.202
		(0.013)	(0.028)	(0.054)	(0.090)	(0.134)	(0.013)	(0.027)	(0.051)	(0.087)	(0.138)
	Trans-SAIC	0.055	0.183	0.411	0.752	1.188	0.054	0.173	0.401	0.733	1.162
		(0.013)	(0.028)	(0.052)	(0.087)	(0.130)	(0.012)	(0.027)	(0.050)	(0.085)	(0.134)
	LSE-Tar	0.030	0.031	0.030	0.030	0.030	0.038	0.037	0.038	0.038	0.039
		(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
	LSE-All	0.172	0.347	0.616	0.997	1.456	0.153	0.317	0.579	0.958	1.408
		(0.040)	(0.066)	(0.107)	(0.164)	(0.224)	(0.036)	(0.059)	(0.101)	(0.160)	(0.218)
	Trans-Lasso	0.123	0.124	0.123	0.125	0.123	0.130	0.130	0.131	0.132	0.134
		(0.014)	(0.015)	(0.014)	(0.016)	(0.013)	(0.014)	(0.014)	(0.015)	(0.015)	(0.016)
	Trans-GLM	0.125	0.125	0.124	0.126	0.124	0.133	0.133	0.133	0.134	0.134
		(0.018)	(0.018)	(0.017)	(0.019)	(0.016)	(0.018)	(0.019)	(0.019)	(0.018)	(0.019)
	Uplift Rate	15.38%	19.23%	15.38%	11.11%	15.38%	11.76%	8.82%	11.76%	8.57%	11.43%
$n_0 = 300$	Trans-SMAP	0.013	0.013	0.013	0.013	0.013	0.021	0.021	0.020	0.021	0.021
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
	Trans-SimpMA	0.103	0.224	0.415	0.685	1.013	0.092	0.204	0.388	0.646	0.966
		(0.018)	(0.028)	(0.041)	(0.058)	(0.085)	(0.016)	(0.025)	(0.039)	(0.057)	(0.080)
	Trans-SBIC	0.044	0.172	0.406	0.757	1.204	0.043	0.164	0.391	0.729	1.171
		(0.009)	(0.020)	(0.042)	(0.072)	(0.116)	(0.008)	(0.020)	(0.041)	(0.075)	(0.110)
	Trans-SAIC	0.043	0.169	0.398	0.741	1.179	0.042	0.161	0.383	0.714	1.147
		(0.008)	(0.020)	(0.042)	(0.070)	(0.114)	(0.008)	(0.020)	(0.040)	(0.074)	(0.108)
	LSE-Tar	0.014	0.015	0.014	0.014	0.014	0.022	0.022	0.022	0.023	0.022
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
	LSE-All	0.146	0.305	0.555	0.908	1.336	0.132	0.279	0.521	0.861	1.296
		(0.027)	(0.044)	(0.071)	(0.116)	(0.171)	(0.026)	(0.038)	(0.068)	(0.111)	(0.178)
	Trans-Lasso	0.109	0.110	0.109	0.109	0.110	0.117	0.117	0.117	0.118	0.117
		(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)	(0.008)
	Trans-GLM	0.109	0.109	0.109	0.108	0.109	0.117	0.116	0.116	0.117	0.116
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.009)	(0.008)
	Uplift Rate	7.69%	15.38%	7.69%	7.69%	7.69%	4.76%	4.76%	10.00%	9.52%	4.76%
$n_0 = 500$	Trans-SMAP	0.008	0.008	0.008	0.008	0.008	0.016	0.015	0.015	0.015	0.015
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	Trans-SimpMA	0.097	0.221	0.408	0.675	1.008	0.085	0.198	0.383	0.639	0.965
		(0.014)	(0.022)	(0.038)	(0.052)	(0.078)	(0.012)	(0.021)	(0.035)	(0.052)	(0.072)
	Trans-SBIC	0.039	0.165	0.392	0.728	1.168	0.038	0.156	0.379	0.710	1.137
		(0.006)	(0.016)	(0.036)	(0.065)	(0.104)	(0.006)	(0.016)	(0.033)	(0.062)	(0.094)
	Trans-SAIC	0.039	0.165	0.391	0.726	1.165	0.037	0.155	0.378	0.707	1.134
		(0.006)	(0.015)	(0.036)	(0.064)	(0.103)	(0.006)	(0.016)	(0.033)	(0.061)	(0.094)
	LSE-Tar	0.009	0.009	0.009	0.008	0.009	0.017	0.016	0.016	0.016	0.016
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	LSE-All	0.129	0.261	0.458	0.739	1.087	0.114	0.235	0.430	0.696	1.030
		(0.020)	(0.031)	(0.055)	(0.084)	(0.127)	(0.018)	(0.029)	(0.052)	(0.085)	(0.118)
	Trans-Lasso	0.104	0.104	0.104	0.104	0.104	0.112	0.112	0.112	0.111	0.112
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
	Trans-GLM	0.103	0.103	0.103	0.103	0.104	0.111	0.111	0.111	0.110	0.111
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
	Uplift Rate	12.50%	12.50%	12.50%	0.00%	12.50%	6.25%	6.67%	6.67%	6.67%	6.67%

Table 6: The averaged MSE of out-of-sample prediction in homogeneous dimension settings for $M = 3$. The standard errors are given in parenthesis.

		Correct Target Model					Misspecified Target Model				
	δ_3, δ_6	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
$n_0 = 150$	Trans-SMAP	0.026	0.025	0.025	0.025	0.025	0.034	0.034	0.034	0.032	0.034
		(0.010)	(0.009)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
	Trans-SimpMA	0.077	0.211	0.437	0.756	1.169	0.069	0.195	0.411	0.719	1.117
		(0.015)	(0.026)	(0.039)	(0.063)	(0.089)	(0.014)	(0.024)	(0.039)	(0.061)	(0.088)
	Trans-SBIC	0.052	0.192	0.447	0.824	1.316	0.051	0.184	0.432	0.794	1.269
		(0.011)	(0.023)	(0.043)	(0.076)	(0.126)	(0.012)	(0.022)	(0.045)	(0.079)	(0.119)
	Trans-SAIC	0.051	0.186	0.433	0.797	1.271	0.050	0.179	0.418	0.768	1.226
		(0.011)	(0.023)	(0.042)	(0.074)	(0.122)	(0.012)	(0.022)	(0.043)	(0.077)	(0.116)
	LSE-Tar	0.031	0.029	0.030	0.030	0.030	0.039	0.039	0.039	0.037	0.038
		(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.013)	(0.013)	(0.012)	(0.011)	(0.012)
	LSE-All	0.141	0.335	0.669	1.146	1.748	0.127	0.312	0.637	1.083	1.677
		(0.025)	(0.049)	(0.081)	(0.145)	(0.233)	(0.023)	(0.043)	(0.085)	(0.138)	(0.221)
	Trans-Lasso	0.122	0.121	0.123	0.124	0.124	0.130	0.131	0.132	0.132	0.133
		(0.013)	(0.013)	(0.014)	(0.015)	(0.015)	(0.015)	(0.014)	(0.015)	(0.014)	(0.015)
	Trans-GLM	0.125	0.122	0.125	0.126	0.125	0.134	0.133	0.133	0.133	0.133
		(0.017)	(0.015)	(0.017)	(0.018)	(0.018)	(0.019)	(0.018)	(0.018)	(0.018)	(0.017)
	Uplift Rate	19.23%	16.00%	20.00%	20.00%	20.00%	14.71%	14.71%	14.71%	15.63%	11.76%
$n_0 = 300$	Trans-SMAP	0.012	0.013	0.012	0.012	0.013	0.020	0.020	0.020	0.020	0.020
		(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)
	Trans-SimpMA	0.066	0.200	0.427	0.748	1.156	0.058	0.184	0.397	0.705	1.110
		(0.010)	(0.020)	(0.035)	(0.056)	(0.082)	(0.009)	(0.018)	(0.033)	(0.052)	(0.077)
	Trans-SBIC	0.040	0.177	0.427	0.791	1.265	0.040	0.168	0.404	0.753	1.231
		(0.007)	(0.017)	(0.036)	(0.063)	(0.103)	(0.006)	(0.016)	(0.035)	(0.058)	(0.096)
	Trans-SAIC	0.040	0.175	0.422	0.781	1.249	0.039	0.166	0.399	0.744	1.215
		(0.007)	(0.017)	(0.035)	(0.063)	(0.102)	(0.006)	(0.015)	(0.035)	(0.057)	(0.095)
	LSE-Tar	0.014	0.015	0.014	0.014	0.015	0.022	0.022	0.022	0.022	0.022
		(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)
	LSE-All	0.116	0.279	0.557	0.953	1.451	0.105	0.260	0.525	0.902	1.396
		(0.016)	(0.033)	(0.062)	(0.106)	(0.155)	(0.015)	(0.031)	(0.057)	(0.097)	(0.142)
	Trans-Lasso	0.109	0.109	0.109	0.109	0.110	0.116	0.117	0.117	0.117	0.117
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)
	Trans-GLM	0.108	0.109	0.108	0.109	0.109	0.116	0.116	0.116	0.116	0.116
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)
	Uplift Rate	16.67%	15.38%	16.67%	16.67%	15.38%	10.00%	10.00%	10.00%	10.00%	10.00%
$n_0 = 500$	Trans-SMAP	0.008	0.008	0.008	0.008	0.007	0.016	0.015	0.015	0.015	0.015
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	Trans-SimpMA	0.061	0.197	0.424	0.739	1.143	0.054	0.178	0.395	0.702	1.112
		(0.008)	(0.017)	(0.032)	(0.052)	(0.071)	(0.007)	(0.015)	(0.030)	(0.049)	(0.073)
	Trans-SBIC	0.036	0.173	0.420	0.776	1.240	0.036	0.162	0.398	0.746	1.217
		(0.005)	(0.015)	(0.032)	(0.060)	(0.085)	(0.005)	(0.014)	(0.031)	(0.054)	(0.090)
	Trans-SAIC	0.036	0.172	0.418	0.772	1.233	0.036	0.161	0.395	0.742	1.210
		(0.005)	(0.015)	(0.031)	(0.059)	(0.085)	(0.005)	(0.014)	(0.031)	(0.053)	(0.090)
	LSE-Tar	0.009	0.009	0.009	0.009	0.009	0.017	0.016	0.016	0.016	0.016
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	LSE-All	0.106	0.260	0.518	0.873	1.334	0.097	0.239	0.485	0.835	1.299
		(0.012)	(0.027)	(0.050)	(0.079)	(0.118)	(0.011)	(0.023)	(0.044)	(0.073)	(0.112)
	Trans-Lasso	0.103	0.104	0.104	0.104	0.104	0.111	0.112	0.111	0.112	0.112
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)
	Trans-GLM	0.103	0.103	0.103	0.103	0.103	0.111	0.111	0.111	0.111	0.111
		(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
	Uplift Rate	12.50%	12.50%	12.50%	12.50%	28.57%	6.25%	6.67%	6.67%	6.67%	6.67%

Table 7: The averaged MSE of out-of-sample prediction in homogeneous dimension settings for $M = 6$. Set the values of δ_3 and δ_6 be equal but vary from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The standard errors are given in parenthesis.

		Correct Target Model			Misspecified Target Model		
	Method	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$
$M = 3$	Trans-SMAP	0.086	0.043	0.027	0.091	0.051	0.035
		(0.051)	(0.020)	(0.014)	(0.043)	(0.021)	(0.016)
	Trans-SimpMA	0.282	0.246	0.234	0.261	0.228	0.212
		(0.063)	(0.038)	(0.033)	(0.059)	(0.039)	(0.030)
	Trans-SBIC	0.243	0.205	0.195	0.228	0.195	0.180
		(0.075)	(0.050)	(0.039)	(0.073)	(0.047)	(0.038)
	Trans-SAIC	0.239	0.202	0.194	0.224	0.192	0.179
		(0.074)	(0.050)	(0.038)	(0.072)	(0.047)	(0.038)
	LSE-tar	0.104	0.049	0.031	0.113	0.058	0.040
		(0.074)	(0.026)	(0.018)	(0.063)	(0.028)	(0.021)
	LSE-All	0.368	0.311	0.259	0.336	0.287	0.234
		(0.083)	(0.052)	(0.040)	(0.076)	(0.052)	(0.036)
	Uplift Rate	20.93%	13.95%	14.81%	24.18%	13.73%	14.29%
$M = 6$	Trans-SMAP	0.039	0.025	0.076	0.088	0.047	0.032
		(0.020)	(0.012)	(0.046)	(0.046)	(0.020)	(0.013)
	Trans-SimpMA	0.221	0.208	0.246	0.233	0.203	0.190
		(0.032)	(0.026)	(0.047)	(0.051)	(0.032)	(0.023)
	Trans-SBIC	0.209	0.194	0.236	0.229	0.195	0.181
		(0.041)	(0.031)	(0.059)	(0.061)	(0.041)	(0.030)
	Trans-SAIC	0.204	0.191	0.226	0.219	0.190	0.178
		(0.040)	(0.030)	(0.058)	(0.060)	(0.040)	(0.030)
	LSE-tar	0.050	0.031	0.101	0.113	0.059	0.038
		(0.028)	(0.017)	(0.069)	(0.065)	(0.027)	(0.018)
	LSE-All	0.267	0.239	0.335	0.313	0.247	0.218
		(0.040)	(0.031)	(0.064)	(0.063)	(0.038)	(0.028)
	Uplift Rate	28.21%	24.00%	32.89%	28.41%	25.53%	18.75%

Table 8: The averaged MSE of out-of-sample prediction in heterogeneous dimension and heteroscedastic settings. The standard errors are given in parenthesis.

		Correct Target Model			Misspecified Target Model		
	Method	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$	$n_0 = 150$	$n_0 = 300$	$n_0 = 500$
$M = 3$	Trans-SMAP	0.085	0.045	0.026	0.094	0.055	0.034
		(0.050)	(0.023)	(0.013)	(0.048)	(0.027)	(0.016)
	Trans-SimpMA	0.280	0.247	0.233	0.258	0.228	0.212
		(0.064)	(0.042)	(0.033)	(0.059)	(0.041)	(0.031)
	Trans-SBIC	0.242	0.210	0.193	0.228	0.199	0.180
		(0.078)	(0.052)	(0.039)	(0.072)	(0.050)	(0.038)
	Trans-SAIC	0.236	0.207	0.192	0.222	0.196	0.179
		(0.078)	(0.052)	(0.039)	(0.071)	(0.050)	(0.038)
	LSE-Tar	0.102	0.051	0.031	0.112	0.064	0.039
		(0.068)	(0.031)	(0.016)	(0.065)	(0.034)	(0.022)
	LSE-All	0.386	0.324	0.271	0.354	0.301	0.251
		(0.087)	(0.055)	(0.040)	(0.076)	(0.053)	(0.038)
	Trans-Lasso	0.175	0.136	0.121	0.185	0.147	0.129
		(0.063)	(0.029)	(0.017)	(0.062)	(0.033)	(0.019)
	Trans-GLM	0.173	0.135	0.120	0.184	0.146	0.128
		(0.057)	(0.027)	(0.017)	(0.059)	(0.032)	(0.019)
	Uplift Rate	20.00%	13.33%	19.23%	19.15%	16.36%	14.71%
$M = 6$	Trans-SMAP	0.082	0.040	0.025	0.091	0.047	0.032
		(0.052)	(0.021)	(0.011)	(0.053)	(0.022)	(0.012)
	Trans-SimpMA	0.255	0.218	0.208	0.236	0.203	0.189
		(0.057)	(0.032)	(0.025)	(0.050)	(0.029)	(0.022)
	Trans-SBIC	0.237	0.199	0.190	0.227	0.191	0.175
		(0.066)	(0.040)	(0.030)	(0.058)	(0.036)	(0.027)
	Trans-SAIC	0.232	0.197	0.189	0.222	0.189	0.174
		(0.065)	(0.040)	(0.030)	(0.058)	(0.036)	(0.027)
	LSE-Tar	0.107	0.051	0.033	0.117	0.060	0.039
		(0.074)	(0.029)	(0.017)	(0.074)	(0.032)	(0.017)
	LSE-All	0.377	0.297	0.269	0.351	0.278	0.247
		(0.075)	(0.041)	(0.031)	(0.067)	(0.040)	(0.029)
	Trans-Lasso	0.174	0.134	0.122	0.186	0.143	0.129
		(0.058)	(0.026)	(0.018)	(0.074)	(0.029)	(0.019)
	Trans-GLM	0.180	0.137	0.122	0.191	0.144	0.129
		(0.061)	(0.029)	(0.017)	(0.065)	(0.028)	(0.018)
	Uplift Rate	30.49%	27.50%	32.00%	28.57%	27.66%	21.88%

Table 9: The averaged MSE of out-of-sample prediction in homogeneous dimension and heteroscedastic settings. The standard errors are given in parenthesis.

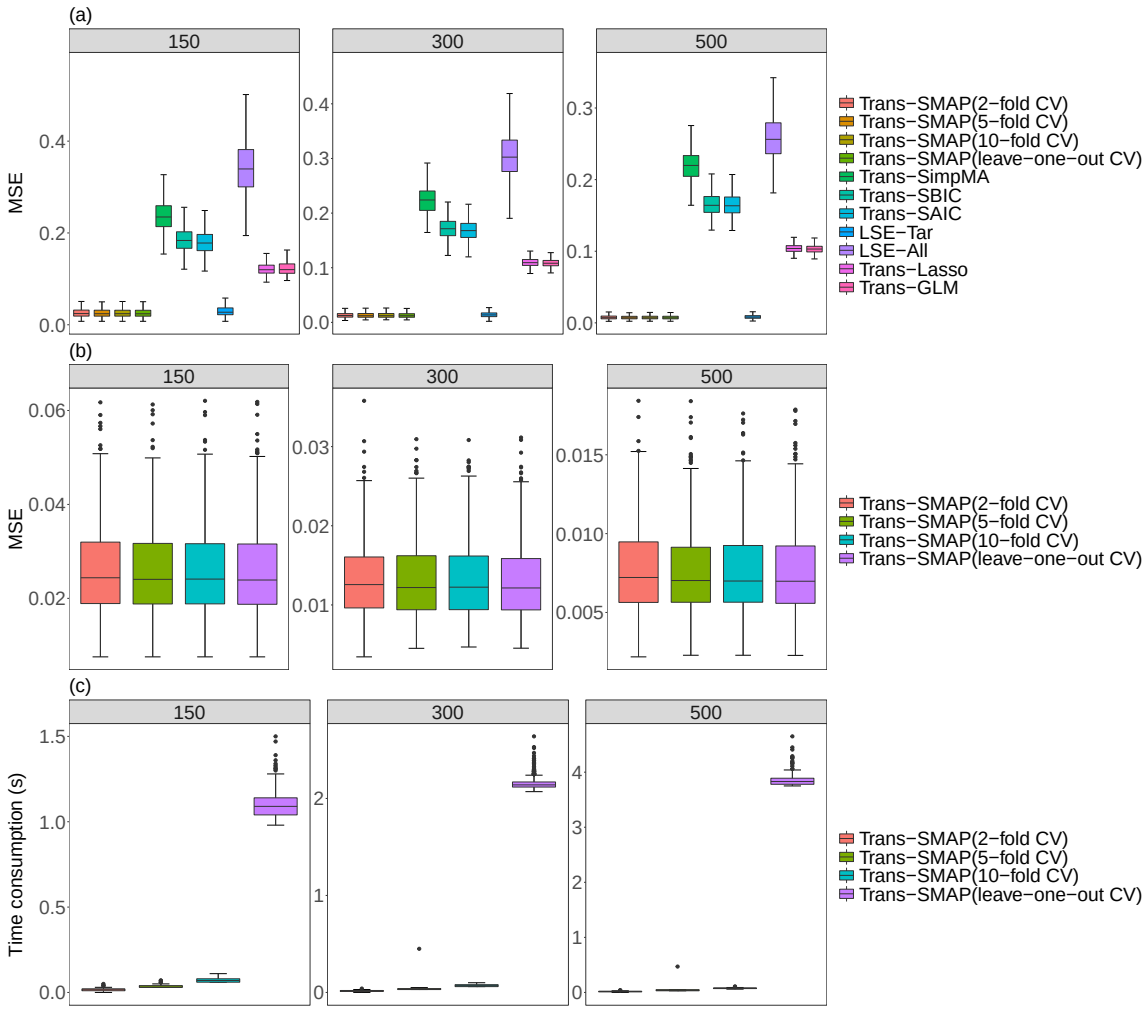

Figure 6: Comparison of various CV criteria for $M = 3$. (a) MSE of out-of-sample prediction for all the methods. (b) MSE of out-of-sample prediction for different CV criteria. (c) Time consumption of our method based on different CV criteria. The numerical computation executes on a regular PC with an intel core i7-10700 2.90 GHz CPU.

Variable	Description	Data Range
$Y^{(m)}$	the natural logarithm of the monthly rent	$[700, 110000]$
$X_1^{(m)}$	the number of rooms	$\{1, 2, 3, 4, 5, 6, 7, 9\}$
$X_2^{(m)}$	the number of restrooms	$\{0, 1, 2, 3, 4, 5, 9\}$
$X_3^{(m)}$	the number of living rooms	$\{0, 1, 2, 3, 4\}$
$X_4^{(m)}$	total area	$[17, 600]$
$X_5^{(m)}$	have or not a bed	$0 = \text{no}, 1 = \text{yes}$
$X_6^{(m)}$	have or not a wardrobe	$0 = \text{no}, 1 = \text{yes}$
$X_7^{(m)}$	have or not a air conditioner	$0 = \text{no}, 1 = \text{yes}$
$X_8^{(m)}$	have or not a fuel gas	$0 = \text{no}, 1 = \text{yes}$
$Z_1^{(m)}$	total floor	$[1, 39]$
$Z_2^{(m)}$	the number of schools within 3 km	$[0, 132]$

Table 10: Description of variables in the housing rental information data.

leave-one-out CV to reduce the computational burden. The more convictive procedure to select J adaptively with theoretical guarantees deserves deeply study in the future.

C.6 More Details of Real Data Analysis

Table 10 provides more details of covariates in our data analysis. In Figure 7, we mark all the specific locations of rental houses in the map to visualize our data source. We can see that these areas are relatively far away from downtown with relatively low rental prices, so these houses are more likely to be the first choice for young people who have just started working.

Figures 8-12 visualize the marginal relationships between the natural logarithm of the monthly rent and predictors for different target data. For the data sets of Daxing, Fangshan, Mentougou, and Shijingshan, eight covariates (the number of rooms, the number of restrooms, the number of living rooms, total area, have or not a bed, have or not a wardrobe, have or not a air conditioner, and have or not a fuel gas) are confirmed to have linear effects on the dependent variable, and two covariates (total floor and the number of schools within 3 km) have nonlinear effects. All the covariates have linear effects for the data set of Fengtai.

To be more intuitive, we display the transfer network based on weight assignments for different target domains in Figure 13 as a supplement of Table 2. In the transfer network, the nodes correspond to different data sources, and the directed edges indicate knowledge transfer from source domain to target domain.

C.7 Simulation Study in Real Data Settings

In this section, we conduct an additional simulation study to compare the performance of different methods. To mimic the real data structure, we consider $M = 5$ data sources denoted as Domain 1, ..., Domain 5, with ten covariates and sample sizes of $(291, 247, 339, 263, 269)$,

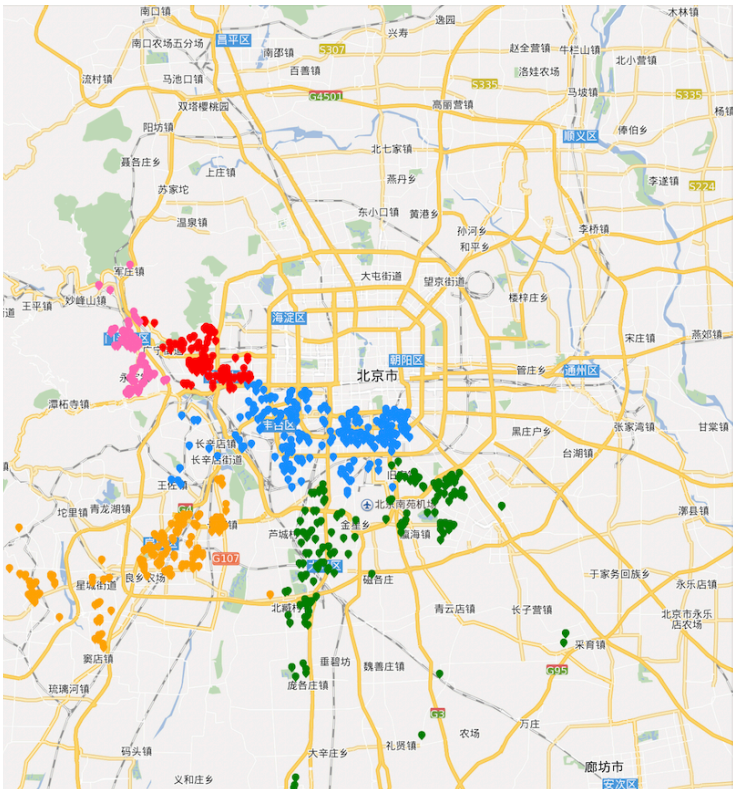

Figure 7: Visualization for the locations of rental houses in Daxing, Fangshan, Fengtai, Mentougou, and Shijingshan in the empirical data anaysis.

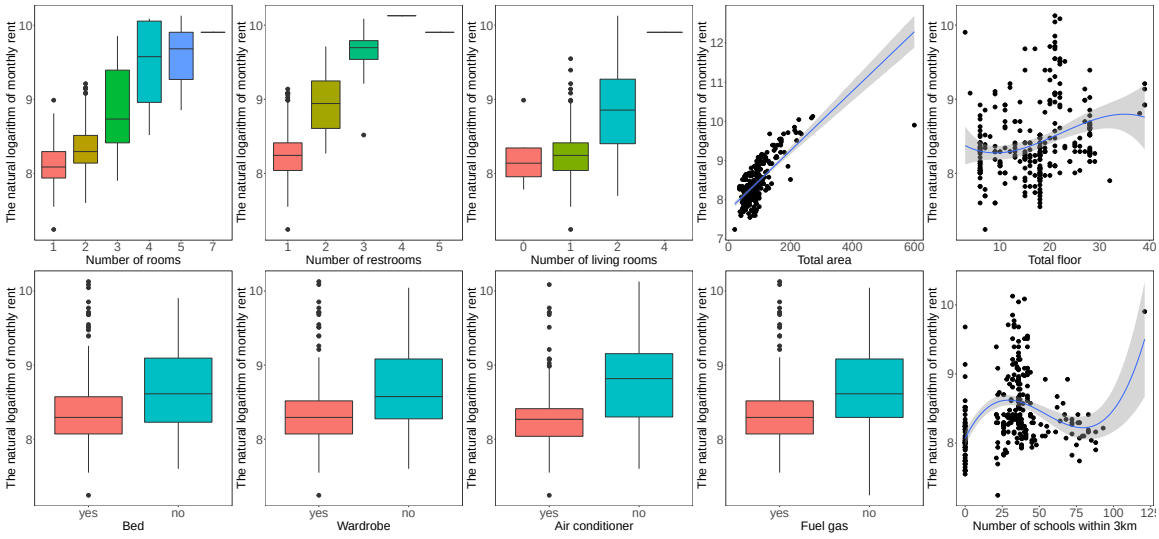

Figure 8: The marginal relationship between the natural logarithm of the monthly rent and ten predictors for the data set of Daxing.

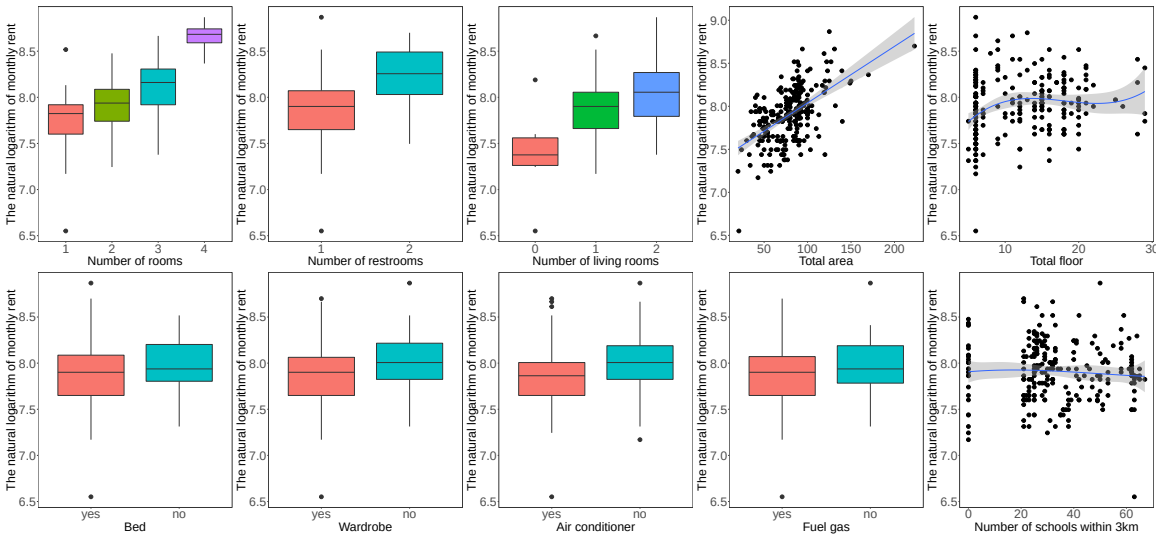

Figure 9: The marginal relationship between the natural logarithm of the monthly rent and ten predictors for the data set of Fangshan.

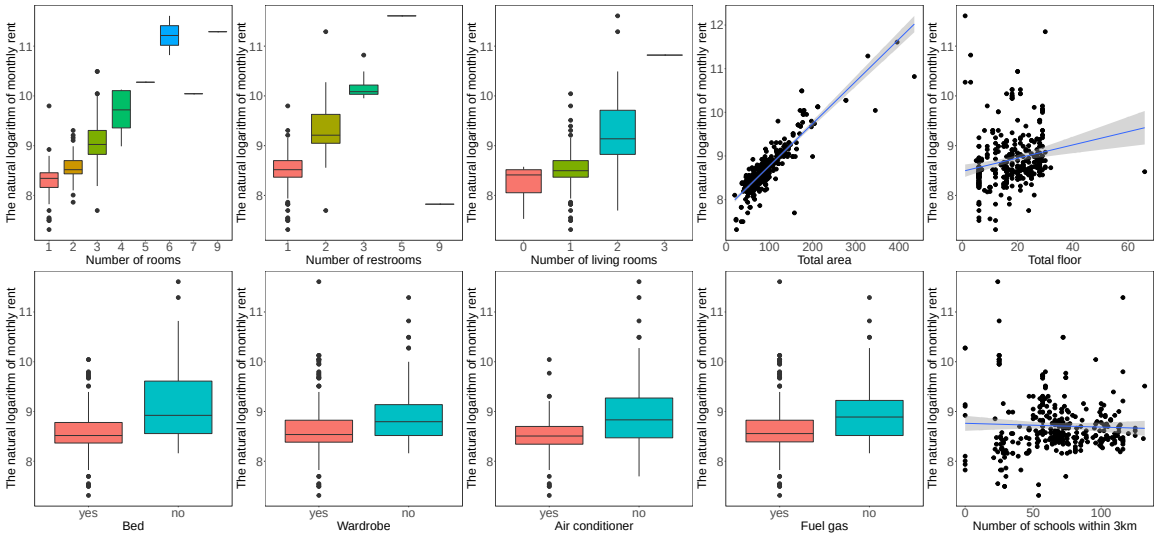

Figure 10: The marginal relationship between the natural logarithm of the monthly rent and ten predictors for the data set of Fengtai.

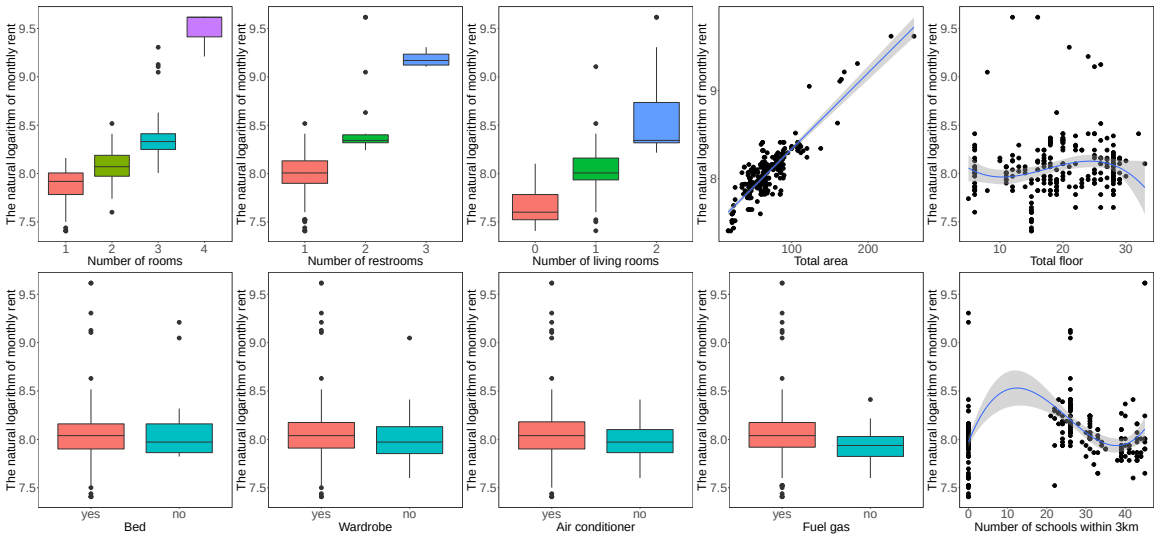

Figure 11: The marginal relationship between the natural logarithm of the monthly rent and ten predictors for the data set of Mentougou.

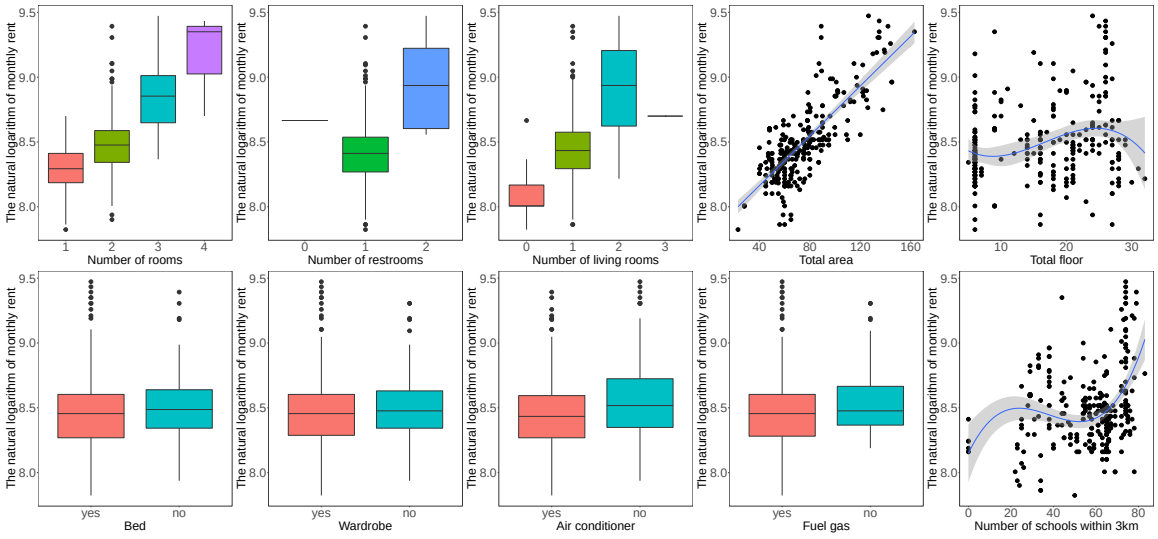

Figure 12: The marginal relationship between the natural logarithm of the monthly rent and ten predictors for the data set of Shijingshan.

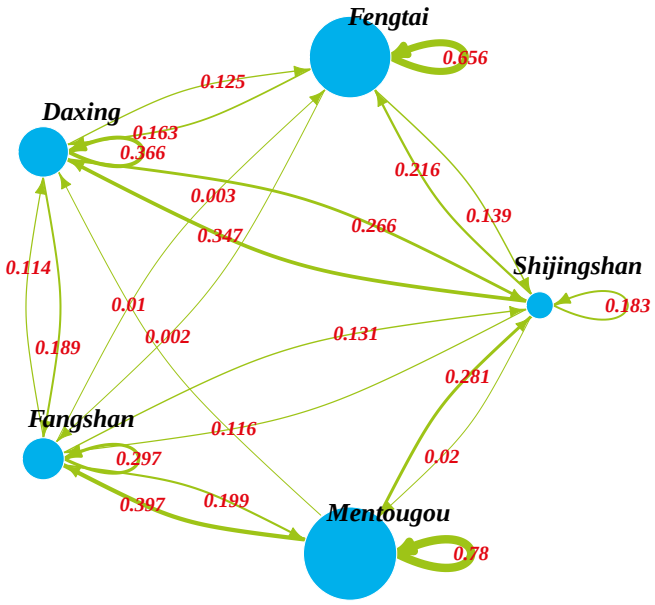

Figure 13: Transfer network based on our weight assignments for different target domains. The size of nodes is proportional to the weight of the target model, and the thickness of edges is proportional to the weights of source models.

which are the same as our real data structure. We set Domain 3 to be the data set generated from a linear regression model, while the other domains are generated from additive partial linear models. For Domain 3, the ten covariates are generated from a 10-dimensional multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with $\Sigma = [0.5^{|a-a'|}]_{10\times 10}$, and the parametric coefficient vector is set to be $\boldsymbol{\beta}^{(3)} = (1.42, -1.18, 1.02, -0.78, 0.67, 0.32, -0.48, 1.02, 0.8, 0.7)^T$. For the other domains, we set the dimensions of parametric and nonparametric components for each model are equally eight and two, respectively. The parametric components for each model are generated from a same 8-dimensional multivariate normal distribution, and the parametric coefficient vectors of different models are set to be $\boldsymbol{\beta}^{(1)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3, -0.5, 1)^T$, $\boldsymbol{\beta}^{(2)} = (1.42, -1.18, 1.02, -0.78, 0.67, 0.32, -0.48, 1.02, 0.12)^T$, $\boldsymbol{\beta}^{(4)} = (1.7, -0.9, 1.3, -0.5, 0.95, 0.6, -0.2, 1.3)^T$, and $\boldsymbol{\beta}^{(5)} = (1.4, -1.2, 1, -0.8, 0.65, 0.3, -0.5, 1)^T$. Similarly, here we consider parametric coefficient vectors may have varying degrees of similarity among different domains. We further assume that the model of Domain 2 is misspecified with the same form as that in Section 4.1. The nonparametric variables $\boldsymbol{u}$ are generated from a uniform distribution $U(0, 1)$, and we consider the following nonlinear functions for different models: $g^{(1)}(\boldsymbol{u}) = 2(u_1 - 0.5)^3 + \sin(\pi u_2)$, $g^{(2)}(\boldsymbol{u}) = 2(u_1 + 0.5)^3 + \cos(\pi u_2)$, $g^{(4)}(\boldsymbol{u}) = (1.8u_1 + 0.3)^3 + \cos(\pi u_2)$, and $g^{(5)}(\boldsymbol{u}) = 1.5(u_1 + 0.5)^3 + \cos(2\pi u_2)$. For simplicity, we only consider the homoscedastic setting and let the random error follow a normal distribution $N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon \in \{0.5, 1.5\}$. We let each data source take turns as the target domain and the other data sources as source domains, and then we evaluate different procedures in five scenarios.

	Method	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5
$\sigma_\varepsilon = 0.5$	Trans-SMAP	0.011	0.021	0.006	0.017	0.051
		(0.004)	(0.006)	(0.003)	(0.007)	(0.006)
	Trans-SimpMA	0.104	0.058	0.052	1.081	0.143
		(0.012)	(0.009)	(0.008)	(0.075)	(0.014)
	Trans-SBIC	0.105	0.059	0.053	1.077	0.145
		(0.014)	(0.010)	(0.009)	(0.079)	(0.016)
	Trans-SAIC	0.107	0.060	0.054	1.072	0.146
		(0.014)	(0.010)	(0.009)	(0.079)	(0.016)
	LSE-Tar	0.014	0.025	0.009	0.017	0.056
		(0.005)	(0.007)	(0.004)	(0.007)	(0.008)
	LSE-All	0.102	0.099	0.052	1.168	0.225
		(0.017)	(0.013)	(0.012)	(0.097)	(0.021)
	Trans-Lasso	0.111	0.241	0.009	0.829	0.649
		(0.009)	(0.017)	(0.004)	(0.052)	(0.034)
	Trans-GLM	0.112	0.241	0.009	0.826	0.657
		(0.009)	(0.017)	(0.004)	(0.051)	(0.038)
	Uplift Rate	27.27%	19.05%	50.00%	0.00%	9.80%
$\sigma_\varepsilon = 1.5$	Trans-SMAP	0.089	0.110	0.048	0.134	0.132
		(0.036)	(0.047)	(0.027)	(0.050)	(0.041)
	Trans-SimpMA	0.165	0.130	0.083	1.155	0.208
		(0.041)	(0.045)	(0.031)	(0.116)	(0.045)
	Trans-SBIC	0.163	0.130	0.082	1.161	0.206
		(0.042)	(0.045)	(0.031)	(0.117)	(0.046)
	Trans-SAIC	0.165	0.130	0.083	1.156	0.208
		(0.042)	(0.045)	(0.032)	(0.117)	(0.046)
	LSE-Tar	0.123	0.156	0.077	0.135	0.174
		(0.045)	(0.060)	(0.036)	(0.051)	(0.052)
	LSE-All	0.153	0.161	0.079	1.237	0.278
		(0.042)	(0.046)	(0.030)	(0.131)	(0.048)
	Trans-Lasso	0.186	0.334	0.075	0.916	0.736
		(0.042)	(0.057)	(0.036)	(0.076)	(0.066)
	Trans-GLM	0.213	0.351	0.095	0.920	0.771
		(0.059)	(0.065)	(0.057)	(0.080)	(0.088)
	Uplift Rate	38.20%	18.18%	56.25%	0.75%	31.82%

Table 11: The averaged MSE of out-of-sample prediction for different target domains in real data settings. The standard errors are given in parenthesis.

To evaluate the prediction performance, we similarly generate 500 testing samples from the target model and calculate the corresponding prediction MSE based on 500 replications. The results are summarized in Table 11. From the results, we can observe that our Trans-SMAP outperforms all competitive methods in most cases, especially for Domain 1 and Domain 3. When Domain 4 is the target domain, our procedure has similar performance to the best alternative method, LSE-Tar, which reflects the influence of parameter similarity between different models on the improvement of parameter-transfer approach. Overall, the proposed Trans-SMAP remains effective in the simulation experiments with the real data structure.

C.8 Simulation Study in High-dimensional Settings

To consider a relatively fair comparison with Trans-Lasso, we conduct a high-dimensional simulation study. Since our framework can not be directly applied to high-dimensional data, we simply replace the least squares estimation (2) in the original step of Algorithm 1 with the following Lasso estimation

$$\widehat{\boldsymbol{\beta}}^{(m)} = \arg\min_{\boldsymbol{\beta}^{(m)}} \left\{ \frac{1}{2n_m} \sum_{i=1}^{n_m} \{y_i^{(m)} - (\boldsymbol{x}_i^{(m)})^T \boldsymbol{\beta}^{(m)}\}^2 + \lambda^{(m)} \|\boldsymbol{\beta}^{(m)}\|_1 \right\}, \quad m = 0, \ldots, M, \quad (26)$$

and the other steps remain unchanged. Specifically, we set $p = 300$ and $(n_0, \ldots, n_M) = (100, 100, 150, 100, 150, 100, 100, 100, 150, 150, 150)$ for $M = 10$. The covariates in each model are generated from the same multivariate normal distribution with an identity covariance matrix followed by Li et al. (2021). For the coefficient vector of the target model, we set $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_p^{(0)})^T = (0.5 \cdot \mathbf{1}_s^T, \mathbf{0}_{p-s}^T)^T$ for $s = 5$. For the coefficient vectors of the source models, we set $\beta_j^{(m)} = \beta_j^{(0)} + \psi_j \cdot \mathrm{I}(j \in \{s+1, \ldots, 5s\})$ for $m \in \{1, 2, 5, \ldots, 10\}$, $\beta_j^{(3)} = \beta_j^{(0)}$, and $\beta_j^{(4)} = \beta_j^{(0)} + 0.2 \cdot \mathrm{I}(j \in \{1, \ldots, 50\})$, where ψ_j is a binary variable with values -1 and 1 following the binomial distribution with parameter 0.5. In addition, we also consider the scenario where the target model and the second source model are misspecified. In this case, the corresponding coefficient vectors are $(p+1)$-dimensional, and we set the $(p+1)$th coefficients of $\beta_{p+1}^{(0)}$ and $\beta_{p+1}^{(2)}$ to 0.5. Similarly, we exclude the $(p+1)$th components of $\boldsymbol{x}_i^{(0)}$ and $\boldsymbol{x}_i^{(2)}$ when fitting the corresponding models. The random error for all $M+1$ models follows a standard normal distribution. To differentiate between the two proposed methods, we refer to the modified method as **Trans**fer learning for **H**igh-dimensional **M**odel **A**veraging **P**rediction (Trans-HMAP). The tuning parameters $\lambda^{(m)}$ for $m = 0, \ldots, M$ are chosen by 8-fold cross-validation suggested by Li et al. (2021).

To evaluate the performance of our method and other competitive methods, we generate $n^* = 100$ testing samples from the target model and calculate the mean squared prediction errors (MSPE). In addition, we report the sum of squared estimation errors (SSE), $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)}\|^2$, for different estimators. All results are based on 200 replications. To accommodate the high-dimensional settings, we use Lasso estimation for LSE-Tar and LSE-All, denoted as Lasso-Tar and Lasso-All, instead of least squares estimation. The corresponding results are reported in Table 12. The results show that our Trans-HMAP still outperforms other methods in both estimation and prediction, demonstrating the effectiveness of our framework in high-dimensional scenarios.

	Correct Specification		Misspecification	
	SSE	MSPE	SSE	MSPE
Trans-HMAP	0.277	0.295	0.308	0.572
	(0.083)	(0.101)	(0.098)	(0.120)
Trans-SimpMA	1.090	1.096	1.093	1.357
	(0.236)	(0.285)	(0.236)	(0.277)
Trans-SBIC	0.910	0.919	0.938	1.202
	(0.192)	(0.236)	(0.202)	(0.233)
Trans-SAIC	1.240	1.247	1.264	1.526
	(0.264)	(0.316)	(0.287)	(0.314)
Lasso-Tar	0.427	0.446	0.516	0.772
	(0.172)	(0.191)	(0.203)	(0.232)
Lasso-All	1.557	1.583	1.562	1.815
	(0.379)	(0.445)	(0.356)	(0.416)
Trans-Lasso	0.388	0.404	0.385	0.641
	(0.264)	(0.267)	(0.251)	(0.276)
Trans-GLM	0.326	0.340	0.401	0.655
	(0.179)	(0.185)	(0.241)	(0.249)
Uplift Rate	17.69%	15.25%	25.00%	12.06%

Table 12: The averaged SSE and MSPE for different methods in high-dimensional settings. The standard errors are given in parenthesis.

References

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11):1817–1853, 2005.

Tomohiro Ando and Ker-Chau Li. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265, 2014.

Tomohiro Ando and Ker-Chau Li. A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, 45(6):2654–2679, 2017.

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics*, 46(3):1352–1382, 2018.

Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(1):46–100, 2021.

Steven T Buckland, Kenneth P Burnham, and Nicole H Augustin. Model selection: an integral part of inference. *Biometrics*, 53:603–618, 1997.

Bertrand Clarke. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research*, 4(Oct):683–712, 2003.

Carl De Boor. *A Practical Guide to Splines*. Springer-Verlag New York, 2001.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, Seattle, Washington, 2004.

Yan Gao, Xinyu Zhang, Shouyang Wang, Terence Tai-leung Chong, and Guohua Zou. Frequentist model averaging for threshold models. *Annals of the Institute of Statistical Mathematics*, 71(2):275–306, 2019.

Nils Lid Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

Rongjie Jiang, Liming Wang, and Yang Bai. Optimal model averaging estimator for semifunctional partially linear models. *Metrika*, 84(2):167–194, 2021.

Mikhail K Kozlov, Sergei Pavlovich Tarasov, and Leonid Genrikhovich Khachiyan. Polynomial solvability of convex quadratic programming. In *Doklady Akademii Nauk*, volume 248, pages 1049–1051. Russian Academy of Sciences, 1979.

Prosenjit Kundu, Runlong Tang, and Nilanjan Chatterjee. Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3):567–585, 2019.

Jialiang Li, Xiaochao Xia, Weng Kee Wong, and David Nott. Varying-coefficient semiparametric model averaging prediction. *Biometrics*, 74(4):1417–1426, 2018.

Jialiang Li, Jing Lv, Alan TK Wan, and Jun Liao. Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association*, 117(537):495–509, 2022a.

Sai Li, T. Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):149–173, 2021.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association, Forthcoming*, 2022b.

Yang Li, Rong Li, Cunjie Lin, Yichen Qin, and Shuangge Ma. Penalized integrative semiparametric interaction analysis for multiple genetic datasets. *Statistics in Medicine*, 38 (17):3221–3242, 2019.

Jun Liao, Alan TK Wan, Shuyuan He, and Guohua Zou. Frequentist model averaging for the nonparametric additive model. *Statistica Sinica, Forthcoming*, 2021.

Chu-An Liu. Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186(1):142–159, 2015.

Jin Liu, Shuangge Ma, and Jian Huang. Integrative analysis of cancer diagnosis studies with composite penalization. *Scandinavian Journal of Statistics*, 41(1):87–103, 2014.

Shishi Liu and Jingxiao Zhang. Model averaging by cross-validation for partially linear functional additive models. *arXiv preprint arXiv:2105.00966*, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, pages 97–105, Lille, France, 2015.

Shuangge Ma, Jian Huang, Fengrong Wei, Yang Xie, and Kuangnan Fang. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine*, 30(28):3361–3371, 2011.

Yanyuan Ma and Liping Zhu. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):305–322, 2013.

Yanyuan Ma, Jeng-Min Chiou, and Naisyin Wang. Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, 93(1):75–84, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

Weike Pan, Evan W. Xiang, Nathan N. Liu, and Qiang Yang. Transfer learning in collaborative filtering for sparsity reduction. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 230–235, Atlanta, Georgia, 2010.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020.

Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

Charles J Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14(2):590–606, 1986.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association, Forthcoming*, 2022.

Kazuo Toraichi, Kazuki Katagishi, Iwao Sekita, and Ryoichi Mori. Computational complexity of spline interpolation. *International Journal of Systems Science*, 18(5):945–954, 1987.

Alan TK Wan, Xinyu Zhang, and Guohua Zou. Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2):277–283, 2010.

Li Wang, Xiang Liu, Hua Liang, and Raymond J Carroll. Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics*, 39(4):1827–1851, 2011.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of The Econometric Society*, 50(1):1–25, 1982.

Minge Xie, Kesar Singh, and William E Strawderman. Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493):320–333, 2011.

Xinyu Zhang. *Model Averaging and Its Applications*. PhD thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2010.

Xinyu Zhang and Hua Liang. Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39(1):174–200, 2011.

Xinyu Zhang and Chu-An Liu. Inference after model averaging in linear regression models. *Econometric Theory*, 35(4):816–841, 2019.

Xinyu Zhang and Chu-An Liu. Model averaging prediction by K-fold cross-validation. *Journal of Econometrics*, 235(1):280–301, 2023.

Xinyu Zhang and Wendun Wang. Optimal model averaging estimation for partially linear models. *Statistica Sinica*, 29(2):693–718, 2019.

Xinyu Zhang, Dalei Yu, Guohua Zou, and Hua Liang. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111(516):1775–1790, 2016.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

Rong Zhu, Alan TK Wan, Xinyu Zhang, and Guohua Zou. A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, 114(526):882–892, 2019.