Multivariate Calibrations with Auxiliary Information

Pei Feng Tong[1], Song Xi Chen[1,2] and Cheng Yong Tang[3]

[1]*Guanghua School of Management, Peking University, Beijing 100871, China*

[2] *Center for Statistical Science, Peking University, Beijing 100871, China*

[3]*Department of Statistics, Operations and Data Science,*

Temple University, Philadelphia, PA 19122-6083, USA

Abstract:

We investigate multivariate calibrations from a modern perspective with a focus on incorporating auxiliary variables and handling complex data dependencies with random effects. By introducing auxiliary variables, the roles of the variables in multivariate calibration problems are no longer restricted to being either response or explanatory, which offers much flexibility and adaptability to a broader range of practical problems. Our analysis reveals that a new shrinkage approach, that connects the conventional generalized least squares and the inverse regression approaches, offers much improved performance. To accommodate complex dependence in contemporary studies, we develop a computationally efficient expectation-maximization algorithm for solving multivariate calibration problems with random effects. The shrinkage approach shows promising performance in numerical simulations and an empirical study.

Key words and phrases: Inverse regression; Linear mixed-effect models; Multi-

variate calibration; Multivariate response variables; Shrinkage estimation.

1. Introduction

Regression analysis is a class of foundational statistical tools for modeling and predictions. In regression analysis, the response variables are modeled by a collection of explanatory variables. Practically, interests may arise from the opposite direction, concerning some or all of the variables whose roles are explanatory rather than response in some established regression models. Such problems, referred to as calibrations or inverse predictions in the literature, are often seen in practical investigations in areas including economics, sociology, earth sciences and analytical chemistry (Marden et al., 2018; Yun et al., 2019; Wei et al., 2021). In this study, the term "calibration" particularly refers to inversely predicting some or all of the explanatory variables in the framework of regression models. It is worth noting that the term "calibration" is employed in other contexts such as to describe the adjustment that correcting the coverage probabilities in interval estimations and the alignment of parameters in generic models like regression quantiles (Kuleshov et al., 2018; Fasiolo et al., 2021).

Multivariate calibration problems have been extensively investigated in the literature; see Sundberg (1999) for an overview. Two kinds of ap-

proaches – the generalized least squares (GLS) estimator and the inverse regression (IR) estimator – are popular, and they are often compared to each other (Krutchkoff, 1967; Hoadley, 1970; Brown, 1982; Sundberg, 1985).

In this study, we present a novel approach to multivariate calibrations and inverse regression through a renewed methodological framework. Our approach involves augmenting the regression model with additional variables to increase flexibility and estimation accuracy, and enabling a wider range of practical problems. Additionally, we integrate multivariate calibrations with random effect models to account for complex and structured data dependencies such as repeated measurements and clustered data. This integration provides a more comprehensive framework for solving practical problems and enhances its applicability to the real-world settings.

Our setting is commonly encountered in practice, such as in the fields of healthcare and chemometrics. For instance, in the analysis of blood samples, the response variables near infrared spectra (NIR) (Jiang et al., 2020) are explained by a set of variables, including the blood sugar levels, temperature, and others. While some of these explanatory variables, such as the blood sugar level, may be of central interest in certain studies, accurate measurements may be difficult to obtain due to cost or patient availability constraints, thus motivating interest in the opposite direction

of the previous regression model. In contrast, the measurements of NIR of blood samples are often more readily available, albeit with a lower level of accuracy. Moreover, a set of additional variables may be available during the data collection process such as the temperature of the samples or surrounding environmental condition, which is suitable as auxiliary variables.

Practical investigations often yield data that exhibit structured correlations. For instance, in the health care example mentioned earlier, measurements from the same patient are likely to be repeated, and data may be clustered due to measurements taken in the same lab. Incorporating such data dependencies is a challenging task, and increased attention is being given to this issue in fields such as spatial econometrics and statistics (Katzfuss and Cressie, 2011; MacKinnon et al., 2023). Our new development in multivariate calibrations, which incorporates random effects, offers a promising approach to addressing these challenges.

Our investigation makes several contributions. Methodologically, we first demonstrate that in a cross-sectional setting, the inverse regression estimator is a shrinkage version of the generalized least squares estimator, which was previously unknown. Furthermore, we show that the IR estimator can be treated as a Bayesian estimator with appropriate priors. Building on these findings, we propose an optimal shrinkage estimator

that minimizes the limiting mean squared errors. These results provide a fresh perspective on conventional multivariate calibrations and inverse regressions. To address the computational challenges associated with estimating the parameters of the random effects, we develop a highly efficient expectation-maximization algorithm. Our algorithm is capable of handling considerably large-scale problems. Together, these contributions enhance our understanding of multivariate calibrations and inverse regressions and provide a powerful tool for researchers and practitioners.

By treating the target variables in our study as missing, our study connects to the area of missing data analysis. The methodology using imputations is commonly applied upon assuming some models respectively on the data generation process (DGP) and the data missingness, e.g, missing-at-random (MAR); see (Little and Rubin, 2002). When MAR assumption does not hold, one may resort to some specific model settings on the DGP and data missingness; see, for example, Hernández-Lobato et al. (2014). Recently, imputation methodology has been used in conjunction with generative machine learning techniques, and have shown promising performance in some settings with MAR; see Jarrett et al. (2022) and the reference therein. Different from the approaches in typical missing data analysis, our study does not necessitate the specification of the mechanism govern-

ing data missingness. Instead, a pivotal aspect of our study is rooted in exploiting the specification of regression models, accommodating the aforementioned practical considerations.

This paper is organized as follows. Section 2 highlights the multivariate calibration model and the main results under homogenous and heterogeneous scenarios. Section 3 presents model performance in simulation. Section 4 gives a case study and Section 5 is the conclusion. Further technical details are relegated to the supplementary material (SM).

2. Methods

2.1 Setting

We adopt a notation system that uses $\boldsymbol{A}$ for a matrix, $\boldsymbol{a}$ for a vector, and a for a scalar; $\boldsymbol{A}^T$ denotes the transpose of $\boldsymbol{A}$. For a matrix $\boldsymbol{A}$, we denote by $\mathbb{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ and $\mathbb{M}_{\boldsymbol{A}} = \boldsymbol{I} - \mathbb{P}_{\boldsymbol{A}}$ be the projection matrices to its column space and the orthogonal complement, respectively.

Suppose that we have a sample of n observations $\{(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^n$, where $\boldsymbol{y}_i \in \mathbb{R}^q$ is the multivariate response with covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{z}_i \in \mathbb{R}^{p'}$. We consider the *observation model*

$$\boldsymbol{y}_i = \boldsymbol{\alpha} + \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{D}^T\boldsymbol{z}_i + \boldsymbol{e}_i, \tag{2.1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^q$, $\boldsymbol{B} \in \mathbb{R}^{p\times q}$, $\boldsymbol{D} \in \mathbb{R}^{p'\times q}$ are the parameters of appropriate sizes, $\boldsymbol{e}_i \in \mathbb{R}^q$ is the model error. We use the notation $\boldsymbol{z}_i$ to represent the auxiliary variables, differentiating their role from that of $\boldsymbol{x}_i$, which typically serves as the explanatory variable in conventional calibration problems.

In a calibration problem, repeated measurements $\{\boldsymbol{y}'_j\}_{j=1}^l$ are available together with only $\{\boldsymbol{z}'_j\}_{j=1}^l$. The un-observable counterpart of $\boldsymbol{x}'_j$ is considered as shared by these l measurements, and is considered as a standalone parameter denoted by $\boldsymbol{\xi} \in \mathbb{R}^p$. Then, one attempts to work with the *calibration model* of the sames structure as (2.1)

$$\boldsymbol{y}'_j = \boldsymbol{\alpha} + \boldsymbol{B}^T\boldsymbol{\xi} + \boldsymbol{D}^T\boldsymbol{z}'_j + \boldsymbol{e}'_j. \tag{2.2}$$

In a conventional setting, (2.1) is cross-sectional with n independent replications with no $\boldsymbol{z}_j$, while (2.2) allows l repeated measurements from the same subject, to recover the un-observable variables with accuracy. We begin with this conventional cross-sectional setting and then extend the calibration problem to cover dependent $\boldsymbol{y}_j$s as shown in Section 2.4.

Collectively, we write the two models in their matrix forms as

$$\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\alpha}^T + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{D} + \boldsymbol{E}, \tag{2.3}$$

$$\boldsymbol{Y}' = \mathbf{1}_l \boldsymbol{\alpha}^T + \mathbf{1}_l \boldsymbol{\xi}^T \boldsymbol{B} + \boldsymbol{Z}'^T \boldsymbol{D} + \boldsymbol{E}', \tag{2.4}$$

where $\mathbf{1}_m \in \mathbb{R}^m$ is a vector of 1s, $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T \in \mathbb{R}^{n \times q}$, $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{Y}', \boldsymbol{Z}', \boldsymbol{E}'$, of appropriate dimensions, are defined in the same fashion.

We make the following assumptions throughout the analysis.

Assumption 1. (i). The residuals $\{\boldsymbol{e}_i\}_{i=1}^n$, $\{\boldsymbol{e}'_j\}_{j=1}^l$ are IID satisfying $\mathbb{E}(\boldsymbol{e}_i) = \mathbb{E}(\boldsymbol{e}'_j) = 0$, $\mathbb{E}(\boldsymbol{e}_i \boldsymbol{e}_i^T) = \mathbb{E}(\boldsymbol{e}'_j \boldsymbol{e}_j'^T) = \boldsymbol{\Gamma}$ for a $q \times q$ positive definite matrix $\boldsymbol{\Gamma}$, and $E||\boldsymbol{e}_i||_2^{2+\delta} = E||\boldsymbol{e}_j||_2^{2+\delta} < \infty$ for a $\delta > 0$.

(ii). The design matrix $[\mathbf{1}_n, \boldsymbol{X}, \boldsymbol{Z}]$ has full column rank, and without loss of generality, each column of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are standardized so that $\sum_i x_{ij} = \sum_i z_{ij'} = 0$ and $n^{-1}\sum_i x_{ij}^2 = n^{-1}\sum_i z_{ij'}^2 = 1$, for $j = 1, \ldots, p$ and $j' = 1, \ldots, p'$.

2.2 Existing methods

The setting with no auxiliary variables has been investigated in the literature, i.e., when $\boldsymbol{z}_i$ and $\boldsymbol{z}'_i$ are not present. Let $\bar{\boldsymbol{y}}' = (l^{-1}\mathbf{1}_l^T \boldsymbol{Y}')^T$. The so-called generalized least squares (Brown, 1993) estimator is

$$\hat{\boldsymbol{\xi}}_{gls} = (\hat{\boldsymbol{B}}\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{B}}^T)^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{\Gamma}}^{-1}(\bar{\boldsymbol{y}}' - \hat{\boldsymbol{\alpha}})$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}$ are the ordinary least squares (OLS) estimators of (2.3) with a null $\boldsymbol{Z}$, $\hat{\boldsymbol{\Gamma}}$ is some weight matrix.

By reversely regressing $\boldsymbol{X}$ on $\boldsymbol{Y}$ based on a working model $\boldsymbol{X} = \mathbf{1}_n\boldsymbol{\theta}^T + \boldsymbol{Y}\boldsymbol{\Phi} + \tilde{\boldsymbol{E}}$ with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$, the inverse regression estimator is:

$$\hat{\boldsymbol{\xi}}_{ir} = \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\Phi}}^T\bar{\boldsymbol{y}}' \tag{2.5}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Phi}}$ are the OLS estimators. It is known that upon assuming Gaussian distributions, $\hat{\boldsymbol{\xi}}_{ir}$ is a Bayesian estimator with appropriate priors on the unknown model parameters; see Hoadley (1970) and Brown (1982).

As a new result, an implication from our Theorem 1 is that $\hat{\boldsymbol{\xi}}_{ir}$ is a shrinkage of $\hat{\boldsymbol{\xi}}_{gls}$ in that

$$\hat{\boldsymbol{\xi}}_{ir} = (\boldsymbol{I} + \boldsymbol{H}_1)^{-1}\hat{\boldsymbol{\xi}}_{gls}$$

where $\boldsymbol{H}_1 = (\hat{\boldsymbol{B}}\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{B}}^T)^{-1}(n^{-1}\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is a positive definite matrix.

In random designs when the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ is normal, both conditional means of $\boldsymbol{X}|\boldsymbol{Y}$ and $\boldsymbol{Y}|\boldsymbol{X}$ are linear functions. This means that their roles as response or explanatory variables are exchangeable in a linear model; so that both the GLS and inverse regression estimators are valid. Furthermore, the inverse regression estimator is found empirically

performing better when $l = 1$ as first discussed by Krutchkoff (1967). Additionally, when the design of $\boldsymbol{X}$ is fixed, Sundberg (1985) showed that the GLS estimator $\hat{\boldsymbol{\xi}}_{gls}$ often performs worse than $\hat{\boldsymbol{\xi}}_{ir}$, except when $\boldsymbol{\xi}$ is far from the mean of $\boldsymbol{X}$ in the observation model.

2.3 Main results

Our aim is generalizing GLS and inverse regressions to a new setting with auxiliary variables $\boldsymbol{Z}$. To this end, let $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{B}}$, and $\hat{\boldsymbol{D}}$ be the OLS estimators of (2.3). Then, we show that the GLS estimator that minimizes the weighted squared error loss of the calibration model (2.4) is

$$\hat{\boldsymbol{\xi}}_{gls} = (\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}\hat{\boldsymbol{B}}^T)^{-1}\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}}' - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{D}}^T\bar{\boldsymbol{z}}'), \tag{2.6}$$

where $\bar{\boldsymbol{z}}' = (l^{-1}\mathbf{1}_l^T\boldsymbol{Z}')^T$, $\boldsymbol{S} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$, and $\hat{\boldsymbol{Y}} = \mathbf{1}_n\hat{\boldsymbol{\alpha}}^T + \boldsymbol{X}\hat{\boldsymbol{B}} + \boldsymbol{Z}\hat{\boldsymbol{D}}$ is the fitted/predicted responses of the model (2.3).

Let $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{\Psi}}$ be the OLS estimators of a working model $\boldsymbol{X} = \mathbf{1}_n\boldsymbol{\theta}^T + \boldsymbol{Y}\boldsymbol{\Phi} + \boldsymbol{Z}\boldsymbol{\Psi} + \tilde{\boldsymbol{E}}$. The IR estimator is then

$$\hat{\boldsymbol{\xi}}_{ir} = \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\Phi}}^T\bar{\boldsymbol{y}}' + \hat{\boldsymbol{\Psi}}^T\bar{\boldsymbol{z}}'. \tag{2.7}$$

We need the following assumptions on the prior distributions to estab-

lish a Bayesian connection to $\hat{\boldsymbol{\xi}}_{ir}$.

Assumption 2. (i). The parameters in (2.3) follow the noninformative invariant Jefferys prior, namely $P(\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Gamma}) \propto |\boldsymbol{\Gamma}|^{-(q+1)/2}$.

(ii). The prior distribution of $\boldsymbol{\xi}$ given $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ is the multivariate t-distribution $T_{\nu-p}\big(\frac{1}{2}\boldsymbol{C}_1^{-1}\boldsymbol{c}_2, \frac{1}{\nu-p}(l^{-1}+n^{-1}+c_3-\frac{1}{4}\boldsymbol{c}_2^T\boldsymbol{C}_1^{-1}\boldsymbol{c}_2)\boldsymbol{C}_1^{-1}\big)$, where $\nu = n - p - p' - q$, $\boldsymbol{C}_1 = (\boldsymbol{X}^T\mathbb{M}_z\boldsymbol{X})^{-1}$, $\boldsymbol{c}_2 = ((\boldsymbol{X}^T\mathbb{M}_z\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\mathbb{M}_x\boldsymbol{Z})^{-1})\bar{\boldsymbol{z}}'$ and $c_3 = \bar{\boldsymbol{z}}'^T(\boldsymbol{Z}^T\mathbb{M}_x\boldsymbol{Z})^{-1}\bar{\boldsymbol{z}}'$.

(iii). The priori of $\boldsymbol{\xi}$ is conditionally independent of the other parameters given $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{Z}'$, namely $P(\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Gamma}, \boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Z}') = P(\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Gamma}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Z}') \times P(\boldsymbol{\xi}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Z}')$.

(iv). The residuals $\{\boldsymbol{e}_i\}_{i=1}^n$ and $\{\boldsymbol{e}'_j\}_{j=1}^l$ in (2.1) and (2.2) are IID from $\mathcal{N}(0, \boldsymbol{\Gamma})$, respectively.

Let $\boldsymbol{H}_{ir} = (\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}\hat{\boldsymbol{B}}^T)^{-1}(\boldsymbol{X}^T\mathbb{M}_Z\boldsymbol{X})^{-1}$ and $\boldsymbol{\zeta} = \boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\bar{\boldsymbol{z}}'$. Our main results on the properties of the GLS and inverse regression estimators, and a new shrinkage estimator are given in the following theorems.

Theorem 1. *(i) Under Assumption 1,*

$$\hat{\boldsymbol{\xi}}_{ir} = (\boldsymbol{I} + \boldsymbol{H}_{ir})^{-1}(\hat{\boldsymbol{\xi}}_{gls} - \boldsymbol{\zeta}) + \boldsymbol{\zeta}. \tag{2.8}$$

(ii) Under Assumptions 1 and 2, $\hat{\boldsymbol{\xi}}_{ir}$ is the Bayesian estimator, namely the posterior mean that minimizes the Bayesian risk.

Here, $\boldsymbol{\zeta}$ is the OLS estimate of $\boldsymbol{\xi}$ using a marginal model with only $\boldsymbol{X}$ and $\boldsymbol{Z}$. It serves as a baseline candidate for predicting the unobserved $\boldsymbol{\xi}$ using only the auxiliary variables, and we assume it to be the mean of the prior distribution of $\boldsymbol{\xi}$. Remarkably, $\hat{\boldsymbol{\xi}}_{ir}$ is seen as a shrinkage estimator towards $\boldsymbol{\zeta}$. In the special case when $\boldsymbol{Z}$ is null (the dimension of $\boldsymbol{Z}$ is 0), Theorem 1 implies (2.5), which is a new finding elucidating the connections between the commonly used GLS and IR approaches.

We give the key properties of a class of shrinkage estimators as follows.

Theorem 2. *For the class of estimators*

$$\hat{\boldsymbol{\xi}} = (\boldsymbol{I} + \boldsymbol{H})^{-1}(\hat{\boldsymbol{\xi}}_{gls} - \boldsymbol{\zeta}) + \boldsymbol{\zeta} \tag{2.9}$$

with a positive definite $\boldsymbol{H} \in \mathbb{R}^{p\times p}$, under Assumption 1, as $n \to \infty$,

$$\|\mathbb{E}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}) - \{(\boldsymbol{I} + \boldsymbol{H})^{-1} - \boldsymbol{I}\}(\boldsymbol{\xi} - \boldsymbol{\zeta})\|_2 \to 0,$$

$$\|\mathrm{var}(\hat{\boldsymbol{\xi}}) - l^{-1}(\boldsymbol{I} + \boldsymbol{H})^{-1}(\boldsymbol{B}\boldsymbol{\Gamma}^{-1}\boldsymbol{B}^T)^{-1}(\boldsymbol{I} + \boldsymbol{H}^T)^{-1}\|_2 \to 0,$$

where $\boldsymbol{\Gamma}$ is the variance-covariance matrix of $\boldsymbol{e}_i$ in (2.1).

When $\boldsymbol{H} = 0$, $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_{gls}$; Theorem 2 implies that the GLS estimator is asymptotically unbiased. When $\boldsymbol{H} = \boldsymbol{H}_{ir}$, $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_{ir}$; Theorem 2 gives the bias and variance of the inverse regression estimator.

Theorem 2 shows that the shrinkage estimators benefit from a reduced variance at a cost of the associated bias. This trade-off is more beneficial when l is smaller, resulting in a greater reduction in variance. This phenomenon helps to explain the better performance of the inverse regression estimators observed in previous studies.

From Theorem 2, as $n \to \infty$, the limiting predictive mean square error (MSE) of the $\hat{\boldsymbol{\xi}}$ is

$$G(\boldsymbol{\xi}, \boldsymbol{\zeta}) = \|\{(\boldsymbol{I}+\boldsymbol{H})^{-1}-\boldsymbol{I}\}(\boldsymbol{\xi}-\boldsymbol{\zeta})\|_2^2+l^{-1}\mathrm{tr}\{(\boldsymbol{I}+\boldsymbol{H})^{-1}(\boldsymbol{B}\boldsymbol{\Gamma}^{-1}\boldsymbol{B}^T)^{-1}(\boldsymbol{I}+\boldsymbol{H}^T)^{-1}\}.$$

Upon treating the design as random, we define the averaged MSE as

$$\mathrm{AMSE} = \mathbb{E}\{G(\boldsymbol{\xi}, \boldsymbol{\zeta})\}, \tag{2.10}$$

where the expectation is taken with respect to the joint distribution of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\xi}, \boldsymbol{Z}')$. Inspired by Theorem 2, we propose to find the optimal shrinkage estimator that minimizes the AMSE as given next.

Theorem 3. *If $\boldsymbol{\xi}$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent and identically distributed (IID) from a distribution having finite $(2+\delta)$-th moment for a $\delta > 0$, let the covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Z}$ be $\boldsymbol{\Gamma}_x$, $\boldsymbol{\Gamma}_z$ and $\boldsymbol{\Gamma}_{xz}$, under Assumption 1, the AMSE (2.10) is minimized at*

$$\tilde{\boldsymbol{H}}_{opt} = l^{-1}(\boldsymbol{B}\boldsymbol{\Gamma}^{-1}\boldsymbol{B}^T)^{-1}(\boldsymbol{\Gamma}_x - \boldsymbol{\Gamma}_{xz}\boldsymbol{\Gamma}_z^{-1}\boldsymbol{\Gamma}_{xz}^T)^{-1}.$$

Furthermore, as $n \to \infty$,

$$\boldsymbol{H}_{opt} = l^{-1}(\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}\hat{\boldsymbol{B}}^T)^{-1}(\boldsymbol{X}^T\mathbb{M}_Z\boldsymbol{X})^{-1}$$

is consistent to $\tilde{\boldsymbol{H}}_{opt}$.

Here $\tilde{\boldsymbol{H}}_{opt}$ in Theorem 3 is an unknown hyper-parameter; and we propose to implement its empirical counterpart $\boldsymbol{H}_{opt}$ for empirically obtaining the optimal shrinkage estimator:

$$\hat{\boldsymbol{\xi}}_{opt} = (\boldsymbol{I} + \boldsymbol{H}_{opt})^{-1}(\hat{\boldsymbol{\xi}}_{gls} - \boldsymbol{\zeta}) + \boldsymbol{\zeta}. \tag{2.11}$$

In summary, the relationship between these estimators is depicted in Figure 1, which shows the spectrum of trade-offs between bias and variance in the class of shrinkage estimators.

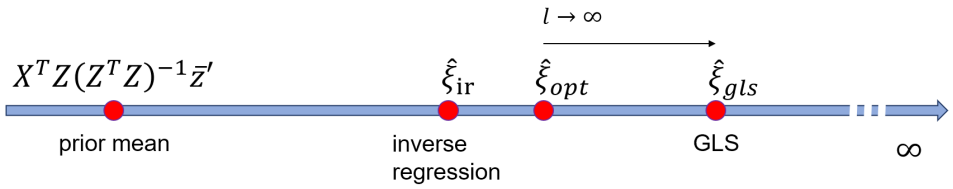

Figure 1: The relationship among GLS, inverse regression and optimal shrinkage estimators. When $l = 1$, the optimal shrinkage estimator coincides with the inverse regression estimator. The optimal shrinkage estimator converges to the GLS estimator as $l \to \infty$.

In the classical calibration setting where there is no $\boldsymbol{Z}$, the shrinkage interpretation of the inverse regression approach presented in our Theorem 1 remains valid; the optimal shrinkage estimator, developed based on our Theorem 3, remains applicable. As shown in the next theorem, incorporating auxiliary variables $\boldsymbol{Z}$ indeed provides an opportunity for improving the accuracy in predicting $\boldsymbol{\xi}$.

Theorem 4. *If $\boldsymbol{\xi}$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ are IID from a distribution having finite $2+\delta$-th moment for a $\delta > 0$, then*

(i) As $n \to \infty$, $\hat{\boldsymbol{\xi}}_{opt}$ minimizes the AMSE (2.10) among all estimators of the form $\hat{\boldsymbol{\xi}} = (\boldsymbol{I} + \boldsymbol{H})^{-1}(\hat{\boldsymbol{\xi}}_{gls} - \boldsymbol{\zeta}) + \boldsymbol{\zeta}$;

(ii) The AMSE (2.10) is non-increasing upon including $\boldsymbol{Z}$ in Models (2.3) and (2.4).

Remarkably, as shown in theorem 5, the optimal shrinkage estimator,

the GLS and IR estimators link to a class of the penalized estimations:

$$\min_{\boldsymbol{\xi}} \operatorname{tr}[\{\boldsymbol{Y}' - \hat{\boldsymbol{Y}}'(\boldsymbol{\xi})\}\boldsymbol{S}^{-1}\{\boldsymbol{Y}' - \hat{\boldsymbol{Y}}'(\boldsymbol{\xi})\}^T] + \lambda\sigma^2(\boldsymbol{\xi}), \tag{2.12}$$

where $\hat{\boldsymbol{Y}}'(\boldsymbol{\xi}) = \mathbf{1}_l\hat{\boldsymbol{\alpha}}^T + \mathbf{1}_l\boldsymbol{\xi}^T\hat{\boldsymbol{B}} + \boldsymbol{Z}'\hat{\boldsymbol{D}}$, $\sigma^2(\boldsymbol{\xi}) = 1/l + 1/n + \boldsymbol{\xi}^T\boldsymbol{C}_1\boldsymbol{\xi} - \boldsymbol{c}_2^T\boldsymbol{\xi} + c_3$, with $\boldsymbol{C}_1$, $\boldsymbol{c}_2$ and c_3 defined in Assumption 2 (b).

Theorem 5. *Under Assumption 1, (i)* $\hat{\boldsymbol{\xi}}_{ir}$ *is the solution for* $\lambda = l$*; (ii)* $\hat{\boldsymbol{\xi}}_{opt}$ *is the solution for* $\lambda = 1$*; (iii)* $\hat{\boldsymbol{\xi}}_{gls}$ *is the solution for* $\lambda = 0$.

2.4 Calibrations with random effects

So far, our method treats the n measurements in $\boldsymbol{Y}$ of model (2.3) as independent. Practically, it is realistic that there are some correlations between them. For example, repeated or clustered measurements are common. For incorporating such scenarios, we extend the framework to accommodate the mixed or random effects in handling broader situations.

Specifically, we consider an extended observation model

$$\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\alpha}^T + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{D} + \boldsymbol{\Psi}\boldsymbol{A} + \boldsymbol{E}, \tag{2.13}$$

where $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)^T \in \mathbb{R}^{k\times q}$, $\boldsymbol{a}_j \in \mathbb{R}^q$ $(j = 1, \ldots, k)$ are k random effects, and $\boldsymbol{\Psi} \in \mathbb{R}^{n\times k}$ is a known model matrix.

Corresponding to (2.2), the calibration model with random effect is

$$\boldsymbol{y}'_j = \boldsymbol{\alpha} + \boldsymbol{B}^T\boldsymbol{\xi} + \boldsymbol{D}^T\boldsymbol{z}'_j + \boldsymbol{A}^T\boldsymbol{\psi}' + \boldsymbol{e}'_j \quad \text{for } j = 1, 2, \ldots l, \tag{2.14}$$

where the extra $\boldsymbol{A}^T\boldsymbol{\psi}'$ reflects the reality that some of the random effects may be shared as reflected by the design $\boldsymbol{\psi}'$.

The random effect $\boldsymbol{A}$ and residual $\boldsymbol{E}$ are assumed to follow the Gaussian assumption as follows.

Assumption 3. The random effect $\{\boldsymbol{a}_i^T\}_{i=1}^k$ are IID generated from $\mathcal{N}(0, \boldsymbol{\Gamma}_1)$, while the residuals $\{\boldsymbol{e}_i\}_{i=1}^n$, $\{\boldsymbol{e}'_j\}_{j=1}^l$ are IID from $\mathcal{N}(0, \boldsymbol{\Gamma}_2)$. In addition, we assume that $\boldsymbol{A}$ and $\{\boldsymbol{e}_i\}_{i=1}^n$, $\{\boldsymbol{e}'_j\}_{j=1}^l$ are mutually independent.

The vector form of (2.13) is

$$\text{vec}(\boldsymbol{Y}^T) = \text{vec}(\boldsymbol{\alpha}\boldsymbol{1}_n^T) + \text{vec}(\boldsymbol{B}^T\boldsymbol{X}^T) + \text{vec}(\boldsymbol{D}^T\boldsymbol{Z}^T) + (\boldsymbol{\Psi}\otimes\boldsymbol{I})\text{vec}(\boldsymbol{A}^T) + \text{vec}(\boldsymbol{E}^T). \tag{2.15}$$

Model (2.13) together with Assumptions 1 and 3 imply that

$$\text{var}\{\text{vec}(\boldsymbol{Y}^T)\} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T \otimes \boldsymbol{\Gamma}_1 + \boldsymbol{I}_n \otimes \boldsymbol{\Gamma}_2.$$

Both $\boldsymbol{A}$ and $\boldsymbol{\Psi}$ can be constructed, adapting to various practical situations.

For example, observations $\{(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^n$ may contain measurements from k clusters. Then $\boldsymbol{a}_j$ is specified as the jth ($j = 1, \ldots, k$) cluster-specific random effect, and $\boldsymbol{\Psi}$ is set as a matrix whose ith row ($i = 1, \ldots, n$) identifies the corresponding cluster where the measurement is taken from. In spatial statistics, the design matrix can be constructed upon using some basis function reflecting broad cluster effects such as from locations, e.g., the inverse of Euclidean distance from the location of the i-th observation to a given point; see Cressie and Johannesson (2008).

The unknown parameters of (2.13) are $\boldsymbol{\alpha}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$. Directly maximizing the likelihood is known to be computationally hard. To meet this challenge, we develop an expectation-maximization (EM) algorithm that is computationally highly efficient, whose detail is given in SM A.

Let $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\Gamma}}_1$, and $\hat{\boldsymbol{\Gamma}}_2$ be the estimators using the EM algorithm. Then, the GLS estimator incorporating the random effects is

$$\hat{\boldsymbol{\xi}}_{gls} = (\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}\hat{\boldsymbol{B}}^T)^{-1}\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}}' - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{D}}^T\bar{\boldsymbol{z}}' - \hat{\boldsymbol{A}}^T\boldsymbol{\psi}'), \tag{2.16}$$

where $\hat{\boldsymbol{A}} = \mathbb{E}(\boldsymbol{A}^T|\boldsymbol{Y}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\Gamma}}_1, \hat{\boldsymbol{\Gamma}}_2)$ and

$$\boldsymbol{S} = l^{-1}\hat{\boldsymbol{\Gamma}}_2 + \boldsymbol{\psi}'^T\boldsymbol{\psi}' \otimes \hat{\boldsymbol{\Gamma}}_1 - (\boldsymbol{\psi}'^T\boldsymbol{\Psi}^T \otimes \hat{\boldsymbol{\Gamma}}_1)(\boldsymbol{\Psi}\boldsymbol{\Psi}^T \otimes \hat{\boldsymbol{\Gamma}}_1 + \boldsymbol{I}_n \otimes \hat{\boldsymbol{\Gamma}}_2)^{-1}(\boldsymbol{\Psi}\boldsymbol{\psi}' \otimes \hat{\boldsymbol{\Gamma}}_1)$$

is the conditional variance of $\bar{\boldsymbol{y}}' - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{D}}^T \bar{\boldsymbol{z}}' - \hat{\boldsymbol{A}}\boldsymbol{\psi}'$ given $(\boldsymbol{Y}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\Gamma}}_1, \hat{\boldsymbol{\Gamma}}_2)$.

Inspired by (2.11), we can develop an optimal shrinkage estimator as

$$\hat{\hat{\boldsymbol{\xi}}}_{opt} = (\boldsymbol{I} + \boldsymbol{H})^{-1} \hat{\hat{\boldsymbol{\xi}}}_{gls}, \tag{2.17}$$

where $\boldsymbol{H} = (\hat{\boldsymbol{B}}\boldsymbol{S}^{-1}\hat{\boldsymbol{B}}^T)^{-1}(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The detailed derivations of (2.17) are left in Section M of the SM.

Furthermore, we can develop an IR approach based upon a working random effect model

$$\text{vec}(\boldsymbol{X}^T) = \text{vec}(\boldsymbol{\omega}\boldsymbol{1}_n^T) + \text{vec}(\boldsymbol{\Phi}^T\boldsymbol{Y}^T) + \text{vec}(\boldsymbol{\Omega}^T\boldsymbol{Z}^T) + (\boldsymbol{\Psi} \otimes \boldsymbol{I}_p)\text{vec}(\boldsymbol{A}^T) + \text{vec}(\boldsymbol{E}^T).$$

Upon estimating the unknown parameters using the proposed EM algorithm in SM, an IR estimator can be developed as

$$\hat{\hat{\boldsymbol{\xi}}}_{ir} = \hat{\boldsymbol{\omega}} + \hat{\boldsymbol{\Phi}}^T \bar{\boldsymbol{y}}' + \hat{\boldsymbol{\Omega}}^T \bar{\boldsymbol{z}}' + \hat{\boldsymbol{A}}^T \boldsymbol{\psi}'. \tag{2.18}$$

With extra parameters involved, the connections between the GLS, IR, and Bayesian estimator become more complicated. So, the implications from Theorem 5 no longer hold, due to incorporating the random effects.

The biases of the calibrations in this setting with random effects may

be of interests. Here we can see that the limiting bias of $\hat{\boldsymbol{\xi}}_{gls}$ goes to 0 as $n \to \infty$, provided that the model (2.13) is correctly specified so that the parameters are estimated consistently by the EM algorithm; see Balakrishnan et al. (2017). As a result, the limiting bias of $\hat{\boldsymbol{\xi}}_{opt}$ is dominated by the shrinkage effect introduced by $(\boldsymbol{I} + \boldsymbol{H})^{-1}$.

3. Simulations

3.1 A baseline setting

We evaluated and compared the performances of the three estimators $\hat{\boldsymbol{\xi}}_{gls}$, $\hat{\boldsymbol{\xi}}_{ir}$, and $\hat{\boldsymbol{\xi}}_{opt}$ by simulations. As a benchmark, we implemented a random forest approach for comparison (Breiman, 2001). We set dimensions $q = 10$, $p = 2$, $p' = 10$ for Models (2.3) and (2.4). To compare the methods, we used the prediction MSE from a 10-fold cross-validation as the criterion. For each setting, we repeated the simulation 1,000 times to calculate the mean and standard deviation of the prediction MSE.

We generated N individuals $\{\boldsymbol{x}_i\}_{i=1}^N$, $\{\boldsymbol{z}_i\}_{i=1}^N$ and coefficients $\boldsymbol{\alpha}$, $\boldsymbol{B}$, $\boldsymbol{D}$ from the standard normal distribution. For each observation i, we then built the l repeated measurements of $\{\boldsymbol{z}_{ij}\}_{j=1}^l$ by adding l independent disturbances following $\mathcal{N}(0, \sigma_z^2 \boldsymbol{I}_{p'})$ to $\boldsymbol{z}_i$ with $\sigma_z = 0.1$. Then, $\{\boldsymbol{y}_{ij}\}_{j=1}^l$ were

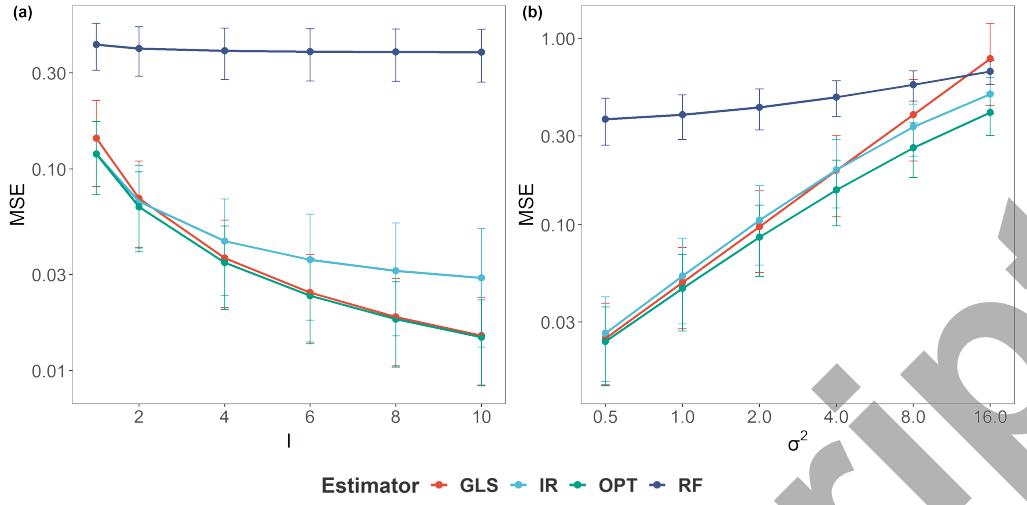

Figure 2: Empirical means and the 10th, 90th quantiles of the MSEs (in log-scale) calculated from 1000 simulation replications for the GLS, the inverse regression (IR), the optimal shrinkage (OPT) and the inverse regression with the random forest (RF) estimators with respect to numbers of repeated measurements l (a) and the variances of the noises σ^2 (b).

generated from

$$\boldsymbol{y}_{ij} = \boldsymbol{\alpha} + \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{D}^T \boldsymbol{z}_{ij} + \boldsymbol{e}_{ij},$$

with $\boldsymbol{e}_{ij}$ IID from $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_q)$. We randomly selected $n = 0.9N$ individuals to train Model (2.3), while the remaining $0.1N$ individuals with repeated measurements are used for validation.

Furthermore, we compared the methods to assess the impact of multiple effects: the number of repeated measurements, different signal-to-noise ratio by varying the model variance parameter σ^2, the robustness of the methods by using data generated from mis-specified models. In this subsection, we

show and compare results with respect to l and σ^2, while the next subsection will focus on the performance under model mis-specification.

Repeated Measurements. In Fig. 2 (a), we presented the mean squared errors of all the four estimators as l was varied. Specifically, we set $l \in 1, 2, 4, 6, 8, 10$ while keeping $N = 2500$ and $\sigma^2 = 1$. Our primary finding was that the optimal shrinkage estimator consistently outperformed the other methods, and the approaches that properly incorporated the model structure significantly outperformed the random forest approach. Additionally, as l increased, the MSE decreased, indicating that even incomplete measurements could provide valuable information. However, as l grew larger, the margin of improvement became smaller.

Scale of noise. In Fig. 2 (b), we considered a comparison under different noise scales. We fixed $l = 3, N = 2500$ with the values of σ^2 selected from 2^{-1} to 2^4. We see that the GLS method was more sensitive to larger error variances, though it had the second best MSE when such variances were relatively small. If σ was large, the estimated $\hat{\boldsymbol{B}}$ was close to zero, which led to an ill-posed inverse problem, and the GLS solution was unstable. In contrast, other methods were quite stable for all the scenarios. The optimal shrinkage still had the least MSE for all the choices of variance.

3.2 Sensitivity analysis

We examine the performance of the concerned estimators under more challenging settings deviated from the baseline setting as in (3.1) with $N = 2500$, $l = 3$, $\sigma^2 = 1$, $q = 10$, $p = 2$ and $p' = 10$.

Omitted dependence variable. We began by considering the case of fitting with incomplete dependence variables. In this setting, data were generated from the baseline model. When fitting it, we interested in the same $\boldsymbol{X}$, with the same set of variables $\boldsymbol{Z}$, but with a "smaller" model involved fewer response variables $\boldsymbol{Y}$. When dropping some columns of $\boldsymbol{Y}$, we correspondingly drop some coefficients $\boldsymbol{B}$ and $\boldsymbol{D}$ to match the dimension.

As shown in Table 1 (a), we trained and validated the calibration model on the first q^* columns of $\boldsymbol{Y}$, where $q^* \in \{0.2q, 0.4q, \ldots, q\}$ controlled the number of accessible dependence variables. In fact, the true model still held, while the problem become one of using less information ($q^* < q$) from the original model to evaluate the calibration accuracy with less number of dependent parameters. It was shown that the optimal shrinkage estimator was still the most accurate and robust against not observing the entire $\boldsymbol{Y}$, while the GLS estimator was sensible to the changes of q^*. For the case of $q^* = 0.2q$, the GLS became extremely unstable as a few outliers could dominate the averaged MSEs.

Omitted covariates $\boldsymbol{Z}$. Similar to the previous setting, the data were generated from the baseline model, but we fit it by omitting some components of $\boldsymbol{Z}$, so it was a case with misspecified models. We considered

Table 1: Empirical means and standard deviations of the MSEs from 1000 simulations for the GLS, the inverse regression (IR), the optimal shrinkage (OPT) and the inverse regression with the random forest (RF) estimators with respect to (a) the number of dependence variables q^* used for training, (b) the number of covariates p'^* used for training, (c) the sample size N and (d) the explanation power of $\boldsymbol{X}$ as represented by shrinking $\boldsymbol{B}^*$ to $\gamma\boldsymbol{B}^*$.

(a) with respect to number of dependence variables

q^*	$0.2q$	$0.4q$	$0.6q$	$0.8q$	q
GLS	3922±60948	0.298±0.568	0.117±0.207	0.069±0.045	0.049±0.022
IR	0.438±0.166	0.199±0.112	0.112±0.066	0.074±0.040	0.053±0.024
OPT	0.385±0.165	0.159±0.094	0.090±0.052	0.061±0.030	0.045±0.018
RF	0.746±0.120	0.599±0.114	0.489±0.103	0.432±0.094	0.391±0.085

(b) with respect to number of covariates

p'^*	$0.2p'$	$0.4p'$	$0.6p'$	$0.8p'$	p'
GLS	0.315±0.255	0.187±0.171	0.110±0.086	0.070±0.044	0.049±0.024
IR	0.211±0.085	0.149±0.069	0.103±0.052	0.072±0.036	0.054±0.025
OPT	0.214±0.097	0.141±0.072	0.092±0.049	0.063±0.030	0.046±0.019
RF	0.413±0.092	0.410±0.089	0.389±0.087	0.388±0.084	0.389±0.083

(c) with respect to sample size

N	30	100	300	1000	2500
GLS	0.461±0.314	0.079±0.038	0.056±0.026	0.050±0.024	0.049±0.023
IR	0.418±0.245	0.079±0.033	0.059±0.026	0.054±0.025	0.053±0.024
OPT	0.441±0.279	0.074±0.030	0.052±0.021	0.047±0.019	0.045±0.018
RF	0.843±0.163	0.666±0.099	0.545±0.086	0.447±0.083	0.390±0.081

(d) with respect to explaining power of $\boldsymbol{X}$

γ	0.1	0.2	0.33	0.5	1
GLS	4.471±1.832	1.157±0.471	0.420±0.171	0.187±0.076	0.047±0.019
IR	0.863±0.040	0.608±0.076	0.356±0.080	0.193±0.061	0.051±0.021
OPT	0.792±0.054	0.498±0.078	0.276±0.067	0.150±0.046	0.044±0.016
RF	0.999±0.022	0.934±0.034	0.819±0.058	0.676±0.077	0.383±0.082

using the first p'^* columns of $\boldsymbol{Z}$ with $p'^* \in \{0.2p', 0.4p', \ldots, p'\}$, and the results were listed in Table 1 (b). Similarly, as more covariates were omitted, the MSE increased for all the estimators, which suggested the important role played by the auxiliary variables $\boldsymbol{Z}$. In this case, since both GLS and the optimal shrinkage estimators were under wrong models, they no longer performed better than the inverse regression estimator when p'^* was small. However, as p'^* got larger, the optimal shrinkage estimator became more accurate along with the other estimators, and was the best performing estimators among the four methods.

Effect of sample size. Here we considered a range of sample size N in the simulation, where N was chosen from $\{30, 100, 300, 1000, 2500\}$. Since the optimal shrinkage estimator is obtained by minimizing the AMSE, it is of the great interest for evaluating how many observations are required in practice. Table 1 (c) showed that the optimal shrinkage estimator performed the best if $N > 30$, while the RF estimator incurred the largest calibration errors for all the sample size considered, and was especially the case in smaller sample size cases.

Weak explanation power on $\boldsymbol{X}$. It is well known that the GLS estimator suffers from weak regressors $\boldsymbol{X}$ in the existed calibration studies without additional covariates $\boldsymbol{Z}$, say, the coefficient $\boldsymbol{B}$ is near zero. We consid-

ered shrinking the coefficient $\boldsymbol{B}$ by a factor γ ranging in $\{0.1, 0.2, 0.33, 0.5, 1\}$, while keeping all other settings the same. Our goal is to see the effect of the magnitude of $\boldsymbol{B}$ on the four estimators in the general calibration problem. As shown in Table 1 (d), the GLS estimator was the most sensitive towards γ, while the optimal shrinkage was robust although the linear effect $\boldsymbol{B}^*$ on $\boldsymbol{X}$ was not that significant.

3.3 Spatial-temporal settings

We demonstrate the applications of the methods in Section 2.4 for calibrations, incorporating additionally the random effects. Inspired by the motivating practical scenario, we designed a setting where the between-observation correlations had some spatial structures. In particular, the distances of the observations play a key role. Due to the known difficulty, there is no common parametric structure available to adequately account for such structures. In the proposed method, the random effect ideally handles such a challenging situation in a nonparametric manner, taking advantages from the flexibility of the basis functions.

In the simulations, we generated $N = 2,500$ individuals from the uniform distribution in a two-dimensional square region $[0, 1] \times [0, 1]$, as demonstrated in Fig. 3 (b). The covariates $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{z}_{ij}\}$ with $i = 1, \ldots, n$,

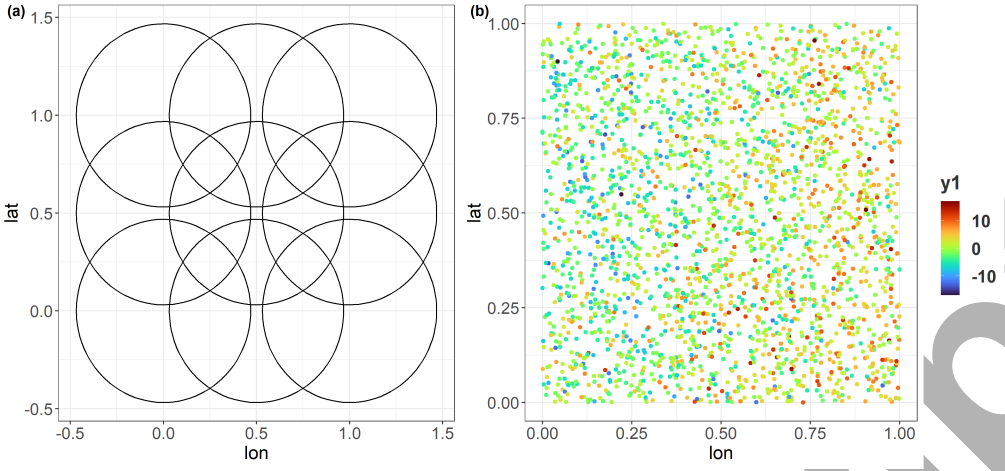

Figure 3: (a). Nine "bisquare" basis functions whose center points are evenly located in the square region $[0,1] \times [0,1]$. (b). Locations of the 2500 individuals, where the colors indicate the value of the first dependent variable y_1. It is clear that y_1 has a correlation structure among the individuals.

$j = 1, \ldots, l$ and coefficients $\boldsymbol{\alpha}$, $\boldsymbol{B}$ and $\boldsymbol{D}$ were generated the same as what in Section 3.1. The dimensions were defined to be $q = 5$, $p = 2$, $p' = 10$. To reflect the closeness between the sites where the observations were from, the "bisquare" basis functions were generated using the newly developed R package FRK (Zammit-Mangion and Cressie, 2021) with $k = 9$ (Fig. 3 (a)). Finally, $\{\boldsymbol{y}_{ij}\}_{j=1}^{l}$ were generated from

$$\boldsymbol{y}_{ij} = \boldsymbol{\alpha} + \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{D}^T \boldsymbol{z}_{ij} + \boldsymbol{A}^T \boldsymbol{\psi}_i + \boldsymbol{e}_{ij},$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k]^T$, $\{\boldsymbol{a}_i\}_{i=1}^{k}$ IID from $\mathcal{N}(0, \boldsymbol{I}_q)$ and $\boldsymbol{e}_{ij}$ IID from $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_q)$.

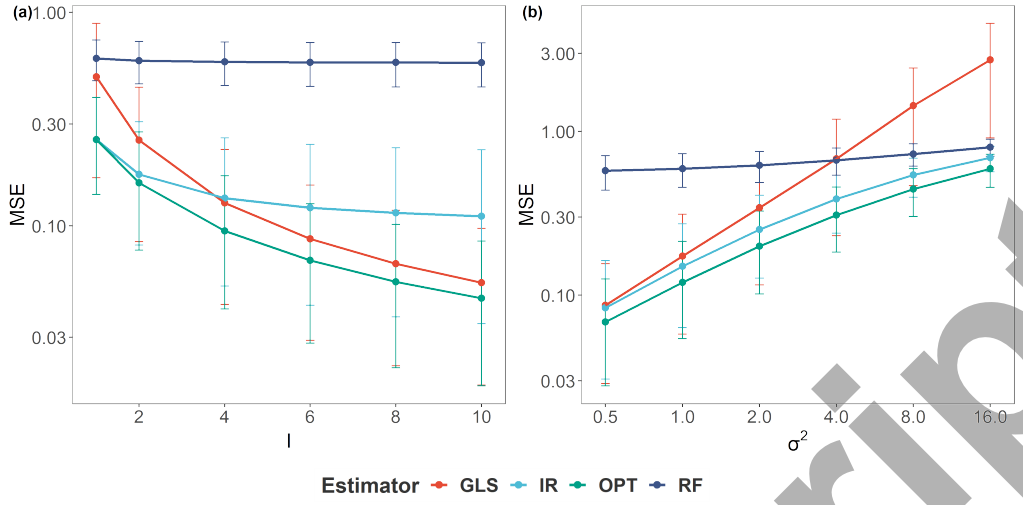

Figure 4: Empirical means and the 10th, 90th quantiles of the MSEs (in the log-scale), in the presence of the random effects, calculated from 1000 simulation replications for the GLS, the inverse regression (IR), the optimal shrinkage (OPT) and the inverse regression with the random forest (RF) estimators with respect to the numbers of repeated measurements l (a) and different scales of noises σ^2.

Fig. 4 reports the mean, and the quantiles of the calibration MSEs in the presence of the random effects with respect to the number of repeated observations l and the signal-to-noise ratio by changing σ^2. It shows that with the number of repeated measurements increased, the performance of the GLS estimator soon improved to outperform the inverse regression estimator, while the optimal shrinkage estimator kept being the best method. On the other side, the GLS estimator was sensitive to the signal-to-noise ratio. When the noise was high, GLS would be unstable and showed a poor performance. As a comparison, the RF seemed to be improper for the linear

model setting, which emphasized that when we have a true model, using this true model instead of a general non-parametric model can help us to calibrate the wanted variable.

4. Real data analysis

4.1 Aqueous Glucose data

Diabetes Mellitus (DM) is one of the common chronic diseases throughout the globe. A continuous blood glucose monitoring is important for DM patients to understand their disease progression and adjust their lifestyle management. However, such blood glucose measurements are often invasive and expensive, thus not proper for regular use. Nowadays, NIR is a promising technique for continuous blood glucose monitoring, since it is cheap, non-invasive and easy to deploy on some wearable devices.

We applied the calibration methods we considered to an open-sourced NIR spectroscopy dataset of aqueous glucose (Fuglerud, 2021). This dataset contained 127 samples measured under laboratory conditions. The glucose concentration was treated as the variable of interest $\boldsymbol{X}$, and five covariates $\boldsymbol{Z}$ (lactates, ethanol, caffeine, acetaminophen, and temperature) were introduced to the dextrose water, along with 4200 NIR wavelength channels ranging from 400nm to 2500nm. The NIR spectra were collected in

triplicate in each setting, sampled every 0.5nm frequency resolution with a bandwidth of 8.75nm. We removed those NIR bands which were highly absorbing water peaks (1900-2100nm, 2300-2500nm), the overlap of the two spectrometer detectors (1090-1110nm), and the fringes of the detectors (<500 and >2300nm), according to Fuglerud et al. (2021). After the data cleaning, we selected the remaining NIR bands every 25nm, resulting in a dimension of $\boldsymbol{Y}$ of 127×60. All the data were preprocessed to have a mean of 0 and a variance of 1.

4.2 Calibration methods comparison

The same four methods in the simulations were applied here to predict the glucose concentration. Ten-fold cross validation was applied to split the training and validation datasets.

Table 2 presents the calibration performance using the glucose data. We can see that the optimal shrinkage method consistently outperformed the other methods, regardless of whether there were repeated observations or not. It is worth noting that the inverse regression and optimal shrinkage estimators were the same when $l = 1$. On the other hand, the GLS estimator performed worse than the inverse regression when $l = 1$. With an increase in the number of repeated measurements from one to three ($l = 3$), both

Table 2: MSE and Bias for glucose prediction under different calibration methods via the ten-fold cross-validation

	MSE		Bias	
	l=1	l=3	l=1	l=3
GLS	0.0912	0.0683	0.2038	0.1563
IR	0.0891	0.0714	0.2008	0.1604
OPT	0.0891	0.0676	0.2008	0.1555
RF	0.4411	0.4690	0.5511	0.5668

the GLS estimator and optimal shrinkage showed significant improvement in their performance, which can be attributed to the reduction in measurement error due to more observations. This result was not surprising since similar results were obtained in the simulation section for such settings. On the other hand, the RF method performed the worst for each model setting. This may be due to the limited sample size relative to the large number of NIR features, which hindered the performance of the RF method.

5. Conclusion

In this study, we generalize the traditional calibration problem with additional covariates $\boldsymbol{Z}$, under both the IID and heterogeneity scenarios. We derived the GLS and inverse regression estimators, and established the equivalence between the inverse regression estimator and a Bayesian estimator, whose prior distribution is the multivariate t-distribution involving information of both $\boldsymbol{X}$ and $\boldsymbol{Z}$. Moreover, we developed a novel optimal

shrinkage estimator, theoretically showed that it has the minimum expected AMSE, and verified its performance in both the simulation and case study. Our study is a comprehensive summary for the properties and relationships among the calibration estimators.

In particular, for many learning problems, one may encounter the reversed causality, say $\boldsymbol{Y}$ is the cause of $\boldsymbol{X}$. Our study provided a direct way to handle the reversed causality, and is the most helpful in the studies which has repeated measurements. Compared with the inverse regression, which is perhaps the most widely used calibration method, the optimal shrinkage estimator is accurate, robust, and still easy to be implemented, which shows more advantages with more repeated measurements being obtained.

Supplementary Materials

Further technical details and proofs are available with this paper at the Statistica Sinica website. The example codes can be found at `https://github.com/tongpf/Inverse-Regression`.

Acknowledgements

We thank three reviewers for constructive comments which have improved the presentation of the work. The research was partially supported by Na-

tional Natural Science Foundation of China Grants 12292983 and 12026607.

References

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics 45*(1), 77–120, 44.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Brown, P. J. (1982). Multivariate calibration. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(3), 287–321.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*. Clarendon Press.

Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Fasiolo, M., S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association 116*(535), 1402–1412.

Fuglerud, S. S. (2021). Aqueous glucose measured by NIR spectroscopy.

Fuglerud, S. S., R. Ellingsen, A. Aksnes, and D. R. Hjelme (2021). Investigation of the effect of clinically relevant interferents on glucose monitoring using near-infrared spectroscopy. *Journal of Biophotonics 14*(5).

Hernández-Lobato, J. M., N. Houlsby, and Z. Ghahramani (2014). Probabilistic matrix factor-

ization with non-random missing data. In *International conference on machine learning*, pp. 1512–1520. PMLR.

Hoadley, B. (1970). A bayesian look at inverse linear regression. *Journal of the American Statistical Association 65*(329), 356–369.

Jarrett, D., B. C. Cebere, T. Liu, A. Curth, and M. van der Schaar (2022). Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, pp. 9916–9937. PMLR.

Jiang, H., T. Liu, and Q. Chen (2020). Quantitative detection of fatty acid value during storage of wheat flour based on a portable near-infrared (nir) spectroscopy system. *Infrared Physics & Technology 109*, 103423.

Katzfuss, M. and N. Cressie (2011). Spatio-temporal smoothing and em estimation for massive remote-sensing data sets. *Journal of Time Series Analysis 32*(4), 430–446.

Krutchkoff, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics 9*(3), 425–439.

Kuleshov, V., N. Fenner, and S. Ermon (2018, 10–15 Jul). Accurate uncertainties for deep learning using calibrated regression. In J. Dy and A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR.

Little, R. J. and D. B. Rubin (2002). *Statistical analysis with missing data*, Volume 793. John Wiley & Sons.

MacKinnon, J. G., M. Ørregaard Nielsen, and M. D. Webb (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*(2), 272–299.

Marden, J. R., L. Wang, E. J. T. Tchetgen, S. Walter, M. M. Glymour, and K. E. Wirth (2018). Implementation of instrumental variable bounds for data missing not at random. *Epidemiology 29*(3), 364–368.

Sundberg, R. (1985). When is the inverse regression estimator mse-superior to the standard regression estimator in multivariate controlled calibration situations? *Statistics & probability letters 3*(2), 75–79.

Sundberg, R. (1999). Multivariate calibration — direct and indirect regression methodology. *Scandinavian Journal of Statistics 26*(2), 161–207.

Wei, J., Z. Li, A. Lyapustin, L. Sun, Y. Peng, W. Xue, T. Su, and M. Cribb (2021). Reconstructing 1-km-resolution high-quality pm2.5 data records from 2000 to 2018 in china: spatiotemporal variations and policy implications. *Remote Sensing of Environment 252*, 112136.

Yun, Y.-H., J. Bin, D.-L. Liu, L. Xu, T.-L. Yan, D.-S. Cao, and Q.-S. Xu (2019). A hybrid variable selection strategy based on continuous shrinkage of variable space in multivariate calibration. *Analytica Chimica Acta 1058*, 58–69.

Zammit-Mangion, A. and N. Cressie (2021). Frk: An r package for spatial and spatio-temporal prediction with large datasets. *Journal of Statistical Software 98*(4), 1 – 48.