Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ubes20

Model Averaging for Prediction With Fragmentary Data

Fang Fang, Wei Lan, Jingjing Tong & Jun Shao

To cite this article: Fang Fang, Wei Lan, Jingjing Tong & Jun Shao (2019) Model Averaging for Prediction With Fragmentary Data, Journal of Business & Economic Statistics, 37:3, 517-527, DOI: 10.1080/07350015.2017.1383263

To link to this article: https://doi.org/10.1080/07350015.2017.1383263

Model Averaging for Prediction With Fragmentary Data

Fang FANG
School of Statistics, East China Normal University, Shanghai, China (*ffang@sfs.ecnu.edu.cn*)

Wei LAN
Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, Sichuan, China (*lanwei@swufe.edu.cn*)

Jingjing TONG
School of Statistics, East China Normal University, Shanghai, China (*1083587797@qq.com*)

Jun SHAO
School of Statistics, East China Normal University, Shanghai, China, and Department of Statistics, University of Wisconsin-Madison, WI 53706 (*shao@stat.wisc.edu*)

One main challenge for statistical prediction with data from multiple sources is that not all the associated covariate data are available for many sampled subjects. Consequently, we need new statistical methodology to handle this type of "fragmentary data" that has become more and more popular in recent years. In this article, we propose a novel method based on the frequentist model averaging that fits some candidate models using all available covariate data. The weights in model averaging are selected by delete-one cross-validation based on the data from complete cases. The optimality of the selected weights is rigorously proved under some conditions. The finite sample performance of the proposed method is confirmed by simulation studies. An example for personal income prediction based on real data from a leading e-community of wealth management in China is also presented for illustration.

KEY WORDS: Asymptotic optimality; Cross-validation; Heteroscedastic errors; Linear regression models; Multiple data sources.

1. INTRODUCTION

The advancement of information technology and prevalence of mobile internet in recent years have made it possible and necessary for both scientific researchers and business analysts to utilize data from many different sources. While the rich information brought by multiple data sources brings opportunities for predicting people's behaviors with potential social and commercial benefits, it also poses challenges to statistical modeling. One major challenge is that not all the sampled subjects have the same predictors (which are called covariates in the rest of this article). For example, many online small loan companies in China need to predict customers' personal incomes based on all possibly available data sources including credit card statements, online shopping records, mobile phone bills, and so on. Unfortunately, usually only a small fraction of the customers have all the information/covariates available; some people may not have credit cards; some do not shopping online at all; probably most people have mobile phones but many of them are not willing to authorize the company to reach their bill information. Consequently, the dataset typically looks like Table 1, in which we just use 10 subjects and 8 covariates for illustration and "*" means a datum is available. Only the first two subjects have data available on all the eight covariates, while other subjects only have partial covariate data available. Unlike the traditional datasets from designed experiments or unified data collection procedures, the data are fragmentary and new statistical methodology is needed to use them for prediction.

Consider a random sample of n subjects with a response variable Y and a covariate set $D = \{X_j, j = 1, \ldots, p\}$. Each subject i only has covariate data available for a subset $D_i \subseteq D$. For the example in Table 1, $D_1 = D_2 = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$, $D_3 = \{X_1, X_2, X_3\}$, and so on. We assume $p < n$ but p may increase when n increases, and X_1 is the intercept effect, that is, $X_1 = 1$ for all subjects. Let $\{\Delta_k, k = 1, \ldots, K\}$ be the set of all different patterns for D_i, $i = 1, \ldots, n$. Note that $1 \le K \le 2^{p-1}$ since X_1 is always available, but K may be much smaller than 2^{p-1}. For Table 1, $K = 7$ but $2^{p-1} = 2^{8-1} = 128$. For notation simplicity, throughout the article we also use D_i or Δ_k to denote the set of indices of the covariates in D_i or Δ_k, for example, $D_i = \{X_1, X_2, X_3\} = \{1, 2, 3\}$ or $\Delta_k = \{X_1, X_4, X_5, X_6\} = \{1, 4, 5, 6\}$.

Our target is to make predictions given the fragmentary data $\{(y_i, x_{ij}), j \in D_i, i = 1, \ldots, n\}$, where y_i and x_{ij} are observations of Y and X_j whenever they are observed. Specifically, for a new subject with available covariate data $D^* \in \{\Delta_k, k = 1, \ldots,$

Journal of Business & Economic Statistics
July 2019, Vol. 37, No. 3
DOI: 10.1080/07350015.2017.1383263

Table 1. An illustrative example for fragmentary data

Subject	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*	*
3	*	*	*	*					
4	*	*	*	*				*	*
5	*	*			*	*	*		
6	*	*			*	*	*		
7	*	*							
8	*	*				*		*	*
9	*	*	*					*	*
10	*	*	*					*	*

*: the datum is available.

$K\}$, how do we estimate the conditional mean $\mu^* = E(Y|D^*)$ as accurately as possible? A naive approach, which is often used in practice because of its simplicity, is to use complete cases (CC) as described in the following. First, choose all the subjects i with $D_i \supseteq D^*$ to form $\mathcal{S}^* = \{1 \le i \le n : D_i \supseteq D^*\}$. Second, build a model M^* based on data $\{(y_i, x_{ij}), j \in D^*, i \in \mathcal{S}^*\}$. Third, use the fitted model M^* and D^* to make the prediction for the new subject. Take Table 1 as an example: if a new subject has available covariates $D^* = \{X_1, X_2, X_3, X_4, X_5, X_6\}$, then we need to choose subjects 1 and 2 to build a model based on data $\{(y_i, x_{ij}), i = 1, 2, j = 1, \ldots, 6\}$ for prediction. The CC method only uses a small fraction of the data and the prediction accuracy could be very low especially for fragmentary data, in which $|\mathcal{S}^*|/n$ could be pretty small, where $|\mathcal{A}|$ denotes the cardinality of any set $\mathcal{A}$ throughout this article.

Alternatively, we may treat the problem as a missing covariate data problem, although some covariate data are not missing but do not exist in our problem, and apply techniques such as imputation and inverse propensity weighting (e.g., Little and Rubin 2002; Kim and Shao 2013). However, such techniques require an additional model for imputation or propensity, which is hard to build because the percentage of unavailable data in fragmentary data is typically very high. Furthermore, the number of covariates with unavailable data could be very large and not all of the covariates are necessarily useful for prediction, which is seldom addressed in the missing covariate data literature.

In this article, we study the approach of frequentist model averaging for prediction with fragmentary data. There is a rich literature for frequentist model averaging when all the covariate data are available, for example, Buckland, Burnham, and Augustin (1997), Yang (2001), Hjort and Claeskens (2003), Yuan and Yang (2005), Hansen (2007), Wan, Zhang, and Zou (2010), Zhang and Liang (2011), Hansen and Racine (2012), Liu (2015), Zhang (2015), Zhang, Wan, and Zou (2013), Zhang, Zou, and Carroll (2015), Zhang et al. (2016), and Gao et al. (2016). The main theoretical justification for model averaging is its asymptotic optimality, a concept developed in Li (1987) for variable selection and described in detail in Section 3. In the context of missing covariate data, Schomaker, Wan, and Heumann (2010), Dardanoni, Modica, and Peracchi (2011), and Dardanoni et al. (2015) proposed model averaging methods based on imputation, but no theory about asymptotic optimality was developed. For covariate missing completely at random and homoscedastic errors, Zhang (2013) proposed a model averaging method and showed its superiority over the previous methods.

In the context of fragmentary data, we propose a new model averaging method, which builds candidate models similar to Zhang (2013), but uses a novel way to select optimal weights and a different idea for response prediction. The new way of selecting weights allows us to naturally extend the homoscedasticity set-up in Zhang (2013) to heteroscedasticity that is often encountered for data from different sources in business, social or medical studies. We prove the asymptotic optimality of our proposed method under an assumption on fragmentary data similar to covariate-dependent missing at random in the missing data literature (Little and Rubin 2002), which is more practical than the assumption of missing completely at random. The proposed method has better prediction performance than Zhang's method and some other methods in simulations and a real-data analysis.

The rest of the article is organized as follows. Our motivation and details of the methodology are presented in Section 2. Asymptotic optimality of the proposed method is established in Section 3. Empirical results of simulation studies and a real-data analysis are presented in Sections 4 and 5, respectively. Concluding remarks including a discussion on more applications of our method are given in Section 6. All the proofs are in the online Appendix.

2. MODEL AVERAGING APPROACH

2.1 Idea and Procedure

For notation simplicity, we focus on the case where $D^* = D$, that is, the available covariate set D^* for prediction is the largest possible covariate set D. Other situations can be handled in a similar manner by ignoring the covariates not contained in D^*. Without loss of generality, let $\Delta_1 = D$. Assume that $\{(y_i, x_{ij}), i = 1, \ldots, n, j = 1, \ldots, p\}$ is a random sample generated from the following linear regression model

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is an unknown regression coefficient vector, ε_i's are independently distributed random noises with $E(\varepsilon_i|x_i) = 0$ and finite variance $E(\varepsilon_i^2|x_i) = \sigma_i^2$, and $x_i = (x_{i1}, \ldots, x_{ip})^\top$. Denote $\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}$ as the conditional mean of y_i given all the covariates, $\mu = (\mu_1, \ldots, \mu_n)^\top$, and $y = (y_1, \ldots, y_n)^\top$.

Since the variance σ_i^2 may vary with i, the errors are heteroscedastic. In a study of a common characteristic based on data from different sources, such as the personal income data from loan companies in China discussed in Section 1 and other datasets discussed in Section 6, data often exhibit heteroscedasticity.

What we observe is a fragmentary dataset $\{(y_i, x_{ij}), i = 1, \ldots, n, j \in D_i\}$, which is our training dataset since model fitting is based on it. A fragmentary dataset exhibits another type of heterogeneity, that is, different covariates are measured from different sources.

Our key idea can be described as follows. The CC method implicitly assumes that all the covariates in D are useful and excludes subjects with $D_i \neq D$, which results in a small-sample size for model fitting. In an application, it is often the case that some of the p covariates are much more useful than the others for prediction although we do not know which covariates are useful. If we exclude some less useful covariates, the sample size for modeling increases, which probably increases the prediction accuracy. For the example in Table 1, the CC method fits a model M_1 based on dataset $\{(y_i, x_{ij}), i = 1, 2, j = 1, \ldots, 8\}$. If covariates X_4 to X_8 are not that useful, we can fit a new model M_2 based on $\{(y_i, x_{ij}), i = 1, 2, 3, 4, j = 1, 2, 3\}$ with a larger sample size. Further, if we know that X_3 is not useful either, we can exclude X_3 from M_2 and fit a model M_3 based on $\{(y_i, x_{ij}), i = 1, 2, 3, 4, 9, 10, j = 1, 2\}$ with an even larger sample size. Formally, we can fit K possible candidate models $\{M_k, k = 1, \ldots, K\}$ with different levels of tradeoff between number of covariates and sample size, where M_k is based on data $\{(y_i, x_{ij}), D_i \supseteq \Delta_k, j \in \Delta_k\}$. However, we do not know which candidate model should be used. Without "putting all our inferential eggs in one unevenly woven basket" (Longford 2005), we propose to apply the idea of model averaging and use a weighted average of the predictions from all the candidate models as our final predictor.

For any $k = 1, \ldots, K$, let $\mathcal{S}_k = \{i : D_i \supseteq \Delta_k\}$ be the set of subjects that can be used to fit a linear model with covariate set Δ_k, $n_k = |\mathcal{S}_k|$, and $p_k = |\Delta_k|$. We assume $n_1 \geq p$, that is, the sample size of CC is no less than p. Consequently, $n_k \geq p_k$ since $n_k \geq n_1$ and $p_k \leq p$. For any $\mathcal{S}_k$ and Δ_k, define $y_{\mathcal{S}_k} = (y_i : i \in \mathcal{S}_k)^\top \in \mathbb{R}^{n_k \times 1}$, $\varepsilon_{\mathcal{S}_k} = (\varepsilon_i : i \in \mathcal{S}_k)^\top \in \mathbb{R}^{n_k \times 1}$, $\beta_{\Delta_k} = (\beta_j : j \in \Delta_k)^\top \in \mathbb{R}^{p_k \times 1}$, and $\mathbb{X}_{\mathcal{S}_k \Delta_k} = (x_{ij} : i \in \mathcal{S}_k, j \in \Delta_k) \in \mathbb{R}^{n_k \times p_k}$.

For each $k = 1, \ldots, K$, with data $\{(y_i, x_{ij}), i \in \mathcal{S}_k, j \in \Delta_k\}$, we fit the following candidate model denoted by M_k,

$$y_{\mathcal{S}_k} = \mathbb{X}_{\mathcal{S}_k \Delta_k} \beta_{\Delta_k} + \varepsilon_{\mathcal{S}_k}. \tag{2}$$

Although we assume (1), model M_k in (2) is not necessarily correct. The least-square estimator of β_{Δ_k} under model M_k is $\hat{\beta}_{\Delta_k} = (\mathbb{X}^\top_{\mathcal{S}_k \Delta_k} \mathbb{X}_{\mathcal{S}_k \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_k \Delta_k} y_{\mathcal{S}_k} \in \mathbb{R}^{p_k \times 1}$. For a new subject with covariate vector $x^* = (x_1^*, \ldots, x_p^*)^\top$, the least-square prediction under model M_k is

$$\hat{\mu}_k^* = x_{\Delta_k}^{*\top} \hat{\beta}_{\Delta_k} = x_{\Delta_k}^{*\top} (\mathbb{X}^\top_{\mathcal{S}_k \Delta_k} \mathbb{X}_{\mathcal{S}_k \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_k \Delta_k} y_{\mathcal{S}_k},$$

where $x^*_{\Delta_k} = (x_j^*, j \in \Delta_k)^\top \in \mathbb{R}^{p_k \times 1}$. Let the K-dimensional weight vector $w = (w_1, \ldots, w_K)^\top$ come from the $\mathbb{R}^K$ unit hypercube

$$\mathcal{H}_K = \left\{ w \in [0, 1]^K : \sum_{k=1}^K w_k = 1 \right\}.$$

A weighted averaging prediction over all the K candidate models for $\mu^* = E(Y|x^*)$ is

$$\begin{aligned} \hat{\mu}^*(w) &= \sum_{k=1}^K w_k \hat{\mu}_k^* = \sum_{k=1}^K w_k x_{\Delta_k}^{*\top} \hat{\beta}_{\Delta_k} \\ &= \sum_{k=1}^K w_k x_{\Delta_k}^{*\top} (\mathbb{X}^\top_{\mathcal{S}_k \Delta_k} \mathbb{X}_{\mathcal{S}_k \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_k \Delta_k} y_{\mathcal{S}_k}. \end{aligned}$$

2.2 Weight Optimization

For linear models with heteroscedastic errors, Hansen and Racine (2012) used delete-one cross-validation to obtain the optimal weight vector w. However, the cross-validation cannot be directly applied to our problem since many covariate data are unavailable. We propose to use the cross-validation based on CC data ($i \in \mathcal{S}_1$) to determine the optimal weight vector. Consider model M_k in (2) restricted to CC data, that is,

$$y_{\mathcal{S}_1} = \mathbb{X}_{\mathcal{S}_1 \Delta_k} \beta_{\Delta_k} + \varepsilon_{\mathcal{S}_1}. \tag{3}$$

The least-square estimator under (3) is $\tilde{\beta}_{\Delta_k} = (\mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} \mathbb{X}_{\mathcal{S}_1 \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} y_{\mathcal{S}_1} \in \mathbb{R}^{p_k \times 1}$. The prediction of μ_i for any $i \in \mathcal{S}_1$ is $\tilde{\mu}_{ki} = x^\top_{i\Delta_k} \tilde{\beta}_{\Delta_k} = x^\top_{i\Delta_k} (\mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} \mathbb{X}_{\mathcal{S}_1 \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} y_{\mathcal{S}_1}$, where $x_{i\Delta_k} = (x_{ij}, j \in \Delta_k)^\top \in \mathbb{R}^{p_k \times 1}$. Define $\tilde{\mu}_{k\mathcal{S}_1} = (\tilde{\mu}_{ki} : i \in \mathcal{S}_1)^\top \in \mathbb{R}^{n_1 \times 1}$. Then, the model averaging prediction of $\mu_{\mathcal{S}_1} = (\mu_i : i \in \mathcal{S}_1)^\top \in \mathbb{R}^{n_1 \times 1}$ can be written as

$$\tilde{\mu}_{\mathcal{S}_1}(w) = \sum_{k=1}^K w_k \tilde{\mu}_{k\mathcal{S}_1} = \sum_{k=1}^K w_k \mathbb{X}_{\mathcal{S}_1 \Delta_k} (\mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} \mathbb{X}_{\mathcal{S}_1 \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} y_{\mathcal{S}_1}.$$

Define $\tilde{P}_k = \mathbb{X}_{\mathcal{S}_1 \Delta_k} (\mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} \mathbb{X}_{\mathcal{S}_1 \Delta_k})^{-1} \mathbb{X}^\top_{\mathcal{S}_1 \Delta_k} \in \mathbb{R}^{n_1 \times n_1}$. Then, we have

$$\tilde{\mu}_{\mathcal{S}_1}(w) = \sum_{k=1}^K w_k \tilde{P}_k y_{\mathcal{S}_1} = \tilde{P}(w) y_{\mathcal{S}_1}$$

with $\tilde{P}(w) = \sum_{k=1}^K w_k \tilde{P}_k$.

For each subject $i \in \mathcal{S}_1$, let $\tilde{\mu}_k^{(-i)}$ be the predicted value of μ_i from (3) after deleting data from subject i. Denote $\tilde{\mu}_k^{cv} = (\tilde{\mu}_k^{(-i)}, i \in \mathcal{S}_1)^\top \in \mathbb{R}^{n_1 \times 1}$. The cross-validation estimator is $\tilde{\mu}^{cv}_{\mathcal{S}_1}(w) = \sum_{k=1}^K w_k \tilde{\mu}_k^{cv}$ and we select the optimal w by minimizing the following cross-validation criterion,

$$\mathrm{CV}(w) = \|y_{\mathcal{S}_1} - \tilde{\mu}^{cv}_{\mathcal{S}_1}(w)\|^2.$$

Cross-validation is often time-consuming. Fortunately, in linear regression models, it is much easier to calculate. To see this, denote $\tilde{e}^{cv} = (\tilde{e}_1^{cv}, \cdots, \tilde{e}_K^{cv}) \in \mathbb{R}^{n_1 \times K}$, where $\tilde{e}_k^{cv} = y_{\mathcal{S}_1} - \tilde{\mu}_k^{cv}$. Following the discussions in Li (1987) and Hansen and Racine (2012), we have $\tilde{e}_k^{cv} = \mathcal{Q}_k (y_{\mathcal{S}_1} - \tilde{\mu}_{k\mathcal{S}_1})$, where $\mathcal{Q}_k = \mathrm{diag}\{(1 - m_1^k)^{-1}, \ldots, (1 - m_{n_1}^k)^{-1}\}$ and m_i^k is the ith diagonal element of $\tilde{P}_k$. Then

$$\mathrm{CV}(w) = \left\| \sum_{k=1}^K w_k (y_{\mathcal{S}_1} - \tilde{\mu}_k^{cv}) \right\|^2 = \|\tilde{e}^{cv} w\|^2 = w^\top \tilde{e}^{cv\top} \tilde{e}^{cv} w.$$

We then choose the weight vector w that minimizes $\mathrm{CV}(w)$ over the set $\mathcal{H}_K$ as

$$\hat{w} = \underset{w \in \mathcal{H}_K}{\mathrm{argmin}}\, \mathrm{CV}(w). \tag{4}$$

Since $\mathrm{CV}(w)$ is a quadratic function of w, it can be easily minimized using the `quadprog` package in R or the quadprog command in Matlab.

Remark 1. Basically, our idea is to use CC data ($i \in \mathcal{S}_1$) to construct the weights, but use all available data ($i \in \mathcal{S}_k$) to fit each candidate model. This weight selection method is different from that in Zhang (2013), which selects weights by applying Mallows' criterion to the entire training dataset

with unavailable covariate data filled by zeros. Because the optimality of Mallows' criterion requires homoscedastic errors, Zhang's method can only be justified under homoscedasticity. Using cross-validation to the CC data, our method naturally handles heteroscedasticity and has wider application scope. Furthermore, Zhang's method uses one weight vector for all the predictions, whereas our method flexibly uses different weight vectors for prediction with different D^*. Finally, when doing prediction, Zhang's method replaces unavailable covariate data by zeros, which is somewhat arbitrary and may affect the prediction accuracy. Our method makes prediction based on the available covariate set D^* and does not use any artificial values. The covariates not in D^* are ignored when we make predictions to subjects with available covariate set D^*. We do this to avoid imputation which may bring more uncertainty. This tradeoff works fine as shown in the empirical results.

3. THEORETICAL RESULTS

To evaluate the optimality of $\hat{w}$ in (4), following Li (1987), Hansen (2007), and Hansen and Racine (2012), we define the following loss function to measure the in-sample fit,

$$L_n(w) = \|\mu_{\mathcal{S}_1} - \hat{\mu}_{\mathcal{S}_1}(w)\|^2, \tag{5}$$

where $\hat{\mu}_{\mathcal{S}_1}(w) = (\hat{\mu}_i(w), i \in \mathcal{S}_1)^\top$ is the in-sample prediction of $\mu_{\mathcal{S}_1}$ based on M_k in (2) and the weight w. More specifically, for any $i \in \mathcal{S}_1$, the prediction of μ_i based on M_k is $\hat{\mu}_{ki} = x_{i\Delta_k}^\top \hat{\beta}_{\Delta_k} = x_{i\Delta_k}^\top (\mathbb{X}_{\mathcal{S}_k\Delta_k}^\top \mathbb{X}_{\mathcal{S}_k\Delta_k})^{-1} \mathbb{X}_{\mathcal{S}_k\Delta_k}^\top y_{\mathcal{S}_k}$. Define $\hat{\mu}_{k\mathcal{S}_1} = (\hat{\mu}_{ki}, i \in \mathcal{S}_1)^\top \in \mathbb{R}^{n_1 \times 1}$. Then, the model averaging predictor $\hat{\mu}_{\mathcal{S}_1}(w)$ with weight w can be written as

$$\hat{\mu}_{\mathcal{S}_1}(w) = \sum_{k=1}^{K} w_k \hat{\mu}_{k\mathcal{S}_1} = \sum_{k=1}^{K} w_k \mathbb{X}_{\mathcal{S}_1\Delta_k} (\mathbb{X}_{\mathcal{S}_k\Delta_k}^\top \mathbb{X}_{\mathcal{S}_k\Delta_k})^{-1} \mathbb{X}_{\mathcal{S}_k\Delta_k}^\top y_{\mathcal{S}_k}. \tag{6}$$

The weight $\hat{w}$ determined by (4) is asymptotically optimal if

$$\frac{L_n(\hat{w})}{\inf_{w\in\mathcal{H}_K} L_n(w)} \to_p 1, \tag{7}$$

where $\to_p$ denotes convergence in probability as $n \to \infty$. If (7) holds, then the squared error obtained using the weight determined by delete-one cross-validation based on CC data is asymptotically equivalent to the infeasible optimal weight vector.

Remark 2. Note that $D^* = D$ is assumed, which means that we consider the prediction of Y given all the p available covariates. So, the loss function $L_n(w)$ in (5) defined in CC sample is natural. Zhang (2013) considered asymptotic optimality of his procedure under the loss function $\|\mu - \tilde{\mathbb{X}}\hat{\beta}(w)\|^2$, where $\tilde{\mathbb{X}}$ is the full covariate matrix $\mathbb{X} = (x_1, \ldots, x_n)^\top$ with unavailable covariate data filled by zeros, $\hat{\beta}(w) = \sum_{k=1}^{K} w_k \Pi_k^\top \hat{\beta}_{\Delta_k}$, and Π_k is the projection matrix mapping β to the subvector $\Pi_k\beta = \beta_{\Delta_k}$. However, the loss function $\|\mu - \tilde{\mathbb{X}}\hat{\beta}(w)\|^2$ is somewhat artificial and is reasonable only when replacing unavailable covariate values by zeros is justifiable.

To show (7), we first prove an intermediate result. Note that $\hat{w}$ actually is the optimal weight if we just focus on CC data. Define

$$\tilde{L}_n(w) = \|\mu_{\mathcal{S}_1} - \tilde{\mu}_{\mathcal{S}_1}(w)\|^2 \quad \text{and}$$
$$\tilde{R}_n(w) = E\{\tilde{L}_n(w)|x_1, \ldots, x_n\}.$$

Similar to Hansen and Racine (2012) and Zhang, Wan, and Zou (2013), we can establish the following Lemma. The proof is given in the online Appendix.

Lemma 1. Assume model (1). Let $\tilde{\xi}_n = \inf_{w\in\mathcal{H}_K} \tilde{R}_n(w)$ and w_k^0 be a weight vector in which the kth element is one and the others are zeros. Suppose that there exist constants $G \geq 1$ and $\Lambda > 0$ such that

(C1) $\sup_{i\geq 1} E(\varepsilon_i^{4G}|x_i) < \infty$, a.s.,
(C2) $0 < \inf_{i\geq 1} \sigma_i^2 \leq \sup_{i\geq 1} \sigma_i^2 < \infty$, a.s.,
(C3) $\sup_k \frac{1}{p_k} \bar{\lambda}(\tilde{P}_k) \leq \Lambda n_1^{-1}$,
(C4) $K\tilde{\xi}_n^{-2G} \sum_{k=1}^{K} \{\tilde{R}_n(w_k^0)\}^G \to_{\text{a.s.}} 0$,
(C5) $p/n_1 \to 0$ and $p = O(\tilde{\xi}_n)$,

where $\bar{\lambda}(\cdot)$ denotes the maximum diagonal element of a matrix. Then,

$$\frac{\tilde{L}_n(\hat{w})}{\inf_{w\in\mathcal{H}_K} \tilde{L}_n(w)} \to_p 1. \tag{8}$$

Condition (C1) is a moment bound and (C2) excludes unbounded heteroscedasticity. Condition (C3) is the same as condition (5.2) by Li (1987) and condition (5) in Ando and Li (2014). As pointed out in Li (1987), it excludes extremely unbalanced design matrices as candidate regression models.

Condition (C4) is the same as condition (8) in Wan, Zhang, and Zou (2010) and condition (13) in Zhang, Wan, and Zou (2013). First, it requires $\tilde{\xi}_n \to \infty$, which is condition (A.3′) in Li (1987) and condition (15) in Hansen (2007). It is usually achieved if all the candidate models are underfitted or p goes to infinity as n increases. Second, it excludes candidate models with extremely large expected losses. To see this, we denote $\eta_n = \max_{w\in\mathcal{H}_K(0)} \tilde{R}_n(w)$, where $\mathcal{H}_K(0)$ is the set comprising the vectors $w_k^0, k = 1, \ldots, K$. Then, $K^2(\eta_n\tilde{\xi}_n^{-2})^G \to 0$ is a sufficient condition for (C4). In practice, the rates of $K \to \infty$ and $\eta_n \to \infty$ can be reduced by removing the very poor models at the outset prior to model combining as suggested in Wan, Zhang, and Zou (2010). Third, K, the number of candidate models, is allowed to increase, although condition (C4) places a restriction on the rate at which K increases with n. If K is diverging, condition (C4) is stronger than condition (21) of Zhang, Wan, and Zou (2013). However, Zhang, Wan, and Zou (2013) imposed another condition (22) in their Theorem 2.2 to restrict the growth rate of number of regressors. Wan, Zhang, and Zou (2010) and Zhang, Wan, and Zou (2013) provided more detailed discussions on this condition. Specially, Wan, Zhang, and Zou (2010) provided two explicit examples under which this condition holds.

Condition (C5) requires that p is bounded by $\tilde{\xi}_n$ and the diverging rate of p is slower than n_1. Since $\tilde{\xi}_n \to \infty$, (C5) actually is a weak condition. For example, when p is assumed to be fixed, condition (C5) holds.

Lemma 1 plus several more assumptions lead to the following main result. The proof is given in the online Appendix.

Theorem 1. Assume the conditions of Lemma 1 and

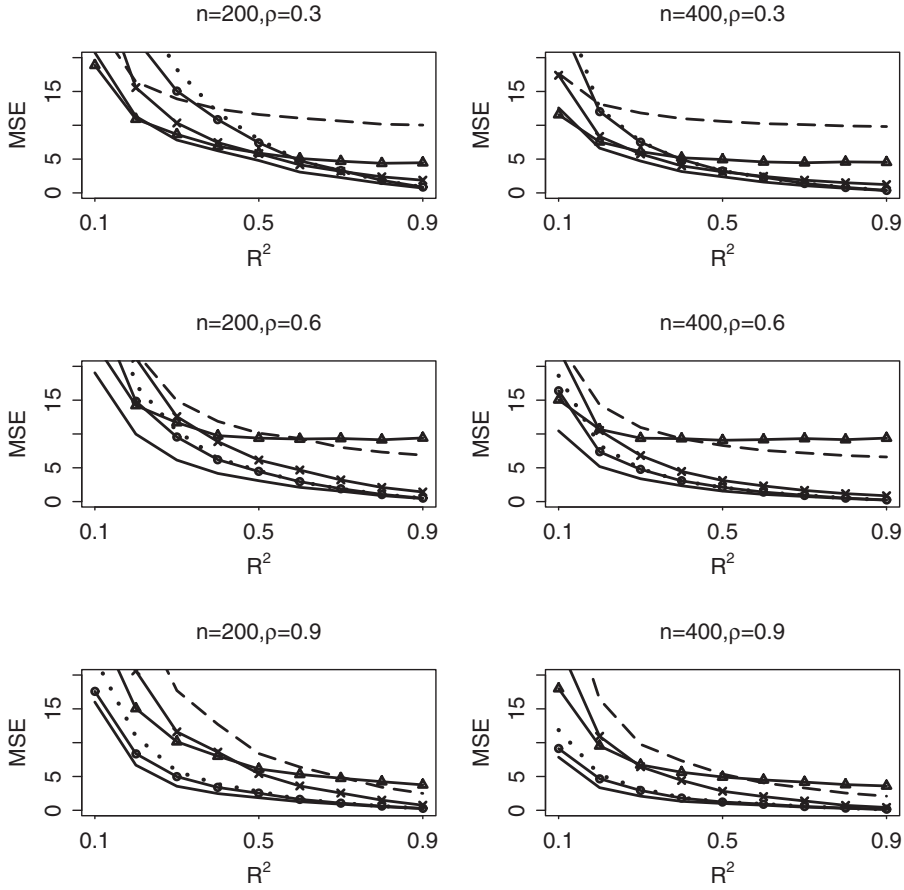

Figure 1. Simulation results when $\beta = (1, 1, \ldots, 1)$. The median of MSEs based on 1000 simulation runs: the proposed method (solid), CC (dotted), G1 (dashed), GLASSO (solid with circle), Zhang (solid with triangle), IMP-MA (solid with cross).

(C6) $\max_{1\le k\le K} \lambda_{\max}(n_1^{-1}\mathbb{X}^{\top}_{\mathcal{S}_1\Delta_k}\mathbb{X}_{\mathcal{S}_1\Delta_k}) = O(1)$ a.s., where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix,

(C7) some of the covariates are always available for all the subjects and the probability of the availability of other covariates depends only on the always available covariates,

(C8) $E(X_{\Delta_k^c}|X_{\Delta_k})$ is linear in X_{Δ_k}, where $X_{\Delta_k} = \{X_j, j \in \Delta_k\}$ and $X_{\Delta_k^c} = \{X_j, j \notin \Delta_k\}$.

Then, (7) holds, that is, the weight $\hat{w}$ determined by (4) is asymptotically optimal.

Condition (C6) is a commonly used regularity condition in linear models. See, for example, condition (A1) of Zou and Zhang (2009) and condition (6) of Zhang (2013).

Condition (C7) is about how the covariate data are available, or equivalently, the missing data mechanism. The assumption is similar to the covariate-dependent missing at random (Little and Rubin 2002), in which the missing data probability only depends on the observed covariates. It usually happens in practice when one main data source is fully accessible and several supplementary data sources are only available for part of the subjects. The example in Section 5 that motivated our research belongs to this case. Condition (C8) is satisfied when the covariate vector X is normal, but it is more general than normality assumption because the conditional linearity only needs to hold for groups of covariates. A similar condition was used by Zhong et al. (2012) and Yu, Dong, and Zhu (2016). Conditions (C7) and (C8) together make sure the difference between the least-square estimator $\tilde{\beta}_{\Delta_k}$ based on CC data ($i \in \mathcal{S}_1$) and $\hat{\beta}_{\Delta_k}$ based on all available data ($i \in \mathcal{S}_k$) is asymptotically negligible, which is the real technical condition we need.

4. SIMULATION STUDIES

Some simulation studies were conducted to examine the finite sample performance of the proposed method and compare it with some other methods.

4.1 General Comparison

The training dataset was generated as follows. A sample was generated from model (1) with $n = 200$ or 400, $p = 13$, $\beta = (1, 1, \ldots, 1)$, $(1, 1/2, \ldots, 1/p)$ or $(1/p, \ldots, 1/2, 1)$, where $x_{i1} = 1$, $(x_{i2}, \ldots, x_{ip})$ was generated from a multivariate normal distribution with $E(x_{ij}) = 1$, $\mathrm{var}(x_{ij}) = 1$, and $\mathrm{cov}(x_{ij_1}, x_{ij_2}) = \rho$ for $j_1 \neq j_2$, $\rho = 0.3, 0.6$, or 0.9, $\varepsilon_i \sim N(0, \sigma_i^2)$, $\sigma_i = \sigma \sum_{j=1}^p x_{ij}^2 / E(\sum_{j=1}^p x_{ij}^2)$, and σ^2 was set so that $R^2 = \mathrm{var}(x_i^{\top}\beta)/\{\mathrm{var}(x_i^{\top}\beta) + \sigma^2\} = 0.1, 0.2, \ldots, 0.9$. The 12 covariates other than the intercept were divided into four groups. The sth group consisted of $X_{3(s-1)+2}$ to X_{3s+1}, $s = 1, 2, 3, 4$. The covariates in the first group were always available and the covariates in the sth group, $s = 2, 3, 4$, were available if $X_s < 1$, which resulted in $K = 8$. Under this setting, the percentages of

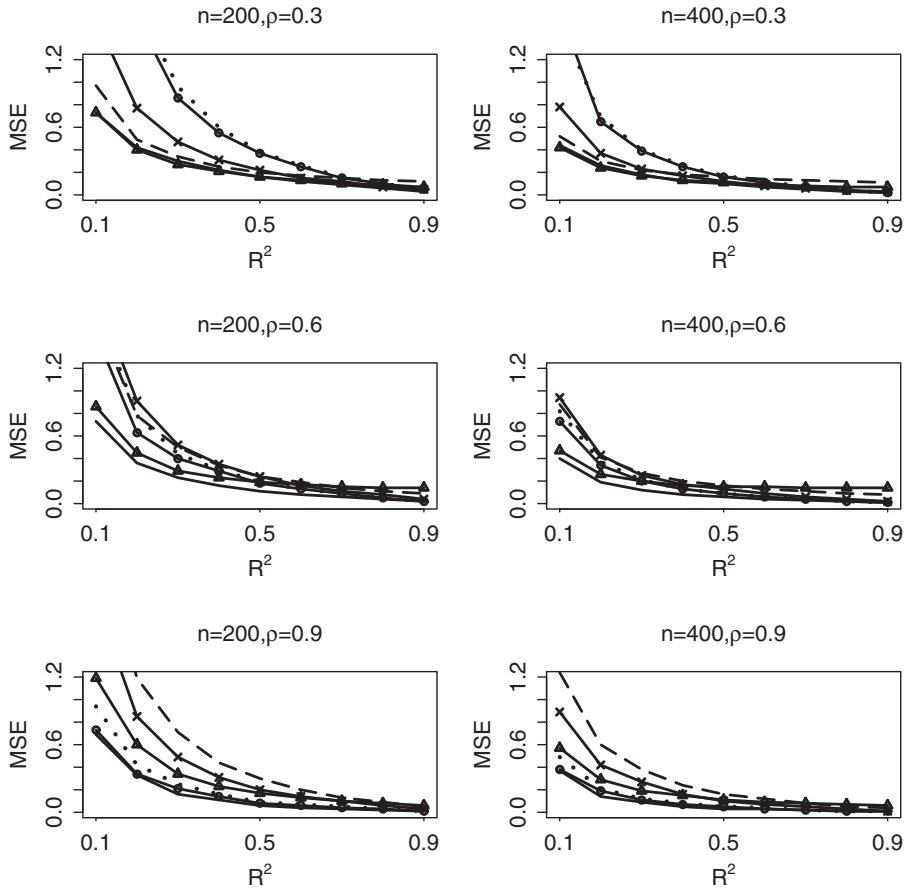

Figure 2. Simulation results when $\beta = (1, 1/2, \ldots, 1/p)$. The median of MSEs based on 1000 simulation runs: the proposed method (solid), CC (dotted), G1 (dashed), GLASSO (solid with circle), Zhang (solid with triangle), IMP-MA (solid with cross).

CC data were 19.8%, 27.9%, and 39.3%, respectively for $\rho =$ 0.3, 0.6, and 0.9.

We considered the prediction when $D^* = D$, and compared the following six methods:

- CC: the method using subjects that all the covariates are available.
- G1: the method using all the subjects and only the covariates in the first group.
- The proposed method.
- GLASSO: the method using CC data and group lasso of Yuan and Lin (2006) to select covariates and fitting a model with the subjects that have all the selected covariates available.
- Zhang: the model averaging method in Zhang (2013).
- IMP-MA: the method in Schomaker, Wan, and Heumann (2010), that is, imputing first then model averaging.

To evaluate the performance of these six methods, we generated a test dataset $\{(\mu_l, x_{lj}), l = 1, \ldots, L, j = 1, \ldots, p\}$ following exactly the same procedure as the one for training data except that L equals one million. Although the training data were regenerated at each simulation, the test data was only generated once and remained the same over the simulation runs.

The number of simulation runs was 1000. In each simulation run, we used the training data to fit the models for the six methods and applied them to the test data to get $\hat{\mu}_l, l = 1, \ldots, L$. The performance of each method was evaluated by MSE, the averaged squared differences between μ_l and $\hat{\mu}_l$, where the average is over the CC sample of the test data.

For each simulation setting, Figures 1–3 display the median of MSEs based on 1000 simulation runs. The G1 curves do not appear in some panels because the MSEs of G1 are much higher than the other methods in these panels. The main conclusions are as follows.

1. When R^2 increases and sample size n increases, the MSEs of all methods generally decrease, which is expected. The figure shapes with $n = 200$ and $n = 400$ are similar.
2. The CC method performs quite bad when $\rho = 0.3$, in which case the sample size of CC data is small, and R^2 is small. The GLASSO method improves the performance a little bit, but overall it performs similar to the CC method.
3. The G1 methods can lead to much higher MSEs than the other methods when the regression coefficients in the first group are relatively small. Note that the CC and G1 methods represent the two extreme cases for the tradeoff between sample size and number of covariates in the model. The CC makes fully usage of the covariates but has the smallest sample size, while the G1 makes fully usage of the sample but only uses the covariates in the first group. The simulation results show that they both could perform quite bad.

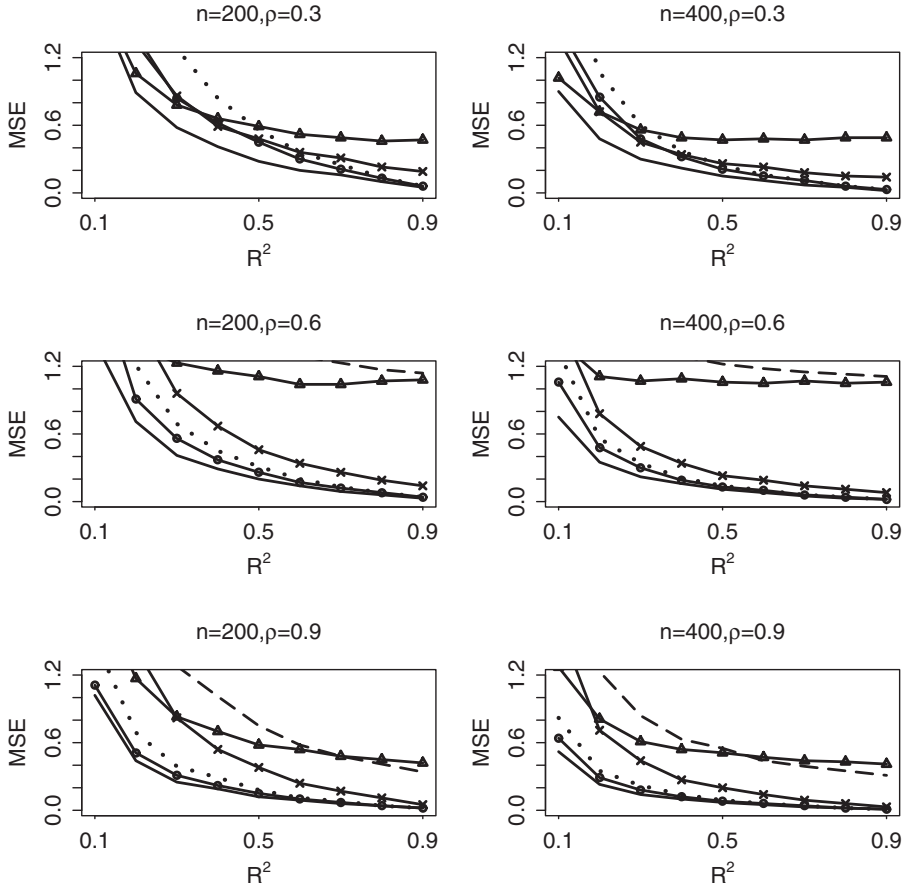

Figure 3. Simulation results when $\beta = (1/p, \ldots, 1/2, 1)$. The median of MSEs based on 1000 simulation runs: The proposed method (solid), CC (dotted), G1 (dashed), GLASSO (solid with circle), Zhang (solid with triangle), IMP-MA (solid with cross).

4. The proposed method produces the lowest MSEs in most situations.
5. Zhang's method generally performs not as well as the proposed method. Some exceptions happen when R^2 is very small (for example, the first panel in Figure 1) or the regression coefficients for the unavailable covariates are relatively small (the first two panels in Figure 2). In these situations, the usage of zeros to replace unavailable covariates in Zhang's method has relatively small effect to the prediction.

4.2 Comparison to Zhang's Method

Is the superiority of our proposed method over Zhang's method due to the new weight selection and prediction framework, or just due to the ability to handle heteroscedasticity by using cross-validation in selecting weights? To answer this question, we conducted a simulation to compare the proposed method and Zhang's method under two situations, one with homoscedastic errors and the other with heteroscedastic errors.

We generated the training dataset the same as in the previous simulation except that two types of errors were considered. In the first case, the errors were heteroscedastic and generated in the same way as before. In the second case, the errors were homoscedastic with $\varepsilon_i \sim N(0, \sigma^2)$. The sample size $n = 400$. We generated a test dataset $\{(\mu_l, x_{lj}), l = 1, \ldots, L, j = 1, \ldots, p\}$ following exactly the same procedure as the one for training data except that L equals one million. The test data were only generated once and remained the same over the simulation run. At each simulation run, the prediction performance of the proposed method and Zhang's method was evaluated by $\text{MSE}=L^{-1}\sum_{l=1}^{L}(\mu_l - \hat{\mu}_l)^2$. Note that the average was over the whole test data. For the prediction of subjects in the test data with $D^* \neq D$, we ignored the covariates not in D^* for modeling and prediction when implementing the proposed method.

Based on 1000 simulation replications, Figure 4 displays the results of "MSE ratio," which is defined as the median of MSEs of the proposed method divided by the median of MSEs of Zhang's method. The solid lines are results with heteroscedastic errors. The dotted lines are the results with homoscedastic errors. If the MSE ratio is less than 1 (below the horizontal line), the proposed method performs better than Zhang's method. The main observations are as follows.

1. The solid lines are always below 1, indicating that the proposed method performs better than Zhang's method in all the considered simulation settings with heteroscedastic errors.
2. When $\rho = 0.3$ or 0.6, the dotted lines are below 1 in most situations. Some exceptions happen when R^2 is very small. When $\rho = 0.9$, the dotted lines are below 1 when R^2 is reasonably large, which is the situation we care about more in practice.

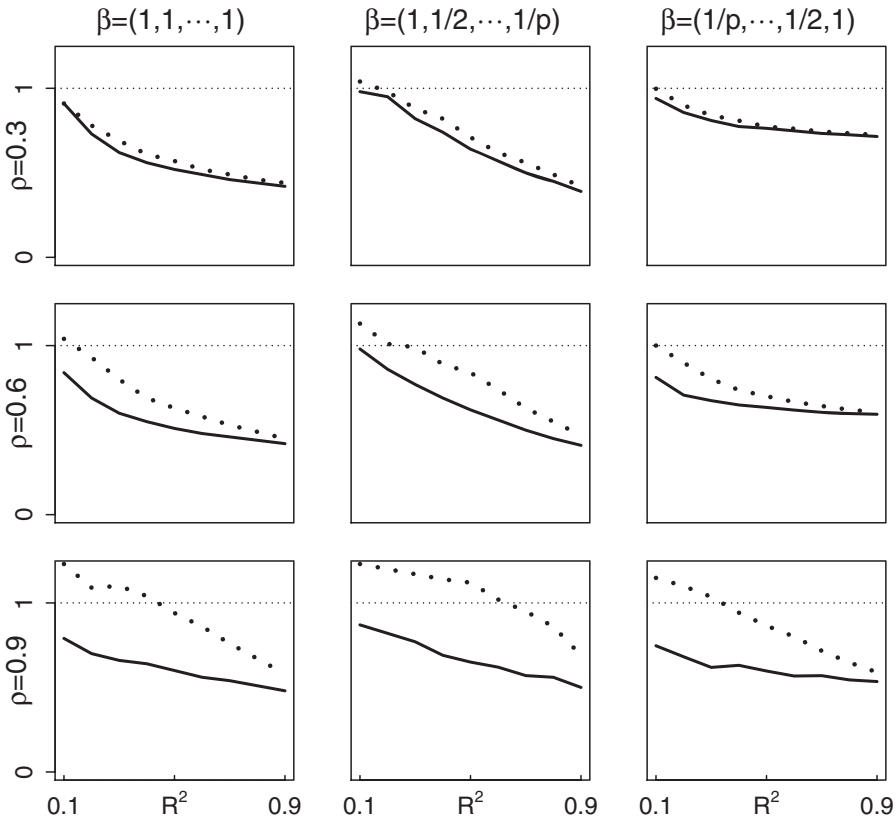

Figure 4. Simulation results of MSE ratio, defined as the median of MSEs (over the 1000 simulation runs) of the proposed method divided by the median of MSEs of Zhang's method, under different settings with heteroscedastic errors (solid) or homoscedastic errors (dotted). The horizontal line is the threshold of 1.

3. Overall speaking, the proposed method still performs better than Zhang's method most of the time even with homoscedastic errors, although the advantage is greater with heteroscedastic errors.

Our conclusion is: the proposed method shows advantage over Zhang's method mainly because of the new weight selection and prediction framework, while the stronger ability to handle heteroscedasticity brings us more benefit when the errors are truly heteroscedastic.

5. APPLICATION TO PERSONAL INCOME PREDICTION

To illustrate the application of our proposed method, we considered a real dataset from a leading e-community of wealth management in China, which actually motivated our study. Recently, the company launched its online personal loan product to provide access to instant cash loans via the internet. To evaluate the repayment ability of each loan applicant, it is critical to know the applicant's personal income. However, unlike traditional personal loan application process in commercial banks, the proof of income is not required in the online loan application since it is time consuming for the applicant to prepare it and also technically complicated for the credit officers to validate it in a short time. Instead, the applicant is required to provide personal ID number, credit card information, and online shopping record in *taobao.com*, the most popular online shopping website in China, if applicable. Based on the personal ID, the company is able to get the applicant's mobile phone usage record if authorized. Also, the company can check the applicant's credit bureau information from the central bank of China and fraud record in an anti-fraud platform if the applicant has bureau and/or fraud record. The company wants to predict applicant's income based on 25 key covariates (so $p = 26$ including the intercept) from the five data sources mentioned above: credit card information, online shopping record, mobile phone usage, bureau information, and fraud record.

Table 2. Covariate data availability patterns and sample sizes

	Data source					
k	Card	Shopping	Mobile	Bureau	Fraud	Sample size
1	*	*	*	*	*	115
2	*	*	*	*		29
3	*	*	*			220
4	*	*		*	*	232
5	*	*		*		113
6	*	*				222
7	*		*			11
8	*			*	*	38
9	*			*		102
10	*					302
					Total	$n = 1384$

*: Covariate data from the source are available.

Table 3. Candidate models and weights for predictions with $D^* =$ {Card, Shopping, Mobile, Bureau, Fraud}

Model	Δ_k	p_k	$\hat{w}_k$	$\hat{w}_k^{\text{zhang}}$
M_1	{Card, Shopping, Mobile, Bureau, Fraud}	26	0.023	0.035
M_2	{Card, Shopping, Mobile, Bureau}	22	0.028	0.019
M_3	{Card, Shopping, Mobile}	15	0.054	0.007
M_4	{Card, Shopping, Bureau, Fraud}	21	0.021	0.054
M_5	{Card, Shopping, Bureau}	17	0.023	0.142
M_6	{Card, Shopping}	10	0.244	0.044
M_7	{Card, Mobile}	11	0.055	0.001
M_8	{Card, Bureau, Fraud}	17	0.117	0.041
M_9	{Card, Bureau}	13	0.118	0.019
M_{10}	{Card}	6	0.316	0.638

Table 4. Candidate models and weights for predictions with $D^* =$ {Card, Shopping, Mobile, Bureau}

Model	Δ_k	p_k	$\hat{w}_k$
M_1	{Card, Shopping, Mobile, Bureau}	22	0.016
M_2	{Card, Shopping, Mobile}	15	0.025
M_3	{Card, Shopping, Bureau}	17	0.020
M_4	{Card, Shopping}	10	0.201
M_5	{Card, Mobile}	11	0.038
M_6	{Card, Bureau}	13	0.287
M_7	{Card}	6	0.412

To this end, the company gathered the real income data of $n =$ 1384 applicants to develop a prediction procedure. The response Y is defined as the logarithm of the income. The credit card information was available for all the applicants since it was required for application. Covariate data from the other four sources were available only for part of the applicants. There were a total of 10 patterns for the covariate data availability as shown in Table 2. Only 115 (8.3%) applicants had all the covariate data available, and 302 (21.8%) applicants only had covariate data from the first source. This is a very typical fragmentary dataset and our proposed method can be applied.

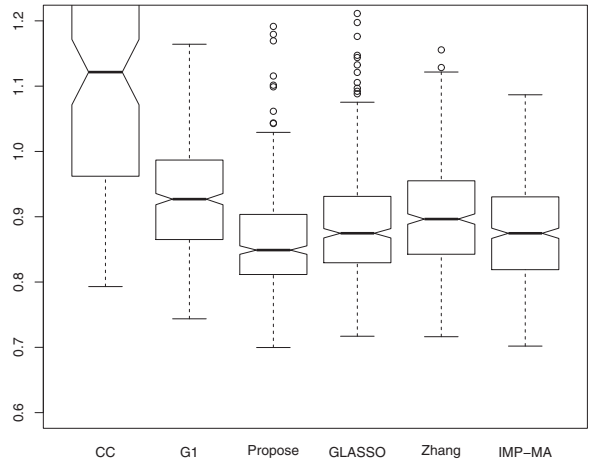

Figure 5. The prediction errors of each method after 500 replications for the real data.

To compare the prediction performances of the six methods studied in the first simulation, we randomly selected 50% of the applicants with response patterns $k = 1, \ldots, 10$ to form training data for model fitting, and use the rest of the applicants to form test data for performance evaluation.

For each of the six methods, we used the training data to fit the model, applied it to the test data, and computed the MSE of the predictions on test data. We repeated this procedure independently 500 replications.

For predictions with $D^* =$ {Card, Shopping, Mobile, Bureau, Fraud}, both the proposed method and Zhang's method considered 10 candidate models. For each candidate model M_k, Table 3 reports the covariate set Δ_k used to fit the model, number of covariates p_k, and the average weight ($\hat{w}_k$ for the proposed method and $\hat{w}_k^{\text{zhang}}$ for Zhang's method) over 500 replications. Zhang's method puts much more weight on M_{10} than the proposed method.

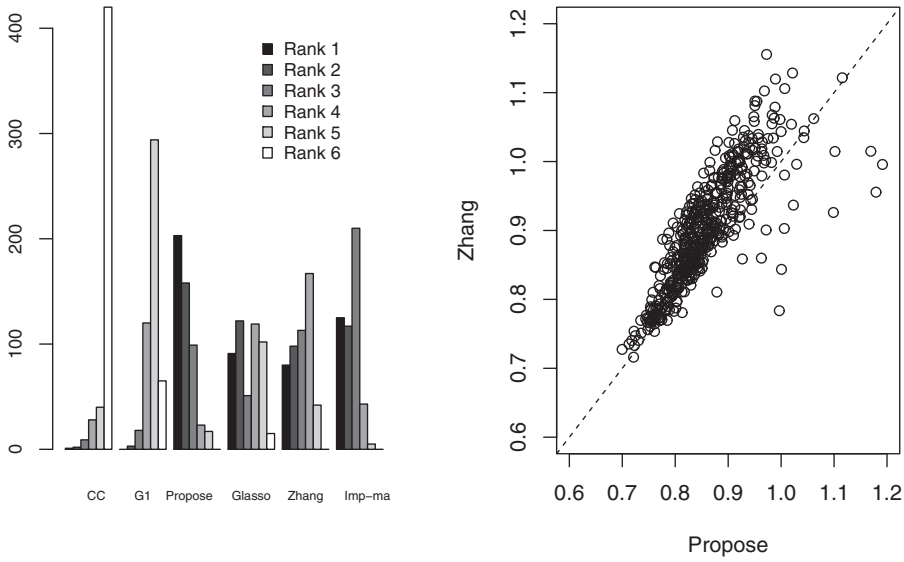

Figure 6. Left: absolute frequencies of ranks for the MSEs over 500 replications. Rank 1 means best. Right: comparison of the MSEs, proposed vs.. Zhang.

For predictions with $D^* \neq$ {Card, Shopping, Mobile, Bureau, Fraud}, Zhang's method still used the 10 candidate models and weights listed in Table 3 and replaced the unavailable covariate data by zeros when needed. The proposed method ignored the covariates not in D^* for modeling and prediction. For example, when $D^* =$ {Card, Shopping, Mobile, Bureau}, the covariates from "Fraud" were ignored and 7 candidate models were considered. Table 4 reports the corresponding Δ_k, p_k and average weight $\hat{w}_k$. Such results for the other D^*s are omitted for simplicity.

For the prediction performance, Figure 5 displays boxplots of the MSEs over 500 replications for each method. Overall speaking, the proposed method performs best among the six methods.

To obtain an idea of the performance in each individual replication, we ranked the MSEs of the six methods at each replication and display the ranks in the left panel of Figure 6. Among the 500 replications, the proposed method ranks 1 or 2 at 361 (72%) replications and never ranks 6. The right panel of Figure 6 compares the MSEs of the proposed method and Zhang's method. The proposed method has smaller MSE than Zhang's method at most of the replications.

6. CONCLUDING REMARKS

One main challenge for statistical prediction with data from multiple sources is that not all the associated covariate data are available for many sampled subjects. For this type of "fragmentary data," we propose a novel method based on frequentist model averaging that fits some candidate models using all available covariate data and selects optimal weights by delete-one cross-validation based on the data from complete cases. The weight selection method is proved to be asymptotically optimal. The proposed method shows superiority over several existing methods in simulations and a real-data analysis.

Although throughout the article we use the personal income prediction for online loan companies as the motivating and illustrating example, the proposed method has wide applications. First, fragmentary data are quite common nowadays especially when internet data is involved since internet users often provide their information in an arbitrary manner. For example, many companies try to predict people's behavior using their information in social websites such as LinkedIn or Twitter. However, different users of such websites usually have different covariates available. Second, although not discussed in details, the proposed method can be extended to binary response by replacing the linear models by logistic models. It has wide applications in credit risk management for the rapidly growing internet finance business in which companies need to predict whether the customers will default in the future but the available covariates are typically fragmentary. Also, it can be used for customer relationship management or marketing departments to predict whether the customers will like one product design or respond to a campaign. Third, an important potential application of the method is in individual participant data meta-analysis for medical studies. When combining patient data from different clinical trials, non-trivial proportions of missing covariate data are often encountered since different trials collect some common covariates as well as some different covariates. So far the available methods in meta-analysis all focus on imputation (Resche-Rigon et al. 2013; Jolani et al. 2015; Quartagno and Carpenter 2016). Our method provides a new idea to address this problem. Furthermore, internet data and individual participant data in meta-analysis are often heteroscedastic and our proposed method can handle it without explicitly modeling the heteroscedasticity.

ACKNOWLEDGMENTS

Fang Fang's research was partially supported by Shanghai Nature Science Foundation 15ZR1410300, Shanghai Rising Star Program (16QA1401700), National Scientific Foundation of China (11601156), and the 111 Project (B14019). Jun Shao's research was partially supported by the 111 Project (B14019) and the US National Science Foundation grant DMS-1305474.

[Received December 2016. Accepted September 2017.]

REFERENCES

Ando, T., and Li, K.-C. (2014), "A Model Averaging Approach for High Dimensional Regression," *Journal of American Statistical Association*, 109, 254–265. [520]

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [518]

Dardanoni, V., Luca, G. D., Modica, S., and Peracchi, F. (2015), "Model Averaging Estimation of Generalized Linear Models With Imputed Covariates," *Journal of Econometrics*, 184, 452–463. [518]

Dardanoni, V., Modica, S., and Peracchi, F. (2011), "Regression With Imputed Covariates: A Generalized Missing Indicator Approach," *Journal of Econometrics*, 162, 362–368. [518]

Gao, Y., Zhang, X., Wang, S., and Zou, G. (2016), "Model Averaging Based on Leave-Subject-Out Cross-Validation," *Journal of Econometrics*, 192, 139–151. [518]

Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [518,520]

Hansen, B. E., and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [518,519,520]

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators," *Journal of American Statistical Association*, 98, 879–899. [518]

Jolani, S., Debray, T. P. A., Koffijberg, H., Buuren, S. V., and Moons, K. G. M. (2015), "Imputation of Systematically Missing Predictors in an Individual Participant Data Meta-Analysis: A Generalized Approach Using MICE," *Statistics in Medicine*, 34, 1841–1863. [526]

Kim, J. K., and Shao, J. (2013), *Statistical Methods for Incomplete Data Analysis*, New York: Chapman & Hall. [518]

Li, K.-C. (1987), "Asymptotic Optimality for Cp, Cl, Cross-Validation and Generalized Cross-Validation: Discrete Index Sex," *The Annals of Statistics*, 15, 958–975. [518,519,520]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [518,521]

Liu, C.-A. (2015), "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, 186, 142–159. [518]

Longford, N. T. (2005), "Editorial: Model Selection and Efficiency is 'Which Model... ?' the Right Question?" *Journal of the Royal Statistical Society*, 168, 469–472. [519]

Quartagno, M., and Carpenter, J. R. (2016), "Multiple Imputation for IPD Meta-Analysis: Allowing for Heterogeneity and Studies With Missing Covariates," *Statistics in Medicine*, 35, 2938–2954. [526]

Resche-Rigon, M., White, I. R., Bartlett, J. W., Peters, S. A. E., Thompson, S. G. on behalf of the PROG-IMT Study Group (2013), "Multiple Imputation for Handling Systematically Missing Confounders in Meta-Analysis of Individual Participant Data," *Statistics in Medicine*, 32, 4890–4905. [526]

Schomaker, M., Wan, A. T. K., and Heumann, C. (2010), "Frequentist Model Averaging With Missing Observations," *Computational Statistics and Data Analysis*, 54, 3336–3347. [518]

Wan, A. T. K., Zhang, X., and Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283. [518,520]

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of American Statistical Association*, 96, 574–588. [518]

Yu, Z., Dong, Y. X., and Zhu, L. X. (2016), "Trace Pursuit: A General Framework for Model-Free Variable Selection," *Journal of American Statistical Association*, 111, 813–821. [521]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [522]

Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of American Statistical Association*, 100, 1202–1214. [518]

Zhang, X. (2013), "Model Averaging With Covariates That are Missing Completely at Random," *Economics Letters*, 121, 360–363. [518,519,520,521]

——— (2015), "Consistency of Model Averaging Estimators," *Economics Letters*, 130, 120–123. [518]

Zhang, X., and Liang, H. (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200. [518]

Zhang, X., Wan, A. T. K., and Zou, G. (2013), "Model Averaging by Jackknife Criterion in Models With Dependent Data," *Journal of Econometrics*, 174, 82–94. [518,520]

Zhang, X., Yu, D., Zou, G., and Liang, H. (2016), "Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models," *Journal of American Statistical Association*, 111, 1775–1790. [518]

Zhang, X., Zou, G., and Carroll, R. (2015), "Model Averaging Based on Kullback–Leibler Distance," *Statistica Sinica*, 25, 1583–1598. [518]

Zhong, W., Zhang, T., Zhu, M., and Liu, J. S. (2012), "Correlation Pursuit: Forward Stepwise Variable Selection for Index Models," *Journal of Royal Statistical Society*, Series B, 74, 849–870. [521]

Zou, H., and Zhang, H. H. (2009), "On the Adaptive Elastic-Net With a Diverging Number of Parameters," *The Annals of Statistics*, 37, 1733–1751. [521]